PURDUE
KRANNERT
SCHOOL OF MANAGEMENT

UF Department of Information Systems and Operations Management
Warrington College of Business Administration
UNIVERSITY *of* FLORIDA

22nd Workshop on Information Technologies and Systems

December 15-16, 2012

Orlando Florida

SMU
SINGAPORE MANAGEMENT UNIVERSITY
School of Information Systems

MANAGEMENT
INFORMATION SYSTEMS
(MIS)

ÖZYEĞİN UNIVERSITY GSB
GRADUATE SCHOOL OF BUSINESS

ARIZONA | Eller

UCMERCED

UNIVERSITY OF ALBERTA

CARL H.
LINDNER
College of Business

University of Connecticut | School of Business

FOREWORD

Welcome to sunny Orlando, and the 22nd Workshop on Information Technologies and Systems (WITS). The purpose of WITS has always been to create a forum for the exchange of ideas among researchers interested in information technologies and systems. This year the theme of the workshop is "Digital Innovation for Sustainability**.**"

Sustainability is the capacity of systems to endure over time. Even though this definition is intended mostly for ecosystems, it obviously extends to systems created and impacted by humans such as information systems. Unconstrained human economic development has lead not only to destruction of delicate ecosystems, but in the end has contributed to volatile economic cycles and increased economic damage from natural disasters. Information technology is both a contributor and a potential solution to the challenges of managing sustainable development. Better decision support for public policy planners, businesses and consumers can lead to better long-term decisions. Examples are numerous and include inefficiencies in power systems which can be identified by data mining tools and solutions which can be generated by optimization. Virtualization, used properly, can reduce the carbon footprint of existing information systems. Understanding the economic and environmental impact of information systems, can lead to greater sustainable adoption, use, and retirement of information systems and technologies.

This year we had 90 research paper submissions. Thirty-nine were accepted for an acceptance rate of 43%. With the help of the WITS community we had a total of 373 reviews. Seven papers had three reviews, 64 had four, 18 had five and one had six reviews. A total of 13 research prototypes/demos and three instructional technology submissions were selected for presentation during the workshop. The WITS community again was heavily involved. We had 213 co-authors submitting papers, with 86 of them international and 127 from the US. The WITS 2012 Program Committee consists of 39 international members and 86 members from the US.

This year's workshop would not be possible without the efforts of many individuals: The local Arrangements Chair – Joni Jones; Prototype and Instructional Technology Chair – Kaushik Dutta; International Program Liaisons - Guoqing Chen, Jae Kyu Lee, Rahul Roy, Chih-Ping Wei and Leon Zhao for Asia; Cecil Chua and Ron Weber for Australia and New Zealand; Wil van

der Aalst and Matti Rossi for Europe; and Paulo Goes for South America. We would like to acknowledge our Best Paper Award Committee: Jeffery Parsons, Sal March, and Ron Weber and the Best Demo Award Committee (research prototype and instructional technology): Jeffery Parsons, Steve Miller and Akhil Kumar.

In addition to faculty members making this event possible, we have four colleagues in the making who volunteered their time: Brent Kitchens and Soohyun Cho from the University of Florida, and Amy Connolly and Shankar Prawesh from the University of South Florida.

We would like to thank again our gold and silver sponsors – Purdue University Krannert School of Management, University of Florida Warrington School of Business, Singapore Management University School of Information Technology, as well as our Bronze Sponsors – Ozyegin University Graduate School of Management, University of Arizona Eller College of Management, University of California Merced, University of Alberta, University of Cincinnati Carl H. Lindner College of Business, and University of Connecticut School of Business. The generous contributions of our sponsors pay for the awards and help us keep registration fees low for our students as well as faculty members, keeping attendance high.

Many additional thanks to all that helped make this event happen and to all attendees.

Sincerely,
Jackie Rees and Haldun Aytug
WITS 2012 Co-Chairs, on behalf of the Organizing Committee

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF BUSINESS

NUS National University of Singapore
Department of Information Systems
School of Computing

PURDUE
KRANNERT
SCHOOL OF MANAGEMENT

UF
Department of Information Systems
and Operations Management
Warrington College of Business Administration
UNIVERSITY of FLORIDA

WORKSHOP & PROGRAM CO-CHAIRS

Haldun Aytug, University of Florida
Jackie Rees Ulmer, Purdue University

LOCAL ARRANGEMENT CHAIR

Joni L. Jones, University of South Florida

PROTOTYPE DEMO CHAIR

Kaushik Dutta, National University of Singapore

PROGRAM LIASONS

ASIA:

Guoqing Chen, Tsinghua University, Beijing
Jae Kyu Lee, Korea Advanced Institute of Science and Technology
Rahul Roy, Indian Institute of Management, Calcutta
Chih-Ping Wei, National Taiwan University
Leon Zhao, City University of Hong Kong

AUSTRALIA AND NEW ZEALAND

Cecil Chua, University of Auckland
Ron Weber, Monash University

EUROPE:

Wil van der Aalst, Eindhoven University of Technology
Matti Rossi, Aalto University

SOUTH AMERICA:

Paulo Goes, Univeristy of Arizona

PROGRAM COMMITTEE

Ahmed Abbasi, University of Virginia
Gediminas Adomavicius, University of Minnesota
Manish Agrawal, University of South Florida
Gove Allen, Brigham Young University
Sameh Al-Natour, University of British Columbia
Kemal Altinkemer, Purdue University
Ofer Arazy, University of Alberta
Anteneh Ayanso, Brock University
Haldun Aytug, University of Florida
Akhilesh Bajaj, University of Tulsa
Subhajyoti Bandyopadhyay, University of Florida

PROGRAM COMMITTEE (continued)

Reza Barkhi, Virginia Tech University
Sudip Bhattacharjee, University of Connecticut
Siddhartha Bhattacharyya, University of Illinois, Chicago
Henry Bi, Willamette University
Jesse Bockstedt, George Mason University
Indranil Bose, Indian Institute of Management, Calcutta
Fidan Boylu, University of connecticut
Hasan Cavusoglu, University of British Columbia
Hock Chan, National University of Singapore
Kaushal Chari, University of South Florida
Samir Chatterjee, Claremont Graduate University
Jianqing Chen, University of Texas, Dallas
Robert Chiang, Fordham University
Roger Chiang, University of Cincinnati
Cecil Chua, University of Auckland
Sabah Currim, Florida State University
Faiz Currim, University of Arizona
Joseph Davis, University of Sydney
Amitava Dutta, George Mason University
Kaushik Dutta, National University of Singapore
Joerg Evermann, Memorial University of Newfoundland
Xiao Fang, University of Utah
Mark Gaynor, St. Louis University
Andrew Gemino, Simon Fraser University
Xianjun Geng, University of Texas, Dallas
Ram Gopal, University of Connecticut
Zhiling Guo, City University of Hong Kong
Liwen Hou, Shanghai Jiaotong university
Marta Indulska, University of Queensland
Wael Jabr, University of Calgary
Varghese Jacob, University of Texas at Dallas
Hemant Jain, University of Wisconsin, Milwaukee
Matthias Jarke, RWTH Aachen University
Yonghua Ji, University of Alberta
Zhengrui Jiang, Iowa State University
Joni Jones, University of South Florida
Keumseok Kang, Florida International University
Wolfgang Ketter, Erasmus University
Vijay Khatri, Indiana University
Henry Kim, York University
John Krogstie, Norwegian University of Science and Technology
Subodha Kumar, Texas A&M University
Akhil Kumar, Pennsylvania State University
Jae Kyu Lee, Korea Advanced Institute of Science and Technology
Thomas Lee, University of California, Berkeley
Xiaobai Li, University of Massachusetts, Lowell
Zhangxi Lin, Texas Tech University
Rong Liu, International Business Machines Corporation

PROGRAM COMMITTEE (continued)

Pericles Loucopoulos, Loughborough University
Michael Mannino, University of Colorado, Denver
Leo Marland, International Business Machines Corporation
James Marsden, University of Connecticut
Syam Menon, University of Texas at Dallas
Simon Milton, University of Melbourne
Barry Mishra, University of California, Riverside
Saby Mitra, Georgia Tech
Radha Mookerjee, University of Texas, Dallas
Sri Narasimhan, Georgia Tech
Balaji Padmanabhan, University of South Florida
Gautam Pant, University of Iowa
Jinsoo Park, Seoul National University
Praveen Pathak, University of Florida
Raymond Patterson, University of Alberta
Selwyn Piramuthu, University of Florida
Nicolas Prat, École Supérieure des Sciences Économiques et Commerciales
Srinivasan Raghunathan, University of Texas, Dallas
Mohammad Rahman, University of Calgary
Jayantha Rajapakse, Monash University
Narayan Ramasubbu, University of Pittsburgh
Ram Ramesh, State University of New York, Buffalo
Sam Ransbotham, Boston College
Jackie Rees, Purdue University
Erik Rolland, University of California, Merced
Matti Rossi, Aalto University
Rahul Roy, Indian Institute of Management, Calcutta
Aditya Saharia, Fordham University
Pallab Sanyal, George Mason University
Serpil Sayin, Koç University
Arun Sen, Texas A&M University
Thomas Setzer, Technical University of Munich
Benjamin Shao, Arizona State University
Raj Sharman, State University of New York, Buffalo
Keng Siau, Missouri University of Science and Technology
Atish Sinha, University of Wisconsin, Milwaukee
Sanjukta Smith, State University of New York, Buffalo
Yong Tan, University of Washington
Giri Kumar Tayi, State University of New York, Albany
Monica Tremblay, Florida International University
Arvind Tripathi, University of Auckland
Alexander Tuzhilin, New York University
Debra Vandermeer, Florida International University
Harry Wang, University of Delaware
Jingguo Wang, University of Texas Arlington
Stanislaw Wrycza, University of Gdańsk
Yinping Yang, Agency for Science, Technology and Research
Lisa Yeo, University of Alberta

PROGRAM COMMITTEE (continued)

Victoria Yoon, Virginia Commonwealth University
Daniel Zeng, University of Arizona
Juheng Zhang, Oakland University
John Zhang, University of Connecticut
Dongsong Zhang, University of Maryland, Baltimore County
J. Leon Zhao, City University of Hong Kong
Huimin Zhao, University of Wisconsin, Milwaukee
Dmitry Zhdanov, University of Connecticut
Rong Zheng, Hong Kong University of Science and Technology
Eric Zheng, University of Texas, Dallas
Lina Zhou, University of Maryland, Baltimore County
Dan Zhu, Iowa State University
Hongwei Zhu, Old Dominion University

SUB-REVIEWERS

Panagiotis Adamopoulos, New York University
Fernando Beltran, University of Auckland
Alfred Castillo, Florida International University
Soohyun Cho, University of Florida
Howard Hao-Chun Chuang, Texas A&M University
Chao Ding, University of Florida
Shaokun Fan, University of Delaware
Fang Fang, National University of Singapore
Philipp Herrmann, University of Calgary
Ian Ho, University of Washington
Yi-Chun Ho, University of Washington
Ke-Wei Huang, Agency for Science, Technology and Research
Arie Jacobi, University of Alberta
Huanhuan Ji, Purdue University
Irfan Kanat, Arizona State University
Brent Kitchens, University of Florida
Youngok Kwon, University of Minnesota
Shengli Li, University of Florida
Zhi Li, University of Florida
Yixin Lu, Erasmus University
Christian Meier, University of Calgary
Seyedreza Mousavi, Arizona State University
Toan Ong, University of Colorado, Denver
Nargis Pervin, National University of Singapore
Markus Peters, Erasmus University
Zhe Shan, University of Delaware
Di Shang, Agency for Science, Technology and Research
Haoyan Sun, University of Washington
Konstantina Valogianni, Erasmus University
Jianquing Wu, Purdue University
Weifang Wu, Hong Kong University of Science and Technology

SUB-REVIEWERS (continued)

Soyoung Yang, Purdue University
Yu-Chen Yang, University of Florida
Niam Yaraghi, State University of New York, Buffalo
Yang Yu, Texas Tech University
Shuai Yuan, State University of New York, Buffalo
Dawei Zhang, University of Calgary
Jingjing Zhang, University of Minnesota
Kang Zhao, Agency for Science, Technology and Research
Steffen Zimmermann, University of Calgary

Statistics by Country

Country	Authors	Submitted	Accepted	Acceptance rate	PC members
Australia	5	2.67			3
Canada	8	4	0.33	0.08	11
China	19	6.38	3.22	0.5	1
Finland					1
France	1	0.33	0.33	1	1
Germany	2	0.67	0.67	1	2
Hong Kong	8	2.83	1.33	0.47	3
India	8	2.53	1.33	0.53	2
Ireland	1	1.33			
Israel	10	3.75	1.75	0.47	
Japan	2	1			
Republic of Korea					2
Malaysia					1
Netherlands	6	2.5	1	0.4	1
New Zealand	1	0.5			2
Norway	1	0.33			1
Poland					1
Saudi Arabia	2	1	1	1	
Singapore	9	3.6	1.8	0.5	3
Taiwan	1	0.5	0.5	1	1
Turkey					1
United Arab Emirates	1	1			1
United Kingdom	1	0.5			1
United States	127	54.57	25.73	0.47	86
Total	**213**	**89.99**	**38.99**		**125**

Saturday, December 15, 2012		
7:30 AM – 8:15 AM	**Continental Breakfast (Salon 8A)**	
8:15 AM – 8:30 AM	**Welcoming Remarks (Salon 8A)**	
8:30 AM- 10:00 AM	**1A: Health Care Information Systems** **Location: Salon 13 Chair: Raj Sharman**	**1B: E-Commerce** **Location: Salon 14 Chair: Anuj Kumar**
	Healthcare Information Exchange: A Game-theoretic Analysis. *Emre Demirezen, Subhodha Kumar, and Arun Sen.*	Merchant's Second Battlefield: Optimal Profit-Sharing Scheme on Cashback Platforms. *Yi-Chun Ho and Yi-Jen Ho*
	Modeling Practice Efficiencies due to Healthcare Information Exchanges and Implications for HIE Network Growth. *Niam Yaraghi, Anna De Yu, Raj Sharman, Ram Gopal, and Ram Ramesh.*	An Information Stock Model of Customer Behavior in Multichannel Customer Support Services. *Anuj Kumar, Kinshuk Jerath, and Serguei Netessine*.
	Extracting and Releasing Information from Clinical Documents with Guaranteed Privacy. *Xiaobai Li and Jialun Qin.*	"Showrooming" and the Competition between Store and Online Retailers. *Amit Mehra, Subodha Kumar, and Jagmohan Raju*.
10:00 AM- 10:30 AM	**Coffee Break (Salon 8A**)	
10:30 AM- 12:00 PM	**2A:Text mining in Finance** **Location: Salon 13 Chair: Praveen Pathak**	**2B:Auctions** **Location: Salon 14 Chair: Huimin Zhao**
	Does the Market Believe in Marketing? A Text Mining Based Informational Value Perspective. *Joseph Johnson, Brent Kitchens, Debanjan Mitra, and Praveen Pathak.*	Optimization-based Decision Support for Scenario Analysis in Sourcing Markets with Economies of Scale and Scope. *Christian Haas, Martin Bichler, and Kemal Guler.*
	Firm Clustering Based on Financial Statement. *Dazhi Chong and Hongwei Zhu.*	Budget Planning for Coupled Campaigns in Search Auctions. *Yanwu Yang, Rui Qin, Jie Zhang, Xin Li, and Daniel Zeng.*
	Exploring the Informativeness of Forward-Looking Statements in Public Firm Disclosures. *Zhimin Hua, Xiaoyu Wu, and J. Leon Zhao.*	Polarization Trend in Sponsored Search Auctions. *Yong Yuan, Daniel Zeng, Huimin Zhao, and Jiesi Cheng.*
12:00 PM- 1:30 PM	**Lunch and Keynote Speech (Salon 8B**) Dan Azevedo, Senior Manager, Critical Facilities at The Walt Disney Company	
1:30 PM- 2:45 PM	**Panel I – Research to Startup (Salon 14)**	
2:45 PM- 3:15 PM	**Coffee Break (Salon 8A)**	
3:15 PM- 5:15 PM	**3A:Recommendation Systems** **Location: Salon 13 Chair: Roger Chiang**	**3B: Software Development and Implementation I** **Location: Salon 14 Chair: Dan Zhu**
	Manipulation Resistance in Feedback Models of Top-N Recommenders. *Shankar Prawesh and Balaji Padmanbhan*.	Group Reputation in an Open Source Software Community. *Yuanfeng Cai and Dan Zhu.*
	Social Media-Based Social TV Recommendation System. *Shawndra Hill, Adrian Benton, and Jin Xu.*	A Case for a Workflow Driven Workflow Execution Engine. *Shubhangi Sharma, Kamalakar Karlapalem, and Radha Krishna Pisipati.*
	Product reputation manipulation: Exploring the linguistic characteristics of shill reviews. *Toan Ong and Michael Mannino.*	Creating a Repository for the Design and Delivery of Web Services. *John Delano, Atish Sinha, and Hemant Jain.*
	Recommender Systems Position and Orientation Study in E-commerce Websites. *Noura Alhakbani and Abdulrahman Mizra.*	
5:00 PM – 6:00 PM	WITS Board Meeting **(Salon 14)**	
6:15 PM	Leave hotel for Dinner at Funky Monkey I-Drive (Buses depart from hotel main entrance – Reception at 6:45, Dinner at 7:30 (**Restaurant Address**: Pointe Orlando, 9101 International Drive, Suite 1208, 407-418-9463)	

Sunday, December 16, 2012		
7:00 AM – 8:00 AM	**Continental Breakfast (Salon 8A)**	
8:00 AM- 10:00 AM	**4A: Machine Learning** **Location: Salon 13** **Chair: Gautam Pant**	**4B: Internet Service Delivery and Mobile Commerce** **Location: Salon 14** **Chair: Robert Easley**
	Online Review Sentiment Classification Using a Dictionary-Enriched Text Mining Model. *Thomas Ngo-Ye and Atish Sinha.*	Quality of Service Tiering: Implications for Content Innovation and Broadband Coverage. *Hong Guo and Robert Easley.*
	Improving Sentiment Analysis by Dimension Reduction and Parameter Optimization in Support Vector Machines. *Xinmiao Li, Yilu Zhou, Jing Li, and Pengzhu Zhang.*	Service Consolidation and Interconnection among Internet Service Providers. *Robert Chiang and Jhih-Hua Jhang-Li.*
	A Statistical Model for Hierarchical Topic Modeling and User Interest Discovery. *Xuning Tang, Christopher C. Yang, and Mi Zhang.*	Surviving Hyper-Competitive, Unforgiving Platform Ecosystems: Examining Developer Strategies in iOS and Android Marketplaces. *Narayan Ramasubbu, Kajanan Sangaralingam, Nargis Pervin, Kaushik Dutta, and Anindya Datta.*
	Network Effects Comparison in Large Social Network -- An Empirical Investigation. *Bin Zhang, Ramayya Krishnan, and David Krackhardt.*	The Economics of Shared Data Plan. *Soumya Sen, Carlee Joe-Wong, and Sangtae Ha.*
10:00 AM- 11:00 AM	**Poster Session** **Coffee Break (Salon 8A)**	
11:00 AM- 12:15 PM	**Panel II (Salon 14)** Working on Mars while Living on Earth	
12:15 PM - 1:30 PM	**Lunch and Award Ceremony (Salon 8B)**	
1:30 PM - 3:00 PM	**5A: Healthcare Applications** **Location: Salon 13** **Chair: Shawndra Hill**	**5B: Software Development and Implementation II** **Location: Salon 14** **Chair: Vijay Khatri**
	Assessing the contribution of HIE systems to medical decision-making. *Ofir Ben-Assuli, Moshe Leshno, Itamar Shabtai, and Shawndra Hill.*	Managing Agile Software Development: A Control-Theoretic Approach. *Subodha Kumar, Yonghua Ji, and Vijay Mookerjee .*
	Mining Patient Orders to Rank Point of Care Tests in Emergency Department Operations. *Thomas Lee and Esther Chen.*	Optimal Coordination in Distributed Software Development. *Hao Xia, Milind Dawande, and Vijay Mookerjee.*
	Evolving Decision Strategies for Dynamic Environments: A Genetic Programming Approach. *G. Meyer, P. Johnson, G. Adomavicius, P. O'Connor, J. Sperl-Hillen, W. Rush, M. Elidrisi, S. Bandyopadhyay.*	Uncertainty, Switching Cost, and Competition in the Software-As-a-Service Market. *Dan Ma and Robert J. Kauffman.*
3:00 PM – 3:30 PM	**Coffee Break (Salon 8A)**	
3:30 PM - 5:00 PM	**6A: Social Networks** **Location: Salon 13** **Chair: Lisa Yeo**	**6B: Analytics/ Big Data** **Location: Salon 14** **Chair: Keumseok Kang**
	Online Opinion Formation and Social Interactions. *Lu Yan, Roch Guerin, Karthik Hosanager, Yong Tan, and Santosh Venkatesh.*	Exploring Crowds' Mean Belief in Fixed Odds Betting for Event Prediction. *Weiyun Chen, Xin Li, and Daniel Zeng.*
	News Article Propagation on Twitter based on Network Measures - An Exploratory Analysis. *Devipsita Bhattacharya and Sudha Ram.*	Learning Sparse Heterogeneous User Preferences. *Marcus Peters and Wolfgang Ketter.*
	Improving robustness of scale-free networks to message distortion. *Arie Jacobi, and Ofir Ben-Assuli.*	Service Systems with Postponable Acceptance and Assignment: A Dynamic and Stochastic Programming Approach. *Keumseok Kang, J. George Shanthikumar, and Kemal Altinkemer.*
5:00 PM	WITS 2013 Planning Meeting **(Salon 14)**	

Poster and Prototype Session

Sunday, December 16, 10:00 AM - 11:00 AM (Salon 8A)

Cluster	Poster/Prototype
Instructional Technology	Using Virtual Worlds as an Innovative Technology for Teaching Rui Huang[1] and Rebecca Jestice[2] ([1] Binghamton University, [2] Earlham College)
	It's Not Just a Class, It's an Adventure: Teaching Web Development Through Game Creation George M. Wyner and Benjamin Lubin, Boston University
	Interactive Classroom Modeling of Key Information & Technology Management Issues Abraham Seidmann, University of Rochester
Healthcare	Wearing Your Heart on Your Sleeve: The Effects of Forums and Search on Sales Tomer Geva[1], Gal Oestreicher-Singer[2], Niv Efron[3], and Yair Shimshoni[4] ([1,3,4] Google, Tel Aviv, [2] Tel Aviv University)
	Virtual World for Advanced Cardiac Life Support Training Prabal Khanal[1], Akshay Vankipuram[2], Aaron Ashby[3], Karen Josey[4], Ashish Gupta[5], Marshall Smith[6] ([1,2,3] Arizona State University Scottsdale, [4,6] Banner Health SimET Center, Phoenix, [5] Minnesota State University Moorhead)
	Research Prototype: A Knowledge Management System to track the Evaluation of the Implementation of a Statewide Health Information Exchange Monica C. Tremblay, Arturo Castellanos, and Gloria Deckard, Florida International University
	A Prototype of a Patient Safety Knowledge Management System (PSKMS) Srikanth Parameswaran, Rohit Valecha, Raj Sharman, H. Raghav Rao, Ranjit Singh, Gurdev Singh,State University of New York at Buffalo
	A Prototype of a Local Emergency Response System based on a Multi-agent conceptual Modeling Language Rohit Valecha, Swati Subhedar, Kshitj Agrawal, Raj Sharman, Raghav Rao, and Shambhu Upadhyaya, State University of New York at Buffalo
Business Analytics	An Analytics Platform for Mobile Applications Anindya Datta and Sangaralingam Kajanan, National University of Singapore
	Demonstration of an Analytical Software Feature in Generalized Audit Software: Use of Benford's Law for Fraud Detection Tasks Hyo-Jeong Kim and Michael Mannino, University of Colorado Denver
Virtual world, Gamification	Face Recognition Enabled Avatar Doug Derrick and Aaron Read, University of Nebraska at Omaha
	Gamifying Collaborative Decision Making Mohammad A. Moradian[1], Kelly Lyons[2], Maaz Nasir[3], and Rock Leung[4] ([1,2,3] UIniversity of Toronto, [4] SAP Canada)
Economics	The Role of Product Variety and Maturity in the Market Valuation of IT Intensive Firms Wael Jabr[1] and Eric Zheng[2] ([1] University of Calgary, [2] University of Texas at Dallas)
	Fixed, Spot and/or Flexi Pricing: An Integrated Prototype for Examining Alternate Pricing Mechanisms in Cloud Computing Yang Yinping[1], Richard Shang[2], and Huang Jianghui[3] ([1,2] Institute of High Performance Computing, A*STAR, [1,3] Singapore Management University)
	Buy it now or later: the Impact of Mari on Multi-unit Sequential Dutch Auctions Yixin Lu[1], Paul van Iterson[2], Alok Gupta[3], Wolfgang Ketter[4], Jan van Dalen[5], and Eric van Heck[6] ([1,2,4,5,6] Rotterdam School of Management, Netherlands, [3] University of Minnesota)
	Increasing Social Welfare and Individual Savings using Economic Incentives for Electric Vehicles Konstantina Valogianni[1], Wolfgang Ketter[2], Mathijs de Weerdt[3], and John Collins[4] ([1,2] RSM Erasmus University, [3] Delft University of Technology, [4] University of Minnesota)

Panel I
Research to Startup

Alok Chaturvedi, Professor, Krannert Graduate School of Managementand the Founder, Chairman, and the CEO of Simulex Inc.
Don Berndt, Associate Professor, Univerity of South Florida, *Co-Founder and Chief Technology Officer*, Medegy (Healthcare Information Management Company)
Anindya Datta, Associate Professor, School of Computing, National University of Singapore, Founder of Chutney Technologies, Chairman of Mobilewalla
Kartik Hosanagar, Associate Professsor, Wharton, Co-founder of Yodle Inc,
David G. Schwartz, Vice-chairman and Professor of Information Systems, Bar-Ilan University

The internet has created a platform where ideas can turn into businesses very quickly. Today anyone with an idea can be an application provider for mobile computing platforms. With many in the WITS community engaging in research on recommender systems, online auctions, mobile computing, healthcare exchanges and many other current topics, it is inevitable that some of these research ideas are also excellent opportunities for starting a business. This panel will discuss what it takes to convert an idea to a business, and what it means for one's [academic] career.

Panel II

Working on Mars while Living on Earth - Balancing Demands across Disciplinary Boundaries

Kevin C. Desouza, Associate Dean for Research and Associate Professor, College of Public Programs, Arizona State University
Sandeep Purao, Professor, College of Information Sciences and Technology, Penn State University (Chair)
Steve Sawyer, Associate Dean for Research and Professor, School of Information Studies, Syracuse University
Ajay Vinze, Associate Dean for International Programs and Earl and Gladys Davis Distinguished Professor, W.P. Carey School of Business, Arizona State University

With the desire to work on a broad range of topics, a number of IS scholars are choosing to make disciplines other than Business and MIS as their home disciplines. Recently, with the advent of Schools of Information and allied information disciplines such as HCI, CSCW, Computer Science, Software Engineering and others such as Service Science and Public Policy, the trend seems to be accelerating. As these IS scholars engage in research inquiries at different levels of analyses that span individual to societal - they face several demands for balancing demands across disciplinary boundaries. What are the concerns these scholars face? What strategies do these scholars employ to overcome these concerns? And how do these scholars contribute to the IS discipline? The panel will explore these concerns by bringing together researchers who have chosen to work with research themes that are relevant to the IS discipline while continuing to live and act within schools, institutes and centers outside IS.

Healthcare Information Exchange: A Game-theoretic Analysis

Emre M. Demirezen, Subodha Kumar, Arun Sen
Texas A&M University, {edemirezen, skumar, asen}@mays.tamu.edu

Abstract

In the last few years, the U.S. government has been aggressively promoting the establishment of regional healthcare information exchanges (HIEs). HIEs facilitate electronic health information exchange among healthcare providers (HPs). In order to incentivize HIE providers (HIEPs) and HPs, federal government and other parties provide funds to HIEPs and participating HPs. Moreover, HPs pay fees that are determined by the HIEPs in order to join these networks. Using a game-theoretic approach, we analyze two settings in this domain. In the first one, we derive the conditions in order for the HPs to join a newly established HIE. Since the well-established HIEs also offer other value added services, in the second setting, we study the conditions for the HPs to stay in these networks and derive the equilibrium service levels. In this case, the HIE maximizes its profits considering the cost of different services and deciding on their prices. The analysis of the second model also reveals the conditions to sustainably increase the network size. In addition, we present some interesting insights that would be useful for both HIEPs and HPs.

Keywords: Healthcare management, HIE networks, game theory, Nash equilibrium.

1. Introduction

Health information exchange (HIE) is usually defined as an information sharing mechanism that automates the transfer and sharing of health-related information typically stored in multiple organizations, while maintaining the context and integrity of the information being exchanged (HIMSSS 2009). An HIE provides access to and retrieval of patient information to authorized users in order to provide safe, efficient, effective, and timely patient care. HIEs are typically formed by a group of participants from a specific area to facilitate the electronic exchange of health-related information between providers (hospitals, physicians, clinics, labs, etc.).

HIE is not only about moving clinical information to the right place, it also affects the clinical workflows and makes the data available to doctors and nurses when they need them to make decisions. It provides improved patient safety by sharing their medical data. In addition, many cost reductions are possible such as elimination of duplicate tests, recovery of missing patient health data, elimination of paper, ink, and manual document printing, and reduction of phone calls and follow-ups with labs. HIE also helps provide accurate feedback to public health registries and assures a strong chain of custody of the patient data and their movements.

Great progress is being made to make HIE a reality to benefit the patients (AHRQ 2006). However, questions and challenges remain – and we study the following questions: (i) How would the HIE be established and which HPs would participate? (ii) Once it is initiated, what additional services it can offer for improved sustainability, and what would be the pricing of these service offerings. Accordingly, we analyze two HIE settings in this paper. In the first one, an HIE is not established and the price set by the HIEP determines the number of HPs joining the network. In this setting, we determine the equilibrium price of the service and the characteristics of the participating HPs. In the second setting, there is an established HIE and it can offer additional services to its members. We determine the requested service levels and the price of these services in this setting, and analyze the conditions for the expansion of the network. We also glean several useful managerial insights in both settings.

2. Problem Definition

We would like to emphasize that our models are based on our experience with different HIEPs.

Hence, before we present our models, we describe a real life HIE scenario (Southeast Texas Health Systems) that forms the underpinning of our first model. The details of the Integrated Care Collaboration (motivation for our second model) are omitted due to the space limitation.

The Southeast Texas Health Systems (SETHS; http://www.seths.info/index.php/sophie) is a hospital association in southeast Texas. One of the main objectives of SETHS is to implement and sustain an HIE among hospitals in order to ultimately improve the health status of the population in a region. SETHS's plan is to develop and operate a high-quality, cost-effective HIE for small and rural HPs. According to the CEO of SETHS, such an HIE will create incentives for its members to share data with other "like" HPs, and the ability to negotiate reimbursements by demonstrating evidence-based clinical decision support.

As in the case of SETHS, younger HIEPs, especially those that are not yet sharing data, tend to measure success in terms of the number of participants. This is because, the HIE will have only base functionality at the beginning, and the expansion of the participating HP base will benefit the sustainability of the network in its earlier stages. Hence, we formulate our Model 1 based on this scenario. Some HIEPs offer other value-added services on top of base functionality (Covich et al. 2011). Therefore, we study Model 2 that represents this environment. Before introducing the models, we would like to note that the HIEPs are sometimes considered as for-profit and sometimes non-profit organizations (Berry and Johnson 2012). However, they all strive for maximizing their profits in order to stay alive. Most HIEPs rely on fees collected from the HPs. Therefore, independent of their nature, we regard them as profit maximizing organizations in our models. We begin with the discussion of Model 1 in the next section.

3. Model 1: Fixed Service Level Problem

In this model, we consider that an HIEP is about to be established within a group of HPs. The total number of potential HPs that consider joining to the HIE is N. As discussed before, because the HIE is new, the service offerings it will have is limited in the short-medium planning horizons. Hence, the HPs decide to join or not but cannot request additional services from HIEP.

3.1. Model Description

If an HP i joins the network, it is able to estimate the net benefit it will get from the network using the methodologies mentioned in the HIE value analysis literature (e.g., Walker et al. 2005). We denote this benefit by r_i. This parameter is the difference between the gains of the HIE and the fixed costs needed to join the HIE network. Gains include operational and quality benefits as well as the fixed portion of the funds and reimbursements awarded by the federal government and other agencies and parties (Dixon et al. 2010). On the other hand, costs include the purchase of new equipment, IT infrastructure including the cost of interface construction for its members, and personnel training. The information about the net benefit r_i is private, i.e., it is unknown to the HIEP and the other HPs. On the other hand, the distribution of r_i is common knowledge. In this setting, it is reasonable to consider that r_i is independent and identically distributed (i.i.d.), and follows uniform distribution with lower and upper bounds $\underline{r}$ and $\overline{r}$, i.e., $r_i \sim U(\underline{r}, \overline{r})$. The HIEP sets the subscription price p that is generally a monthly fee. The price determines the number of HPs that join the network. Let us denote this number by K and its expected value by $\mathrm{E}(K)$. On the other hand, federal government, state appropriations, state Medicaid funds, and funds from various foundations reimburse a portion g of the cost of the HPs that join the network (Dixon et al. 2010). Hence, the objective function of HP i can be written as $r_i - p(1 - g)$. The HP i joins the network only if this value is positive.

The HIEP collects fees for its service from all the HPs that join. Hence, the higher the price it sets, the more money it collects per HP that join. However, high price for the HIE service deters the HPs as they will benefit less from joining the HIE network. Therefore, the HIE needs to balance the price accordingly that will determine the most beneficial network size. In addition to the subscription fees collected from the HPs, the HIEP receives funds from government and other third parties that are usually proportional to the network size (Covich et al. 2011). Hence, we model it as Kf, where $f \geq 0$. On the other hand, the costs for maintaining the HIE include implementation costs, hosting and data services costs, and administrative and operational costs (Covich et al. 2011). According to Mostashari et al. (2009), the cost for maintaining the network is linearly proportional to network size. Therefore, we model it as cK, where $c \geq 0$. We also solve the problem with quadratic increasing form, i.e., cK^2. However, we do not report the results due to brevity. In summary, given the parameters and variables above, the objective function of the HIEP and each HP, and the constraints can be written as:

$$\operatorname*{Max}_{p} \mathrm{E}[K(f+p) - cK]$$

$$\mathrm{Max}\,[0, r_i - p(1-g)] \quad \forall i$$

$$\text{s.t.} \quad r_i - p(1-g) > 0;\ \mathrm{E}[K(f+p) - cK] > 0$$

3.2. Discussion and Insights

We begin with deriving the expected network size for any given price set by the HIEP (i.e., p). This result is presented in the following lemma. Due to space limitation, we omit all proofs.

***Lemma* 1:** *The expected network size (i.e, $\mathrm{E}(K)$) equals to $Max\left\{0, \frac{\overline{r}-(1-g)p}{\overline{r}-\underline{r}}N\right\}$.*

It is easy to observe in Lemma 1 that the expected size of the network linearly decreases in p. Also, using the result of Lemma 1, it is easy to show that p is bounded as follows: $\frac{\underline{r}}{1-g} \leq p \leq \frac{\overline{r}}{1-g}$. Reducing the price lower than this bound in not Pareto-optimal. Likewise, the upper bound Pareto-dominates the prices that are higher. Next, we present the equilibrium price.

***Proposition 1*:** *Equilibrium price of the network service (i.e., p) is*
(a) $\frac{1}{2}\left(c - f + \frac{\overline{r}}{1-g}\right)$ if $\underline{r} < (c-f)(1-g) < \overline{r}$, *(b)* $\frac{\underline{r}}{1-g}$ if $\underline{r} \geq (c-f)(1-g)$, *(c)* $\frac{\overline{r}}{1-g}$ if $\overline{r} < (1-g)(c-f)$.

It is interesting to observe that the price in part (a) increases as the upper bound on the gain for the HPs (i.e., $\overline{r}$) increases. HIEP increases the price in such a case in order to extract more value from the HPs. Using the above results, we can determine when (i) there are no participants in the network, (ii) every HP joins the network, and (iii) a partial network is formed. We summarize these results in the following proposition.

***Proposition* 2:**

(a) There are no participants in the network when $\overline{r} < (1-g)(c-f)$,

(b) Every HP joins the HIE when $\underline{r} \geq (c-f)(1-g)$,

(c) A partial network is established, where only some HPs join, when $\underline{r} < (c-f)(1-g) < \overline{r}$.

Based on Propositions 1 and 2, we present the equilibrium network size below.

***Proposition 3*:** *The expected network size in the equilibrium is given by*
(a) $\frac{\overline{r}-(c-f)(1-g)}{2(\overline{r}-\underline{r})}N$ if $\underline{r} < (c-f)(1-g) < \overline{r}$, *(b)* N if $\underline{r} \geq (c-f)(1-g)$, *(c)* 0 if $\overline{r} < (c-f)(1-g)$.

The rationale behind the funding provided by the government and the other parties to HIEs is to improve the overall healthcare for the society and to increase the efficiency of the healthcare system. It is obvious that these benefits increase when there are more participants in the HIEs.

From Proposition 3, it is easy to see that the incentives for both HIEP and HPs (i.e., f and g) encourage a larger network. Further, we find that the incentives need to be above a given threshold in order to have full participation of the HPs. This result is presented in Proposition 4.

Proposition 4: *If* $\underline{r} < \mathrm{Min}\left(\frac{\overline{r}}{2}, (c - f)(1 - g)\right), f < c$, *and* $g < 1$, *then full participation is not possible.*

This result shows that if the lower bound of the benefits gained by HPs is low, then the incentives need to be high enough in order to induce full participation. Finally, in the following proposition, we present the objective function values of HPs and the HIEP.

Proposition 5: *At the equilibrium, the objective function values of HP* i *is* $\mathrm{Max}\left\{0, r_i - \frac{1}{2}(\overline{r} + (c - f)(1 - g))\right\}$, *and the HIEP is* $\frac{(\overline{r}+(c-f)(1-g))^2}{4(\overline{r}-\underline{r})(1-g)}N$.

It is easy to see from Proposition 5 that there is a threshold value that determines which HPs will join the HIE. If the benefit for HP i (i.e., r_i) is more than $\frac{1}{2}(\overline{r} + (c - f)(1 - g))$, then HP i joins the network. From this threshold, we observe that if $\overline{r}$ increases, the HPs with low benefit levels might choose not to join the HIE. In other words, if the upper bound of the benefits gained by HPs is high, then it is beneficial for the HPs with low benefits to stay out of the HIE network.

On the other hand, the value of the HIEP increases in $\overline{r}$ and $\underline{r}$. This implies that the HIEP extracts more value from its network if the HPs are expecting higher benefits from the network. It is also worth mentioning that the values of both the HIEP and the HPs that join the network increase as any of the government reimbursement parameters (i.e., f or g) increases. Next, we present our second model.

4. Model 2 - Variable Service Level Problem

In this scenario, the HIEP serves a total of K HPs. This model and its analysis are applicable to environments where the HIEP is a well-established network and offers its participants additional value-added services ranging from providing master patient index to longitudinal patient record viewers, and from ePrescribing to patient management tools and quality reporting (Covich et al. 2011). We derive the conditions for the HPs that are already in the network to stay in the HIE. The solution of this model also helps us study when expanding for the HIEP would be beneficial, and the conditions for the outside HPs to join the HIE and select service levels. The analysis in this section holds for any set of $K \subseteq N$ HPs present in the network.

4.1. Model Description

Each participating HP i requests additional services on top of base HIE functionality that we denote as s_i. As discussed earlier, federal funds, state appropriations, state Medicaid funds, and funds from various foundations reimburse a fraction of the cost g_i if HP i joins. Clearly, $g_i = g \; \forall i$ is a special case of our model. The benefit of receiving service level s_i by HP i increases at a decreasing rate and modeled as $r_i {s_i}^{\alpha_i} K^{\beta_i}$ where $0 < \alpha_i < 1$ (Covich et al. 2011). This implies that as the HP increases the service level it requests, additional services will become less beneficial than the earlier services. Similarly, the benefit of network size (due to the increase in the amount of shared information) increases at a decreasing rate which is depicted by $0 < \beta_i < 1$. The HIEP can charge different prices for different HPs depending on the characteristics of the HPs, i.e., price p_i is different for each i. Here, p_i can be thought of as the price per service level (Walker et al. 2005, Dey 2009). Therefore, the total price an HP pays is $p_i s_i$ before receiving the funds from government and other parties. The benefit and cost difference discussed so far, i.e., $r_i {s_i}^{\alpha_i} K^{\beta_i} - p_i(1 - g_i)s_i$, needs to be more than a threshold in order for HP i to join the HIE. We use Δ_i to represent this net reservation utility or threshold level. For HPs that are already in the network, this threshold can be calculated using indirect costs of staying in the network, costs of switching out of network, and additional endowments that are received for staying in the network. Similarly, for the HPs that consider joining, it can be calculated as benefits (including indirect network benefits and additional endowments) minus fixed costs of joining the network (including training, extra equipment, IT interface construction, etc.).

As mentioned, the HIEP sets the price per service level p_i for each HP differently depending on the characteristics of these HPs (Covich et al. 2011, Private Communication 2012). With the assumptions stated so far, it is easy to see that the problem of the HIEP is separable in HPs. This means that the HIEP can decide on the pricing of its services for a specific HP independent of other HPs. The cost for keeping HP i in the network is $c_i {s_i}^{\gamma_i} K^{\delta}$ (Walker et al. 2005, Covich et al. 2011). Besides, it is costlier for the HIEP to provide the same level of service in larger networks, i.e., $\delta > 0$. Parameter c_i can be thought of as the base level of costliness per service level in order to accommodate HP i in the network. This cost is HP specific and might depend on factors such as its distance from the HIEP headquarters, the size of the HP, etc. Moreover, it is costlier for the HIEP to offer a higher level of service for any HP, and therefore $\gamma_i > 0$. As can be seen, our model is flexible enough to represent different environments. If the IT capabilities of the HIEP is high, then we expect to see a low value for δ. Likewise, if the HP i has sophisticated IT in place, then we would expect to see a low γ_i value (Walker et al. 2005, Covich et al. 2011, Dey 2009). Without loss of generality, we normalize the reservation utility of the HIEP to zero. This utility term includes additional endowments and the fixed cost of establishing the HIE. In summary, given the parameters and variables above, the objective function of the HIEP and each HP, and the constraints can be written as the following:

$$\text{Max}_p \sum_{i=1}^{K} p_i s_i - \sum_{i=1}^{K} K^{\delta} c_i {s_i}^{\gamma_i}$$
$$\text{Max}\, r_i {s_i}^{\alpha_i} K^{\beta_i} - p_i(1-g_i)s_i \quad \forall i$$
$$\text{s.t.} \quad \sum_{i=1}^{K} p_i s_i - \sum_{i=1}^{K} K^{\delta} c_i {s_i}^{\gamma_i} > 0; \; r_i {s_i}^{\alpha_i} K^{\beta_i} - p_i(1-g_i)s_i \geq \Delta_i$$

4.2. Discussion and Insights

The solution of the game provides us the price set by the HIEP in the equilibrium that is presented in the following proposition.

Proposition 6: *Equilibrium price per service level is* $p_i = K^{\frac{\delta(1-\alpha_i)+\beta_i(\gamma_i-1)}{\gamma_i-\alpha_i}} \left(\frac{c_i\gamma_i}{\alpha_i}\right)^{\frac{1-\alpha_i}{\gamma_i-\alpha_i}} \left(\frac{r_i\alpha_i}{1-g_i}\right)^{\frac{\gamma_i-1}{\gamma_i-\alpha_i}}$.

It is easy to observe in this proposition is that the equilibrium price increases with network size (i.e., K) only if $\frac{\delta(1-\alpha_i)+\beta_i(\gamma_i-1)}{\gamma_i-\alpha_i} > 0$. Hence, if this term is negative for an HP, then it pays less per service level as K increases. Now, given this price structure, each HP selects its service level as presented below.

Proposition 7: *Equilibrium service level for HP* i *is* $K^{\frac{\beta_i-\delta}{\gamma_i-\alpha_i}} \left(\frac{r_i(\alpha_i)^2}{(1-g_i)c_i\gamma_i}\right)^{\frac{1}{\gamma_i-\alpha_i}}$.

Here, the service level s_i increases with K iff $\frac{\delta-\beta_i}{\gamma_i-\alpha_i} < 0$. Therefore, it is apparent that increasing the network size would result in lower service levels to some HPs. Hence, it is not always beneficial for the HIEP to increase the network size, because the additional service fees collected from new HPs that join might not cover the loss in the fees collected from earlier members. In addition, if s_i increases with r_i and g_i, and decreases with c_i iff $\gamma_i > \alpha_i$. Therefore, the cost/gain dynamics pertaining the service level (i.e., γ_i and α_i) have an important effect on the behavior of the service level with respect to other parameters. Proposition 7 also shows that

s_i increases with the value sensitivity to the service level for the HP (i.e., α_i) iff $\alpha_i^2 e^{\frac{2\gamma_i}{\alpha_i}} > e^2 K^{\delta-\beta_i} \frac{c_i\gamma_i(1-g_i)}{r_i}$, i.e., when α_i is large enough. This condition states that an HP will increase its service level with α_i only if the base level of α_i is high. The last result obtained from Proposition 7 is that s_i increases with the IT sophistication parameter of HP i (i.e., γ_i) iff $\gamma_i e^{\frac{\alpha_i}{\gamma_i}} > eK^{\beta_i-\delta} \frac{r_i\alpha_i^2}{c_i(1-g_i)}$. This implies if the IT sophistication of an HP increases, it requests higher service.

Using Proposition 6 and 7, we can now calculate the objective function values of the HPs and the HIEP. These expressions are not provided because of the space limitation. However, we would like to discuss an interesting result. We find that the value for the HIEP contributed by HP i increases in the network size only if $\frac{\delta\alpha_i-\beta_i\gamma_i}{\alpha_i-\gamma_i} > 0$. This implies that, under certain conditions, increasing the network size may reduce the values received from the HPs already in the network. Hence, in these situations, if an additional HP wants to join, the subscription fees the HIEP collects from this HP must cover any losses from the existing HP base in the network.

5. Conclusions

The HIEs facilitate electronic health information exchange among HPs and have several benefits to the participating HPs and to the society. In this paper, we study the decisions made by the HPs and the HIEP in two HIE settings. In the first setting, we derive the conditions in order for the HPs to join a newly established HIE. Other results include the equilibrium network size and subscription fee set by the HIEP. In the second setting, we consider that the HIEP can offer other value added services. In this case, we study the conditions for the HPs to stay in these networks and derive the requested service levels. We also derive the equilibrium pricing of the services. Next, we analyze the effects of increasing the network size on prices, service levels, and the value to the HIEP. One important finding is that it is not always beneficial for the HIEP to increase the network size. We plan to derive more results before the conference and discuss them in our presentation.

References

Agency for Healthcare Research and Quality (AHRQ). 2006. *Evolution of state health information exchange: A study of vision, strategy, and progress*. Contract# HHSP233200500259P, Publication # 06-0057.

Berry, K. and G. Johnson. 2012. Introduction to HIE landscape. *National eHealth Collaborative Seminar Series*.

Covich, J., G. Morris, and M. Bates. 2011. Determining the path to HIE sustainability. *Thomson Reuters & EHealth Initiative*.

Dey, A. S. 2009. Managing information technology integration and implementation in health care supply chains. *PhD Thesis*, University of Minnesota.

Dixon, B., A. Zafar, and J. M. Overhage. 2010. A framework for evaluating the costs, effort, an value of nationwide health information exchange. *Journal of American Medical Informatics Association* 17 295–301.

HIMSS. 2009. Defining health information exchange. *HIMSS HIE Guide Work Group* **November**.

Mostashari, F., M. Tripathi, and M. Kendall. 2009. A tale of two large community electronic health record extension projects. *Health Affairs* 28 **(2)** 345–356.

Private Communications. 2012. *Shannon S. Calhoun*, Executive Director, Southeast Texas Health System.

Walker, J., E. Pan, D. Johnston, J. Adler-Milstein, W. Bates, and B. Middleton. 2005. The value of health care information exchange and interoperability. *Health Affairs* **January** 5–18.

Modeling Practice Efficiencies in Using Healthcare Information Exchanges and Implications for HIE Network Growth

Niam Yaraghi[1], Anna Ye Du[1], Raj Sharman[1], Ram Gopal[2]and Ram Ramesh[1]

[1]Department of Management Science and Systems, State University of New York at Buffalo
Buffalo, New York 14260, {niamyara, yedu, rsharman, rramesh}@buffalo.edu

[2]Department of Operations and Information Management, University of Connecticut,
Storrs,Connecticut 06269, ram@business.uconn.edu

Abstract

Health Information Exchange (HIE) systems may not deliver their promised benefits unless used in a meaningful and carefully adopted way. We develop an analytical model to measure the enhancement in the efficiency of healthcare providers enabled through the use of HIE systems. We test our model based on a panel data set of HIE system logs over a period of 30 months. We show that the relative value of HIE services, as well as competition and labor inputs are highly affecting the use of HIE among different practices. Our proposed method for measuring efficiency has direct implications on studies on HIE adoption, usage and outcomes. Our research findings provide important guidelines on triggers that enhance the overall use of HIE.

1. Introduction

Health Information Exchanges (HIE) enable online sharing of medical records by different healthcare providers. Like other information systems, the high expectations on HIE cannot be realized unless the systems are fully implemented and the assimilation gap between adoption and actual usage is significantly narrowed. Sharing medical records can be done either offline by exchanging documents through carriers such as patients and previous healthcare providers[1], or online through HIE systems. HIE systems tend to operate in a *self-service* mode where the providers actively participate in the sharing process. In contrast, the off-line systems tend to operate in a *full-service* mode where the sharing is enabled through a variety of carriers such as patients, available interoperable healthcare systems and other mechanisms of information exchange The degree to which practices use and benefit from HIE systems is a function of efficiencies enabled through HIE, capital and labor inputs required of both practices and Regional Health Information Exchange Organizations (RHIO) that serve as the main providers of the HIE services.

In this paper, we define and investigate the factors that affect the use of HIE by different healthcare providers. Following the method proposed by Xue et al. (2007), we develop an analytical model that is a variation of the Cobb-Douglas production function which considers HIE usage as a function of not only firm inputs but also customer inputs and characteristics. We show how it can be applied to measure the efficiencies of practices in using HIE system as a latent variable which otherwise cannot be measured directly. Our key findings on the impacts of

[1] In this paper, "Practice" and "Health care provider" are used interchangeably and refer to an independent organization which consists of at least one healthcare professional and provides health care services to patients.

practice efficiency, labor and the relative values of sharing medical records through HIE provide the basis for the design of strategies to enhance effective and relevant use of HIE by the providers and hence, lead to sustainable business models for such platforms.

2. Background and Data

Our research is derived from the adoption characteristics of the participating members of HEALTHeLINK, a regional health information exchange organization in western New York. Currently, HEALTHeLINK has 2053 healthcare professionals within 433 practices subscribed to their services. These members can access three different types of medical records - lab reports, radiology reports and hospital transcriptions from 21 sending facilities including different hospitals, laboratories and radiology centers. The members can access these records either through a fully automatic web interface[2] that links the HIE to interoperable Electronic Medical Record (EMR) systems or through a web portal called Virtual Health Records (VHR) that requires them to manually search for specific records and download them into their own local systems. The HIE database studied in this research consists of 503,300 observations from system logs of HIE access through the time period of March 2009-August 2011. The logs identify the users who have accessed specific records, the record types, the access dates and the patients of medical records that were accessed[3]. We reconstructed this data set into a panel data structure which shows the use of each of the three different types of HIE services, for each practice per month.

3. HIE-Enabled Practice Efficiency Model

We propose that variations in the use of HIE across different access channels depends on a series of practice-related factors including cost of labor, practice capital and efficiency, and relative practice value of the service. In addition, there are also RHIO factors that can affect practice channel choice. Some RHIO choices have effects that can vary by practice, such as the location of medical record providers relative to each practice location, while others affect all customers equally (e.g., the design of the system interfaces). Finally, significant differences in channel use could arise due to enabled practice efficiencies, some of which could be due to observable factors and the rest unobservable, either directly or indirectly. Our principal empirical task is to construct suitable proxies for each so that we can: (a) isolate practice efficiency from other factors that affect channel choice and (b) provide support to the claim that our definition of practice efficiency is measuring what we expect by demonstrating that it is correlated with factors that we believe to be associated with efficiency in a plausible way.

In this analysis, we denote each of the three channels of access (iHUB, VHR, offline) by C_c in which $c \in \{1,2,3\}$. Each of the three services (lab reports, radiology reports, and hospital transcriptions) are denoted by $J \in \{1,2,3\}$. The cost of labor for a practice is w (e.g., time opportunity cost per unit of input labor ($\mathrm{L_{cj}}$) and the value of the service is v per unit of output($\mathrm{O_{cj}}$). Service value is independent of the channel through which the service is acquired but differs by service (e.g., lab reports may be more important to a set of specific practices, as

[2] This channel is called iHUB

[3] The name and other information of the patients are de-identified and replaced with a unique code in our data set.

compared with hospital transcripts). Thus, the utility to a practice in using channel c to receive $O_{cj}$of type j medical records is

$$(1) U_{cj} = v_j O_{cj} - wL_{cj}$$

Let the production inputs for channel and service type be indicated as practice-invested capital(R), practice labor(L), RHIO-invested capital(K), and RHIO employee labor(H), respectively. Assuming that the effects of practice inputs and RHIO inputs are in multiplicative form, this yields an overall production function for service of type j in channel c (or output O_{cj}) of the form

$$(2) O_{cj} = g_c(R_c, L_{cj}) f_c(K_c, H_c)$$

We now model the practice portion of the production function using a Cobb-Douglas format as follows:

$$(3) g_c(R_c, L_c) = R_c^{\alpha_c} (A_c L_{cj})^{\beta_c}$$

where α_c and $\beta_c$are the output elasticity of practice capital and practice labor, respectively, and $A_c$is a practice-specific factor that affects the practice's productivity of labor when using channel c. Note that the firm portion of the production function $f_c(K_c, H_c)$ is likely to be slow changing and does not vary across practices. Therefore, from the perspective of a practice, this term is quasi-fixed. A practice chooses an effort level and a capital level for each channel that maximizes its overall utility. Thus, the optimum level of $L_{cj}$is computed as follows:

$$(4) L_{cj}^{*} = \left(L_{cj} \middle| \frac{\partial O_{cj}}{\partial L_{cj}} = \frac{w}{v_j} \right)$$

Now, differentiating O_{cj} with respect to L_{cj} and substituting the L_{cj}^{*}in the derivative, the optimal labor choice for channel c is obtained. Substituting this optimal labor choice back into the original production function yields O_{cj}. Since $O_c^{*} = \sum_j O_{cj}$, we can write O_c^{*} (as the optimal level of HIE usage through channel c) in the logarithmic format as

$$(5) \log O_c^{*} = \frac{\alpha_c}{1-\beta_c} \log R_c + \frac{\beta_c}{1-\beta_c} \log A_c + \frac{1}{1-\beta_c} \log f_c(K_c, H_c) + \frac{1}{\beta_c - 1} \log w + \frac{1}{1-\beta_c} \log \beta_c + \log \sum v_j^{1/1-\beta_c}$$

If we run a regression of the number of medical records received by each practice on measures of practice capital R_c, practice effort costs w and medical record value v_j then we could retrieve the efficiency measures as the residual of that equation.

$$(6) \log O_c^{*} = \frac{\alpha_c}{1-\beta_c} \log R_c + \frac{1}{1-\beta_c} \log f_c(K_c, H_c) + \frac{1}{\beta_c - 1} \log w + \frac{1}{1-\beta_c} \log \beta_c + \log \sum v_j^{1/1-\beta_c} + \varepsilon$$

In the regression equation above, $\varepsilon = \frac{\beta_c}{1-\beta_c} \log A_c$ and thus $A_c = e^{\frac{1-\beta_c}{\beta_c}}$

The regression equation is derived for each channel separately. Further, greater precision may be obtained if we aggregate the residuals from all of the channels. The unobservable practice characteristics may affect its choice of channel and thus the observed residual may include an error term s which can be practice-specific or fixed or random effect and $\acute{\varepsilon} = \frac{\beta_c}{1-\beta_c} \log A_c + s$. To address these issues, we construct a customer efficiency measure as the difference between full service (offline) channels C'' and self-service (online; HIE) channels C'.

$$(7) \qquad CE = \theta_{C'}\varepsilon_{C'} - \theta_{C''}\varepsilon_{C''}$$

The optimal weights are theoretically related to the variance of each residual and are proportional to the marginal product of labor in each type of channel (through $\frac{\beta_c}{1-\beta_c}$). Since we only have access to HIE data use through VHR and iHUB channels, to empirically test our model, we consider VHR channel use as self-service channel and iHUB as full service channel. Table 1 shows how we have operationalized different covariates of equation (5) in a panel data structure and states our testable hypotheses accordingly.

Table 1. Definition of parameters in equation (5) and research hypotheses

Variable	Definition	Hypotheses	References
		Correlates of practice efficiency	
Tenure Tenure^2	The number of months since the practice has adopted the HIE system. Tenure^2 is the squared term of tenure	Training and familiarity with the system can affect customer efficiency. ***H1: Practices who have adopted HIE sooner, use HIE more than others.***	(Jasperson et al. 2005; Jha et al. 2009)
		Correlates of cost of practice effort	
Nurse Primary Care	The number of nurses and physician assistants in a practice Primary care is the number of permanents primary care physicians affiliated with a practice	Time saving benefits of electronic health records is realized by nurses rather than physicians, moreover, nurses are the main end users of the system as the ones who are responsible for preparing patient's documents before a physician's visit. ***H2: Practices with more number of nurses use HIE more than others***	(Poissant et al. 2005)
		Correlates of service value	
Lab Radiology Transcription Degree centrality Market Share	Lab, Radiology and Transcription are the log value of total number of accessing each of these records in the prior month by each practice Degree centrality is the number of other practices that have common patients with each practice Market share is the log value of ratio the total population of the city that practice is located in to the total number of practitioners in that city	We argue that the number of access to each of the three types of records in prior month is a proxy for the actual relative value of that type of records for practice. The practices that have more common patients with other absorb a higher value from accessing medical documents. Practices with a larger market share will benefit more from HIE system by saving more overall time/cost they are also more open in using HIE due to lower competition with other smaller practices ***H3: Practices with larger market share use HIE more than others.*** ***H4: Value of HIE services directly affects the use of HIE.***	(Grossman et al. 2006; Xue et al. 2007)
		Correlates of channel availability	
Rural Senders	Rural is a binary variable indicating whether the practice is located in a rural area Seders is the number of sending facility that each practice receives	Availability of alternative channels such as proximity to physical location of data providers negatively affects the use of new information systems. The practices located in rural areas are farther away from main	(Boyer et al. 2002)

medical records from.	sources of data providers (labs, radiology centers, hospitals). Practices will benefit more if they receive data from more facilities. ***H5: Practices located in rural areas use HIE more than others.*** ***H6: Practices who can access medical records from more sending facilities use HIE more than others***

4. Results

Table 2 shows the results of regressing the log value of HIE access times on correlates of practice efficiency, cost of practice effort, and service value and alternative channels. Neither tenure nor its squared term is a significant factor in using HIE systems. Although there is plenty of support in literature about the significance of training and experience in using information systems, HIE seems to be fairly user-friendly and mastering this technology does not require substantial training or experience and thus tenure would not affect the level of use. The number of nurses is positively affecting the use of HIE. As expected, this implies that practices with higher number of nurses use HIE services more frequently as compared with the practices with fewer nurses. This is mainly due to the fact that nurses are the main end users of HIE systems. They are the main personnel in each healthcare facility that are responsible for preparing medical records of patients before they are seen by the physicians. Our results confirm previous findings, and show that while the number of nurses is significant in increasing the use of HIE, the number of physicians will not affect the level of HIE use in a practice. The value of HIE services as captured by the log values of total count of access to each type of services in prior month by each practice, is highly significant. Moreover, the results confirm that the number of patients shared among practices is affecting the use of HIE. The practices with high levels of degree centrality tend to use HIE systems more than the ones that are isolated or do not significantly share patients with others. Competition as captured by market share is playing an important role in the HIE usage. Practices with higher market shares are the ones that face lower competition and thus have a higher tendency of using HIE services. The effects of competition can also be explained with the locations of practices. The practices in dense urban areas face higher competition and hence, tend to use HIE less than the ones in rural areas. Finally, the number of sending facilities from which practices receive medical records is positively affecting the use of HIE. In order to measure the efficiencies of practices as discussed in Section 3, we ran the model in equation (6) on iHUB and VHR use separately. The difference between the residuals represents the efficiency of practices as shown in equation (7). Table 3 shows the regression results of measured efficiencies on the variables that we posited as affecting efficiencies. We have used mean centered tenure in this analysis since we suspected the possibility of a multi-collinearity problem between tenure and tenure^2. Tenure and tenure^2 significantly affect the practice efficiencies in our analysis. The interesting observation in this analysis is the negative estimate of tenure^2which shows the diminishing effect of tenure on efficiencies.

Table 2.Parameter estimates for equation (5)

Variable	estimate	Std. Err.	t-value	p-value
Tenure	-0.0156	0.0175	-0.89	0.3736
Tenure2	-0.00031	0.000372	-0.83	0.4078
Nurse	0.075096	0.0365	2.06	0.0399
Primary	0.008366	0.0213	0.39	0.694
Lab value	0.293147	0.027	10.87	<.0001
Radio. Value	0.37437	0.0416	9	<.0001
Trans. Value	0.372104	0.0471	7.89	<.0001
Degree centrality	0.039091	0.00916	4.27	<.0001
Market share	0.082922	0.0278	2.99	0.0029
Rural	0.411231	0.1896	2.17	0.0302
Senders	0.042833	0.0188	2.28	0.0225

Table 3.Parameter estimates of regressing practice efficiency on its covariates

Variable	estimate	Std. Err.	t-value	p-value	VIF
Intercept	-0.04797	0.08868	-0.54	0.5887	0
tenure	0.08092	0.00904	8.95	<.0001	1.15465
tenure2	-0.00128	0.000536	-2.38	0.0173	1.15465

5. Discussion

Unless HIE systems are used properly, their benefits will not be fully realized. Our analysis reveals of set of variables that will affect the use of HIE by practices. Of these variables, practices efficiency is a latent variable which cannot be measured directly. We develop an analytical model based on the Cobb-Douglas production function which encompasses effects of both practice and RHIO inputs on HIE use and thus enables us to measure the latent efficiency of practices in using HIE systems. Our future research will attempt to investigate the effects of practice efficiencies on HIE adoption.

References

Boyer, K. K., Hallowell, R., and Roth, A. V. 2002. "E-services: operating strategy–a case study and a method for analyzing operational benefits," *Journal of Operations Management* (20:2), pp. 175–188.

Grossman, J. M., Bodenheimer, T. S., and McKenzie, K. 2006. "Hospital-Physician Portals: The Role Of Competition In Driving Clinical Data Exchange," *Health Affairs* (25:6), pp. 1629–1636.

Jasperson, J. S., Carter, P. E., and Zmud, R. W. 2005. "A comprehensive conceptualization of post-adoptive behaviors associated with information technology enabled work systems," *Mis Quarterly* (29:3), pp. 525–557.

Jha, A. K., DesRoches, C. M., Campbell, E. G., Donelan, K., Rao, S. R., Ferris, T. G., Shields, A., Rosenbaum, S., and Blumenthal, D. 2009. "Use of electronic health records in US hospitals," *New England Journal of Medicine* (360:16), pp. 1628–1638.

Poissant, L., Pereira, J., Tamblyn, R., and Kawasumi, Y. 2005. "The impact of electronic health records on time efficiency of physicians and nurses: a systematic review," *Journal of the American Medical Informatics Association* (12:5), pp. 505–516.

Xue, M., Hitt, L. M., and Harker, P. T. 2007."Customer efficiency, channel usage, and firm performance in retail banking," *Manufacturing & Service Operations Management* (9:4), pp. 535–558.

Extracting and Releasing Information from Clinical Documents with Guaranteed Privacy

Xiao-Bai Li and Jialun Qin
Department of Operations & Information Systems, University of Massachusetts Lowell
xiaobai_li@uml.edu; jialun_qin@uml.edu

Abstract: This study concerns privacy-preserving extraction and release of information from clinical text documents. Existing studies on privacy-preserving data mining/publishing focus mostly on structured data. We propose a novel approach to privately extract and release patients' demographic, health and medical data from clinical documents. The extracted data is represented in a semi-structured, set-valued data format, which can be stored in a health information system for query and association analysis. The privacy preserving mechanism is based on the cutting-edge idea of differential privacy, which offers rigorous privacy guarantee.

1. Introduction

Medical documents and other unstructured data, such as clinical narratives and discharge summaries, are essential for documenting interactions between patients and healthcare providers. They are typically embedded in an electronic health records (EHR) system. These clinical and medical texts also contain rich information useful for improving clinical decision support and for medical and healthcare research. Traditionally, extraction of information from clinical text into a form suitable for analysis and research is done manually by domain experts. In recent years, there have been significant developments in using natural language processing for information extraction from clinical documents (Murphy et al. 2010; Savova et al. 2010).

In order to make patient data available for research and analysis, it is vital to ensure that patient privacy is appropriately protected. To this end, the Health Insurance Portability and Accountability Act (HIPAA) has established the "Safe Harbor" privacy rule, which specifies 18 categories of identifying attributes – called Protected Health Information (PHI) – that cannot be released to a third party. The vast majority of existing studies on data privacy focus on structured data (Aggarwal and Yu 2008). For unstructured data like clinical documents, essentially all the current privacy approaches attempt to implement the HIPAA rule directly, by following a de-identification framework that focuses on accurately identifying PHI fields from the text (Uzuner et al. 2007; Meystre et al. 2010). However, a strict implementation of the HIPAA rule may be inadequate for protecting privacy or preserving data utility. Studies have shown that the HIPAA rule lacks the flexibility to adequately meet the diverse needs of data users; it can be under-protective in some cases and over-protective in others (Meystre et al. 2010; Uzuner et al. 2007). Recognizing this limitation, HIPAA also provides guidelines that enable a scientific assessment of privacy disclosure risk in order to determine if the data is appropriate for release. This study focuses on this aspect of the HIPAA principle.

The scientific approaches currently used for privacy protection in the healthcare domain, such as *k*-anonymity (Sweeney 2002) and noise-based perturbation (Aggarwal and Yu 2008), typically depend on some assumptions about the privacy intruder's auxiliary information regarding individual subjects. When the assumptions are violated, these approaches may not work well (Dwork 2011). To overcome this limitation, Dwork (2006, 2011) introduces the notion of differential privacy. Intuitively, differential privacy ensures that the released information about a dataset is essentially the same whether or not an individual's data was included in the dataset. In other words, there is virtually no additional privacy disclosure risk if the individual opts in to the dataset. Differential privacy is defined independent of any auxiliary information assumption.

Thus, it provides the most rigorous privacy guarantee among existing approaches. On the other hand, differential privacy requirements often result in significant information loss in the released data, which limits its applicability. A recent survey found that application of differential privacy to the healthcare domain remains an unexplored research area (Dankar and Emam 2012).

In this paper, we propose a novel approach to extract and release patients' demographic, health and medical data from clinical text. The extracted data is represented in a semi-structured, set-valued data format, which is then used for privacy-preserving query and analysis. Our privacy-protection mechanism is designed based on the differential privacy framework.

2. Privacy and Information Quality in Healthcare Data

From a privacy standpoint, information contained in clinical text can be classified into three categories: (1) Explicit identifiers (EID), which are PHI attributes that can be used to directly identify an individual, such as name, social security number, and phone number. HIPAA requires that EIDs be removed or encrypted in the released data. (2) Quasi-identifiers (QID), which are not explicit identifier but can be used to identify individuals by matching their values from different data sources (Sweeney 2002). QIDs include some PHI attributes such as date of birth, admission/discharge date, and zip code; they also include some non-PHI attributes such as age, gender, race and marital status. (3) Health and medical information (HMI), such as symptoms, test results, diagnosis, disease, medications and procedures. HIPAA does not restrict the release of patients' HMI as long as the related PHI fields are protected.

Given a set of patient clinical documents, our task is to extract the above three categories of information and store them in a semi-structured scheme for query and analysis. Following HIPAA, EIDs will be removed or encrypted in the scheme. The QID data will be loaded into a standard table. The HMI data will be stored in a set-valued format. Such a scheme is supported by many health information systems that enable the use of EHR data for decision support and healthcare research (e.g., the i2b2 system described in Murphy et al. 2010).

Original Text	Visit Month	Visit Year	Age	Gender	Zip Code	HMI
1. Visited on 4/5/2009. Male, 24 year old. Feeling sore throat, fever, headache, fatigue…	April	2009	24	Male	12301	sore throat, fever, headache, fatigue
2. Mr. Brown's daughter is 9 year old. Visited on 4/13/2009…Having runny nose, sore throat, fever, headache…	April	2009	9	Female	12301	runny nose, sore throat, fever, headache
3. Admitted on 4-21-2009, patient is a 9 year old female. Having runny nose, sore throat, diarrhea, fever.	April	2009	9	Female	12301	runny nose, sore throat, diarrhea, fever
4. Amy is 17 year old. Having fever, joint pain, nausea, sore throat…Visited 5/14/2009.	May	2009	17	Female	12302	fever, joint pain, nausea, sore throat
5. Admitted on 6/7/2009, the 88 year old man is complaining chills, body pain, sore throat, fatigue, fever…	June	2009	88	Male	12302	chills, body pain, sore throat, fatigue, fever

Figure 1. An Illustrative Example

To describe the idea, consider a set of five patient records shown in Figure 1, taken from a hypothetical community hospital. The left panel shows the simplified original clinical notes. The right panel illustrates extracted information and its representation. The first five columns follow a relational database table format (where the additional Zip Code data is obtained from the patient registration). The last column contains a set of HMI terms/values. To comply with

HIPAA privacy rule, the hospital can only release the data in Visit Year, Age, Gender and the first three digit of Zip Code, as well as the HMI data. However, this HIPAA-compliant release can be over-protective. For example, because only the visit year can be released, the important "season" information is lost, which could be crucial for detecting an epidemic disease outbreak. For the same reason, releasing the 3-digit zip code (e.g., 123**), instead of the 5-digit zip code, also causes significant information loss. On the other hand, the HIPAA-compliant release may be inadequate for privacy protection. For example, it may not be difficult to identify the 88-year-old man (record 5) in the region who has been hospitalized in 2009, using publically available data.

The proposed mechanism releases information as an output response to a search query using HMI terms. We focus on count query only in this preliminary study. Even with this restriction, the output based on our approach can provide much more useful information. For example, for search terms {sore throat, fever}, the output can show a perturbed count for **<**VisitMonth=April, VisitYear=2009, Age=9, Gender=Female, ZipCode=12301, HMI={sore throat, fever}**>** (without loss of clarity, hereafter written as **<**April, 2009, 9, F, 12301, {sore throat, fever}**>**). Since there are 2 matching records in the example set, the perturbed count will be 2 plus a noise (to be discussed in the next section). Using different predicates, we can also get a perturbed count for **<**April, 2009, 12301, {sore throat, fever}**>**, which will be 3+noise. Besides, we can also get a perturbed count for **<**April~June, 2009, 1230*, {sore throat, fever}**>**, which will be 5+noise. These outputs provide useful information about a flu-like disease that may be spreading in the area during the period (assuming many similar records are found in the entire patient database). Note that it is also possible for the proposed mechanism to output the perturbed count for the records with QID values matching those of record 5 (even if the match is unique), but the noise for the count is likely to be very large compared to the original count.

3. Differentially Private Data Release

Our proposed approach is based on the notion of differential privacy (Dwork 2006, 2011), defined bellow:

Definition 1. *Given any two datasets D_1 and D_2 that differ in only one record, a perturbation mechanism M provides ε-**differential privacy** if for any set of possible outputs S of M (i.e., $S \subseteq Range(M)$),*

$$\Pr[M(D_1) \in S] \le e^{\varepsilon} \times \Pr[M(D_2) \in S]. \tag{1}$$

The parameter ε represents disclosure risk, which is usually controlled to be small so that e^{ε} is close to one. As such, differential privacy guarantees, in a probabilistic sense, that the outputs will be essentially the same with or without any specific individual's participation. This property has a very appealing implication. For example, if the dataset were to be used by a healthcare provider to analyze the demographics of its patient population, then the presence or absence of a patient's record in the dataset will not significantly change the results of the analysis. In this sense, the participating patient's demographic information is well hidden.

For frequency queries (e.g., count and histogram), there is a straightforward way to construct a perturbation mechanism that satisfies ε-differential privacy. The mechanism is based on the notion of sensitivity defined below (Dwork 2006):

Definition 2. *For a function f over dataset D with numeric output, the **sensitivity** of f is*

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|_1 \tag{2}$$

for all D_1, D_2 differing in at most one record.

In other words, the sensitivity is the maximum change in the value of f when any single record of D changes. It has been shown (e.g., in Dwork 2011) that *for a numeric function f, a perturbation mechanism that adds noise with a Laplace*($\Delta f/\varepsilon$) *distribution to the output satisfies* ε*-differential privacy* (the *Laplace*(σ) distribution has a density function of $p(x \mid \sigma) = (1/2\sigma)e^{-|x|/\sigma}$). To be rigorous, for an integer-valued output, a geometric distribution should be used for perturbation (Ghosh et al. 2009); but this subtle difference is not considered important in the literature.) When f represents a count query, sensitivity $\Delta f = 1$ since the count can differ at most by one due to the addition or removal of one record. Therefore, for a count query f, the perturbation mechanism

$$M(D) = f(D) + Laplace(1/\varepsilon) \tag{3}$$

provides ε-differential privacy.

Given a set of clinical documents, our approach first extracts EID-, QID-, and HMI-related terms. There is no existing information extraction system that can effectively extract all these terms. We have adopted two open-source systems to perform these functions: the Stat De-id system (Uzuner et al. 2008) for extracting EID and QID terms and the cTAKES system (Savova et al. 2010) for extracting HMI terms. After extraction, EIDs are removed. The QID and HMI data are populated into a scheme exemplified in Figure 1. Each row in the scheme can be viewed as a transaction in the context of association rule mining, and each value or term can be viewed as an item. Therefore, the Apriori algorithm can be applied to find frequent itemsets. The minimum support count for the frequent itemsets (i.e., the number of transactions containing the itemsets) can be considered as a privacy parameter (controlled by the data owner but unknown to the data user). It can be set to a relatively small value because the count will be perturbed before releasing.

1. For a set of clinical documents, extract EID, QID and HMI terms and values.
2. Remove EID values. Load QID and HMI values into a table *D* where the HMI field allows a set of multiple terms or values.
3. Run the Apriori algorithm on *D* to find all frequent itemsets that contain at least an HMI value.
4. For a count query *f*(*D*) involving a set of HMI value, obtain the count result from the output of Step 3. Perturb the result using Equation (3).

Figure 2. Computational Procedure

The entire computational procedure for our proposed approach is summarized in Figure 2. Steps 1, 2 and 3 can be preprocessed, so the real time computation for a query is very fast. Also, the Step 3 computation is faster than that of the classical Apriori algorithm because the itemset not containing any HMI term can be removed immediately at each Apriori iteration. So, the computation is efficient even if it is necessary to re-run Step 3 (e.g., due to a change in support).

4. Preliminary Experiment

The Informatics for Integrated Biology and the Bedside (i2b2) project has obtained multiple sets of clinical documents from healthcare organizations and made them available for research (https://www.i2b2.org/NLP/DataSets). We used four of the datasets for the experiment, all of which are medical discharge summaries. The first set is related to a clinical-text de-identification challenge competition. The second set was initially used for evaluating document classification techniques. The third set was used for extraction of medication information from clinical text. The fourth set was used for an organized challenge to extract medical concepts, assertions and relations. Because all of the datasets are medical discharge summaries, the elements of information contained in different datasets are similar, most including patient name, admission and discharge date, age, gender, hospital, symptoms, test result, diagnosis, disease, medications, and so on. Therefore, we merged the four sets into a single set, resulting in 2,867 text records. After extracting the QID and HMI values from the text, we found that there were very few zip code and/or location values in the data that we could use for the experiment. Therefore, we focused on the query results involving the visit year and month data, which appear in nearly all records (the HIPAA "Safe Harbor" rule prohibits releasing patient visit month data).

The privacy protection level is naturally measured by the parameter ε. Clearly, the smaller the ε value, the better the privacy protection the mechanism offers. In terms of data utility, since the count query result can be regarded as an itemset count, we use an itemset-related measure, called *relative error*, defined below (Rizvi and Haritsa 2002, Evfimievski et al. 2004):

$$\text{Relative Error} = \frac{1}{|I|}\sum_{i\in I}\frac{|\tilde{n}_i - n_i|}{n_i}, \tag{4}$$

where I represents the set of all frequent itemsets with support count larger than the specified threshold value; n_i and $\tilde{n}_i$ are respectively the original and perturbed count of the ith frequent itemset. Since an itemset must contain at least an HMI item, the relative error also measures the error rate for the queries having the minimum count when at least an HMI term is involved.

We set parameter ε to five different values: 0.1, 0.2, 0.3, 0.4, and 0.5, which are in general more conservative (i.e., with stronger privacy protections) than commonly used ε values in differential privacy research. To evaluate the performance at different frequency levels, we set minimum support count to five values: 10, 20, 30, 40, and 50. The results of the perturbation algorithm vary slightly with different random number seeds. Therefore, for each scenario the algorithm was run five times, each run using a different seed. The average results are reported.

Privacy Parameter	Support Count				
	10	20	30	40	50
$\varepsilon = 0.5$	0.1489	0.0779	0.0560	0.0423	0.0342
$\varepsilon = 0.4$	0.1836	0.0959	0.0698	0.0545	0.0459
$\varepsilon = 0.3$	0.2462	0.1318	0.0950	0.0743	0.0600
$\varepsilon = 0.2$	0.3498	0.1867	0.1310	0.1044	0.0857
$\varepsilon = 0.1$	0.7312	0.3944	0.2877	0.2233	0.1772
Month-Estimated	0.5097	0.4332	0.4364	0.4610	0.4364

Figure 3. Results of Relative Error

The results of the experiment are shown in Figure 3. It is observed that the error rate decreases as the privacy risk (ε) increases, which is expected. Furthermore, the error rate decreases as the support increases. This also makes sense because the frequency counts for the selected itemsets (the denominator in Equation 4) become larger when the support is increased. The added Laplace noise, however, is independent of the support. When ε is small, its value is approximately the odds that the output results will be different due to the addition or removal of any record (e.g., when $\varepsilon = 0.1$, the odds is about $e^{0.1} - 1 = 0.105$). Note that the results are based on the queries that allow releasing visit month data, which is prohibited in HIPAA. Therefore, the proposed approach provides an additional option to HIPAA for data release, based on well-grounded assessment of disclosure risk. If the data include other HIPAA-restricted QIDs such as zip code, location, and date of birth, similar analyses can be performed based on our approach.

Because the problem we study is new to the literature, there are no existing techniques that can be compared directly. We have assumed a scenario where the output release is HIPAA-compliant (i.e., without month), but the data user attempts to estimate the month value for the query output with a probability proportional to the marginal distribution of the month values. The month values in the dataset are distributed unevenly, ranging from 4% for the least frequent to 13% for the most frequent. The error results under this scenario are shown on the last row of Figure 3 (labeled 'Month-Estimated'). It is observed that the resulting error rates are in general much higher than those from our approach, particularly when the count becomes large. Therefore, if the month information is important, it is worthwhile to consider using the proposed approach.

5. Next Steps

We should note that the proposed perturbation mechanism only applies to query output, not to the original data. It is well-known that output perturbation is vulnerable to the same repeated query because in this case the independently added noises will eventually be averaged out, revealing the true value. Various methods have been proposed to address this problem (Dwork 2011), but they all cause considerable deterioration in output quality. For our problem which deals with unstructured data, however, it is possible to add noise in between the text input and the query output. For example, the noise addition may be performed in the information extraction stage. We will probe viable approaches along this direction. We also plan to work on releasing the extracted data directly in a set-valued format using differential privacy. We have tested some ideas, but they require fairly large amount of noise. It appears that a relaxed notion of differential privacy may be necessary for this task.

References

Aggarwal, C.C., and Yu, P.S. (eds.) 2008. *Privacy-Preserving Data Mining: Models and Algorithms*, New York: Springer.

Dankar, F.K., and Emam, K.E. 2012. "The Application of Differential Privacy to Health Data," In *Proceedings of 5th International Workshop on Privacy and Anonymity in the Information Society*.

Dwork, C. 2006. "Differential Privacy," In *Proceedings of 33rd International Colloquium on Automata, Languages and Programming, Part II* (ICALP 2006), pp. 1–12.

Dwork, C. 2011. "A Firm Foundation for Private Data Analysis," *Communications of the ACM* (54:1), pp. 86-95.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. 2004. "Privacy Preserving Mining of Association Rules," *Information Systems* (29:4), pp. 343-364.

Ghosh, A., Roughgarden, T., and Sundararajan, M. 2009. "Universally Utility-Maximizing Privacy Mechanisms," In *Proceedings of 41st Annual Symposium on Theory of Computing* (STOC 2009), pp. 351-359.

Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., and Samore, M.H. 2010. "Automatic De-identification of Textual Documents in the Electronic Health Record: A Review of Recent Research," *BMC Medical Research Methodology* 10, Article 70, 16 pages.

Murphy, S.N., Weber, G., Mendis, M., Chueh, H.C., Churchill, S., Glaser, J.P., Kohane, I.S. 2010. Serving the Enterprise and beyond with Informatics for Integrating Biology and the Bedside (i2b2)," *Journal of American Medical Informatics Association* (17:2), pp. 124-130.

Rizvi, S.J., and Haritsa, J.R. 2002. "Maintaining Data Privacy in Association Rule Mining," In *Proceedings of the 28th Very Large Data Base Conference*, Hong Kong, China, pp. 682-693.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., and Chute, C.G. 2010. "Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications," *Journal of the American Medical Informatics Association* (17:5) pp. 507–513.

Sweeney, L. 2002. "*k*-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* (10:5), pp. 557-570.

Uzuner, O., Luo, Y., and Szolovits, P. 2007. "Evaluating the State-of-the-Art in Automatic De-identification," *Journal of American Medical Informatics Association* (14:5), pp. 550-563.

Uzuner, O., Sibanda, T., Luo, Y., and Szolovits, P. 2008. "A De-identifier for Medical Discharge Summaries," *Artificial Intelligence in Medicine* (42:1), pp. 13-35.

Affiliate or Not?
A New Price Discrimination Strategy on the Cashback Platform

Yi-Chun Ho[1], Yi-Jen Ho[2]
[1]Foster School of Business, University of Washington
[2]Merage School of Business, University of California, Irvine

Abstract
This paper examines the impact of "cashback" mechanism on merchants' marketing strategy. Through reimbursing a portion of the transacted amount to consumers in a form of cashback, merchants are able to exercise second-degree price discrimination on unique cashback platforms. We develop an analytical framework which endogenizes consumers' preference over brands into utility formation. We derive the market equilibrium and find the optimal profit-sharing scheme under different market configurations. Our finding provides guidance to practitioners on optimal consumer segmentation strategy and cashback percentages. Managerial implications of brand valuation are also discussed.

Keywords: cash back, segmentation, price discrimination, pricing, duopoly

1. Introduction

Companies have for decades built up their business around the traditional brick-and-mortar channel. The rise of the Internet and the surging popularity of online shopping have embodied the emerging click-and-mortar business model. As the focus of market moves away from brick-and-mortar to click-and-mortar, a new *cashback* concept - which allows customers to be rewarded for shopping through certain affiliated websites - has made such sites as Ebates and Mr. Rebates extremely successful.

Ebates, the leading cashback site[1] in the U.S. with 10 million registered users, in 2011 brokered 900 million dollars in merchandise sales for its 1,200 affiliated merchants. Its revenue growth has trended 50 percent higher for the second year in a row since 2010 (San Francisco Business Times 2011). Interestingly, cashback websites are not the only one trying to exploit this new marketing strategy. Software giant Microsoft in 2008 implemented the cashback feature that allows its search engine Live Search to act as a cashback platform. One year later, Google also introduced a way to use its Google Checkout as a platform on rewarding customers.

A cashback site is a two-sided market platform consisting subsidizing and subsidized segments (Rochet and Tirole, 2003). On one side, it collects a commission from the merchant when a consumer makes purchase through the merchant's referral link on the platform. On the other side, it entices consumers into shopping by giving them a portion of the commission collected from merchants as cashback (Wall Street Journal 2005). Unlike mail-in-rebates which is associated with high redemption cost and uncertainty (Lu and Moorthy 2007), cashback can be easily earned by few finger clicks on the webpages.

In a survey by TopCashBack in UK, 32% of shoppers begin their shopping on a cashback site and 17% of them change their purchasing decision and source based on the cashback rate offered. Through affiliation with the cashback platform, merchants can price-discriminate consumers based on their purchasing habit: They can sell a product at the regular price on their own websites, while charge a discounted price on the cashback site by reimbursing cashback to site members. The site is also an advertising solution where merchants, especially no-name

[1] Throughout this paper, we use cashback site, platforms or intermediaries interchangeably.

brands, can effectively promote their brand awareness when accessing the site's substantial user base. The site increases a merchant's visibility through email feeds, on-site advertisements, and merchant search ranks.

While the cashback mechanism has demonstrated significant influences on consumers' purchasing behaviors, less clear is the impact of cashback market on merchants and on competition among them. In particular, this study seeks answers to the following questions:

- Under what conditions should merchants join the cashback site?
- When merchants join, how should they share the profit with the cashback site?
- Given the commission rates set by merchants, how should the platform determine cashback rates to give away?

To answer these questions, we develop a game theoretical model to characterize a unique market where merchants can exercise price discrimination in the presence of cashback mechanism. We take into consideration consumers' heterogeneity in brand preference, price sensitivity, and redemption cost. Our result provides guidance for the merchants and the platform on how to optimally choose commission and cashback rates, respectively. Managerial implications of brand valuation are also discussed.

The rest of this paper is organized as follows. In section 2, we review related literature and position our paper. Model setup is detailed in section 3. Findings on merchants-platform alliance strategy and optimal profit-sharing scheme are provided in section 4. Section 5 gives concluding remarks and future research directions.

2. Literature Review

Existing literature in price discrimination provides the theoretical fundamental for this paper. In the traditional brick-and-mortar channel, price plays a significant role in the competition among merchants who sell homogeneous or substitute products (Moorthy 1988). The significance of click-and-mortar channel and emergence of cashback sites nowadays intensify the competition by opening the second battlefield, a new market where consumers' willingness to pay is reshaped through the cashback rewards. Given this nature, the cashback mechanism can be viewed as an example of the second-degree price discrimination which occurs when consumers are charged different prices (Pigou 1920). The whole consumer base by market segmentation can be divided into two groups: cashback (informed) consumers and loyal (uninformed) consumers (Baye & Mogan 2001, Varian 1980).

Since 1936, Robinson–Patman Act has forbidden direct price discrimination. In practice, cashback websites enable merchant to implement the second-degree price discrimination where each consumer is self-selected to join cashback sites. Holmes (1989) examines the output and profit effects of price discrimination in a symmetric duopoly model. While Holmes focuses on product differentiation, we analyze the effect of intermediary on the price decimation.

Due to cashback mechanism, the cashback platform serves as another channel reintroducing not only merchants' price competition but the second-time price-dispersion. The variation in the cost structures brings the first-time price-dispersion in the traditional channels. Borenstein (2001) uses US airline industry as an example. Complementary to the empirical evidence (Clemons et al., 2002), our model probes into this issue and explains merchants' pricing decision to induce price-dispersion on online intermediary.

The other stream of related literature is about price sensitivity. Price sensitivity is the weight attached to price in a consumer valuation of a product's overall utility. Erden et al. (2002) found that price sensitivity is moderated by the brand credibility and the magnitude of the moderating effect varies across product categories. Empirical studies (Kalwani & Yim 1992, Gupta & Cooper 1992) show that unless the amount of price drop is higher than a certain threshold, consumers in general do not experience any positive transaction utility. However, since the cashback site's users are a group of consumers who highly appreciate savings (Swan 2010), we assume the threshold for utility gain from cashback to be minimal.

This paper contributes to the existing literature in the following two aspects. First, while there is prior literature investigating the role of intermediaries in the market (Baye 2001, Nahm 2003), none of them are studied from a profit-sharing perspective. Second, to capture consumers' heterogeneity, we model consumers' utility based on their brand preference and price sensitivity. This paper aims to examine the significance of the still-nascent cashback market and sheds light on its economic impact on merchants' consumer segmentation strategy.

3. The Model

We first consider a three-stage game theoretic model with three players: merchants, the cashback site, and consumers. The sequence of the game can be summarized as follows: Merchants move first by setting commission rates paid to the cashback site for the transactions made through it. Then, the site determines and reveals cashback rates to consumers. Finally, consumers observe cashback information and make purchasing decision based on the overall utility derived from the products. In the following discussion, we introduce the basic notation and model setup, explain the decision-making process for each player, and derive equilibrium solutions. Our analysis proceeds in the reverse direction. We first analyze merchants' net payoff conditional on their join or no-join decisions. Then, in section 4 we discuss whether the merchant should join at all, given the optimal solution.

Game Setting

Let us assume that there are two merchants selling a homogeneous product or perfect substitute products. One merchant is high-type and the other is low-type, denoted by M_1 and M_2, respectively. Factors determining a merchant's type could be its brand image, level of customer services, or so on. In the stage 1, a merchant j, where $j=\{1, 2\}$, simultaneously chooses the price p_j and decides whether to join the *cashback* platform. If joining, M_j pays the platform owner a commission rate $b_j \in [0, 1]$ for transactions made through referral links on the platform. The marginal cost is constant across two merchants and can be normalized to zero by interpreting the consumers' willingness to pay as net of marginal cost. The merchant j's goal is to find optimal (p_j, b_j) which maximizes its payoff on the cashback site, given by:

$$\max_{p_j, b_j} \Pi_{M_j} = q_j * p_j\left(1 - b_j\right),$$

where q is the quantity of products sold through the platform. Note that b_j=0, if a merchant j does not join. In this case, the merchant's only has one decision variable.

As an transaction broker, the cashback site receives commission rate b_j from merchant j and rewards its members who make purchases from merchant j with cashback rate $a_j \in [0, b_j]$. Earning a commission of $p_j\left(b_j - a_j\right)$ from j per unit sold, the site's problem is to maximize its profit by choosing the optimal (a_1, a_2):

$$\max_{a_1, a_2} \Pi_{CB} = \sum_{j=1,2} \left[q_j * p_j\left(b_j - a_j\right)\right].$$

In the spirit of Salop's model, we assume that consumers are uniformly distributed on a line with infinite distance. This setup allows us to ignore firm's location decisions and to take consumers' heterogeneity in brand preference into consideration. Each consumer has a unitary demand and the location of a consumer identifies the ideal brands she prefers. We denote the distance between two merchants on consumers' preference line by d, which captures the extent of brand competition. The reservation price a consumer would like to pay for her ideal brand is v_j. A brand at distance x away generates a utility of $v_j - cx$, where c is the *brand-fit* coefficient. If c is high, consumers value brand name more than the price. On the contrary, if c is low, consumers are more price-prone (Rao 1991). Figure 1 plots the market shares in our duopoly model with two asymmetric firms.

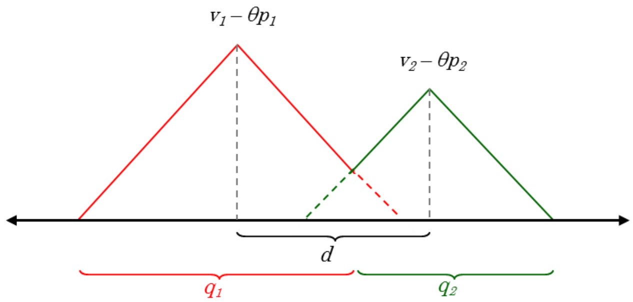


Figure 1. Market shares in a duopoly setting with two asymmetric firms

Consumer segments

To model consumer segments, we further assume that there are two types of consumers: h and l, of size α and $1 - \alpha$, respectively. Consumers of type h have higher sensitivity for the price than consumers of type l, i.e. $c_h < c_l$. Note that the notation here is somehow misleading; c_h is actually lower than c_l. The merchant would like to give discount to consumers of type H. Consumers self-select their types depending on purchasing habit. That is, buyers who make transaction through referral link on the cashback site are identified as highly price-sensitive *bargain hunters*, as opposite to normal online shoppers who don't bother take advantage of the discount. The transaction cost of purchasing through the platform (i.e., redemption cost of cashback) is denoted by r, and we assume $r_h < r_l$. Consumers derive *transaction utility* from net price over the utility obtained from the product (Kalwani & Yim 1992 Mayhew & Winder 1992, and Thaler 1985). The transaction utility of product from merchant j for type i consumers is given by:

$$u_{i,j}(x) = v_j - p_j - c_i x \text{, where } i = \{h, l\}.$$

Finally, we need to add a necessary constraint to the platform's maximization problem to make sure that consumers of type h have incentive to use *cashback* platform while consumers of type l do not.

4. Optimal Profit-sharing Scheme on the Cashback Market

Proposition 1. *Given the merchants' commission rates, the cashback site's optimal choice of cashback rates* $\{a_1^*, a_2^*\}$ *are:*

i. *Market 1* (*competitive market*): where $v_1 \in R_1$

$$\left\{a_{1,c}^*,\ a_{2,c}^*\right\} = \left\{\frac{1}{2}\left(1 + b_1 - \frac{2v_1 + dc_h - 2r_h}{4p_1\theta}\right),\ \frac{1}{2}\left(1 + b_2 - \frac{2v_2 + dc_h - 2r_h}{4p_2\theta}\right)\right\};$$

ii. *Market 2* (*M_2 monopolistic*): where $v_1 \in R_2$,

$$\left\{a_{1,m2}^*, a_{2,m2}^*\right\} = \left\{0,\ \frac{1}{2}\left(1 + b_2 - \frac{3v_2 - v_1 + p_1\theta + dc_h - 2r_h}{6p_2\theta}\right)\right\};$$

iii. *Market 3* (*no market*): where $v_1 \in R_3$, $\left\{a_{1,n}^*, a_{2,n}^*\right\} = \{0, 0\}$;

Proposition 1 shows that there are three feasible pure strategy market outcomes in the equilibrium[2]. The platform owner's optimal *cashback* rate is an increasing function of the merchant's commission rate. In other words, the site will give a higher cashback rate if the commission rate she collects from the merchant is greater.

Proposition 2. *The merchant j's optimal commission rates and price* $\{p_j^*, b_j^*\}$ *are:*

i. *Market 1 (competitive market): where* $v_1 \in R_1$

$$\{p_{1,c}^*, b_{1,c}^*\} = \left\{\frac{17v_1 - 3v_2 + 7dc_l}{35\theta}, \frac{7d(c_l - c_h) + 14r_h}{17v_1 - 3v_2 + 7dc_l}\right\},\ \{p_{2,c}^*, b_{2,c}^*\} = \left\{\frac{17v_2 - 3v_1 + 7dc_l}{35\theta}, \frac{7d(c_l - c_h) + 14r_h}{17v_2 - 3v_1 + 7dc_l}\right\};$$

ii. *Market 2 (M_2 monopolistic): where* $v_1 \in R_2$

$$\{p_{1,m2}^*, b_{1,m2}^*\} = \left\{\frac{t_1}{\theta C_{m2}}, 0\right\},\ \{p_{2,m2}^*, b_{2,m2}^*\} = \left\{\frac{t_2}{6\theta C_{m2}}, \frac{C_{m2}(dc_h - dc_l - 2r_h)}{t_2}\right\};$$

iii. *Market 3 (no market): where* $v_1 \in R_3$

$$\{p_{1,n}^*, b_{1,n}^*\} = \left\{\frac{(17v_1 - 3v_2)C_0 + 7c_l(dc_h - 14\alpha r_h)}{35\theta C_0}, 0\right\},\ \{p_{2,n}^*, b_{2,n}^*\} = \left\{\frac{(17v_2 - 3v_1)C_0 + 7c_l(dc_h - 14\alpha r_h)}{35\theta C_0}, 0\right\};$$

where $C_{m2} = 67\alpha c_l + 70(1-\alpha)c_h, C_0 = \alpha c_l + (1-\alpha)c_h,$

$$t_1 = 2(17v_1 - 3v_2)C_0 + \alpha c_l(3v_2 - v_1 + 14dc_h - 30r_h) + 14dc_h c_l,$$

$$t_2 = 12(17v_2 - 3v_1)C_0 + \alpha c_l(2v_1 - 6v_2 + 67dc_l - 69dc_h - 30r_h) + 84dc_h c_l.$$

Proposition 2 characterizes the subgame perfect Nash equilibrium (SPNE). It details merchant j's optimal pricing strategy $\{p_j^*, b_j^*\}$ under different market outcomes. It can be shown analytically that $p_{j,c}^* > p_{j,0}^* > p_{j,c}^*(1 - b_{j,c}^*)$, suggesting that a merchant will raise the price if she exercises segmentation strategy by joining the platform, compared. This finding is consistent with the principle of price discrimination; unitary price is bounded between two differentiated prices.

The comparative statics of the optimal commission rate under competitive market are summarized in Table 1. We find that if a merchant is able to increase its brand valuation, i.e. v_j is higher, the optimal choice of commission rate would be lower. For the high-type brand, if its competitive advantage is large enough, it would quit the *cashback* market. On the contrast, a merchant will have more incentive to join the market and hence gives a higher b_j, if its rival is able to increase v_{-j}. Left panel of Figure 2 plots two merchants' optimal commission rate as v_1 gets larger, given v_2 fixed. When v_1 is large enough, we will expect b_j=0, implying that premium brand has no incentive to join the platform. If the gap of two consumer segments gets more salient, i.e. Δc is higher, both merchants should take a more aggressive segmentation strategy by choosing a greater b_j. Figure 2(b) plots the optimal choice of b_j over Δt. As we can see, b_2^* increases in a faster rate than b_1^* does. With a higher *d*, which indicates less head-to-head brand competition, merchants will have a higher incentive to price-discriminate consumers. Suppose *d* is large enough such that two merchants will behave like two monopolists by setting price difference at the maximal level. Loosely speaking, practitioners should take a conservative segmentation strategy if consumers' preference over brands is not salient.

	Parameters				
	v_j	v_{-j}	Δc	d	r_h
b_j	−	+	+	+	+

Table 1. Comparative statics of the optimal commission rate

[2] It can be shown that the equilibrium where only premium brand (M_1) joins doesn't exist.

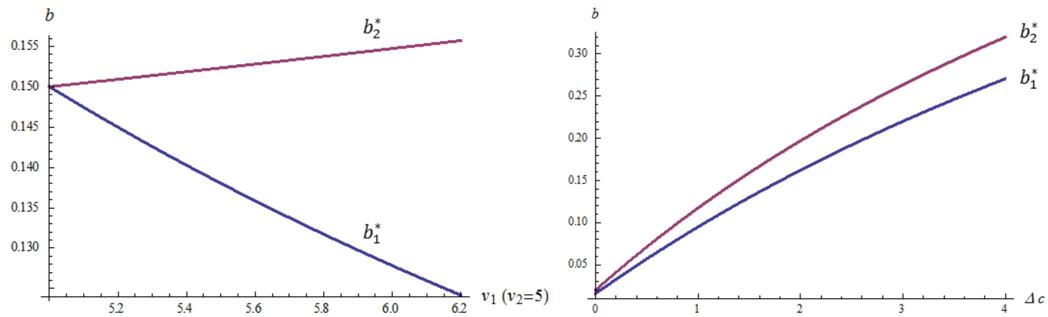


Figure 2. Optimal commission rates as functions of v and Δc

5. Conclusion and Future Research

In this paper, we examine the economic impact of the cashback mechanism on merchants' price-discrimination strategy. Using a game theoretical model, we identify three possible outcomes of the cashback market. We also examine how market configuration affects merchant-platform affiliation strategy and profit-sharing scheme. Our result provides guidance for merchants and the cashback site on how to optimally choose commission and cashback rate, respectively.

We believe the following concerns, which are also the limitations of this paper, could point out directions for the future research. First, since merchants have no *direct* control on the cashback rates, it would be interesting to examine the underlying pricing strategy from a bargaining's perspective. Second, we also want to explore how the optimal profit-sharing scheme would change when multiple cashback sites are involved. We seek to answer the following questions: Under what conditions should a merchant consider multihoming? Does the existence of cashback sites increase or decrease the social welfare?

Reference

Arthur, W. B. (1994). Increasing returns and path-dependence in the economy. *University of Michigan Press, Ann Arbor, MI.*

Baye, M. R. & Morgan, J. (2001). Information gatekeepers on the Internet and the competitiveness of homogeneous product markets. *American Economic Review*, 91, 454-474.

Clemons, E. K., Hann, I., & Hitt., L. (2002). Price dispersion and differentiation in online travel: An Empirical Investigation. *Management Science*, Vol. 48, No. 4, 534-549.

Gupta, S. & Cooper, L. G. (1992). The discounting of discounts and promotion thresholds. *Journal of Consumer Research*, 191 401-411.

Holmes, T. J. (1989). The effects of third-degree price discrimination in oligopoly. *American Economic Review*, 79, 244-250.

Kalwani, M. U. & Yim, C. K. (1992). Consumer price and promotion expectations: an experimental study. *Journal of Marketing Research*, 29 , 90 –100.

Lu, Q. & Moorthy, S. (2007). Coupons versus rebates. *Marketing Science*, 26, No.1, 67-82.

Moorthy, K. (1988). Product and price competition in a duopoly. Marketing Science, 7, 141-168.

Nahm, J. (2003). The gatekeeper's optimal fee structure when sellers can price discriminate. Economics Letters, 80, 9-14

Pigou, A.C. (1920). The economics of welfare. *London: Macmillan.*

Rao, R. (1991). Pricing and promotions in asymmetric duopolies. Marketing Science, 10, No.2, 131-144.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science*, 4, 199–214.

Salop, S. (1978). Monopolistic competition with outside goods. Bell Journal of Economics, 10, 141-156.

Varian, H.R. (1989). Price discrimination. Chapter 10 in *Handbook of Industrial Organization*, Vol. I, by R. Schmalensee and R.D. Willig, Eds. Elsevier Science Publishers B.V.

An Information Stock Model of Customer Behavior in Multichannel Customer Support Services

Kinshuk Jerath, Anuj kumar, and Serguei Netessine
kinshuk@cmu.edu, akumar1@ufl.edu, Serguei.Netessine@insead.edu

Abstract

We propose a novel information stock based framework to understand customer behavior in a multichannel customer support services scenario. We assume that a customer's observed usage behavior is a stochastic function of her latent "information stock," which is determined dynamically by the queries that arise as she uses the product and the assistance she obtains on contacting the firm's support channels. We estimate our model on customer level data obtained from a US based health insurance firm to quantify the efficacy of different support channels in terms of resolving customers' queries, and find that, in our setting, the telephone channel offers significantly greater value than the web channel. Our model can also aid in call center staffing decisions as it provides accurate predictions for future query volumes.

Introduction - Firms predominantly offer customer support via telephone and web portal at their call centers. In this paper, we develop a novel information based framework to assess the relative values of these support channels for customers from transactional data at a call center of a large US health insurance firm (name withheld due to a non disclosure agreement). For a customer, we assume that product usage leads to an information need and support channel usage leads to an information gain to satisfy this need. We model the latent "transactional information stock" for a customer as the sum of her information needs and information gains. The customer's observed channel usage behavior is then a realization of a two stage stochastic process: (1) a Poisson query arrival process and (2) a Bernoulli channel choice process once the query arrives. Thus, we model the query arrival process for a customer as a nonhomogeneous Poisson process with the mean query arrival rate at a given time being a function of her information stock at that time. Once a query is to be made, the customer's choice among the support channels is modeled as a Bernoulli choice process, with the web choice probability at a given time again being a function of her latent information stock at that time. Thus unlike queuing models in OM (Gans et al. 2003), we endogenize the customers query arrival and channel choice process on her information stock. We allow for unobserved heterogeneity across customers in the baseline query arrival rate and the web choice propensity. Besides the "transactional information stock" described above, we also account for a "seasonal information stock" to allow for queries that occur due to a time event, such as renewing the insurance contract or changes in contract terms. Although there is some literature on determinants of customer adoption of self service technologies (SST) but it is done on survey data and not actual transaction data (Meuter et al. (2005), Bobbitt and Dhabolkar (2001)). In the present study, we estimate customer behavior on actual transaction data.

We collected individual level customer data on claims and channel usage from the firm. Our model estimates suggest that the telephone channel in our setting provides, on average, an order of magnitude more information to the customer than the web channel does. We also find that health event related information needs vary with the nature of health events—health events where customers have out of pocket expenses leads to 79% higher information need than a health events with no customer payment liability, and repeated health events (say, events associated with multiple claims for a chronic disease) lead to less than 0.1% of the information need of the original health event. We find that customers prefer the telephone channel for higher health event related information needs but prefer the web portal for simpler seasonal information needs. This result is in line with Kumar and Telang (2011). We also estimates the heterogeneity in customers' propensity for query and web choice. We find a bipolar distribution of web choice probability across customers indicating two distinct customer segments: "web avoiders" and "web seekers."

We compare the predictions of our information stock model with the benchmark model available in literature and find that our model has significantly lower predictive errors for total queries, telephone queries, and web queries in both in sample and out of sample predictions. This finding further validates the information stock based modeling of consumer behavior and it demonstrates that our model can accurately predict the future load on different support channels at call centers. Our information stock model is also able to identify with high precision customers who, with high probability, will be calling in the near future. The firms can make outgoing calls to such identified customers and thus prevent some of its peak time calls to save on the customer service representative costs at their call centers.

Our research makes several contributions to the literature. First and foremost, to the best of our knowledge, ours is the first attempt to formalize customer query arrival and channel choice processes by explicitly modeling a customer's latent information stock. This model provides highly accurate predictions of future load on call centers. Second, our proposed framework utilizes information as a common denominator to understand determinants of customers channel usage, with different sources contributing to the latent construct. This approach allows call center managers to make inferences about the relative values of different factors that influence customers' behavior to seek support from the firm. For the first time, we are able to utilize this approach to provide quantitative evaluation of the customers' information demand and firms' information supply through different channels. Third, we provide a practical framework which allows a company, using only limited transactional data on customer product and channel usage (usually captured in today's business environment) to improve quality of service through better estimation of the query arrival process and, potentially, to even preempt query arrival.[1]

Research Setting and Data Description - We study a multichannel call center of a US health insurance firm. Customers purchase annual health insurance plans from the firm and thereafter utilize it for reimbursements of their medical expenditures. During the health plan usage, customers often have queries regarding their plan coverage, status of claims, etc., for which they contact the firm. The firm offered support to its customers via telephone and web portal. Form web portal, customers can obtain information regarding their plan benefits, claims, health providers, and general information on diseases, procedures, preventive care. We constructed dataset for a random sample of 2462 customers from the web registered customer population of the firm, covering our July'05 Dec'07study period. We collected data on the date of claim filing, the customer out of pocket expenses and the provider charges for each claim from claim processing database of the firm. For a customer, if a claim has the same provider charges as one of her previous claims, we term the subsequent claim as being from a repeated health event. We extracted telephone usage and web portal usage information, specifically, the date of a use, from the Automatic Call Distributor (ACD) of the call center and the web informatics database of the firm respectively.[2]

Model Development - We assume independent Poisson query arrival processes at the two channels in our benchmark model. Let t_{ij} be the time of arrival of the j^{th} telephone call for

[1] Literature review not provided to adhere to page limits

[2] Summary statistics not provided to adhere to page limits

customer i, where j= *1, 2, 3,..* x_{iT} represent the sequence number of telephone calls. If λ_{i0T} is the baseline mean call arrival rate for customer i, then the likelihood of observed x_{iT} call arrivals for the customer is: $L_{iT} = \prod_{k=1}^{x_{iT}} \lambda_{i0T} \mathrm{e}^{-\lambda_{i0T}(\mathrm{t_k}-\mathrm{t_{k-1}})}$. Likewise, the likelihood of x_{iW} web query arrivals for customer i is: $L_{iW} = \prod_{k=1}^{x_{iW}} \lambda_{i0W} \mathrm{e}^{-\lambda_{i0W}(\mathrm{t_k}-\mathrm{t_{k-1}})}$, where λ_{i0W} is the baseline web query arrival rate for customer i.

The benchmark model assumes the query arrival to be exogenous. However, we wish to understand the underlying mechanism that drives the query arrival and channel choice processes for customers. An effective approach to do this, used widely in the marketing and economics literatures, is by modeling the latent construct that drives observed behavior. For instance, McFadden (1973) introduced the idea that the latent construct, "utility," drives the stochastic choice process leading to observed consumer choice. Moe and Fader (2004) examine consumers' dynamic conversion behavior at an e commerce website with a latent visit effect that evolves over visits. In a similar vein, we assume that, at a given time, each customer has an "information stock" which determines her query behavior. We categorize information stock into two broad categories. First category called the "transactional information stock" emanates from the health events faced by a customer. We assume that each claim filed for the customer leads to health event related information need C. To fulfill her information needs, the customer approaches the firm's support channels and receives some information *gain* (W and T respectively) from web and telephone channels. The net of information need and gains results in transactional information *stock*. This information stock varies across customers based on the number of health events they face and the queries they make. The claims and queries for customers can be organized in sequence of claims followed by a query. For instance a customer i makes x_i queries with t_{ij} as the time of the j^{th} query, j=0, 1, 2,.. x_i. Between two queries several claims are filed for the customer. Let n_{ij} be the number of claims that are filed between the j^{th} and $(j+1)^{\text{th}}$ query by customer i and t_{ijk} be the time of arrival of k^{th} claim after j^{th} query with k=0, 1, 2, 3, …, n_{ij}. For customer i, the net transactional information stock after the k^{th} claim after j^{th} query $I_{ijk} = \sum_{z=0}^{j} [n_{iz}C - (y_{iz}W + (1 - y_{iz})T)] + kC$, where y_{iz} = 1 if the z^{th} query for customer i is a web portal visit and 0 otherwise.

The second category of information stock is "seasonal information stock" includes needs and gains that arise due to insurance related events in time. For instance, around insurance contract renewal time, customers make more queries regarding their ID card, web portal login/password reset, insurance forms, etc. Similarly, the firm periodically sends information bulletins to consumers leading to information gains. These seasonal information needs are variable with time but equally applicable to all customers at a point in time. We model seasonal information needs as parameters I_m in month m, where $m \in \{0, 1, 2, \dots, 23\}$. The total stock of information need for customer i after the k^{th} claim after j^{th} query, and where the month is m, is given as the sum of the transactional and seasonal information stocks ($I_{ijk} + I_m$), Note that for a customer C, W and T remain the same for all claims, web visits, and telephone calls respectively. The dynamics in the total information stock for a customer are shown in Figure 1.

The Poisson query arrival process for customer i after the k^{th} claim after the j^{th} query, and where the month is m, is $\lambda_{ijkm} = \lambda_{i0}\exp\left(I_{ijk} + I_m\right)$, where λ_{i0} is the baseline mean query arrival rate for customer i. Therefore, query arrival for a customer is a nonhomogeneous Poisson process with λ_{ijkm}, will keep changing for a customer with arrival of claims or queries or with change in month.[3] After deciding to query, a customer makes a Bernoulli choice between using the web portal and making a telephone call, with a web choice probability p. Since telephone calls are answered by trained representatives of the firm, it is likely that when the information need is high, customers prefer making a telephone call. Likewise, customers may prefer the web portal for structured needs such as seeking insurance contract related information like ID card and password reset. We allow for these possibilities by making the web choice probability at the j^{th} query for customer i as a function of the two types of information stocks: $p_{ijm} = p_{i0} \frac{\exp(\pi_T I_{ij} + \pi_S I_m)}{[1 - p_{i0} + p_{i0}\exp(\pi_T I_{ij} + \pi_S I_m)]}$, where I_{ij} denotes the transactional information stock for customer i at the time of query j, the j^{th} query for the customer arrives in month m, p_{i0} indicates customer i's baseline web choice probability. The parameters π_T and π_S allows the impacts of the transactional and seasonal information stocks, respectively, to be different on the channel choice probability than on the query arrival rate.[4]

Customers are likely to make queries for more severe health events and where they have to pay out of their pockets. Customers often face repeated health events e.g., multiple chemotherapy or dialysis sessions. In such cases, customers may have higher information needs in the first few claims but once they understand their insurance plan coverage, they may have little to ask in further repeated claims. To allow for these possibilities, we assume that for customer i's claim associated with health event h is: $C_{ih} = C_0 \exp(\alpha_{LIAB} \cdot D_{LIAB,ih} + \alpha_{RPT} \cdot D_{RPT,ih})$, where, $D_{LIAB,ih}$ is a dummy variable equal to 1 for claims with positive customer out of pocket expenses and 0 otherwise, $D_{RPT,ih}$ is a dummy variable equal to 1 if the claim pertains to a repeated health event and 0 otherwise, and C_0 is the baseline information need from a claim which is constant across all claims and across all customers. We further allow for gamma distributed heterogeneity in the baseline query arrival rate, $\lambda_{i0} \sim gamma(\gamma, \theta)$, and beta distributed heterogeneity in the baseline web choice probability, $p_{i0} \sim beta(a, b)$ across customers.

Estimation and Results - We use a hierarchical Bayes framework for model estimation, using the following MCMC chains: (1) Draw ($\lambda_{i0}, p_{i0} \mid C_0, W, T, I_m, \pi_T, \pi_S, \alpha_{LIAB}, \alpha_{RPT}, \gamma, \theta, a, b,$ data$_i$) using the MH algorithm, (2) Draw ($C_0, W, T, I_m, \pi_T, \pi_S, \alpha_{LIAB}, \alpha_{RPT} \mid \gamma, \theta, a, b, \lambda_{i0}, p_{i0},$ data$_i$) using the MH algorithm, (3) Draw ($\gamma, \theta, a, b \mid C_0, W, T, I_m, \pi_T, \pi_S, \alpha_{LIAB}, \alpha_{RPT}, \lambda_{i0}, p_{i0}$) using the MH algorithm. The first 30,000 MCMC iterations were used as initial burn in to reach convergence, and the last 10,000 iterations were used to infer the posterior distributions of the parameters. We report the estimation results in Table 1. The lower value of negative log marginal density and a higher value of Bayes factor (1928.76) provide evidence that our model fits the observed data better than the benchmark model. The heterogeneity plot of mean query arrival rate and web choice probability across customers in our sample is reported in Figure 2. Figure 2(b) indicates a polarized distribution of web choice probability that suggests that a relatively larger number of customers have very low web choice probability and can be classified as web avoiders, whereas a relatively smaller number of customers have very high web choice probability and can be classified as web seekers. The median web choice probability is 0.38.

The estimated value of the baseline information need from a claim, C_0, is 2.694, the information gain from a web visit, W, is 0.658, and the information gain from a telephone call, T, is 20.768. The above estimates suggest that, in our setting, a telephone call, on average, provides a large

[3] Mean query arrival rate changes with the customer's information stock – a higher information stock leads to a larger query arrival rate

[4] The web choice probability p_{ijm} for customer changes with the arrival of claims, queries, and change of month.

information gain to the customer in comparison to other channels—an order of magnitude higher than the information need from a claim and almost two orders of magnitude higher than the information gain from a web visit. The negative and significant estimate for π_T (3.930) suggests that the probability of web usage decreases with higher transactional information stock, which implies that customers prefer to use the telephone during high health event related information need. In contrast, the positive and significant estimate for π_S (1.590) suggests that the probability of web usage increases with higher seasonal information stock. The positive and significant estimate of α_{LIAB} (0.583) indicates a 79% higher information need from claims where customers have out of pocket expenses as compared to the claims with no customer out of pocket expenses. A negative and significant estimate of α_{RPT} (11.944) indicates that a repeated health event generates less than 0.1 % of the information need generated by the original health event.

Model Predictions - For both of the estimated models, we predict total queries, telephone queries, and web queries for each customer in our sample for the calibration period (Jul’05 Jun’07) as well as the hold out period (Jul Dec’07). To test these predictions from the models, we aggregate them across the full cohort of customers, and also aggregate them across time up to the monthly level. We find that the information stock model makes significantly superior in sample and out of sample predictions for all types of queries as compared to the benchmark model (see Table 2). We compute the calling probability for each customer for each month in the out of sample period (Jul Dec’07) from the two models. We then compare the percentage of correctly identified calling customers from the two models assuming top 20% of customers based on the calling probability are the calling customers. We find that our model correctly identifies a larger percentage of calling customers than benchmark model (see Table 3). This feature of our model can be utilized by the firm to further reduce call center costs. For instance, with the calling customers identified, the CSRs can make outgoing calls (or use other interventions) to resolve queries of such customers in non peak times of the day, thereby preventing some of the calls that these identified customers would have made at the peak time of the day.

Conclusions - We propose a novel information stock based framework to endogenously model the query generation and channel choice processes in customer support services. Our model not only provides interesting insights in the determinants of the customer channel choice but also outperforms benchmark model in terms of predicting query volumes at different channels. In addition, our model is able to identify with high accuracy the customers who are likely to make queries in the near future. Making advance outgoing calls to these identified customers can help reduce the peak time calls, which can lead to substantial cost savings for the firm. At a time when web portals are becoming a popular choice as a way to reduce the cost of customer service, our results show that there are more subtle and complex phenomena at play. Our estimates indicate the superior informational value of the traditional telephone channel over the self service web portal. This suggests that the assisted telephone channel is still a dominant customer support channel at least for complex services such as health insurance. Our estimates also suggest that web portals are effective for simple, unambiguous tasks (such as seasonal information needs, which are more structured and routine information needs).

References –

Bobbitt, L. M., and P. A. Dabholkar (2001), “Integrating Attitudinal Theories to Understand and Predict Use of Technology Based Self Service: The Internet as an Illustration.” *International Journal of Service Industry Management* 12 (5), 423–450.

Gans, N., G. Koole, and A. Mandelbaum (2003), "Telephone Call Centers: Tutorial, Review, and Research Prospects." *Manufacturing & Service Operations Management* 5 (2), 79–141.
Kumar, A., and R. Telang (2011), "Does Web Reduce Customer Service Cost? Empirical Evidence from a Call Center." *Information Systems Research,* Articles in Advance.
McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers of Econometrics*, ed. P. Zarembka. 105 142. New York, USA: Academic Press.
Moe, W. W., and P. S. Fader (2004), "Dynamic Conversion Behavior at E Commerce Sites." *Management Science* 50(3), 326–335.
Meuter, M. L., M. J. Bitner, A. L. Ostrom, and S. W. Brown (2005), "Choosing Among Alternative Service Delivery Modes: An Investigation of Customer Trial of Self Service Technologies." *Journal of Marketing* 69 (2), 61–83.

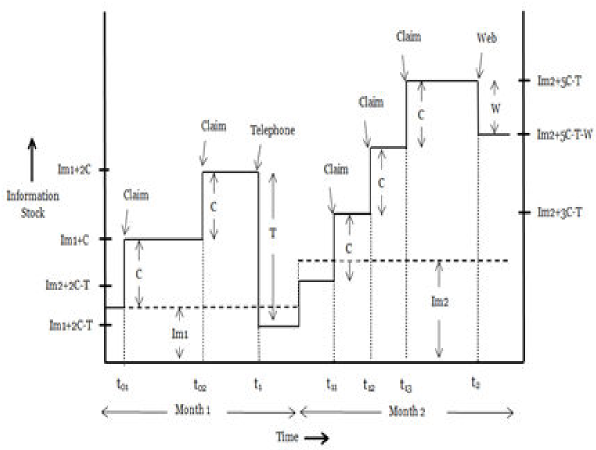


Figure 1: Dynamics in a Customer's Information Stock over Time (*m1* and *m2* denote Month 1 and Month 2, respectively)

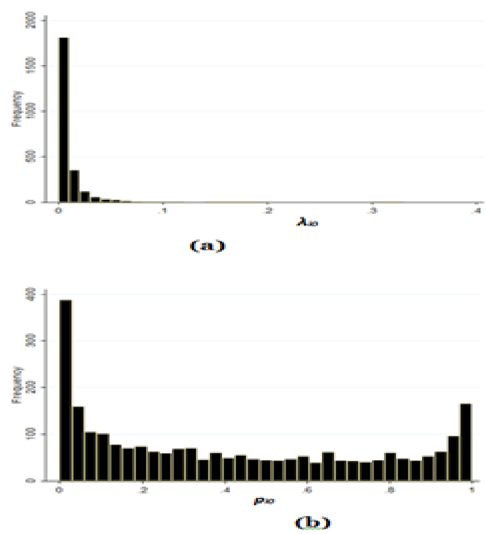


Figure 2: Distribution of baseline query arrival rate and web choice probability

	Benchmark Model		Info. Stock Model
In-Sample Fit Statistics			
-2 Log Marginal density	234263.88		230406.36
Log Bayes factor			1928.76
Parameter Estimates			
γ_T	0.860 (0.032)	γ	0.637 (0.018)
θ_T	175.982 (7.611)	θ	58.797 (3.053)
γ_W	0.283 (0.007)	a	0.470 (0.021)
θ_W	32.126 (1.652)	b	0.658 (0.023)
		C_0	2.694 (0.373)
		W	0.658 (0.241)
		T	20.768 (2.296)
		π_T	-3.930 (0.606)
		π_S	1.590 (0.101)
		α_{LIAB}	0.583 (0.132)
		α_{RPT}	-11.944 (5.783)

Table 1: Estimation Results (posterior mean followed by the standard error in parentheses)

	Tot Queries	Telephone	Web
In-sample Predictions			
Benchmark Model	15.06%	8.63%	22.61%
Info Stock Model	2.22%	7.01%	3.76%
Out-of-sample Predictions			
Benchmark Model	18.54%	12.50%	24.61%
Info Stock Model	12.13%	9.28%	15.89%

Table 2: Prediction Errors (MAPE)

Month	% of customers correctly identified	
	Benchmark Model	Info stock model
Jul-07	46.93%	71.84%
Aug-07	42.31%	73.08%
Sep-07	42.91%	77.33%
Oct-07	37.18%	80.14%
Nov-07	42.00%	77.60%
Dec-07	42.23%	75.73%

Table 3: Identification of Calling Customers

“Showrooming” and the Competition between Store and Online Retailers

Amit Mehra
Indian School of Business
amit_mehra@isb.edu

Subodha Kumar
Texas A&M University
subodha@tamu.edu

Jagmohan S. Raju
The Wharton School
rajuj@wharton.upenn.edu

Abstract

Increasing numbers of customers are using physical store retailers' outlet to evaluate products. Although this helps them to identify their “best fit” product, they often end up buying the product not at the store but rather at a competing online retailer to take advantage of lower prices. This free-riding behavior by customers is referred to as “showrooming.” We study a game theoretic model to analyze the competitive equilibrium where customers free-ride, and show that the profits of the store retailer are reduced compared to the situation when customers do not free-ride. This finding is in line with the current debate in the retailing industry where store retailers are discussing ways and means to improve their profits in the face of such showrooming. We also analyze the impact on the profits of a physical store retailer stocking a different assortment from that of the online retailer and show that this strategy may, under certain situations, improve its profits.

Keywords: online retailing, showrooming, retail competition, pricing, game theory.

1 Introduction

In the context of multichannel retail, it is well known that the customers may use one channel only to research products and buy in another. Thus, for example, customers may use Target's physical stores to evaluate products and then purchase at Amazon's online store if Amazon sells the selected product at lower prices. Customers benefit from using the physical store to evaluate products because many product attributes are of non-digital nature and cannot be properly assessed online (Lal and Sarvary 1999). Obviously, this shopping technique results in the physical store losing potential customers. Some of the recent media articles (Bosman 2011, Zimmerman 2012) document that this trend, referred to as “showrooming,” seems to be on the rise. An MSN Money article (Datko 2012) reports that physical store retailers like Target are reporting a significant reduction of their profits due to this tendency of the customers. The first research question we focus on is for what product categories does showrooming exist in equilibrium and whether it reduces profits of physical store retailers.

Our analysis reveals that showrooming does reduce profits for the physical store retailers in many situations. We then investigate the impact on a physical store's profits of offering a *product assortment* that differs from that of its online competitor. This line of questioning is reminiscent of the ideas presented by Carlton and Chevalier (2001), who show empirically that manufacturer's benefit by limiting the assortment of goods available through online channels so as to mitigate the free-riding of those channels on brick-and-mortar stores. In our situation, too, customers cannot benefit from showrooming if there is no online availability of the product that visiting the physical store has revealed to be the best fit for their needs. We find that this strategy does improve a physical store's profits on products with predominantly digital attributes and those with relatively higher customer valuations. In the next section, we begin with discussing the model.

2 Model

2.1 Customers

Customers need to buy one unit of the product. A visit to the physical store allows customers to evaluate both digital and non-digital attributes of each product in the assortment in order to select the one product that best fits their unique needs. All customers receive utility v from their best-fit product. In contrast, customers who evaluate the product assortment at the online store are unable to accurately assess the non-digital product attributes and so the product they select may not be the best-fit one. Therefore, each customer's expected utility from a product selected online is δv, where $\delta < 1$. The values of δ and v are public information.

Customers are heterogeneous in their channel preference. By this, we mean that a customer incurs different shopping costs when using either of the two retail channels. Further, these shopping costs are different for each customer. We capture this notion by considering customers to be uniformly distributed on the standard Hotelling line between 0 and 1, where a customer's index $x \in [0, 1]$ is proportional to the cost of shopping at the physical store and $(1-x)$ is proportional to the cost of shopping online, respectively. We use t as the proportionality constant here. Note that shopping costs at the physical store are mainly due to the time and effort required to travel to the store and evaluate the products, because the costs of completing the purchase are minimal. On the other hand, it is relatively costless to visit the website and to examine and evaluate the product assortment; in this case, shopping costs are incurred mainly at the purchase stage at which the customer must reveal sensitive credit card information and wait for several days before the product is delivered. Each customer's cost of shopping at the two channels is private information to that customer.

2.2 Retailers

There are two retailers: the physical store retailer and the online retailer. They are represented by the subscripts s and o, respectively. To begin with, we assume that both retailers carry the same assortment of products. We will relax this assumption when we consider the physical retailer's strategy of using a different product assortment to mitigate the reduction in profits due to showrooming. The price of each product in the assortment is set at p_s and p_o by the respective retailers. This setup captures the common situation in which all products in a particular category, for example casual shirts, are priced the same even though they differ in sizes and designs. Each customer prefers a different shirt based on her size and tastes. The prices are common knowledge among firms and customers. We normalize the marginal product costs to be zero for both retailers.

2.3 Game Structure

The game proceeds in three stages. At stage 0, each retailer announces its prices, which are observed by its competitor and by all customers. Next, at stage 1, customers decide to *visit* either the physical store or the online store to evaluate the product assortment and identify the product they want to buy. On the one hand, a customer incurs a shopping cost of xt to visit the physical store but this allows her to identify the best-fit product, which yields a gross utility v. On the other hand, visiting the online store's website to evaluate products is costless but can yield only the reduced gross utility of δv since the customer may be unable to identify her best-fit product. Thus, her ex-ante expected valuation from selecting the product online is lesser than her valuation when she chooses the product in the physical store. Customers decide which channel to visit by comparing the respective expected values to be derived from such a visit.

Finally, at stage 2, customers decide the channel at which to *complete* the purchase. A customer who visits the physical store to evaluate products can buy from that store with no additional shopping costs. In this case, the customer's net expected utility is $v - tx - p_s$. Alternatively, a customer can purchase at the online retailer after first identifying her best-fit product at the brick-and-mortar store. This approach entails a shopping cost of tx to visit the physical store and then a shopping cost of $t(1-x)$ to purchase at the online store, since the customer must now supply credit card information online and wait for delivery. Hence, a customer's net expected utility in this case is $v - t - p_o$. Customers who choose this option are the ones who do showrooming to maximize their surplus from the transaction. Finally, a customer could use the online retailer for both product evaluation and purchase. Their net expected utility under this scenario is then $\delta v - t(1-x) - p_o$. Note that although it is possible for customers to evaluate a product online, and then buy it at the physical store, this never happens in practice because the resulting net expected utility of $\delta v - t - p_s$ is less than $v - tx - p_s$, which is the expected utility when a customer evaluates and buys a product at the physical store.

Figure 1 illustrates stages 1 and 2 of a customer's decision process.

3 Analysis of the Game

Customers may exercise one of the following three options: (a) evaluate and buy at the physical store, (b) evaluate at the physical store but purchase online (showrooming), and (c) evaluate and buy at the online store. All customers who exercise the same option constitute a market segment. Note that the ex-ante surplus from exercising the first option is decreasing in x, that of the second option is

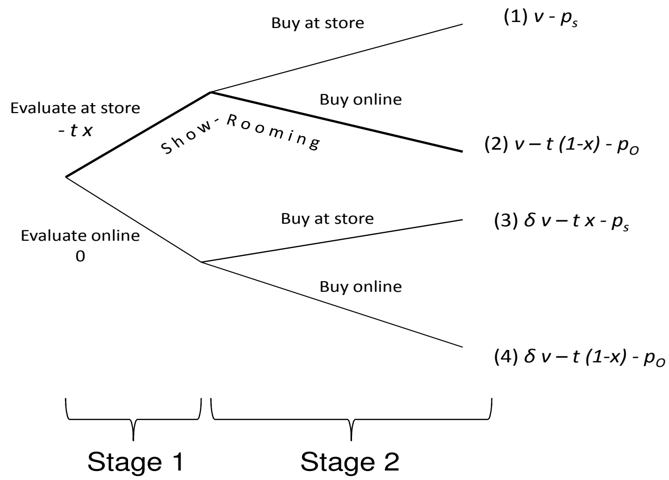


Figure 1: Decision tree for customers

independent of x, and that of the third option is increasing in x. Hence, customers exercising the first to the third option must be positioned from left to right on the Hotelling line. This arrangement is depicted in Figure 2 where the customer surpluses from exercising options (a), (b), and (c) are depicted by the lines P, S, and O, respectively. It is also possible that not all segments exist. This is because that an option that provides lower surpluses to all customers compared to other options will not be exercised by anyone. For example, it is possible that the lines P and O cross each other over the horizontal line S. This implies that option (b) will not be exercised by anyone. These considerations yield several possible arrangements of customer segments, each representing a particular market configuration, and we analyze the equilibrium for each of these.

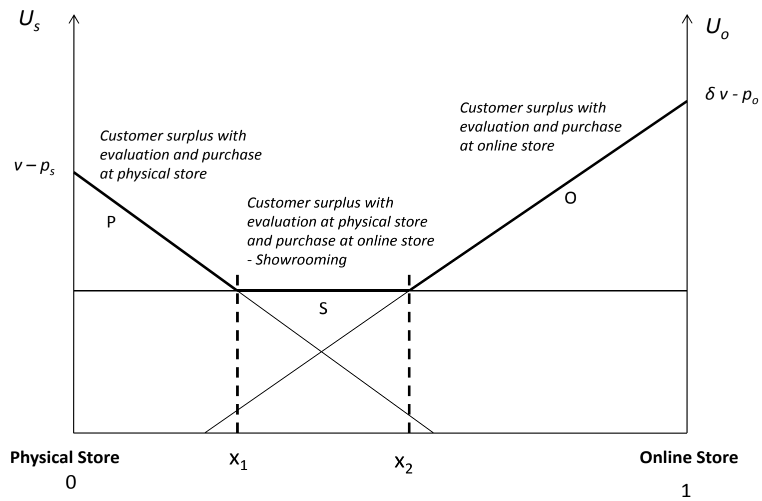


Figure 2: Customer surpluses in the competitive showrooming equilibrium

We now present some arguments that reduce the number of market configurations to be analyzed. First, observe that if segment (a) does not exist, then there are no physical store customers, in which case this store makes zero profits. For the leftmost customer on the Hotelling line, the surplus from exercising option (a) is $v - p_s$, from option (b) is $v - t - p_o$, and from option (c) is $\delta v - t - p_o$. Thus option (b) is always better than option (c) for this customer. If the physical store sets a positive price $p_s = p_o + t - \epsilon$, where ϵ is positive and vanishingly small, the surplus from the option (a) is better than option (b) for the leftmost customer and she will purchase from the store. This means that in equilibrium, there must always be some customers who exercise option (a). Hence we will consider only those market configurations that include this segment.

Second, we assume that the online retailer will always attract some customers who do not engage in showrooming. As a result, the customer at the rightmost end of the Hotelling line secures a higher surplus from exercising option (c) than from option (b). This leads to the requirement $v < \frac{t}{1-\delta}$, and if this is satisfied, the customer segment who exercises option (c) must also exist. Hence, we will further restrict our analysis to market configurations that also include this segment. Thus, we have two configurations, one consisting of only segments (a) and (c); and another (the showrooming configuration) that includes all three segments. The showrooming customers may have either a positive surplus, or zero surplus. Finally, it may also be the case that the market is not fully covered, in other words, that not all customers end up purchasing. This is the local monopoly equilibrium in which, with respect to the Hotelling line, segment (a) of customers are on the left, non-purchasers are in the middle, and segment (c) of customers are on the right. In short, there are four market configurations that need to be studied.

We start by analyzing the case with all three segments. Let the index of the customer who is indifferent between choosing options (a) and (b) be x_1, and let that of the customer who is indifferent between options (b) and (c) be x_2. We first analyze the case where the showrooming customers get a zero surplus. In this case, setting $v - t - p_o^* = 0$ yields $p_o^* = v - t$. The profit function of the physical store retailer is $p_s x_1$, and therefore optimizing this function with respect to store price gives $p_s^* = \frac{v}{2}$. This equilibrium is special because even though the markets of the physical store retailer and the online retailer are contiguous, there is no competitive effect. This is easily seen when we observe that the store price in this equilibrium equals the monopoly price, giving monopoly profits to the physical store. Hence, showrooming in this equilibrium is clearly not a problem for the store's bottom line.

We now focus on the situation when the showrooming customers get positive surpluses, implying a competitive equilibrium. The profit function for the physical store retailer is $p_s x_1$ and that of the online retailer is $p_o(1 - x_1)$. The corresponding Nash equilibrium prices are $p_s^* = \frac{2t}{3}$ and $p_o^* = \frac{t}{3}$. We refer to the this equilibrium as "A." For consistency of customer behavior, we need $x_2^* > x_1^*$ and $v - t - p_o^* > 0$. These inequalities yield the conditions $v > \frac{2t}{3(1-\delta)}$ and $v > \frac{4t}{3}$, respectively.

Now, we verify that the retailers have no incentive to deviate from the equilibrium prices. The analysis above ensures that such deviation confers no advantages provided the market configuration, and hence the demand pattern seen by the two retailers, remains the same. Yet, for instance, it is possible that, given the equilibrium price of the physical store retailer, the online retailer sets a price that differs from its equilibrium price, to enforce a market configuration other than the three-segment one that we assumed when evaluating the equilibrium prices. Such price setting would change the demand pattern, and therefore could enable profitable deviation for the online retailer. In general, both retailers can either lower or raise their prices from the prices in the equilibrium in order to enforce a market configuration different from that in the equilibrium. We exhaustively consider all the different types of market configurations that result due to deviations from equilibrium prices by the physical and the online retailers and then report our first result.

Proposition 1 *There exist a competitive pure-strategy equilibria under which customers practice showrooming in the following parameter space:*

(a) *If $\delta \leq 0.25$, then this equilibrium does not exist.*

(b) *If $0.25 < \delta \leq 0.4879$, then this equilibrium exists in the range $\frac{4t}{3} \leq v < \frac{t}{1-d}$.*

(c) *If $0.4879 < \delta \leq 0.4941$, then this equilibrium exists for $\frac{4t}{3} \leq v \leq \frac{\left(5(2-d)-\sqrt{(-4+d(-4+25d))}\right)t}{6(1+d)}$ and $\frac{\left(5(2+d)+\sqrt{(-4+d(-4+25d))}\right)t}{6(1+d)} \leq v < \frac{t}{1-d}$.*

(d) *If $0.4941 < \delta \leq 0.5$, then this equilibrium exists for $\frac{4t}{3} \leq v \leq \frac{\left(5(2-d)-\sqrt{(-4+d(-4+25d))}\right)t}{6(1+d)}$ and $\frac{(5-2\sqrt{2})t}{3(1-d)} \leq v < \frac{t}{1-d}$.*

(e) *If $\delta > 0.5$, then this equilibrium exists in the range $\frac{(5-2\sqrt{2})t}{3(1-d)} \leq v < \frac{t}{1-d}$.*

We now study the benchmark case when customers do not engage in showrooming even if it provides them with higher surpluses. This will allow us to compare retailers' profits in the two situations. The benchmark case represents the situation when customers were less sophisticated in the past and did not separate their evaluation and purchase decisions across physical and online channels. This analysis helps us see how the store profits may be affected as showrooming by customers becomes the norm.

In a competitive situation, the markets of the two retailers must be contiguous, and therefore we find the indifferent customer, x_b, by solving $v - tx - p_s = \delta v - t(1-x) - p_o$. The profit functions of the physical and the online retailers can be written as $p_s x_b$ and $p_o(1-x_b)$, respectively, and the corresponding Nash equilibrium prices of store and online retailers are $p_s^* = \frac{3t+v-\delta v}{3}$ and $p_o^* = t - \frac{v(1-\delta)}{3}$. For consistency of customer behavior, the surplus of the indifferent customer must be non-negative. This imposes the requirement $v > \frac{3t}{1+\delta}$. We analyze deviations from this equilibrium as well. By comparing profits of the physical store retailer in the benchmark case with its profits in the competitive showrooming equilibrium, we obtain the following result.

Proposition 2

Prices, as well as profits, are lower and market coverage is higher for the store retailer in the competitive showrooming equilibrium as compared to those in the competitive benchmark equilibrium.

This proposition shows that the physical store retailer's survival requires that it adopt profit-improving strategies. In the next section, we evaluate one such strategy.

4 Different Product Assortment Offered by the Physical Store

In this section, we consider the impact on profits of the physical store retailer when it offers an assortment of products that differs from that offered by the online competitor. The idea is that, when a customer visits the brick-and-mortar store to identify her best-fit product, this product may be available *only* at the physical store (and not online). We use a to denote the probability that the product identified in the physical store is available online as well. Thus, the parameter a captures the extent of commonality in the product assortment available from the physical and the online stores. If the assortments are identical, then $a = 1$. This implies that the customer is certain that her best-fit product is also available online. As the overlap in product assortment diminishes, the value of a becomes smaller.

A customer's decision to visit the physical store for product evaluation depends on her ex-ante utility from exercising this option $(av + (1-a)\delta v - t - p_o)$ versus going directly to the online retailer $(\delta v - t(1-x) - p_o)$. Let x_v denote the location of the customer who is indifferent between these two options. Customers to the left of x_v on the Hotelling line visit the physical store for product evaluation whereas customers to the right evaluate and purchase via the online channel.

With probability $(1-a)$, a customer will be unable to find her best-fit product in the online store after completing her evaluation at the physical store. In such a situation, she can buy from the physical store without incurring any additional shopping cost because she is already there and the cost of product evaluation is a sunk cost at this stage. Her surplus from exercising this option is $v - p_s$. Alternatively, she can buy a product from the online store, although her expected utility from doing so is less than if she purchased from the physical store because the product she purchases online may not be her best-fit. Exercising this latter option also requires the customer to incur an additional shopping cost for transacting through the website and then waiting for product delivery. Hence, a customer will exercise this option, which provides a utility of $\delta v - t(1-x) - p_o$, only if online prices are lower than store prices. We use x_n to denote the customer who is indifferent between purchasing from a physical and an online store. Then customers to the left of x_n will purchase from the physical store and those to the right of x_n will purchase online.

With probability a, a customer succeeds in finding her best-fit product at the online store after evaluating products at the physical store. As before, she can buy from the physical store and obtain a surplus of $v - p_s$, or she can buy from the online store and obtain a surplus of $v - t(1-x) - p_o$.

Because the customer in this case succeeds in finding her best-fit products online, her gross utility from purchasing via this channel is v rather than δv for those who could not find her best-fit product online. Using x_f to denote the customer who is indifferent between purchasing online and from the physical store, we can see that customers to the left of x_f will purchase from the physical store whereas those to the right will purchase online. The customer's decision tree is depicted in Figure 3.

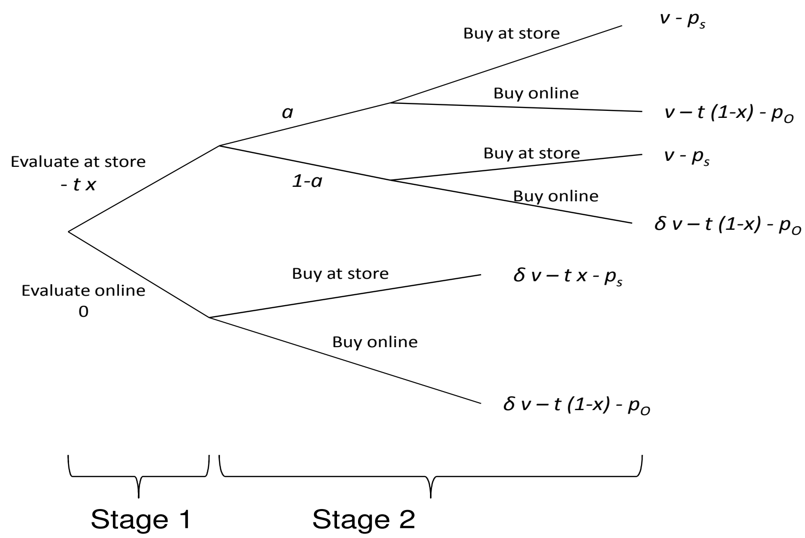


Figure 3: Decision tree for customers in the case of different product assortment

Next, we compare the physical store's profits when implementing this strategy with its profit under competitive showrooming equilibrium. That result is reported next.

Proposition 3 *The physical store's profits under the "different assortment" equilibrium are higher than its profits under competitive showrooming equilibrium.*

5 Conclusions

Competition between physical store retailers and online retailers is now picking up even in those product categories that were previously uncompetitive (or less competitive), such as clothes. The reason for less competition in the past was that the customers could not evaluate products in such categories on the online channel as effectively as they could in the physical store. As a result, the customers ended up making suboptimal purchases if they evaluated products only via the online channel. However, customers are now evaluating products in a physical store, and then ordering the preferred product from an online retailer. This trend is perceived as a threat to the bottom lines of physical store retailers. We report on the extent of this threat for retailers of various product categories and analyze a strategy of maintaining a different product assortment by the physical store retailer as a strategy to combat this challenge. We are currently analyzing some other strategies that the physical store retailers could adopt in addition to this strategy. We plan to present these additional results in conference presentation.

References

[1] Bosman, J. 2011. Book Shopping in Stores, Then Buying Online. *The New York Times*. Dec. 4.

[2] Carlton, D. W. & J. A. Chevalier. 2001. Free Riding and Sales Strategies for the Internet. *The Journal of Industrial Economics*. 49(4), 441-461.

[3] Datko, K. 2012. Target Takes Aim At 'Showrooming'. *MSN Money*. Jan. 23.

[4] Lal, R. & M. Sarvary. 1999. When and How is the Internet Likely to Decrease Price Competition? *Marketing Science*. 18(4), 485-503.

[5] Zimmerman A. 2012. Showdown Over 'Showrooming'. *The Wall Street Journal*. Jan. 23.

Does the Market Believe in Marketing?
A Text Mining Based Informational Value Perspective

Joseph Johnson, Brent Kitchens[1], Debanjan Mitra, and Praveen A. Pathak

Abstract: There is growing belief in the marketing community about the decreasing influence of the marketing department within firms and, in particular, on firms' corporate strategy. We examine this issue from the investors' perspective through estimating the value of marketing information. We argue that if the role of the marketing function within the firm is insignificant, it should be reflected in a small valuation of a firm's marketing information compared to that for other information types. Moreover, if marketing is indeed losing its clout, this value must be getting smaller (or even negative, in the extreme case) over time. We use text mining methods to extract the information contained by business functions in over 26,000 individual news articles between 1985 and 2010 that are related to 23 S&P 500 firms in the food, beverage, and tobacco industries. Based on the efficient markets hypotheses, we ascertain the value of marketing information by examining whether it is related to the firm's abnormal stock returns and if so, how it compares to the value of other types of business information. We find a significantly positive value of marketing information but it is lower than that of financial information. This value varies between firms and over time. Further tests reveal marketing information is more valuable when the variance of the equity market is high and when the firm has low marketing assets as reflected in low stock price to book value and low SGA expense to revenue ratios.

Introduction: Peter Drucker, widely considered as the father of modern management, concludes "Because the purpose of business is to create a customer, the business enterprise has two – and only two – *basic* functions: *marketing* and innovation. *Marketing* and innovation produce results; all the rest are costs" [3] (italics added). Such a distinguishing and unique function of a business should result in marketing departments having a prominent role in every firm. Yet, academic literature and the popular press generally conclude that "in many companies, the marketing function is in steep decline" [13] and that there appears to be a "marked fall-off in the influence, stature and significance of the corporate marketing department" [14].

While dire statements about the diminishing role of marketing in firms are primarily based on anecdotal evidence, there are also a few academic studies that examine this question empirically. Results in these studies are mixed, at best. For example, early studies in this domain used qualitative surveys of senior managers to report growing skepticism with the marketing concept and the value of satisfied customers [14]. Homburg et al. [6] disagree. They examine marketing's influence in a cross-section of US and German firms and find marketing to be "the most influential unit with respect to decisions about the strategic direction of the business unit."

Several studies have made important contributions towards understanding this issue, yet there seems to be certain limitations to this research. First, the surveys used in these studies provide rich data necessary for robust analysis. Yet, survey data "rely on retrospective recall …that is likely to reconstruct the past to make it consistent with … conventional story lines" [7]. To arrive at the actual status of marketing's influence on the firm, there is a need for objective data and analyses. Second, even though several researchers' claims imply a trend, i.e., the *decreasing* influence of marketing, none of the studies are able to track the importance of marketing over time. As a result, the conclusions on the changes in influence of the marketing department are based on conjectures. Third, while there is substantial research that alludes to differences in marketing's importance in specific contexts, these intuitions need to be tested based on empirically observed variations over time and across firms. In particular, can we

[1]Corresponding author

deduce any firm-specific or environment-specific observable characteristics that can be used to validate the extant theory related to marketing's influence on the firm?

These limitations underscore our research agenda. Like prior researchers, our central question relates to the size of marketing's influence on the firm, both in absolute and relative (to that of other business functions) terms. However, we examine this issue from a different perspective. If the marketing function's influence on a firm is insignificant, it should be reflected in a small valuation of the firm's marketing information compared to that for other information types. Moreover, if marketing is indeed losing its clout, this value must be getting smaller over time. We extract marketing information contained in news articles appearing in the Wall Street Journal (WSJ) and Dow Jones News Service (DJNS) related to 23 S&P 500 firms in the food, beverage, and tobacco industries. Based on the efficient markets hypotheses, we ascertain the value of marketing information by examining how this information is related to the firm's abnormal stock returns during the time window when the information becomes public.

Our data also enables us to address several related and pertinent questions, answers to which have so far been unknown. For example, how has the value of marketing information changed over the past few decades? More important, are there any explanations behind the empirically observed changes in the value of marketing information? Likewise, can we predict the firms for which marketing information is likely to be more valuable? To answer these questions, we utilize information retrieval and text analysis techniques to identify the level of marketing information and sentiment in news events for a set of firms, and use these inputs to analyze the impact of marketing information on firm value.

Text Analysis: We collected data on all 23 companies with SIC codes between 2000 and 2100 which were listed on the S&P 500 as of 12/31/2010. Companies in this SIC code range are those in the food, beverage, and tobacco industries. For each of these companies, we collected articles written about them in the WSJ and DJNS between 1/1/1985 – 12/31/2010. Our data was collected from Factiva, and we used Factiva's company search option to find articles about each company during the respective time frame. Once all articles were collected, they were cleaned and processed through the General Inquirer (GI) word disambiguation engine [4], which has been used in other studies relating news events to financial markets [11,12]. GI stems and analyzes words in context to determine their true meaning, removing ambiguity present if considering a word alone. Words were classified into categories, including "positive" and "negative," according to the GI's Harvard IV-4 and Lasswell value dictionaries [5]. After disambiguation, articles were condensed into vectors which represent term frequencies. The total number of words, positive words, and negative words were noted for use in our analysis.

As our intent is to measure the value of marketing information, we further required information regarding the extent to which each article relates to marketing. XML tags in each article provided information as to the subject of each article. One of the identified subjects was indeed "marketing," and 2,199 of the 26,896 collected articles were tagged as such. However, upon inspection of a sample of articles, it was noted that while articles tagged with the marketing subject were related to marketing, other articles discussed marketing as well. The subject tags appear to indicate primary subjects only, and we wished to include all articles related to marketing in our analysis, not only those whose primary subject was marketing. Therefore, we utilized text mining techniques to estimate the extent to which each of our collected articles relates to marketing. In order to do this, we first created a marketing "dictionary" – a vector representing how strongly each word indicates that an article relates to marketing.

To create this dictionary, we utilized the articles tagged as marketing. By comparing the words used in these articles and their frequencies to those used in articles not tagged as marketing, we determined those words which strongly identify marketing. Our method is based on that of Cecchini and colleagues [2]. We first split our articles into a corpus of marketing-tagged articles M and a corpus of non-marketing-tagged articles N. For each term i and document $m \in M$, we define a document score ds_{im} which represents the discriminatory power of word i in distinguishing document m from all documents in N. This document score extends the concept of TF-IDF, or term frequency-inverse document frequency, a frequently used construct in information retrieval research. Term frequency is simply the number of times word i appears in a particular document, normalized by the total number of words in a document. Inverse document frequency is the log of the ratio of the number of documents in a corpus to the number of documents containing word i. In our case, however, the IDF function is not determined using the entire set of documents, but rather only the corpus N, as we wish to distinguish document m not from all other documents, but only from non-marketing documents. Therefore, for each term i and document $m \in M$, our document score function is calculated as

$$ds_{im} = \left(\frac{tf_{im}}{wc_m}\right) * \log\left(\frac{|N|+1}{n_i+1}\right),$$

where tf_{im} is the number of occurrences of word i in document m, wc_m is the total number of words in document m, and n_i is the number of documents $n \in N$ which contain word i. Note that 1 is added to the numerator and denominator of the idf portion of the function to ensure that the function is defined even if there are no documents in N which contain word i.

The document score ds_{im} represents the relative ability of word i to discriminate between document m and all documents in N. However, we want a score that represents the power of word i to discriminate all documents in M from those in N. To attain this, we find the sum of all document scores for each word i, and weight this sum according to m_i, which is the number of documents which contained word i. The marketing dictionary is comprised of word scores, s_i. The formula for these final scores is

$$s_i = \sum_{m \in M} ds_{im} * \log(m_i) .$$

The marketing dictionary is therefore simply a high-dimensional vector of these word scores. Once we have calculated the dictionary, we may use it to determine the extent to which a particular article j is related to marketing, to be represented by the marketing similarity score ms_j. To do this, we simply need a measure of similarity of the marketing dictionary and each article. As both dictionary and article are represented as vectors, with dimension equal to the number of unique words used, we may take the cosine of the angle between these vectors to arrive at a measure of similarity commonly used measure in information retrieval research. The measure ms_j is used as a representative of the marketing information contained in each article for purposes of our analysis. We validated our final measure of marketing information, ms_j, using several common techniques, including correlation with marketing articles, precision, recall, and area under the ROC curve (specific validation results omitted due to space limitations).

Model and Results: To determine the valuation of marketing information, we compare marketing information in company news events to variations in stock price for that company. We focus on the absolute value of cumulative abnormal returns in our analyses because while we expect marketing information (or other types of business information) to cause investors to revise their assessment of the value of a firm, we have no expectation that investors will systematically revise their assessments in a particular direction. In this aspect, we follow other studies that examine the magnitude of market responses to information events where the nature of the news is ill-defined [1]. We implement a fixed effect model across companies to compare absolute abnormal returns to the measure of marketing information described above, as well as a similar measure we created for finance for purposes of comparison. We control for sentiment extremity which is also likely to influence the size of the abnormal return in addition to the information type [11]; we measure this by the absolute value of the difference in positive and negative words in an article divided by the total words (we tested a model using the proportion of negative words only with similar results). Our final model is aggregated at a daily level:

$$|AR_{it}| = \beta_0 + \beta_i + \beta_m MKT_{it} + \beta_f FIN_{it} + \beta_e EXTR_{it} + \varepsilon_{it} \qquad \left(EXTR_t = \frac{|P-N|}{T}\right).$$

Table 1 presents results of the model. In the form including only marketing information and sentiment extremity (model 1), both are found to be positively related to absolute abnormal returns (p<0.001). Once finance information is included, sentiment extremity ceases to have a significant effect (models 2, 3), but the effect of marketing information remains significant and positive. Model fit statistics are quite good for predicting abnormal returns. Results demonstrate that marketing information is valued by and influences financial markets. This influence appears less than that of financial information, as may be expected, but is still important.

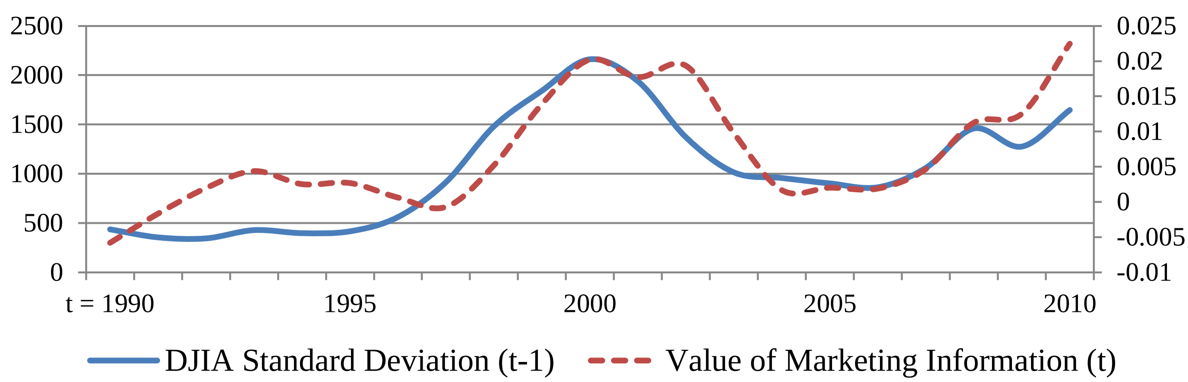
Table 1 – Daily abnormal return firm fixed effect model parameters

	Model 1	**Model 2**	**Model 3**
Intercept (β_0)	0.014***	0.011***	0.010***
Marketing Information (β_m)	0.017***		0.010***
Financial Market Information (β_f)		0.029***	0.027***
Extremity of Information (β_e)	0.0045***	0.00073	0.0032
Model Fit (R square)	0.062	0.078	0.088

*** p<0.001, ** p<0.01

Having confirmed the value of marketing information, we turn to the question of how and why the value of this information might be changing over time. To do this, we again compute our full model for each five year window in our dataset (on a rolling basis, i.e. 1985-1989, 1986-1990, …, 2006-2010). Due to space constraints, we do not provide the full set of results here, but we find that for each window, financial market information is positively related to abnormal returns (p<0.001), while marketing information is only significant in certain periods. The value of marketing information, represented by its regression coefficient, is plotted in Figure 1; values above 0.005 were typically significant (p<0.01). This value shows substantial increase during 1997-2002 and 2005-2010. In other periods, marketing information has a smaller value.

Figure 1 - Comparison of five-year trends



Also included in Figure 1 is a lagged five-year trend of the standard deviation of the Dow Jones Industrial Average (DJIA). It is clear that there is significant correlation between the two trends; in fact, the correlation coefficient between the two is 0.838. No such correlation was found with the value of financial market information. This result echoes contentions in the literature regarding marketing's enhanced value under conditions of market uncertainty [8, 6].

Finally, we turn our attention to differences which may exist between firms with respect to the valuation of marketing information, as well as the factors which might create such differences. To determine differences between companies, we computed individual OLS regression models for each company, similar to the model presented above but without the fixed effect. Due to space considerations we will not present the full results of our models here, but we found that the value of marketing information ranges, finding coefficients for marketing information positive and significant for some companies, and insignificant for others.

These differences may be related to several factors, but we expect that an important influence would be the extent to which a firm is engaged in marketing activities. This arises from the concept of diminishing marginal returns – as a firm expends more and more on marketing activities, the marginal returns on and value of information about these activities decreases. Firms with lower marketing spend should have a higher value of marketing information when compared to those with higher marketing spend. To test this assertion, we regress each firm's marketing value coefficient from its OLS model against the equity price to book value ratio and sales, general, and administrative expense (SGA) to revenue ratio for that firm. Equity price to book value represents the level of marketing assets owned by a firm, while SGA to revenue ratio represents the level of marketing expense [10]. Results in Table 2 show both have negative and significant correlations with the value of marketing information. This supports our assertion that firms with lower marketing spend experience a higher value of marketing information.

Table 2 – Influences of marketing value regression model parameters

	Model 1	**Model 2**	**Model 3**
Intercept	0.029***	0.048***	0.043***
Equity price to book value ratio	-0.0025**		-0.0020**
SGA to revenue ratio		-0.088**	-0.072**
Model fit (R square)	0.298	0.155	0.489

*** $p<0.001$, ** $p<0.01$

Conclusion: Recently both academic research and anecdotal evidence have implied that marketing is losing its importance in firms; however there has been little or no empirical analysis of how the value of marketing has changed over a significant time period, nor investigation into the factors which might influence this value over time and across varied firms. In this paper, we utilize information retrieval and text analysis techniques to identify marketing information in news events, and use this information to study the influence of marketing on abnormal stock market returns over the past 25 years. We apply a word sense disambiguation technique to a large corpus of news articles in order to classify words to their true basic meaning according to context and perform sentiment analysis on the articles. We then implement a modified TF-IDF procedure to create a topic dictionary for marketing, and using a cosine similarity measure, we determine the extent to which news regarding a specific firm at a specific time is marketing related. The outputs of these techniques are used to study the value of marketing information.

We find that marketing information does have a significant and positive value, albeit lower than that of financial information. Further, marketing information is more valuable when the variance of the stock market is high. Not surprisingly, this is reflected through anecdotes and experiences about marketing's low influence on the firm given the long period of stability and prosperity in the US market experienced until recent years. Given that markets are entering a more turbulent phase, our results suggest that marketing's influence is on the rise and may increase further. There are also large differences in this value between firms. Marketing is more valuable for firms with low marketing assets and expenses as reflected in low stock price to book value and low SGA to revenue ratios, respectively. This result has a simple interpretation, in that it reflects the diminishing marginal returns from investments in marketing.

These findings suggest that the simple assertion that marketing is losing its importance may have been true in the past but not necessarily so under current uncertain conditions. Aside from market uncertainty, other factors that affect the value of marketing to a firm include the level of marketing assets and spend. Further research is warranted to more precisely study effects of these and other factors which might influence the dynamics of marketing's value to the firm.

In this study, we have combined information retrieval as well as text and sentiment analysis concepts in a technique to identify the type and sentiment of information contained in news articles. We demonstrate this technique by addressing questions regarding the value of marketing to a firm. However, our technique is quite general in nature, and could be used to address other problems, such as financial event detection, topic identification in social media and other user created content, or evaluation of issues discussed in a political domain. We are exploring applications to additional issues such as these in ongoing research.

References:

[1] Baginski, S. P., J. M. Hassell and M.D. Kimbrough (2004), "Why do managers explain their earnings forecasts?" Journal of Accounting Research, 42 (1), 1-29.

[2] Cecchini, M., H. Aytug, G. J. Koehler, and P. Pathak (2010), "Making words work: Using financial text as a predictor of financial events." Decision Support Systems 50 (1), 164-175.

[3] Drucker, P. (1993), Management: tasks, responsibilities, and practices, Harper Collins: NY.

[4] Hurwitz, R. "Inquirer home page," retrieved 8/1/2012 at wjh.harvard.edu/~inquirer/.

[5] Hurwitz, R. "GI categories," retrieved 8/1/2012 at wjh.harvard.edu/~inquirer/homecat.htm.

[6] Homburg, C., J. P. Workman, Jr., and H. Krohmer (1999), "Marketing's influence within the firm," Journal of Marketing, 63 (2), 1-17.

[7] March, J. G. and R. I. Sutton (1997), "Organizational performance as a dependent variable," Organization Science, 8 (6), 698-706.

[8] Paşa, M. and S. M. Shugan (1996), "The value of marketing expertise," Management Science, 42 (3), 370-388.

[9] Sheth, J. N. and R. S. Sisodia (2005), "Does marketing need reform?" Journal of Marketing, 69 (October), 10-12.

[10] Simon, C. J. and M.W. Sullivan (1993), "The measurement and determinants of brand equity: A financial approach," Marketing Science, 12(1), 28-52.

[11] Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008), "More than words: quantifying language to measure firms' fundamentals." Journal of Finance, 63 (3), 1437-1467.

[12] Tetlock, P. C. (2007), "Giving content to investor sentiment: The role of media in the stock market." The Journal of Finance, 62 (3), 1139-1168.

[13] Verhoef, P. C. and P. S. H. Leeflang (2009), "Understanding the marketing department's influence within the firm," Journal of Marketing, 73 (1), 14-37.

[14] Webster, Jr., Frederick E., A. J. Malter and S. Ganesan (2005), "The decline and dispersion of marketing competence," MIT Sloan Management Review, 46 (4), 35-43.

Firm Clustering based on Financial Statements

Dazhi Chong, Hongwei Zhu
Old Dominion University, Norfolk, VA
Dchong@odu.edu, Hzhu@odu.edu

ABSTRACT

The classification and clustering of firms can help investors, consumers, and stakeholders to collect information, analyze behavior and predict potential market opportunities. Most existing approaches only pay attention to the operation processes and outputs of firms. In this paper, we introduce a firm clustering method that uses XBRL based financial information. With the adoption of XBRL, financial data has become increasingly available to allow for the application of a wide range of clustering methods. In this research, we construct the network of firms using the shared data elements in firms' financial statements. By implementing a spectral clustering method and applying it to the network, we identify clusters and characteristics of firms in the same cluster but classified as different industries according popular classification methods.

Keywords

XBRL, Spectral clustering, Financial Statement, NAICS.

INTRODUCTION

There are several ways to group firms. The most common method is to group them into different industries according to firm's core business. In the U.S., the Standard Industry Classification (SIC) codes, and more recently the North American Industry Classification System (NAICS) codes, have been widely used to classify a firm into a certain industry. These classification codes not only provide an efficient way for investors and policy-makers to collect and analyze economic data but also reveal the structure of economy which is useful in analyzing economic changes (US Census Bureau 2012a). However, the code assignments are somewhat static, not capturing the evolvement of firm's business and industry structure. For example, as firms expand or shift their business fields, the codes assigned to firms often do not accurately reflect the "natural" grouping of firms. Thus more "dynamic" and "efficient" classification approaches are required in today's highly competitive market.

When financial statements were not digitized or were in unstructured format, it was difficult to efficiently process them to derive useful information. This has been fundamentally changed since the recent adoption of eXtensible Business Reporting Language (XBRL) (XBRL International 2006). In the U.S., the Securities and Exchange Commission has adopted the GAAP Taxonomy as a data standard. Specified in XBRL, this Taxonomy defines a set of financial reporting concepts and their relationships. Starting 2009, the SEC mandated that all publicly traded companies must use the Taxonomy to provide their financial statements in XBRL. Firms are allowed to extend the Taxonomy to report items not specified in the GAAP Taxonomy. Earlier studies(Zhu and Wu 2011a; Zhu and Wu 2011b) show that although many firms extended the Taxonomy, on average 87% the reported data are defined in the Taxonomy. The use of the data standard and XBRL makes it possible for us to mine a large quality of financial statements to identify firm clusters.

In this research, we model firms and the GAAP Taxonomy elements used by the firms as a bipartite "social network". We implement a spectral clustering method and apply it to the "social network". Our results demonstrate the feasibility of using financial data to identify firm clusters.

RESEARCH METHOD

Data Collection

The data is collected from the archive of financial statements in XBRL format submitted to the SEC. Larger amounts of companies will facilitate clustering algorithm to achieve more accurate grouping results. In this research, we collect the 10-K annual financial statements from 2009 to 2011. If a firm has more than one 10-K, we choose to use the earliest one. Since there is no reliable method to match custom data elements introduced by the different companies, our analysis focuses on how companies use GAAP Taxonomy elements. In the rest of the paper, we use the term *tag* and GAAP *element* interchangeably. The dataset has 10-K's of 1799 firms, which together use 7021 GAAP elements.

In the financial statement, some commonly used tags such as "Assets", "CostofRevenue" are used by most companies. The strong relationship created by these commonly used tags will make most companies belong to the same cluster. Since our main objective is to cluster companies based on their usage of specific tags, it is necessary to ignore these commonly used tags. Therefore, we delete tags that are used by more than 50 companies. After removing these tags, 5815 elements were used by this research in the final dataset.

Clustering Approach

Clustering is an efficient way to explore information within certain networks or groups. Many clustering algorithms are based on the assumption that the data within the dataset has specific attributes or links (Zha et al. 2001). By converting these attributes or links to *n*-dimensional vectors, we can cluster the dataset based on the relationship among vectors. Our objective is to identify the company clusters by analyzing the links between companies and financial statement items. Each company's 10-K uses a number of tags specified in the GAAP Taxonomy. Likewise, each tag may also be used by several companies. By treating tags as attributes of companies, the company-tag relationship can be represented as an m-by-n matrix A. Here, A_{ij} represents the relationship between the i-th company and the j-th tag.

There are numerous clustering algorithms, such as agglomerative clustering (Voorhees 1986) and k-means algorithm (Dhillon and Modha 2001), that can be used to identify clusters. However, most of them can't guarantee global optimization of clusters (Shi and Malik 2000). Spectral clustering addresses this deficiency. The objective of spectral clustering is to find the partition of a graph so that the linkages between groups are minimized and the linkages within groups are maximized. In graph language, the linkages among groups are called "cuts", which can be computed through the total weight of the edges between connected groups. Shi and Malik (Shi and Malik 2000) suggested a co-clustering algorithm to minimize cuts globally. They argued that second smallest eigenvalue of Laplacian matrix can be used to find the minimum cut vertex partitions in a graph. Dhillon (Dhillon 2001) extended this algorithm. By using the singular value decomposition (SVD) approach, he found that the second left and right singular vectors of a normalized matrix provide an optimal solution for co-clustering problem.

Dhillon's approach is more efficient and can cluster tags and companies simultaneously. We choose to use this approach in this paper. In order to use Dhillon's approach to analyze our dataset, the firm-tag matrix is defined as follows: an undirected bipartite graph is W=(F,T,E) where F={f1,f2....,fm}, T={t1,t2....,tn} are two sets of vertices which represent firms and Taxonomy tags. In our research, there are 1799 firms and 5815 tags, so there are 1799 elements in F and 5815 elements in T. E is the set of edges between F and T, an edge <fi,tj> exists when firm fi uses tj tag. For simplicity, we set the weight of each edge as 1 and the edges are undirected.

Our clustering algorithm uses k-means algorithm on the singular vectors to obtain the desired clusters, thus we need to determine the best value for k. Numerous approaches have been developed to resolve this problem. Based on the various aspects of cluster validity, Schaeffer (2007) considered three measurements to evaluate the fitness of the k value. Among these measurements, relative criterion is used to measure the goodness of inter-cluster density. Local criterion focuses on the goodness of a clustering structure without external information. In order to optimize both local density and global density, Schaeffer (2007) used the product of the local and relative densities to measure the fitness of cluster function. This approach provides an easy way to optimize k value in this research.

FINDINGS

Using the approach mentioned above, we cluster the 1799 companies into 20 clusters. We use Schaffer's method and find it is optimal when k=20. Then we try to reveal potential relationship among firms and compare the clusters with standard classification codes. Despite that NAICS is more up to date and has been used by US statistical agencies to classify businesses and analyze the U.S. economy (US Census Bureau 2012b), firms still use the SIC codes in their SEC filings. We use NAICS in this study. Mappings between SIC and NACICS are used to obtain each firm's NAICS code. Then we use the first two digits of NAICS code to distinguish major industry groups. The firms belong to 24 major industry groups according to NAICS. See Table 1 for the distribution of firms among these industry groups.

Table 1. Firm Distribution among Industry Groups according to NAICS

NAICS	INDUSTRY	# Firms
11	Agriculture, Forestry, Fishing and Hunting	5
21	Mining, Quarrying, and Oil and Gas Extraction	110
22	Utilities	98
23	Construction	28
31	Manufacturing (Good, Beverage and Tobacco Product, Apparel and Leather and Allied Product Manufacturing, Textile, Textile Product Mills)	96
32	Manufacturing (Wood Product, Paper, Petroleum and Coal Products, Chemical, Plastics and Rubber Products and Nonmetallic Mineral Product Manufacturing, Printing and Related Support Activities)	195
33	Manufacturing (Primary Metal, Fabricated Metal Product, Machinery, Computer and Electronic Product, Electrical Equipment, Appliance, and Component , Transportation Equipment , Furniture and Related Product and Miscellaneous Manufacturing)	410
42	Wholesale Trade	54
44	Retail Trade (Motor Vehicle and Parts Dealers, Furniture and Home Furnishings Stores, Electronics and Appliance Stores, Building Material and Garden Equipment and Supplies Dealers, Food and Beverage Stores, Health and Personal Care Stores, Gasoline Stations, Clothing and Clothing Accessories Stores,	61
45	Retail Trade (Sporting Goods, Hobby, Book, and Music Stores, General Merchandise Stores, Miscellaneous Store Retailers, Nonstore Retailers)	29
48	Transportation and Warehousing(Air, Rail, Water, Truck , Transit and Ground Passenger , Pipeline, Scenic and Sightseeing and Support Activities for Transportation)	64
49	Transportation and Warehousing(Postal Service, Couriers and Messengers, Warehousing and Storage)	3
51	Information	82
52	Finance and Insurance	330
53	Real Estate and Rental and Leasing	23
54	Professional, Scientific, and Technical Services	80
55	Management of Companies and Enterprises	9
56	Administrative and Support and Waste Management and Remediation Services	31
61	Educational Services	12
62	Health Care and Social Assistance	23
71	Arts, Entertainment, and Recreation	24
72	Accommodation and Food Services	23
81	Other Services (except Public Administration)	8
92	Public Administration	1
Total		1799

The clusters and distribution of NAICS industry groups are shown in Figure 1.

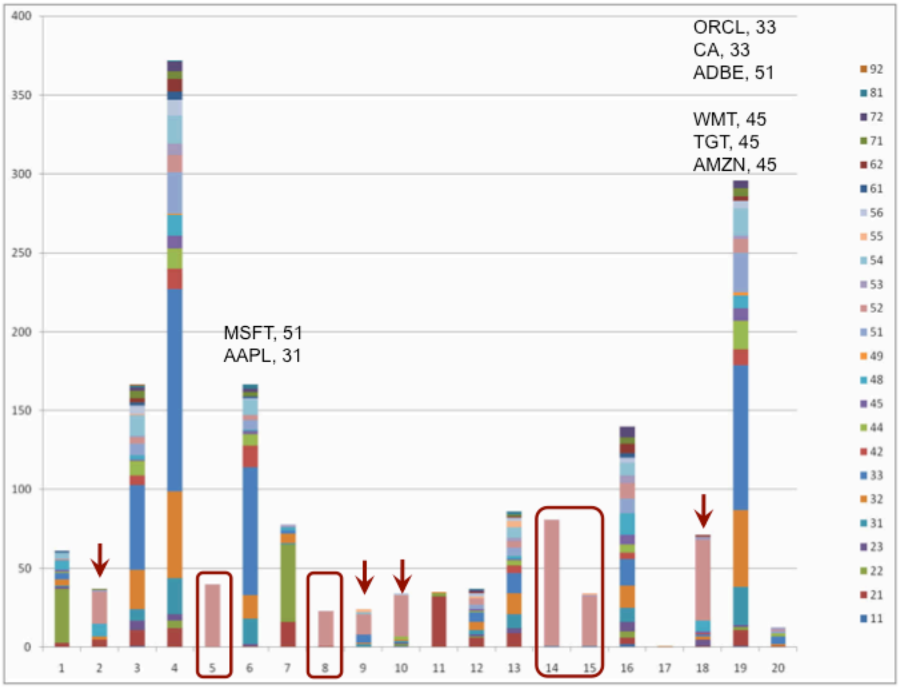


Figure 1. The distribution of major industry in 20 clusters

Cluster 17 contains only one firm, Verde Resources Inc (VRI), an exploration stage company. After removing popular tags, it only has seven tags, among which only five tags are also used by firms in other clusters. In addition to this outlier case, cluster 5 is also "pure", containing only "Finance and Insurance" (NAICS code 52) firms. Clusters 8, 14, and 15 are also nearly pure with predominantly Finance and Insurance firms. Such firms are also the majority of clusters 2, 9, 10, and 18. An examination of the firms in these clusters shows that firms within each cluster have significant similarities in their core businesses. For example, firms in cluster 5 are primarily insurance companies, whereas firms in cluster 8 are primarily commercial banks. The majority of the Finance and Insurance firms in cluster 2 invest in real estate. In addition, Utilities (code 22) firms are dominant in clusters 1 and 7. Similarly, Mining, Quarrying, Oil, Gas Extraction (code 21) firms are dominant in cluster 11.

When clusters and industries according to NAICS codes do not coincide, firms in the same cluster do have certain meaningful similarities. For example, we know that Microsoft and Apple are competitors, and in our result, they are in the same cluster (#6). However, they are in different industries according to their NAICS codes: Microsoft is 51 (Information) and Apple is 33 (Manufacturing). The two companies share five tags. In contrast, despite the fact that Microsoft and Oracle also compete in data management software space, Oracle does not complete in personal computing business and it also has a sizeable server business after its acquisition of Sun. Microsoft and Oracle have 2 shared tags. As a result, they belong to different clusters.

Further research is still needed to understand the clustering result (e.g., why Oracle is in the same cluster as Adobe and a number of retailers).

This case study leads us to hypothesize that certain firm behaviors (such as operation mode, investment strategy) determine the contents and structures of firm financial statements. Conversely, the contests and structures should help us infer firm characteristics. To preliminarily test our hypothesis, we analyze the tags used by firms within each cluster. We rank the tags according to the number of firms that use them. The frequently used tags may indicate major financial behaviors in these clusters. Table 2 lists the top 10 tags in selected clusters (the list does not include the removed elements – those used by more than 50 firms).

Table 2. Frequent Elements in Clusters

Cluster	Frequently Used Elements
1	AccumulatedDeferredInvestmentTaxCredit,PublicUtilitiesDisclosureTextBlock,RegulatoryAssetsCurrent RegulatoryLiabilityCurrent, RegulatoryLiabilities, AdditionalCollateralAggregateFairValue, RegulatoryAssets , ScheduleOfJointlyOwnedUtilityPlantsTextBlock, JointlyOwnedUtilityPlantProportionateOwnershipShare, JointlyOwnedUtilityPlantOwnershipAmountOfPlantAccumulatedDepreciation. DebtInstrumentUnamortizedDiscountPremiumNet, PublicUtilitiesPolicyTextBlock
2	PartnersCapital, GeneralPartnersCapitalAccount, LimitedPartnersCapitalAccount, PartnersCapitalAccountDistributions, LimitedPartnersCapitalAccountUnitsOutstanding NetIncomeLossAllocatedToLimitedPartners , PartnersCapitalNotesDisclosureTextBlock, LiabilitiesAndPartnersCapital, NetIncomeLossPerOutstandingLimitedPartnershipUnit NetIncomeLossAllocatedToGeneralPartners , WeightedAverageLimitedPartnershipUnitsOutstanding
5	IncreaseDecreaseInUnearnedPremiums, IncreaseDecreaseInPremiumsReceivable, PrepaidReinsurancePremiums, ReinsurancePayable IncreaseDecreaseInDeferredPolicyAcquisitionCosts, NetInvestmentIncome SupplementalScheduleOfReinsurancePremiumsForInsuranceCompaniesTextBlock， SupplementaryInsuranceInformationForInsuranceCompaniesDisclosureTextBlock， IncreaseDecreaseInReinsuranceRecoverable , PremiumsReceivableAtCarryingValue
6	ScheduleOfProductWarrantyLiabilityTableTextBlock, FutureAmortizationExpenseAfterYearFive ScheduleOfAccruedLiabilitiesTableTextBlock, ScheduleOfDebtTableTextBlock BusinessAcquisitionsProFormaRevenue, LiabilitiesFairValueDisclosure BusinessAcquisitionsProFormaNetIncomeLoss, ShareBasedCompensationArrangementByShareBasedPayme. ShareBasedCompensationArrangementByShareBasedPayme, StandardProductWarrantyPolicy

The tags in cluster 1 indicate that firms are regulated public utilities, some of which are jointly owned (JointlyOwnedUtilityPlantProportionateOwnershipShare is used by more than 15 firms in cluster 1). The tags in cluster 2 indicate that most firms are partnerships. The tags in cluster 5 indicate that insurance premiums are important to insurance business and the firms are also in the reinsurance business. The tags in cluster 6 indicate that firms sell products with warranty, use elaborate financing and compensation methods, and are engaged in acquisition of other businesses. All these findings are useful but cannot be derived in any way from firms' NAICS codes. These preliminary findings are very promising to support our hypothesis. Since there are hundreds of frequently used tags in 20 clusters, we plan to analyze them all to better understand the clustering results and fully test our hypothesis.

CONCLUSION

Prior research attempted to identify firm groups based on the operation process and output of firms. The interrelation and interdependence among firms were not captured. In this paper, we introduce a spectral clustering method to cluster firms based on their financial statements. These XBRL-based statements are created using a shared data standard, the GAAP Taxonomy. When constructing their financial statements, firms select data elements from the Taxonomy to tag their financial data. Firms that tend to use the same set of tags have similarities in their businesses. Similar to other "social networks", this statement-tag network has natural groupings of firms' financial statements. Our work demonstrates the feasibility of clustering firms based on their financial statements and shows that clusters exhibit interesting common features of the firms within the same cluster. Tag analysis also reveals firm characteristics that cannot be derived from existing industry classification schemes. Since financial statements are generated quarterly, this approach has the potential to produce insights into industry dynamics.

There are several areas that need further work. First, preliminary tag analysis shows that firms in the same cluster have similar characteristics. We still need to systematically examine the frequent tags within each cluster. Second, we have looked at the tags used by firms within the same cluster, but we have not examined the tag clusters. We need to analyze the corresponding tag clusters. Third, in addition to tag analysis, we also need to analyze a number of other characteristics (e.g., revenue, return on asset, and profit) to develop further understanding of the firms in the same cluster. Fourth, customized data elements introduced by firms are not included in the dataset. Future research will include these data elements, which are expected to provide additional information on firm's operational behavior. Nevertheless, this preliminary has revealed great potentials of using clustering methods on financial data to understand and predict firm characteristics. The findings are also useful to data standards development organizations. For example, although the GAAP Taxonomy is currently organized by industry based on SIC and NAICS codes, our findings suggest that it should also examine commonalities of firms and structure the Taxonomy according to similar usage requirements of firms.

REFERENCES

Dhillon, I.S. 2001. "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," *the Seventh ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 269-274.

Dhillon, I.S., and Modha, D.S. 2001. "Concept Decompositions for Large Sparse Text Data Using Clustering," *Machine learning* (42:1), pp. 143-175.

Schaeffer, S.E. 2007. "Graph Clustering," *Computer Science Review* (1:1), pp. 27-64.

Shi, J., and Malik, J. 2000. "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (22:8), pp. 888-905.

US Census Bureau. 2012a. "Conceptual Issues of Economic Classification." Retrieved 02-20, 2012, from http://www.census.gov/eos/www/naics/history/docs/issue_paper_1.pdf

US Census Bureau. 2012b. "Introduction of NAIC." Retrieved 02/20, 2012, from http://www.census.gov/eos/www/naics/

Voorhees, E.M. 1986. "The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval," *Dissertation Abstracts International Part B: Science and Engineering[DISS. ABST. INT. PT. B-SCI. & ENG.]* (47:2).

XBRL International. 2006. "Extensible Business Reporting Language (XBRL) 2.1," XBRL International.

Zha, H., He, X., Ding, C., Simon, H., and Gu, M. 2001. "Bipartite Graph Partitioning and Data Clustering," *the Tenth International Conference on Information and Knowledge Management*, New York, USA: ACM, pp. 25-32.

Zhu, H., and Wu, H. 2011a. "Interoperability of XBRL Financial Statements in the U.S.," *International Journal of E-Business Research* (7:2), pp. 18-33.

Zhu, H., and Wu, H. 2011b. "Quality of Data Standards: Framework and Illustration Using XBRL Taxonomy and Instances," *Electronic Markets* (21:2), pp. 129-139.

Exploring the Informativeness of Forward-Looking Statements in Public Firm Disclosures

Zhimin Hua, Xiaoyu Wu, J. Leon Zhao
Department of Information Systems, City University of Hong Kong
{zmhua2, xiaoyuwu5}@student.cityu.edu.hk, jlzhao@cityu.edu.hk

Abstract: As a major communication medium between firm managers and investors, Forward-Looking Statements (FLS) in Management Discussion and Analysis (MD&A) sections in financial reports play an important role in reducing information asymmetry between firms and investors. However, MD&A section is not audited currently and the FLS in MD&A are subjective to management's incentives and other concerns to some extent. This study examines the informativeness of the FLS in the MD&A sections of public firm disclosures. We extract and classify FLS and Non-Forward-Looking Statements (NFLS) in the MD&A sections of annual reports, i.e., 10-K filings, by means of textual analytical methods and tools. We find that for firms with more MD&A disclosures, smaller relative ratios of forward-looking disclosures to non-forward looking disclosures have better future earnings. This finding shows the firm-level informativeness of forward-looking disclosures signaled in the MD&A sections of SEC filings.

1. Introduction

According to the Securities and Exchange Commission (SEC), all the public-listed firms are required to provide Management Discussion and Analysis (MD&A) in the financial statements. The major components of MD&A include narrative discussions from management's perspective about the firm's past performance, financial conditions, and future prospects. MD&A was believed as an important component of financial reports. Forward-looking statements in MD&A refer to the sentences that discuss the future trends, uncertainties, events, etc. about the firm, industry or market. Accordingly, those discuss past information or facts are usually referred to as non-forward-looking statements (NFLS) (Li 2010; Muslu et al. 2011). As textual disclosures, MD&A is not audited although required by SEC. FLS is not required but encouraged by SEC to reduce the information asymmetry between the firm and investors.

There is extensive research on the MD&A disclosures, especially with the prevalence of textual analytics and text mining techniques in business research. Pioneering work include (Li 2008), which examined the readability of annual reports with computational linguistics techniques; and (Li 2010), which analyzed the tone of FLS in MD&A with Naïve Bayesian Classifier to determine the tone of each FLS sentence. Brown and Tucker (2011) employed Vector Space Model (VSM) and TFIDF to assess the year-to-year modifications in firms' MD&As. Muslu et al. (2011) examined firm's information environment, including the availability of supplemental corporate information and the attainment of earnings thresholds, and its implications on the levels of forward-looking disclosures. A multi-label text classification algorithm (ML-CKNN) was introduced in (Huang and Li 2008) for labeling risk factors in SEC filings. A new negative word list for tone and sentiment analysis in financial contexts was developed in (Loughran and McDonald 2011). They found that about 75% of negative word counts in SEC 10-K filings using the commonly-used Harvard dictionary are not negative in the financial context.

Because MD&A section is not audited and the FLS in MD&A are subjective to management's incentives and political or legal concerns to some extent, the informativeness of forward-looking statements is a big concern of investors (Li 2008; Skinner 1994). Whether the forward-looking statements in MD&A release singles about a company's future performance and whether FLS reduce the information asymmetry between the firm and investors are worth careful studying. Our research is motivated by this concern.

Based on existing research as mentioned above, our research examines the informativeness of the amount of forward-looking information and the relative ratio of FLS to NFLS in a filing to future earnings. Empirically, we first extract MD&A sections from SEC filings and classify forward-looking and non-forward-looking information in them with textual analytical techniques. Then we test the implications of absolute and relative forward-looking disclosures on firm's future earnings. We find that firms with more MD&A disclosures (both forward-looking and non-forward looking), smaller relative ratio of forward-looking information to non-forward looking information have better future earnings. The results suggest the firm-level informativeness of forward-looking disclosures signaled in the MD&A sections of SEC filings.

2. Research Design

Since the MD&A is a quasi-mandatory disclosure (Muslu et al. 2011), keyword-based methods are proved to be effective in previous literature for classifying forward-looking disclosures (Brown and Tucker 2011; Li 2010). Figure 1 illustrates the general framework of our research design.

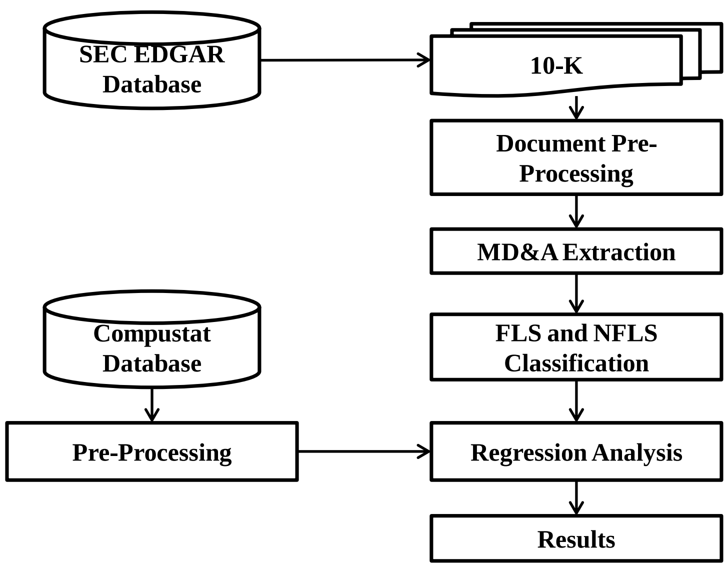


Figure 1. General Framework of the Research Design

In this study, we use customized textual processing techniques, which are implemented with Perl language. First, we downloaded all the 10-K filings from SEC EDGAR Database between 2000 and 2011. Then we extracted the MD&A sections in the filings and removed the HTML tags. This is done with Perl script by searching key words such as "Item 7", "MD&A", "Management Discussion and Analysis", etc.

Next we got sentences from the MD&A sections and classified them in into FLS and NFLS.

Sentences will be categorized as FLS if they meet at least one of the following criteria (Muslu et al. 2011): (1) The sentence includes keywords "will" or "future". (2) The sentence uses adjectives "next", "subsequent", "following", "upcoming", "incoming" or "coming" which is followed by "month", "quarter", "year", "fiscal" or "period". (3) The sentence includes any of the following verbs or their conjugations: "aim", "anticipate", "assume", "commit", "estimate", "expect", "forecast", "foresee", "hope", "intend", "plan", "project", "seek", "target". If sentences do not satisfy any of the criteria, they will be tagged as NFLS. This method is not perfect, but it works effectively in our setting. After the classification, we conduct a manual validation to check the classification effectiveness. We randomly select sentences from 20 MD&A sections and let two Research Assistants do a manual classification. The precision, recall and overall accuracy are 0.71, 0.72 and 0.97, respectively. Then we count the numbers of FLS and NFLS and calculate the ratio of the number of FLS to the number of NFLS in MD&A for each 10-K filing. After that, we stored relevant data of these filings (i.e., Central Index key, Report Date, number of FLS and number of NFLS) into database.

Finally, we downloaded other fundamental annual data of firms from Compustat Database. The following data are included: Central Index Key, Fiscal Year, Earnings Before Interest and Taxes, Total Asset, Operating Cash Flow, Total Current Liabilities, Common Shares, Closing Stock Price (at the end of the fiscal year), Total Liability, Special Items, etc. We then merged the two database tables by matching Central Index Key and Report Date in the two database tables. To form the final data set, we exclude records which have less than five sentences in MD&A or include null value in any item.

3. Data Analysis

Our final dataset include 29161 firm-year records after matching two databases and excluding extreme or null value items. Table 1 shows the descriptive statistics of key variables.

Table 1. Descriptive Statistics

Variable	Mean	STDEV	Minimum	P25	Median	P75	Maximum
FLS	35.890	26.328	0.000	18.000	31.000	48.000	316.000
NFLS	297.430	190.377	3.000	173.000	270.000	387.000	2842.000
FLSRATIO	0.118	0.056	0.000	0.084	0.114	0.148	0.762
SNO	333.320	212.560	5.000	193.000	302.000	435.000	3158.000
EARN	-0.353	13.919	-2106.200	-0.059	0.052	0.109	35.917
EARN2	-0.392	11.042	-954.000	-0.056	0.053	0.108	17.627
CFRATIO	0.288	30.184	-859.000	-0.063	0.305	0.697	4444.000
MTB	8.247	454.488	0.006	1.109	1.532	2.415	68563.000
SI	-0.045	2.936	-338.750	-0.014	0.000	0.000	224.083
SIZE	2473.303	13304.825	0.000	52.216	245.070	1029.631	516887.720

The first column of Table 1 illustrates the variables used in data analysis. FLS measures the number of forward-looking sentences of the MD&A section in a 10-K filing. If a sentence is not a forward-looking sentence, it belongs to non-forward-looking sentence. NFLS is the number of non-forward-looking sentences of the MD&A section in a 10-K report. FLSRATIO is the ratio of FLS to NFLS and SNO measures total number of sentences in MD&A (which is the sum of FLS and NFLS). On average, there are 35.89 FLS and 297.43 NFLS in a MD&A section of a filing (i.e., averagely 333.32 sentences in a MD&A). The mean of FLSRATIO is 0.1176, indicating FLS counts for a small portion of a MD&A. EARN is the earnings (before interest and taxes) of a firm in the fiscal year. EARN2 represents the earnings (before interest and taxes) of the firm in the next fiscal year. Both EARN and EARN2 are scaled by firm's total assets. CFRATIO is the ratio of annual cash flow from operations to book value of current liability. SIZE measures the scale of a company which is the market value of equity at the end of the fiscal year (Common Shares times Annual Closing Price) and LOGSIZE is the natural logarithm of SIZE. MTB is the ratio of the sum of SIZE and total liabilities to total assets. SI is the ratio of the amount of special items reported for the year to total assets.

Table 2 shows the Pearson correlations for the aforementioned variables. There is strong positive correlation between FLS and NFLS (0.882), thereby also strongly correlated with SNO. It is intuitive that more FLS is accompanied with more NFLS discussion, and vice versa. The positive correlation between FLS and FLSRATIO (0.489) shows that when there are more FLS discussions in a MD&A, the FLS-to-NFLS ratio is higher. This suggests that the proportion of FLS content increases at a greater rate than NFLS content, for larger values of SNO.

Table 2. Correlation Matrix (p -values in parentheses)

	FLS	NFLS	EARN	EARN 2	CFRA TIO	SIZE	MTB	SI	LOG SIZE	FLSRA TIO	SNO
FLS	1										
NFLS	0.822 (0.000)	1									
EARN	0.016 (0.008)	0.023 (0.000)	1								
EARN 2	0.025 (0.000)	0.035 (0.000)	0.231 (0.000)	1							
CFRA TIO	-0.009 (0.107)	-0.003 (0.571)	0.003 (0.562)	0.004 (0.484)	1						
SIZE	0.007 (0.22)	0.058 (0.000)	0.006 (0.274)	0.009 (0.139)	0.003 (0.638)	1					
MTB	-0.01 (0.089)	-0.014 (0.017)	-0.175 (0.000)	-0.09 (0.000)	0.000 (0.951)	-0.002 (0.714)	1				
SI	-0.001 (0.799)	0.003 (0.608)	0.094 (0.000)	0.042 (0.000)	0.003 (0.584)	0.002 (0.736)	0.26 (0.000)	1			
LOG SIZE	0.199 (0.000)	.333 (0.000)	0.054 (0.000)	0.069 (0.000)	0.014 (0.015)	0.377 (0.000)	-0.012 (0.046)	0.02 (0.001)	1		
FLSR ATIO	0.489 (0.000)	.086 (0.000)	-0.02 (0.001)	-0.034 (0.000)	-0.021 (0.000)	-0.065 (0.000)	0.025 (0.000)	-0.009 (0.132)	-0.174 (0.000)	1	
SNO	0.861 (0.000)	0.998 (0.000)	0.022 (0.000)	0.034 (0.000)	-0.004 (0.48)	0.053 (0.000)	-0.014 (0.019)	0.003 (0.668)	0.323 (0.000)	0.138 (0.000)	1

We develop our models by extending previous literature findings (Brown and Tucker 2011; Li 2010; Muslu et al. 2011). Table 3 reports the OLS regression results for EARN2 (future earnings) regressed on EARN (current earnings), CFRATIO (liquidity), LOGSIZE (firm size), SI (special items amount), MTB (market-to-book ratio), as well as the FLS, NFLS, SNO, and FLSRATIO. The first three models test the absolute numbers of FLS and NFLS to the future earnings. All the three coefficients are positive and significant at 10% level, which confirms the positive correlations between EARN2 and FLS (NFLS or SNO). These results show that more MD&A disclosures indicate predict more future earnings. Model (4) tests the FLS-to-NFLS ratio with future earnings. The negative coefficient -3.635 with a t-statistic -3.194 indicates that relatively, more forward-looking information disclosures suggest worse future earnings. Model

(5) considers both the absolute amount of MD&A sentences and the relative forward-looking ratio. The result is consistent with previous models, as indicated by the positive SNO coefficient and negative FLSRATIO coefficient, while the significance level improved to 1%. We also examined temporal effects by including years-dummies but turned out to be not significant.

Table 3. Regression Models

	DV: EARN2				
	(1)	(2)	(3)	(4)	(5)
	Coefficient (t-statistic)				
Intercept	-1.910*** (-10.967)	-1.907*** (-11.071)	-1.912*** (-11.062)	-1.299*** (-5.646)	-1.349*** (-5.846)
EARN	0.170*** (36.625)	0.170*** (36.626)	0.170*** (36.626)	.170*** (36.619)	0.169*** (36.610)
CFRATIO	0.001 (.451)	0.001 (.445)	0.001 (.446)	0.001 (.370)	0.001 (.384)
LOGSIZE	0.265*** (9.290)	0.256*** (8.639)	0.257*** (8.687)	0.260*** (9.134)	0.231*** (7.598)
SI	0.136*** (6.085)	0.136*** (6.080)	0.136*** (6.081)	0.135*** (6.043)	0.135*** (6.042)
MTB	-0.001*** (-10.158)	-0.001*** (-10.150)	-0.001*** (-10.151)	-0.001*** (-10.090)	-0.001*** (-10.053)
FLS	0.004* (1.752)				
NFLS		0.001* (1.910)			
SNO			0.001* (1.929)		0.001*** (2.655)
FLSRATIO				-3.635*** (-3.194)	-4.280*** (-3.678)
Observations	29161	29161	29161	29161	29161
Adjusted R^2	0.061	0.060	0.060	0.060	0.061

Note: *** indicates $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

4. Discussions

The empirical tests and analyses in previous section confirm the implications of forward-looking disclosures on future earnings. Specifically, the absolute amount of forward-looking information has positive implication on future earnings. This is consistent with existing findings which examine MD&A and future performance (Brown and Tucker 2011; Li 2010; Muslu et al. 2011).

While the relative measure of FLS vs. NFLS indicates a negative relation with future earnings. MD&A is not audited and subjective to the perceptions and concerns of management. More forward-looking discussions in MD&A are often negative, which is consistent with tone analysis of MD&A (Li 2010) and other “voluntary” disclosure (Skinner 1994). This suggests that higher forward-looking ratio is informative on negative futures and thereby possibly worse future earnings.

5. Concluding Remarks

In this research, we examine the informativeness of forward-looking disclosures as signaled in the MD&A sections of SEC filings. The findings show that firms with more MD&A disclosures (both forward-looking and non-forward looking) have better future earnings. Further, a high relative ratio of forward-looking information to non-forward looking information indicates worse future earnings. This study extends current literature by examining the firm-level informativeness of forward-looking disclosures (both absolute and relative measure) on firm's future earnings.

Next, we will extend our study by analyzing the theoretical implications of this work. More empirical analyses would be conducted on the dataset. The current empirical model could be improved by considering more factors. The goodness of fit of the models can be further improved. Furthermore, we will examine the industry-level and market-level informativeness of the forward-looking disclosures in public filings. In the future, we will further explore the relationship between analyst forecast and forward-looking disclosures. Besides, we will also check market valuation as the barometer to see any insightful results.

Acknowledgement

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 124510).

References

Brown, S.V., and Tucker, J.W. 2011. "Large-Sample Evidence on Firms' Year-over-Year Md&a Modifications," *Journal of Accounting Research* (49:2), pp 309-346.

Huang, K.-W., and Li, Z. 2008. "A Multilabel Text Classification Algorithm for Labeling Risk Factors in Sec Form 10-K," *ACM Trans. Manage. Inf. Syst.* (2:3), pp 1-19.

Li, F. 2008. "Annual Report Readability, Current Earnings, and Earnings Persistence," *Journal of Accounting and Economics* (45:2), pp 221-247.

Li, F. 2010. "The Information Content of Forward-Looking Statements in Corporate Filings—a Naïve Bayesian Machine Learning Approach," *Journal of Accounting Research* (48:5), pp 1049-1102.

Loughran, T.I.M., and McDonald, B. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance* (66:1), pp 35-65.

Muslu, V., Radhakrishnan, S., Subramanyam, K., and Lim, D. 2011. "Firm's Information Environment and Forward-Looking Disclosures in the Md&a, Working Paper Http://Www.Utdallas.Edu/~Vmuslu/Msrl%20(2011).Pdf."

Skinner, D.J. 1994. "Why Firms Voluntarily Disclose Bad News," *Journal of Accounting Research* (32:1), pp 38-60.

Optimization-based Decision Support for Scenario Analysis in Sourcing Markets with Economies of Scale and Scope

Christian Hass*, Martin Bichler*, and Kemal Guler#
*TU München, Germany
#HP Labs, Palo Alto, USA

Abstract

Procurement markets typically exhibit scale economies leading to various types of volume discounts which are in wide-spread use in practice. Scenario analysis describes a process in which procurement managers compute different award allocations as a result of different allocation constraints and parameters that they put in place. This paper presents simulation results considering instance sizes, which can be solved to optimality, imposing a reasonable time limit for scenario analysis. Additionally it discusses an optimization model and computational methods, which allows for effective scenario analysis with allocation problems in the presence of different types of discount policies and allocation constraints. The model we propose reduces the number of parameter settings to explore considerably.

1 Introduction

Volume discounts are ubiquitous in procurement negotiations and a result of economies of scale that a supplier experiences. Such scale economies can arise from lower input costs to reductions in unit cost as the size of a facility and the usage levels of other inputs increase. Whereas economies of scale primarily refer to efficiencies associated with supply-side changes, economies of scope refer to efficiencies primarily associated with demand-side changes, such as increasing or decreasing the scope of marketing and distribution, of different types of products.

There is a growing literature on combinatorial auctions (CA) in the IS literature (Bichler et al. 2011, Adomavicius et al. 2012, Scheffel et al. 2011). CAs can be seen as the most general type of market mechanism as they allow bidders to express general valuations including substitutes and complements. Combinatorial auctions can well be used for procurement decisions with multiple line items and large volumes of each. A typical example would be the procurement of different types of semiconductors for the next two years for a hardware manufacturer. However, the number of possible bundle bids grows super-exponentially. Compact bid languages provide a remedy for such markets.

Davenport and Kalagnanam (2000) were among the first authors to discuss volume discount bids. Their bidding language requires suppliers to specify continuous supply curves for each item, allowing them to express economies of scale. They apply discounts only to the additional units above the threshold of a specific price interval, whereas tiered bids are valid for the entire volume of goods purchased (Goossens et al. 2007).

The problem gets computationally hard as purchasing managers typically need to take multiple allocation constraints into account and the cost minimal solution across multiple line items needs to be found. For example, they want to set lower and upper bounds for the number of suppliers or for the total spend awarded to a single supplier or a group of suppliers. In addition, suppliers often provide rebates for a certain overall volume, such as a 3% rebate if they sell items for more than a million. Lump sum rebates are a possibility to express economies of scope across multiple items. In most real-world procurement negotiations different suppliers use different discount policies including incremental, and total quantity discounts, and lump sum rebates, and a procurement manager needs to take all of them into account in an award decision.

Bichler et al. (2011) present an optimization formulation allowing for incremental and total quantity discounts, as well as lump sum rebates based on total spend or volume on different items. They formulate the allocation problem as a mixed integer program (MIP), and analyze the computational complexity and the empirical hardness of this problem. The formulation allows for a rich bidding language and a large number of discount policies that can be found in procurement practice. This MIP has already been used as decision support in a number of high-stakes procurement events in the field. With the use of such optimization approaches in practice new problems arise, which require decision support for procurement managers.

1.1 Decision Support for Scenario Analysis

Typically, complex procurement decisions are not met with a single optimization. Procurement managers often need to consider various allocation constraints such as quantity constraints to diversify risk, or limits on the number of winners to consolidate the set of suppliers. These side constraints are not always strict. But procurement managers cannot price out the value for such constraints in advance. Therefore they usually conduct an iterative process, which starts with an initial set of constraints. After solving the problem, they reflect on the solution, modify the constraints and resolve it with different parameters until they find a suitable allocation. This process has been referred to as *scenario navigation* or *scenario analysis* (Boutilier et al. 2004) and requires re-optimization of different versions of the problem with different constraints interactively in minutes. Such an interactive usage poses tight time constraints of a few minutes only on the solution time of each optimization run.

One of the problems in scenario analysis is the large number of parameters to explore. Typically, the parameters of interest are the right hand sides of the allocation constraints, more specifically, bounds on the number of winners, the volume awarded per bidder or the demand of the buyer. Procurement managers want to learn about parameters, or yield a total cost below a certain budget constraint. Due to the fact that dual variables cannot readily be interpreted in a mixed integer program, there is little guidance for procurement managers as to which constraints have the biggest impact on total cost or which combination of constraints they should relax in order to achieve a certain total spend. Therefore, scenario analysis often leads to many optimization runs with many different combinations of right-hand side parameters. So far, we do not know of literature on decision support tools for scenario analysis of this sort.

1.2 Contribution

Our contribution is two-fold: First, we present an experimental evaluation about problem-sizes, which can be solved to optimality when we impose a 10 minutes time limit. We will see that only small instances can be solved. Manual scenario analysis is very restricted under these circumstances. Therefore we will introduce a new formulation of the winner determination problem (SQSSA) that allows for effective scenario analysis. The formulation allows specifying an initial set of right-hand side parameters for the winner determination problem and identifying the binding business rules. So, rather than having to solve many optimization problems with different parameter settings, the procurement manager uses optimization to find the best solution with a minimal violation of the initial right hand sides. The model is in the spirit of earlier literature on parametric integer programming (Bradley et al. 1977) and reduces the effort for procurement managers to find appropriate scenarios significantly.

2 Experimental evaluation of problem sizes

For optimization-based decision support of this sort is important to understand problem sizes, which can be solved exactly. In what follows, we examine exact branch-and-cut approaches focusing on instances with selected side constraints. Bichler et al. (2011) performed an extensive evaluation of runtime and MIP gaps, but they placed less emphasis on the impact of allocation constraints. Such allocation constraints often reduce the problem instances that can be solved exactly within 10 minutes to small ones with up to 5 items and bidders.

2.1 Experimental setup

To analyze the impact of allocation constraints we imposed additional constraints putting limits to the number of winners or the overall spend per supplier. In this paper we will focus on limiting the number of winners to 3. Other treatments yielded similar or even worse results. We used the same multi-product cost function as in Bichler et al. (2011), but instead of 5 fixed tier sizes, we allowed bidders to individually specify up to 10 discount intervals. This is motivated by the fact that it is often difficult in procurement practice to predefine the discount intervals for all bidders and bidders want to have flexibility in their discount policy. The generation of bids is then based on minimizing the deviation between bid- and cost-curves. Details of the cost functions and the bid generation process as well as the software and instances used are available on demand. All experiments were conducted on an iMac with a 2.9 GHz Intel Core i7 CPU (Quad Core) and 8 GB RAM using the 64-bit operating system Mac OS X 10.6. We have used the latest version of Gurobi Optimizer (5.0). Some instances were solved with IBM's ILOG CPLEX optimizers (version 12.2) yielding comparable results.

2.2 Results of exact solvers

In what follows, we will provide an overview of instance sizes, which can be solved to optimality within 10 minutes. For this analysis we computed ten instances for each size and report the average runtime. If an instance was not solved to optimality within 10 minutes, we aborted the process and continued with the next run.

Incremental Volume Bids		**Total Quantity Bids**	
Stepwise fixed costs	**Decreasing unit costs**	**Stepwise fixed costs**	**Decreasing unit costs**

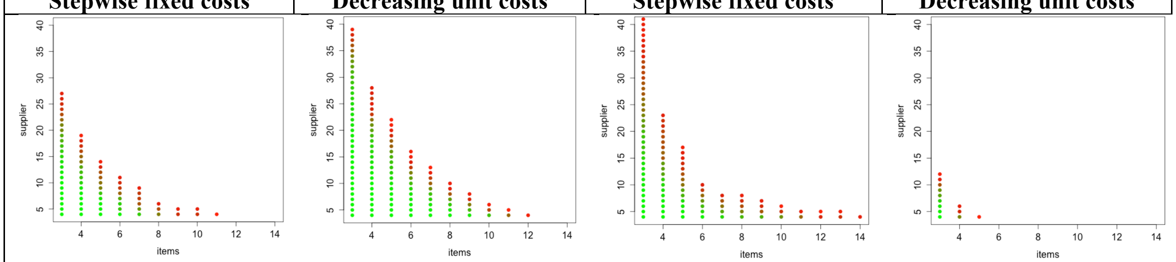


Figure 1: Computation times of problems with different constraints with and without stepwise fixed costs

The colored dots in Figure 1 describe the average time required to solve an instance of this size to optimality. The topmost red dot in each column indicates the problem size where optimality could not be proven for at least 8 out of 10 simulation runs. Green dots describe instance sizes, which could be solved to optimality. In particular, the combination of purely decreasing unit cost functions and total quantity bids turned out to be very hard to solve. Likewise the type of allocation constraint had an influence on runtime. Instances containing limits on spend took longer to solve than instances with limits on the number of winners. For some setups even instances with 5 items and suppliers turned out to be too hard to prove optimality.

3 Mathematical Formulation

In the previous sections we showed that instances, which can be solved exactly, are fairly small. Heuristics may help, but as we found out in other simulations, which are not presented in this paper due to brevity, there efficiency is also limited depending on the setup. Therefore we developed an alternate approach to support decision makers in performing scenario analysis. The following MIP, referred to as *supplier quantity selection for scenario analysis* (SQSSA), is designed in the spirit of parametric linear programs and uses optimization to find solutions minimizing violations on pre-specified bounds in the right-hand sides of SQS (Bichler et al. 2011). SQSSA adds the SQS objective function as a new constraint, and treats the right-hand side parameters of SQS to variables. The SQSSA objective function minimizes the sum of all changes on these new variables.

Table 1 outlines the decision variables used for SQSS. Ψ defines the set of business rules r, while a full description of the remaining parameters and sets can be found in Bichler et al. (2011). The first constraints (1) to (7) reflect the SQS and ensure that the demand is fulfilled and the amount of each product purchased from each supplier does not exceed the maximum quantity provided by the suppliers. They also determine the relevant volume for each discount or markup resp. and enable or disable discounts and markups based on the quantity and spend limits. For a detailed description of these constraints please refer to Bichler et al. (2011).

Variable	Description	Range
$x_{i,s}$	Quantity of item i purchased from supplier s	$\mathbb{R}_0^+$
y_d	Amount active in discount d	$\mathbb{R}_0^+$
y_m	Amount active in markup m	$\mathbb{R}_0^+$
c_d	Indicator for discount d	$\{0,1\}$
c_m	Indicator for markup m	$\{0,1\}$
c_l	Indicator for overall discount (lump sum rebate)	$\{0,1\}$
c_k	Indicator for overall markup	$\{0,1\}$
c_e	Indicator for excluded pricing rule	$\{0,1\}$
j_n	Indicator for condition n	$\{0,1\}$
a_r	Indicator if rule r is active (used for only-if-winning rules)	$\{0,1\}$
$a_{r,s}$	Indicator if supplier s in winner limiting rule r is active	$\{0,1\}$
f_r	Winner limit for rules $r \in \Psi_{WL}$	$\mathbb{N}_0$
t_r^l, t_r^u	Lower / upper limit on spend for all items $i \in I_r$ of suppliers $s \in S_r$ of absolute-spend-based rules $r \in \Psi_Q$	$\mathbb{R}_0^+$
v_r^l, v_r^u	Lower / upper limit on quantity for all items $i \in I_r$ of suppliers $s \in S_r$ of absolute-spend-based rules $r \in \Psi_Q$	$\mathbb{R}_0^+$
z_r	Percentage based limit on quantity or spend for all items I_r of suppliers S_r in relation to all items I_r of all suppliers S of relative-value based rule $r \in \Psi_Q$	$[0..1]$
Γ	Budget constraint	$\mathbb{R}_0^+$

Table 1: Decision Variables in SQSSA

Constraints (8) to (12) describe allocation rules, which are typically being used in scenario analysis. Constraints (8) define absolute resp. relative limits on the allocation of items to suppliers. Constraints (9) have the same purpose, but instead of limits on quantity they define limitations on spend. The variables a_r in these constraints are used as binary indicators to allow "only-if-winning" rules. Instead of specifying fixed lower bounds, which would cause partially predefined allocations, these lower bounds are only valid, if the supplier is winning any amount of the item(s) in that rule. In constraints (10) the aforementioned indicator variables are defined. Additionally, constraints (10) introduce the variables $a_{r,s}$ which count the number of winning suppliers and are then used in constraints (11) to limit them. The final constraint (12) models the individual spend constraint. Its right hand-side Γ is a budget constraint or the optimal value calculated during the first optimization of SQS.

$$\min \sum_{r\in\Psi_{Q_{abs}}} v_r^u + \sum_{r\in\Psi_{S_{abs}}} t_r^u + \sum_{r\in\Psi_{WL}} f_r + \sum_{r\in\Psi_{Q_{rel}}\cup\Psi_{S_{rel}}} z_r - \sum_{r\in\Psi_{Q_{abs}}} v_r^l - \sum_{r\in\Psi_{S_{abs}}} t_r^l$$

s.t.

$$\sum_{s\in S} x_{i,s} \geq W_i \qquad \forall i\in I \qquad (1)$$

$$x_{i,s} \leq E_{i,s} \qquad \forall i\in I, \forall s\in S \qquad (2)$$

$$\sum_{i\in I_d} x_{i,s_d} - D_d c_d \geq y_d \qquad \forall d\in D \qquad (3d)$$

$$\sum_{i\in I_m} x_{i,s_m} + Bc_m - D_m \leq y_m + B \qquad \forall m\in M \qquad (3m)$$

$$Bc_d \geq y_d \qquad \forall d\in D \qquad (4d)$$

$$\sum_{n\in N_d} j_n - \sum_{e\in E_d} c_e \geq (|N_d| + |E_d|)c_d - |E_d| \qquad \forall d\in D\cup L \qquad (5d,l)$$

$$|N_m|^{-1}(\sum_{n\in N_m} j_n) - \sum_{e\in E_m} c_e \leq c_m + 1 - |2N_m|^{-1} \qquad \forall m\in M\cup K \qquad (5m,k)$$

$$\sum_{i\in I_n} P_{i,s_n} x_{i,s_n} - \sum_{d\in D_n} R_d y_d + \sum_{m\in M_n} R_m y_m \geq S_n j_n \qquad \forall n\in N \qquad (6d,l)$$

$$\sum_{i\in I_n} P_{i,s_n} x_{i,s_n} - \sum_{d\in D_n} R_d y_d + \sum_{m\in M_n} R_m y_m \leq Bj_n + S_n \qquad \forall n\in N \qquad (6m,k)$$

$$\sum_{i\in I_n} x_{i,s_n} \geq Q_n j_n \qquad \forall n\in N \qquad (7d,l)$$

$$\sum_{i\in I_n} x_{i,s_n} - Q_n \leq Bj_n \qquad \forall n\in N \qquad (7m,k)$$

$$a_r v_r^l \leq \sum_{i\in I_r}\sum_{s\in S_r} x_{i,s} \leq v_r^u \qquad \forall r\in\Psi_{Q_{abs}} \qquad (8a)$$

$$\sum_{i\in I_r}\sum_{s\in S_r} x_{i,s} - z_r \sum_{i\in I_r}\sum_{s\in S} x_{i,s} \leq 0 \qquad \forall r\in\Psi_{Q_{rel,cap}} \qquad (8rc)$$

$$\sum_{i\in I_r}\sum_{s\in S_r} x_{i,s} - z_r \sum_{i\in I_r}\sum_{s\in S} x_{i,s} \geq 0 \qquad \forall r\in\Psi_{Q_{rel,floor}} \qquad (8rf)$$

$$a_r t_r^l \leq \sum_{i\in I_r}\sum_{s\in S_r} P_{i,s} x_{i,s} - \sum_{d\in D_r} R_d y_{i,d} + \sum_{m\in M_r} R_m y_{i,m} \leq t_r^u \qquad \forall r\in\Psi_{S_{abs}} \qquad (9a)$$

$$\sum_{i\in I_r}\sum_{s\in S_r} x_{i,s} \leq Ba_r \qquad \forall r\in\Psi_{Q_{abs}}\cup\Psi_{S_{abs}} \qquad (10q)$$

$$\sum_{i\in I_r} x_{i,s} \leq Ba_{r,s} \qquad \forall r\in\Psi_{WL} \qquad (10w)$$

$$\sum_{s\in S_r} a_{r,s} \leq f_r \qquad \forall r\in\Psi_{WL} \qquad (11)$$

$$\sum_{i\in I}\sum_{s\in S} P_{i,s} x_{i,s} - \sum_{d\in D} R_d y_d + \sum_{m\in M} R_m y_m - \sum_{l\in L} R_l c_l + \sum_{k\in K} R_k c_k < \Gamma \qquad (12)$$

In constrains (8) and (9) decision variables (a_r and v_r^l or t_r^l as well as z_r and $x_{i,s}$ resp.) are multiplied, which leads to non-linearity. Standard algorithms to solve MIPs cannot handle such non-linear constraints, but SQSSA can be formulate such that non-linear terms are approximated with piecewise linear functions (Bisschop 2012, Bradley et al. 1977). If n defines the number of breakpoints, which describe the number of linear segments less 1, the approximation requires $2 * |I_r| * |S|$ new constraints and $2n * |I_r| * |S|$ additional variables.

The proposed formulation does not change the problem of long solver runtimes itself. But it guides procurement manages and reduces the number of instances which need to be solved. Instead of iteratively solving multiple instances with varying business rules, decision marker can now define budget constraints and SQSSA calculates which business rules need to be relaxed.

4 Discussion

We have shown that combinations of well-known meta-heuristics such as RINS and VNS with exact branch-and-cut approaches are a promising approach. But we also saw, that their efficiency strongly depends on the setup. This paper finally combined different modeling approaches and computational methods from combinatorial optimization to allow for more effective decision support in e-sourcing. This type of decision aid is made possible by the advances in combinatorial optimization in the recent years, and would have been impossible only a few years ago. A systematic and quantitative analysis of SQSSA's efficiency is still missing; but first applications show promising results. Using the proposed approach saves time and guides the search to identifying better solutions. Additionally, we expect that with advances in multi-core hardware architectures and respective parallel implementations of branch-and-cut algorithms larger problems can be solved in only a few years from now.

5 References

Bichler, M., Gupta, A., and Ketter, W. 2010. Designing Smart Markets. Info. Sys. Research 21, 4 (December 2010), 688-699.

Adomavicius, G., Curley, S., Gupta, A., and Sanyal, P. 2012. Effect of Information Feedback on Bidder Behavior in Continuous Combinatorial Auctions. Management Science 58, 4 (April 2012), 811-830.

Scheffel, T., Pikovsky, A., Bichler, M., and Guler. K. 2011. An Experimental Comparison of Linear and Nonlinear Price Combinatorial Auctions. Info. Sys. Research 22, 2 (June 2011), 346-368.

Davenport, A., Kalagnanam, J., 2000. Price negotiations for procurement of direct inputs. In: IMA "Hot Topics" Workshop: Mathematics of the Internet: E-Auction and Markets. Vol. 127. Minneapolis, USA, pp. 27–44

Goossens, D. R., Maas, A. J. T., Spieksma, F., van de Klundert, J. J., 2007. Exact algorithms for procurement problems under a total quantity discount structure. European Journal of Operational Research 178, 603–626.

Bichler, M., Schneider, S., Guler, K., and Sayal, M. 2011. Compact bidding languages and supplier selection for markets with economies of scale and scope. European Journal on Operational Research, 214(1). (2011), 67-77.

Boutilier, C., Sandholm, T., and Shields, R. 2004. Eliciting bid taker non-price preferences in (Combinatorial) Auctions. In Proceedings of the National Conference on Artificial Intelligence.

Mladenovic, N., Hansen, P. 1997. Variable neighborhood search. Computers and Operations Research, 24(11). 1097–1100.

Hansen P., Mladenovic M., Uroševic D. 2006. Variable neighborhood search and local branching. Computers & Operations Research 33. (2006), 3034–3045

Bradley, S. P., Hax, A.C., Magnanti, T.L. 1977. Applied Mathematical Programming. Addison-Wesley.

Bisschop, J. 2012. AIMMS Optimization Modeling.

Budget Planning for Coupled Campaigns in Search Auctions

Yanwu Yang[1], Rui Qin[1], Jie Zhang[1], Xin Li[2] and Daniel Zeng[1,3]

[1]The State Key Lab of Intelligent Control and Management of Complex Systems, Inst. of Automation, CAS, Beijing, China
[2]Department of Information Systems, City University of Hong Kong, Kowloon Tong, Hong Kong
[3]Department of Management Information Systems, The University of Arizona, USA
{yangyanwu.isec, qinrui.isec, zhangjie.isec}@gmail.com, xin.li@cityu.edu.hk, zeng@email.arizona.edu

Abstract: Budget-related decisions in search auctions are recognized as a structured decision problem, rather than a simple constraint. Budget planning over several coupled campaigns remains a challenging but utterly important task in search advertisements. In this paper, we propose a multi-campaign budget planning approach using optimal control techniques, with consideration of the substitute relationship between advertising campaigns. A measure of coupled relationships between campaigns is presented, e.g., the overlapping degree (O) in terms of campaign contents, promotional periods and target regions. We also discuss some desirable properties of our model and possible solutions. Furthermore, computational experiments are conducted to evaluate our model and identified properties, with real-world data collected from logs and reports of practical campaigns. Experimental results show that, (a) coupled campaigns with higher overlapping degrees can reduce the optimal budget level and the optimal revenue, and also arrive the budgeting cap earlier; (b) The advertising effort could be seriously weakened when ignoring the overlapping degree between campaigns.

1 Introduction

Search auctions have become the most successful business model accounting for 47.7% revenues of online advertisements in 2011 (IAB, 2011). On the one hand, more and more advertisers choose search auctions to promote their products or services. On the other hand, search advertisements form the dominating revenue resource for major search engine companies. One of the most difficult tasks for advertisers is how to effectively determine and allocate the optimum level of advertising budgets in search auctions.

Budget is an endogenous factor in search auctions, that heavily influences other advertising strategies. Moreover, budget-related decisions in search auctions are recognized as a structured decision problem, rather than a simple constraint (Yang et al., 2012). Specifically, throughout the entire lifecycle of search advertising campaigns, there exist three interweaving budget decisions and being affected by various factors: allocation across search markets, temporal distribution over a series of slots (e.g. day) and adjustment of the remaining budget (e.g., the daily budget). This work considers the following scenario: an advertiser usually has several campaigns simultaneously executed in a search market (e.g., Google Adwords); then given the advertising budget in a search market is determined, how to make budgeting plans for these campaigns simultaneously over time, in order to maximize the global advertising performance? As shown by some previous research (Tull et al., 1986; Fisher et al., 2011), profit improvement from better allocation strategies is much higher than from improving the overall budget.

There have been some efforts along the line of budget-related decisions in search auctions. Most of them either take the budget as the constraint for other advertising strategies, or allocate the budget over keywords. However these efforts are not operationally suitable to practical paradigms provided by major search engines, because of ignoring the search advertising structure. Budget planning over several campaigns remains a challenging but utterly important task for advertisers in search auctions. First, the search marketing environments are essentially dynamic and uncertain, advertisers usually have no sufficient knowledge and time to track and adjust various advertising decisions. Secondly, an advertiser's campaigns rarely independent. Similar to the case among products (Doyle and Saunder, 1990), there are certain relationships (e.g., complementarity and substitution) between advertising campaigns, which leads to cross elasticities. For example, for a retailing advertiser, one campaign featuring smart phones might have some substitution effects on another featuring cheap cellphones, and vice versa. Thirdly, the complex structure of search advertising markets and campaigns consists of a lot of parameters, which made multi-campaign budgeting decisions not straightforward.

The objective of this research is to to explore the dynamic budget planning problem for several coupled campaigns in terms of substitution relationships in search auctions. In this paper we formulate the multi-campaign budget planning problem as an optimal control process under a finite time horizon. First, we present a measure of coupled relationships between advertising campaigns by considering the overlapping degree (O) in terms of campaign contents, promotional periods and target regions. The overlapping degree refers to the degree to which target markets (or audiences) of these two campaigns overlap each other. Intuitively, it is defined as the probability that search users (e.g., potential customers) issuing with keywords in campaign j can also be reached by campaign $j^{'}$ in search auctions. Secondly, we propose a random walk-based approach for the ad overlapping degree γ (e.g., campaign contents) in the context of a directed keyword graph relevant to a given advertiser. The higher the ad overlapping degree between two campaigns, the more advertising effort is weakened. Thirdly, we provide a feasible solution to our model and study some desirable properties. Furthermore, we also conduct computational experiments to validate and evaluate our budget planning approach, with real-world data collected from logs and reports of practical campaigns. Experimental results show that (a) the overlapping degree (O) between campaigns has serious effects on optimal budgets and the advertising effort, and the advertising effort can be seriously weakened if an advertiser ignore the overlapping degree between campaigns while making advertising decisions; (b) the case with the higher ad overlapping degree (γ) leads to lower optimal budget level and reaches the budgeting cap earlier; (c) the higher the overlapping degree is, the less optimal revenues can be obtained. The rest of this paper is organized as follows. In Section 2, we propose a measure of 3-dimensional relationship between campaigns, and present a budget planning strategy over several campaigns in search auctions. Section 3 studies some desirable properties and provides possible solutions for our model. Section 4 reports some experimental results to validate some normative findings from our model. Finally we conclude this work and discuss future research directions in Section 5.

2 Multi-campaign Budget Planning

2.1 The 3-dimensional Relationship between Campaigns

It is observed that, the overlapping degree between two campaigns j and $j^{'}$ is zero if there is no overlaps from any single one of those aspects (e.g., campaign contents, promotional intervals and target regions). Thus, the overlapping degree $O(j, j^{'})$ is the product of overlaps from three aspects $O(j, j^{'}) = I_t(j, j^{'}) \times I_s(j, j^{'}) \times \gamma(j, j^{'})$, where I_t and I_s represent the temporal indicator function (e.g., promotional intervals) and the spatial indicator function (e.g., target regions), respectively, and $\gamma(j, j^{'})$ is the overlapping degree in terms of campaign contents.

Denote the indicator function $I_A(x)$ as follows: $I_A(x) = 1$ if $x \in A$, otherwise it is 0.

We will give the definition of the temporal indicator and the spatial indicator.

The Temporal Indicator Function: Let T_j and $T_{j'}$ denote promotion intervals of campaign j and campaign $j^{'}$, respectively. The temporal indicator function is given as $I_t(j, j^{'}) = I_{T_j}(t) I_{T_{j'}}(t)$.

The Spatial Indicator Function: The overlaps with respect to target regions can be given in a similar way. Let S_j, $S_{j'}$ be target regions of campaign j and campaign $j^{'}$, respectively. The spatial indicator function is given as $I_s(j, j^{'}) = I_{S_j}(s) I_{S_{j'}}(s)$.

Next, we will discuss the ad overlapping degree (e.g., in terms of campaign contents) in search auctions.

2.2 The Ad Overlapping Degree

We construct a directed graph of keywords (K) relevant to a given advertiser (or her products/services) with the appearance probability as the edge weight. Let K_j and $K_{j'}$ be keyword sets of campaigns j and $j^{'}$, respectively. For each pair of k and $k^{'}$, we can apply a random walk approach (Doyle and Snell, 1984) to compute the appearance probability$\omega_{k,k'}$, given as $\omega_{k,k'} = P(l_k = k^{'}) = \sum_{r:e(k,r)\in E} \beta_{k,r} \mu_{k,r} P(l_r = k^{'})$, where $l_k = k^{'}$ represents starting at keyword k to hit keyword $k^{'}$, $\mu_{k,r} = \omega_{k,r}$ if $\beta_{k,r} = 1$. Then, $\zeta(j, j^{'}) = \frac{1}{|K_j|}\sum_{k\in K_j}\frac{1}{|K_{j^{'}}|}\sum_{k^{'}\in K_{j^{'}}}\omega_{k,k^{'}}$ represents the probability that search users (e.g., potential customers) issuing keywords in campaign j can also be reached by campaign $j^{'}$, and $\zeta(j, j^{'}) \in [0, 1]$. Define $\gamma(j, j^{'}) = [d_j\zeta(j, j^{'}) + d_{j^{'}}\zeta(j^{'}, j)]/(d_j + d_{j^{'}})$, where d_j and $d_{j^{'}}$ represent the potential query demands of the jth and $j^{'}$th campaign, respectively. Then γ represents the ad overlap degree between two campaigns (e.g., in terms of campaign contents).

2.3 The Model

In this section, we establish a budget planning model for coupled campaigns in search auctions. First, suppose the advertiser aims to maximize the total revenue from advertising activities. Let $d_{t,s}$ denote the number of query demand (relevant to an advertiser's promotions in a search market) in region s at time t, $\theta_{j,t,s}$ campaign j's market share in region s at time t. Then the number of potential query demands that might be obtained by the advertiser in region s at time t is $d_{t,s}\theta_{j,t,s}$. Let $c_{j,t,s}$ denote the (average) click-through-rate of campaign j in region s at time t, $v_{j,t}$ the (average) value-per-click of campaign j at time t, and $b_{j,t,s}$ the budget segment for campaign j in region s at time t, then the total revenue for the advertiser can be represented as $\sum_{j=1}^{m}\sum_{s\in S_j}\int_{T_j}e^{-rt}(d_{t,s}\theta_{j,t,s}c_{j,t,s}v_{j,t} - b_{j,t,s})\mathrm{d}t$, where e^{-rt} is the discount factor.

Secondly, due to marketing dynamics in search auctions, an advertiser's market share changes with time. Following (Yang et al., 2011), the response function in search markets can be given as $\mathrm{d}\theta_{j,t,s}/\mathrm{d}t = \rho q u(t, s)\sqrt{1 - \theta_{j,t,s}}$, where ρ is the response constant, δ the decay constant, and q is the quality score. The advertising effort u represents the effective part of advertising budget b.

Thirdly, let B_{market} denote the overall advertising budget allocated to a given search market, then the present value of total advertising budgets (or expenditures) under a finite time horizon should not exceed it. That is, $\sum_{j=1}^{m}\sum_{s\in S_j}\int_{T_j}e^{-rt}b_{j,t,s}\mathrm{d}t \le B_{market}$.

Thus, we can the multi-campaign budget planning problem as follows,

$$\begin{aligned}
\max\quad & \sum_{j=1}^{m}\sum_{s\in S_j}\int_{T_j}e^{-rt}(d_{t,s}\theta_{j,t,s}c_{j,t,s}v_{j,t} - b_{j,t,s})\mathrm{d}t \\
\text{s.t.}\quad & \sum_{j=1}^{m}\sum_{s\in S_j}\int_{T_j}e^{-rt}b_{j,t,s}\mathrm{d}t \le B_{market} \\
& \mathrm{d}\theta_{j,t,s}/\mathrm{d}t = \rho q u(t, s)\sqrt{1 - \theta_{j,t,s}} \\
& b_{j,t,s} \ge 0,
\end{aligned} \tag{1}$$

where $b_{j,t,s}$ is the control variable, and $\theta_{j,t,s}$ is the state variable.

3 The Solution & Properties

In this section, we study some desirable properties of our budget planning model, and provide possible solutions. Note that we focus on the case with two campaigns in this work.

Let us consider that there are two campaigns for an advertiser in a market. First, the objective function of model (1) can be written as $\sum_{j=1}^{2}\sum_{s\in S}\int_0^T e^{-rt}I_{T_j}(t)I_{S_j}(s)(d_{t,s}\theta_{j,t,s}c_{j,t,s}v_{j,t} - b_{j,t,s})\mathrm{d}t$. Secondly, the budget constraint becomes $\sum_{j=1}^{2}\sum_{s\in S}\int_0^T e^{-rt}I_{T_j}(t)I_{S_j}(s)b_{j,t,s}\mathrm{d}t \le B_{market}$. Thirdly, if the two campaigns are mutually independent, the advertising effort can be given as $u(t, s) = \sum_j (b_{j,t,s})^{\alpha_{j,t,s}}$, where $b_{j,t,s}$ represents the budget of campaign j at time t in region s, $\alpha_{j,t,s}$ denotes the advertising elasticity of campaign j at time t in region s. If there are overlaps in terms of campaign contents between j and $j^{'}$, then the advertising effort is weakened, given as, $u(t, s) = \sum_j (b_{j,t,s})^{\alpha_{j,t,s}} - \sum_j (O(j, j^{'})b_{j,t,s})^{\alpha_{j,t,s}}$, where $O(j, j^{'})$ is the proportion of the allocated budget (for these two campaigns) where the advertising effort is weakened.

The optimal solution is $b^{*}_{j,t,s}$. It represents the optimal budget allocated to campaign j in region s at time t. With the optimal control trajectory of budget, we can also obtain the optimal budget allocated to campaign j in a finite time horizon (e.g., T) is $\sum_s\int_0^T e^{-rt}b^{*}_{j,t,s}\mathrm{d}t$.

Next, we study some properties and possible solutions of the model. By introducing a Lagrange multiplier λ, we employ the principle of dynamic programming and obtain the optimal feedback advertising decisions b_1^* and b_2^* satisfy the following conditions: $\alpha_i(I_{T_i}(t)I_{S_i}(s) - O^{\alpha_i})(b_i^*)^{\alpha_i - 1} = \frac{e^{-rt}(1+\lambda)}{\rho q\sqrt{1-\theta_i}\cdot V_{\lambda,\theta}}$.

Then we have, $\lambda\left(B_{market} - \sum_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{\lambda,1}(t,\theta) + I_{T_1}(t)I_{S_2}(s)b^*_{\lambda,2}(t,\theta))dt\right) = 0$

With infinite budget, we have $\lambda = 0$ and $\sum_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{0,1}(t,\theta)+I_{T_2}(t)I_{S_2}(s)b^*_{0,2}(t,\theta))dt < B_{market}$.

Considering the case that the budget is limited, we choose the minimal $\lambda > 0$ so that

$$\sum_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{\lambda,1}(t,\theta) + I_{T_2}(t)I_{S_2}(s)b^*_{\lambda,2}(t,\theta))dt = B_{market}.$$

From the above analysis, we can come to the following theorems.

Theorem 1. *If the total budget B_{market} is less than $\sum\limits_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{0,1} + I_{T_2}(t)I_{S_2}(s)b^*_{0,2})dt$, the optimal budget allocation strategy is:*

$$\begin{aligned} b_1, b_2 &= \underset{b^*_{\lambda,1}, b^*_{\lambda,2}}{\operatorname{argmin}}\, \lambda \\ s.t. \quad & \sum_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{\lambda,1}(t,\theta) + I_{T_2}(t)I_{S_2}(s)b^*_{\lambda,2}(t,\theta))dt = B_{market}, \lambda \geq 0. \end{aligned}$$

Theorem 2. *If $B > \sum_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{0,1} + I_{T_2}(t)I_{S_2}(s)b^*_{0,2})dt$, the optimal way is to invest $\sum_{s\in S}\int_0^T e^{-rt}(I_{T_1}(t)I_{S_1}(s)b^*_{0,1} + I_{T_2}(t)I_{S_2}(s)b^*_{0,2})dt$ in the search advertising market.*

In reality, if we ignore the ad overlapping degrees (γ) between two campaigns, then the optimal revenue will be diminished because the corresponding advertising effort are weakened. This can be guaranteed by the following corollary.

Corollary 3. *Let U^* denote the optimal revenue of the model, and $\bar{U}$ the revenue corresponding to strategies $\bar{b}_1$ and $\bar{b}_2$, where $\bar{b}_1$ and $\bar{b}_2$ are optimal solutions ignoring the overlapping degree in terms of campaign contents between two campaigns, respectively. Then $U^* > \bar{U}$.*

4 Experimental Validation

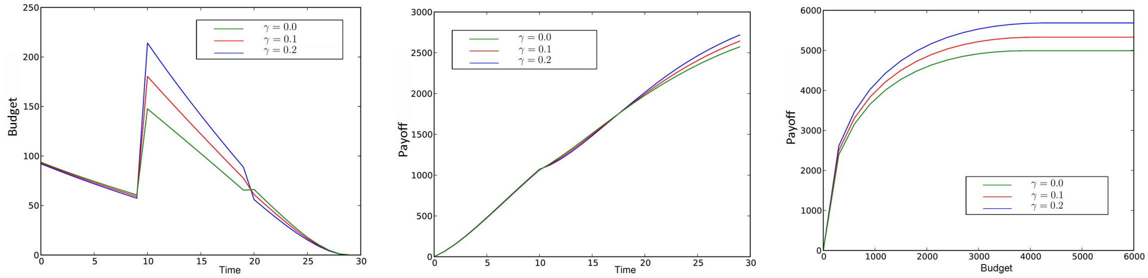
In this section, we design computational experiments to validate the proposed model and properties. We generate experimental datasets from historical advertising logs including operations and effects collected from real-world advertising campaigns of search auctions. In the following experiments we take a search advertising scenario where two campaigns are assigned same target regions, and different promotional intervals: one from Sep. 1st to 20th, 2009, another from Sep. 10th to 30th, 2009. Then we can get the ad overlapping degree (e.g., $\gamma = 0.11$) with the algorithm provided in Section 2.

4.1 The Ad Overlapping Degree (γ)

In the first experiment, we aim to prove the influence of the ad overlapping degree (γ) on the optimal budget and corresponding revenue. For this purpose, we set $B = 3000$, and compute the optimal budget and corresponding revenue with different settings of ad overlapping degrees. That is, the spatial and temporal overlapping degrees are kept unchanged (as described above), and the ad overlapping degrees are assigned different values (e.g., Case 1: $\gamma = 0.0$, Case 2: $\gamma = 0.1$, Case 3: $\gamma = 0.2$). The change of the optimal budget and corresponding revenue over time are illustrated in Figure 1(a) and Figure 1(b), respectively.

From Figure 1(a) and Figure 1(b), we can notice that,

(1) As for the case with the higher ad overlapping degree (γ), the optimal budget is lower during the period when the overlapping degree $O > 0$, and is higher when the overlapping degree $O = 0$ (e.g., there is no overlapping degree), and vice versa. This phenomenon can be explained by the fact that, in the case with higher γ the more advertising effort is weakened, thus the optimal budget and corresponding revenue are less (e.g., it is easier to reach the optimal level).



(a) (b) (c)

Figure 1: (a) The optimal budget over time with different γ; (b) The accumulated revenue over time with different γ; (c) The optimal revenue at different budgeting level with different γ

(2) Concerning the accumulated revenue, the case with larger γ is slightly bigger at the initial period, then the increasing speed becomes lower when the overlapping degree $O > 0$, and vice versa. The reason might be that, the case with larger γ allocates more budgets when $O = 0$, thus it gets a bit more revenue at the initial stage; when $O > 0$ both its optimal budget and revenue are lower, thus the increasing speed of accumulated revenue become slower; then during the period from 21th to 30th, again it allocates more budgets but the accumulated revenue is kept lower due to previous performance.

4.2 The Optimal Revenue at Different Budgeting Levels

The second experiment intends to illustrate the relationship between the optimal budget and corresponding revenue of these two campaigns with different settings of the ad overlapping degree (γ) same as in the second experiment. We set $B \in [0, 6000]$. The experimental result is illustrated in Figure 1(c).

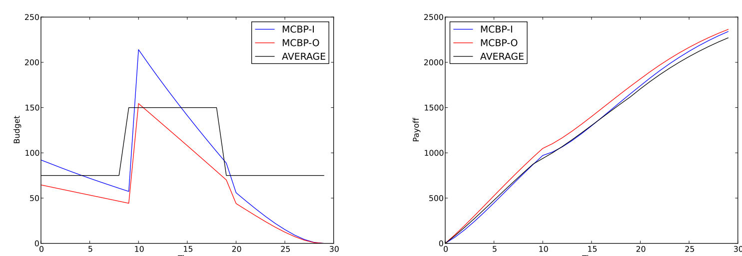
From Figure 1(c), we can see that,

(1) The optimal revenue grows steadily until reaching the cap where the marginal revenue (e.g., the change in additional revenue) is 0, when the total budget increases. In other words, there exists a budgeting cap in the case with unlimited budgets. And the case with larger γ arrives the budgeting cap earlier, and vice versa.

(2) The optimal revenue in the case with larger γ is always less (than that of other cases), and vice versa.

4.3 The Overlapping Degree (O)

The third experiment concerns if and how superior it is to consider the overlapping degree (O) when doing the budget planning for multiple campaigns in a search market. We implement our multi-campaign budget planning approach (MCBP) as provided in Section 3 into two strategies: with (MCBP-O) and without (MCBP-I) consideration of the overlapping degree. We choose the AVERAGE strategy as a baseline strategy, which allocates the budget averagely between each campaign and over time. The optimal budget and revenue are illustrated in Figure 2(a) and Figure 2(b), respectively.



(a) (b)

Figure 2: (a) The optimal budget accumulated over time; (b) The accumulated payoff over time

From Figure 2(a) and Figure 2(b), we can see that,

(1) The optimal (total) budget allocated to these two campaigns is 1847.17, 2459.08 and 3000.00 by the these three strategies (e.g., MCBP-O, MCBP-I and AVERAGE), respectively. And correspondingly the optimal payoff is 2363.70, 2340.06 and 2270.74, respectively.

(2) The MCBP-O strategy and the MCBP-I strategy can obtain 1.280 and 0.952 payoff per unit budget, respectively. In other words, the payoff per unit budget is increased 34.45% by considering the overlapping degree (O) between campaigns. This can be explained by the fact that the advertising effort is weakened when the overlapping degree between campaigns exists (e.g., $O > 0$). The situation might become even worse if the advertiser ignores the overlapping degree between campaigns while making budget planning decisions in sponsored search auctions.

(3) The AVERAGE strategy can obtain 0.757 payoff per unit budget, and both MCBP-O and MCBP-I outperform the AVERAGE strategy from the view of payoff per unit budget (69.09% and 25.76%), which illustrates that our multi-campaign budget planning approach can help advertisers to increase the overall payoff in a certain degree.

4.4 Management Insights

Our work provides critical managerial insights for advertisers to make budgeting decisions over campaigns in search auctions. First, advertisers usually pay less attentions to relationships and cross-effects between their own campaigns in a search market, probably due to the fact that it is not easy to measure and manipulate the overlapping degree. This research indicates that the overlapping degree (O) between campaigns have serious effects on optimal budget strategies at the campaign level. Secondly, for an advertiser, the larger the overlapping degree between campaigns, the more advertising effort is weakened, and thus the optimal revenue is less. Thus it's important for an advertiser to reduce the overlapping degree among campaigns as much as possible, then correspondingly adjust the optimal budgets over campaigns in order to maximize the expected revenue. Thirdly, our normative findings of multi-campaign budget planning can also provide some valuable insights to other similar decision scenarios of advertising budget allocation, such as across several markets, across different medias (or channels).

5 Conclusions and Future Work

In this paper, we present a multi-campaign budget planning approach using optimal control techniques, under a finite time horizon. Our model takes into account the overlapping degree (e.g., the substitute relationship) between campaigns in search auctions, with respect to three dimensions: target regions, promotional periods, and campaign contents. We discuss some desirable properties of our model and possible solutions. Computational experimental studies are made to evaluate our model and identified properties. Experimental results show that the overlapping degree between campaigns has serious effects on budgeting decisions and advertising performance, and higher overlapping degrees weaken the advertising effort and thus diminish optimal budgets and revenues. We are in the process of extending our model in the following directions: (a) spatial heterogeneousness and relationships to capture spatial effects on advertising decisions and performance; (b) stochastic budget planning strategies in uncertain marketing environments of search auctions; (c) the complementary relationship between campaigns and its effects on budgeting decisions.

Acknowledgements

We are thankful to anonymous reviewers who provided valuable suggestions that led to a considerable improvement in the organization and presentation of this manuscript. This work is partially supported by NSFC (71071152, 70890084, 71025001, 71102117, 60875049, 60921061).

References

Doyle, P. and Snell, L. Random Walks and Electrical Networks. Mathematical Association of America, 1984.

Doyle, P. and Saunders, J. Multiproduct Advertising Budgeting Marketing Science Spring 1990 9:97-113.

Fischer, M., Albers, S., Wagner, N., and Frie, M. 2011. Practice Prize Winner—Dynamic Marketing Budget Allocation Across Countries, Products, and Marketing Activities. Marketing Science 30, 4 (July 2011), 568-585.

Interactive Advertising Bureau (IAB), 2011, Internet Advertising Revenue Report. Stable URL: http://www.iab.net/media/file/IAB-HY-2011-Report-Final.pdf

Tull S., Wood R., Duhan D., Gillpatrick T., Robertson R., Helgeson G. Leveraged decision-making in advertising: The flat maximum principle and its implications. Journal of Marketing Research 23 (1 (February)) 25–32, 1986.

Yang, Y., Zhang, J., Liu, B., Zeng, D. Optimal Budget Allocation across Search Advertising Markets, Proceedings of the 21st Workshop on Information Technologies and Systems, pp. 97–102, Shanghai, China, December 3-4, 2011.

Yang, Y., Zhang, J., Qin, R., Li, J., Wang, F., Qi W. A budget optimization framework for search advertisements across markets, IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans, 42 (5): 1141–1151, 2012.

Polarization Trend in Sponsored Search Auctions

Yong Yuan[1], Daniel Zeng[1, 2], Huimin Zhao[3], Jiesi Cheng[2]
[1] State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China,
[2] Department of Management Information Systems, University of Arizona, Tucson, Arizona, USA,
[3] Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Abstract – Quality score (QS) plays a critical role in sponsored search auctions and in practice is closely related to the historical click-through rate (CTR) of an advertisement. Existing research, however, has not explicitly considered this correlation between QS and historical CTR. In this paper, we strive to bridge this gap. Based on a discrete time-dependent optimal control model, which explicitly captures the CTR-QS correlation, we analyze the optimal positioning strategy and the widely-observed greedy positioning strategy for advertisers. We find that both strategies lead advertisers to monotonically increase or decrease their ranks over time, and thus may result in a polarization trend in sponsored search markets. Our findings can help characterize advertisers' behavior dynamics and also offer valuable insights and suggestions to search engines.

1. Introduction

Over the last decade, the online advertising industry has witnessed a rapid growth of sponsored search advertising (SSA), in which advertisers bid for their query-specific advertisements to appear alongside organic search results. The basic economic instrument behind most SSA platforms, such as Google AdWords and Yahoo! Search Marketing, is keyword-based position auction.

Due to its huge commercial success, SSA has recently attracted extensive research interests. Researchers have conducted intensive analyses on the Nash equilibrium continuum of SSA auctions and its refinements [1, 2]. Various strategic behaviors, such as greedy bidding, vindictive bidding and cyclical bidding, have also been identified and analyzed [3]. However, we identified a major limitation in existing research efforts: the quality score (QS) of an advertisement is treated as an exogenous variable and is assumed to be independent of the historical click-through rate (CTR).

In practice, however, most search engines use historical CTR as a dominant factor in their QS measurements so that advertisements receiving more clicks tend to get higher QSs. As our analysis will show, this correlation between QS and historical CTR imposes great influence on advertisers' positioning behavior in SSA auctions. As illustrated in Figure 1, an advertiser can increase its bid to get a higher rank and CTR, which raises the QS of its advertisement and thus decreases its cost-per-click (CPC). With a lowered CPC, the advertiser can bid more aggressively, resulting in a virtuous circle. Analogously, an advertiser decreasing its bid may go into a vicious circle. Therefore, there is a critical need to investigate the influence the CTR-QS correlation imposes on the behavior dynamics of advertisers in SSA markets and identify the optimal positioning strategy for advertisers.

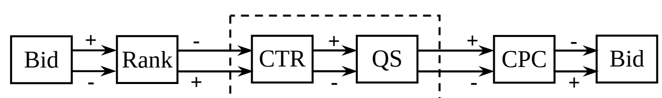


Fig. 1. The virtuous and vicious circles in SSA auctions

In this paper, we take the cumulative effect of CTR on QS into consideration explicitly and formulate an advertiser's decision-making process as a discrete time-dependent optimal control problem. Using policy iteration, we numerically determine an optimal strategy for an advertiser's rank selection in SSA auctions. We also analyze a widely-observed greedy strategy, which maximizes an advertiser's immediate revenue in each stage. With an optimal targeting rank, an advertiser can easily find the optimal bid for that position by experimenting on the SSA markets or using online services that help maintain specific positions. Our key finding is that, due to the CTR-QS correlation, advertisers using either of the two strategies will monotonically adjust their targeting ranks. On one hand, big-brand advertisers with high per-click values can keep increasing their ranks and revenues by bidding aggressively. On the other hand, however, small advertisers with low values have to lower their ranks while suffering from decreasing revenue, or even quit the SSA markets. Our finding indicates that a polarization trend may emerge in the SSA markets. As empirical evidence, it has been witnessed that the average CPC in SSA markets has decreased and increased for advertisers on high and low slots, respectively, during 2009-2011 [4]. Our findings offer a potential explanation for the underlying dynamics behind this phenomenon.

2. An SSA Auction Scenario and the Model

We consider an SSA auction scenario with N advertisers competing for K slots on a given keyword. The CTR of the k^{th} slot is denoted by x_k. We assume $x_1 > \cdots > x_K > 0$ and let $x_t = 0$ for $t > K$. Each advertiser in slot $i \in [1, N]$ assigns a click with a private value v_i denoting its maximum willingness to pay. Its bid and the QS of its advertisement are b_k and q_k, respectively.

The search engine uses generalized second price (GSP) auctions and allocates sponsored advertisements to the slots in descending order by their QS-weighted bids, which are called adranks in Google AdWords. We denote the adrank of the k^{th} slot by $h(k) = q_k b_k$. The payment rule follows the second-price scheme: the advertiser in slot k pays $h(k+1)/q_k$ each time its advertisement is clicked, with its revenue realized as $u_k = (v_k - h(k+1)/q_k)x_k$.

Assumption 1: Advertisers have complete information about the adrank and CTR of each slot.
This is reasonable as it is relatively easy for advertisers to experiment on the market by shifting their bids and submitting dummy search requests to test the rank changes, so as to come up with an approximate estimation of the adrank in each slot [5]. Besides, various analytical tools and services are available online to help estimate the CTRs of all slots.

Assumption 2: The QS of an advertisement is determined by its historical CTR.
Before 2005, Google ranked advertisements purely by their bids multiplied by CTRs. Although other factors, such as keyword relevance and landing page quality, are taken into consideration later on, the historical CTR has always been the dominant factor in QS measurement in most SSA platforms. Since our work focuses on the cumulative effect of CTR on QS, we ignore other factors and assume that the QS of an advertisement is determined by (equals) its historical CTR.

Assumption 3: The volume of search requests (impressions) is stable in each stage.
We do not consider the fluctuations of search volume caused by holidays or sudden events.

We focus on the decision faced by an advertiser. In practice, an advertiser usually checks its QS regularly (e.g., daily). After observing the immediate QS in each stage, the advertiser has to make a decision about the optimal slot and adjust its bid accordingly. Since the QS is determined by the historical CTR, the CTR of the selected slot can affect the QS in the next stage. Hence, an advertiser faces a multi-stage decision optimization problem. Figure 2 illustrates the decision process.

We develop the following discrete time-dependent optimal control model to analyze an advertiser's optimal positioning strategy. The model can be specified by a seven-tuple $(\mathcal{Q}, \mathcal{K}, \mathcal{T}, \sigma, u, f, \delta)$, where

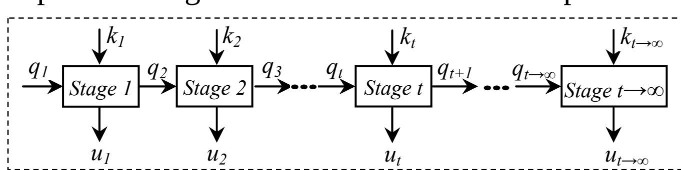


Fig. 2. The multi-stage decision optimization process

1). $\mathcal{Q} \in (0,1)$ is an infinite, continuous state space. A state $q \in \mathcal{Q}$ denotes the QS of the advertisement.

2). $\mathcal{K} = \{0,1,2,\dots,K\}$ is a discrete action space. An action $k \in \mathcal{K}$ denotes bidding for slot k and $k = 0$ means that the advertiser chooses to quit the auctions.

3). $\mathcal{T} = \{1,2,\dots\}$ is an infinite set of stages.

4). $\sigma: \mathcal{Q} \times \mathcal{T} \to \mathcal{K}$ is a strategy determining an action $k \in \mathcal{K}$ for each state $q \in \mathcal{Q}$ in stage $t \in \mathcal{T}$.

5). $u: \mathcal{Q} \times \mathcal{K} \rightarrow \mathcal{R}$ is a one-stage revenue function that specifies an immediate revenue for taking an action $k \in \mathcal{K}$ in a state $q \in \mathcal{Q}$. Following Varian [5], we focus on an exogenous advertiser who needs only to adjust its QS-weighted bid over $h(k)$ to obtain slot k and thus pay $h(k)/q$. Hence,

$$u(q,k) = \left[v - \frac{h(k)}{q}\right] x_k. \tag{1}$$

6). $f: \mathcal{Q} \times \mathcal{K} \times \mathcal{T} \rightarrow \mathcal{Q}$ is the state transition function that specifies the following state in terms of the current stage, state, and action. Note that QS equals the historical CTR of an advertisement. So if an advertiser with QS q_t targets at slot k_t in stage t, then

$$q_{t+1} = \frac{q_t * \sum_{i=0}^{t-1} c_i + x_{k_t} * c_t}{\sum_{i=0}^{t-1} c_i + c_t}, \quad t = 1,2,\dots \tag{2}$$

where c_t is the number of impressions that an advertisement gets in stage t, with c_0 intially. According to Assumption 3, $c_i = c_j$ holds for $\forall i,j \in 0,1,\dots$, and thus

$$q_{t+1} = f(q_t, k_t, t) = \frac{q_t t + x_{k_t}}{t+1}, \quad t = 1,2,\dots \tag{3}$$

7). $\delta \in (0,1)$ is the discount factor.

To maximize the discounted cumulative revenue, an advertiser should find the control law to determine an optimal sequence of actions, each in a stage, after observing its initial state. Formally,

$$argmax_{k_t \in \mathcal{K}} \sum_{t=1}^{\infty} \delta^{t-1} u(q_t, k_t) \text{ Subject to } q_{t+1} = \frac{q_t t + x_{k_t}}{t+1}, \ t = 1,2,\dots \text{ and } q_1 \text{ is known.} \tag{4}$$

3. Strategy Optimization Based on Policy Iteration

It is rather difficult to derive a closed-form solution to the above optimal control problem, since 1) the state transition is explicitly time-dependent; 2) the control variable k_t is implicitly embedded in x_{k_t} in the state transition equation and constrained to only discrete values. As such, we discretize the state space into a finite set of values, and use the cumulative revenue from stage 1 to T to approximate the infinite sum of revenues in all stages. This way, we can numerically approximate the optimal strategy based on policy iteration.

Since our model is time-dependent, the policies will be optimized though iterative search in the discrete state-action-stage space. The optimal value function satisfies the Bellman optimality equations and can be defined as follows.

$$V^*(q,t) = \max_{\substack{k_t \in \mathcal{K} \\ q_t = q}} \{u(q_t, k_t) + \delta V^*(q_{t+1}, t+1)\}. \tag{5}$$

In our model with finite states, actions, and stages, policy iteration is guaranteed to converge to the optimal strategy in a finite number of steps [6]. The detailed algorithm is as follows.

Algorithm 1. SSA Strategy Optimization Based on Policy Iteration

Input: An SSA scenario (including CTR, adrank, discounting factor, and initial QS and per-click value of the advertiser)
Output: The maximum cumulative revenue and the optimal positioning strategy

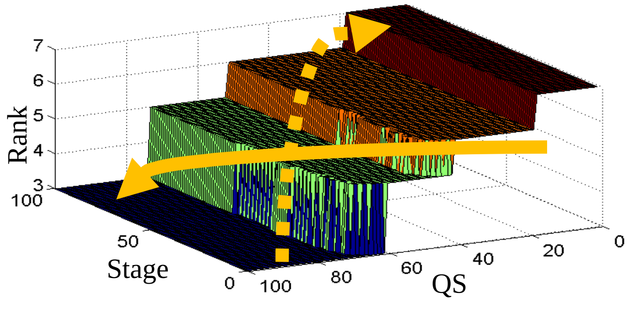
1. %% Step 1: Initialization
2. For each $q \in \mathcal{Q}$ and $t \in \mathcal{T}$
3. Arbitrarily set $V(q,t) \in \mathcal{R}$ and $\sigma(q,t) \in \mathcal{K}$
4. %% Step 2: Policy Evaluation
5 Repeat
6. $\Delta \leftarrow 0$;
7. For each $q \in \mathcal{Q}$ and $t \in \mathcal{T}$
8. $v = V(q,t)$; $q' = (q * t + x_{\sigma(q,t)})/(t+1)$;
9. $V(q,t) \leftarrow u(q, \sigma(q,t)) + \delta * V(q', t+1)$;
10. $\Delta \leftarrow \max(\Delta, |v - V(q,t)|)$;
11. Until $\Delta < \varepsilon$ (a small positive number)
12. %% Step 3: Policy Improvement
13. $PolicyStable \leftarrow True$;
14. For each $q \in \mathcal{Q}$ and $t \in \mathcal{T}$
15. $b \leftarrow \sigma(q,t)$;
16. $\sigma(q,t) \leftarrow \underset{k \in [0,K]}{argmax}[u(q,k) + \delta * V(q'', t+1)]$, where $q'' = (q * t + x_k)/(t+1)$;
17. If $b \neq \sigma(q,t)$, then $PolicyStable \leftarrow false$;
18. If $PolicyStable = false$, then goto 5.
19. %% Step 4: Algorithm Output
20. Set q_1=initial QS and cumulative revenue $u = 0$;
21. For $t = 1$ to T
22. $k_t = \sigma(q_t, t)$;
23. $u = u + u(q_t, k_t)$;
24. $q_{t+1} = (q_t t + x_{k_t})/(t+1)$;
25. Output u and the optimal strategy $(k_1, k_2, \dots, k_T)$.

Using Algorithm 1, we can investigate an advertiser's optimal positioning strategy in various SSA scenarios. Below we present our key conclusion with an illustrative example.

Table 1. Parameter setting of Example 1

Auction Parameters	Values	Algorithm Parameters	Values
# of Slots	6	# of Stages	100
Delta	0.7677	# of States	100
Per-click value	0.3094	ε	0.0001
QS	0.8837		
CTR	[0.6365,0.3951,0.3603,0.2098,0.1609,0.0771]		
Ad Rank	[0.5895,0.2816,0.1753,0.1632,0.1104,0.0571]		

Example 1: Without loss of generality, we use an auction scenario with randomly generated parameters, which are listed in Table 1.



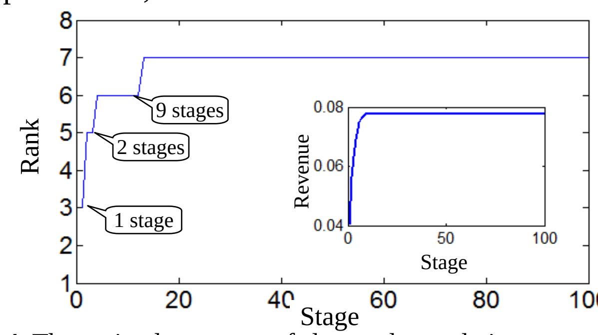


Fig. 3. The optimal strategy generated by Algorithm 1 **Fig. 4**. The optimal sequence of slots and cumulative revenue

Conclusion: An advertiser's optimal positioning strategy is to increase or decrease its rank monotonically. The direction of rank adjustments depends on its per-click value and initial QS.

Figure 3 presents the optimal rank at each state and stage. We can see obvious characteristics of stepped distribution of the optimal ranks. The width of each step is determined by an advertiser's per-click value. An advertiser with a higher value enjoys a wider step of better ranks and is more likely to get a higher rank. If an advertiser's initial QS is lower than the CTR of the optimal slot in stage 1, its QS and rank monotonically rise over time and the advertiser "steps down" following the solid arrow, until stabilizing at the highest possible slot. Otherwise, the advertiser has to lower its rank due to decreasing QS, until finally quitting the auction, as is illustrated by the dotted arrow. Figure 4 shows the optimal sequence of the advertiser's slot adjustment and its cumulative revenue. The advertiser drops from slot 3 to slot 5, then slot 6, and finally quits the auction.

4. Analysis of the Greedy Strategy

In practice, most advertisers do not have sufficient expertise to identify the optimal long-term strategy as our algorithm generates. Instead, they keep monitoring their QSs and make an optimal decision about the slot in the present stage. This is especially prominent for myopic advertisers who value the immediate revenue the most. In this section, we show that, due to the cumulative effect of CTR on QS, advertisers using this greedy strategy also monotonically adjust their ranks.

Definition – Greedy Strategy: A strategy is called a greedy strategy if

$$\sigma_G^*(q_t, t) = argmax_{k_t \in [0,K]}\, u(q_t, k_t), \quad t = 1,2 \ldots \tag{6}$$

Greedy advertisers select the optimal slot to maximize their revenues in each stage, without considering the influence of current actions on future states.

For simplicity, we define the inverse QS (IQS for short) of an advertisement as $L = 1/q$, where $L > 1$. Similarly, $\sigma^*(L, t)$ denotes the optimal slot for an advertiser with QS $q = 1/L$. An advertiser's revenue in slot k can be transformed into a linear function as follows.

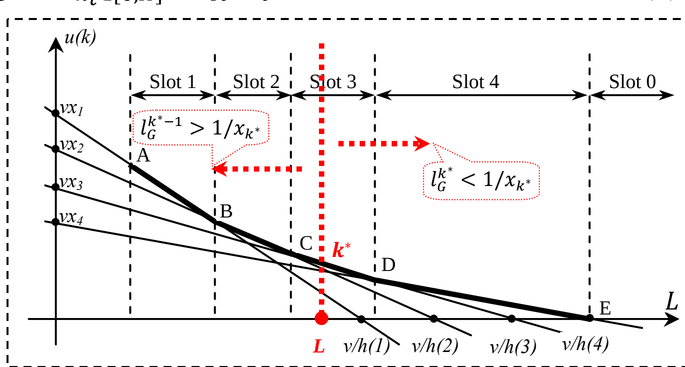


Fig. 5. The behavior dynamics of greedy advertisers

$$u(L,k) = -h(k)x_k L + v x_k. \quad (7)$$

The above equation represents a family of K revenue functions with $k \in [1, K]$, each specifying the revenue in one slot. These revenue functions have incremental slopes $(-1 < -h(1)x_1 < \cdots < -h(K)x_K)$ and decreasing intercepts $(vx_1 > \cdots > vx_K)$. Figure 5 plots these functions. For a specific value of IQS, an advertiser should choose the slot with the maximum revenue, which corresponds to the boldfaced segments A-B-C-D-E. These segments constitute an optimal path.

Below we analyze the optimal path of slot adjustments of a greedy advertiser. For each pair of slots (k, j), we can compute a matrix denoted as follows.

$$LM(k,j) = \frac{v(x_k - x_j)}{h(k)x_k - h(j)x_j}, \quad k, j \in [1, K], k \neq j. \quad (8)$$

Geometrically, $LM(k,j)$ is the value of the IQS corresponding to the intersection of revenue functions $u(L,k)$ and $u(L,j)$ in Figure 5. Using the LM matrix, we present the optimal path, the behavior dynamics, and the stable state of a greedy advertiser's slot adjustments, with three theorems. Due to the page limit, we omit the detailed proofs.

Lemma 1: The QS of an advertiser staying in slot k converges to x_k, i.e., $lim_{t\to\infty} q_t = x_k$.

Theorem 1: For any $q \in Q$ and $\sigma_G^*(q,t) = k^*$, a greedy advertiser's next higher targeting slot is $j^* = argmax_{j<k^*} LM(k^*, j)$, while the next lower targeting slot is $j^* = argmin_{j>k^*} LM(k^*, j)$.

Using Theorem 1, we can recursively determine the optimal path for a greedy advertiser's slot adjustment, starting from the top slot. The optimal path can be defined as follows.

Definition 2: The Greedily Optimal Path. The greedily optimal path $P = (k_G^1, k_G^2, \ldots, k_G^{|P|})$ is an increasing sequence of slots, where

$$k_G^i = argmin_{j>k_G^{i-1}} LM(k_G^{i-1}, j), k_G^1 = 1, i, j \in \mathcal{K}. \quad (9)$$

We also compute a vector $l = (l_G^1, l_G^2, \ldots, l_G^{|P|})$, where

$$l_G^i = min_{j>k_G^{i-1}} LM(k_G^{i-1}, j), \quad \forall i, j \in \mathcal{K}. \quad (10)$$

Each element l_G^i denotes the corresponding value of the advertiser's IQS for slot switching from k_G^i to k_G^{i+1}, and $l_G^{|P|} = v/h(K)$. It is worth noting that advertisers might not traverse all slots in their optimal path, so that $|P| \leq K + 1$. The reason lies in that a slot, say k, might not be an optimal action for all possible IQSs, or formally, $\forall L, \sigma_G^*(L,t) = k$ does not hold. Thus, the realized optimal path, denoted as P^*, is only a subset of P.

Theorem 2: For any $L > 0$ and $\sigma_G^*(L,t) = k^*$, a greedy advertiser will raise its slot if $l_G^{k^*-1} > 1/x_{k^*}$ and lower its slot if $l_G^{k^*} < 1/x_{k^*}$.

As is illustrated in Figure 5, Theorem 2 shows that if $l_G^{k^*-1} > 1/x_{k^*}$, a greedy advertiser can keep improving its QS and target at higher slots with more revenue realized. Otherwise, if $l_G^{k^*} < 1/x_{k^*}$, a greedy advertiser suffers from a decreasing QS and thus has to lower its slot with less revenue realized, and even finally quit SSA auctions (if $L > v/x_K$).

We can draw an important conclusion from Lemma 1, Theorems 1 and 2: The optimal positioning strategy for a greedy advertiser is to monotonically increase or decrease its ranks along the optimal path. The direction of the rank adjustments depends on its per-click value and the initial QS of its advertisement. Note from Figure 5 that the intersections of revenue functions with both axes are linear functions of advertisers' per-click values. So, advertisers' values actually serve as a scale controller that can proportionately zoom in or out the landscape of all revenue functions. Advertisers with higher values enjoy a proportionately higher LM matrix and hence a higher opportunity of raising their ranks. On the contrary, advertisers with lower values are more likely to be forced to target at lower slots with QSs further decreasing. So, due to the CTR-QS correlation, such divergent behavior dynamics is expected to cause a polarization trend in SSA markets.

Theorem 3: Greedy advertisers stabilize at the first slot in the optimal path that satisfies

$$l_G^{k^*-1} < \frac{1}{x_{k^*}} < l_G^{k^*}. \tag{11}$$

We define the slot k^* as an absorbing slot and the IQS value $L = 1/x_{k^*}$ as a stable state.

Theorem 3 determines the long-term stable state of a greedy advertiser's slot adjustments. It can be proven that at least one absorbing slot exists if advertisers monotonically increase their ranks. However, absorbing slots might not exist when advertisers decrease their ranks. In this case, advertisers will finally find that the optimal action is to quit the SSA auctions.

Example 2: We use the auction scenario in Table 2 to illustrate the above findings.

Table 2. Parameter Setting of Example 2

Auction Parameters	Values	Auction Parameters	Values
# of Slots	8	Per-click value	0.5168
Delta	0.98	QS	0.005
CTR	[0.8531,0.7622,0.6228,0.5217,0.3787,0.3015,0.2335,0.1600]		
Ad Rank	[0.4215,0.1611,0.0247,0.0140,0.0072,0.0058,0.0029,0.0011]		

We can recursively compute the greedily optimal path, $P = (1, 2, 3, 4, 5, 7, 8)$, and the IQSs for slot switching, $l = (0.20, 0.67, 6.46, 16.23, 36.31, 75.12, 469.82)$. Clearly, the 6th slot will never be an optimal action for any realized value of QSs and the top two slots will be unreachable, because $l_G^1 = 0.20 < 1/x_1 = 1.17$ and $l_G^2 = 0.67 < 1/x_2 = 1.31$. We can further verify that there is one unique absorbing slot, i.e., the 3rd slot, and the stable state is $L = 1/x_3 = 1.61$. Since the initial QS is $L = 200$, the realized optimal path is $P^* = \{3, 4, 8\}$, indicating that the advertiser starts from the 8th slot, increases its rank to the 4th slot, and stabilizes at the 3rd slot.

5. Conclusion

In this paper, we studied the decision optimization problem in SSA auctions, with a focus on the correlation between the QS and historical CTR of advertisements. Our finding has important implications for both search engines and advertisers. For advertisers, our work indicates that advertisers should monotonically adjust their ranks to better exploit the CTR-cumulative effect on QS for maximized revenue. We also present the optimal and greedy strategies to help formulate their positioning decisions. For search engines, our work shows that the CTR-QS correlation may result in a polarization trend. As a long-run consequence, search engines might lose their small advertisers and SSA will lose its superiority on cost efficiency as a "long-tail" advertising format. Our work highlights the needs for search engines to suppress this polarization trend by 1) setting a lowered weight for historical CTRs in their QS measurements to eliminate the CTR-QS correlation; or 2) removing positional bias in their CTR measurements through rank-normalization to break the causal effect between rank and CTR in Figure 1. Our work can be extended to include advertisers' game-theoretic interactions in incomplete information settings.

6. Acknowledgements

The authors gratefully acknowledge funding support from the National Natural Science Foundation of China (71102117, 70890084, 71025001, 71202169, 71071152, and 60921061).

References

[1] Y. Kiho. Optimal Quality Scores in Sponsored Search Auctions: Full Extraction of Advertisers' Surplus. *The B.E. Journal of Theoretical Economics*. 10(1):28-46, 2010.

[2] L. Li, D. Zeng, and H. Zhao. Pure-strategy Nash Equilibria of GSP Keyword Auction. *Journal of the Association for Information Systems*, 13(2):57-87, 2012.

[3] X. Zhang and J. Feng. Cyclical Bid Adjustments in Search-Engine Advertising. *Management Science*, 57(9):1703-1719, 2011

[4] http://www.adgooroo.com/has_google_changed_their_cpc_formula.php.

[5] H.R. Varian. Position Auction. *International Journal of Industrial Organizations*, 25(6):1163-1178, 2007.

[6] R.S. Sutton and A.B. Barto. Reinforcement Learning, MIT Press, Cambridge, MA, 1998.

MANIPULATION RESISTANCE IN FEEDBACK MODELS OF TOP-N RECOMMENDERS

Shankar Prawesh **Balaji Padmanabhan**
Information Systems and Decision Sciences
College of Business, University of South Florida
{sprawesh, bp}@usf.edu

Abstract. Top-N recommendation lists have been shown to have some undesirable properties such as a self-reinforcing nature and susceptibility to manipulation. One solution is probabilistic selection of articles into such lists. While this addresses some key problems of Top-N systems it has some practical limitations in use. We show that a simple *feedback model* can be an elegant solution here to model a continuum of selection mechanisms ranging from traditional Top-N selection to probabilistic variants. In this paper we study the issue of manipulation under feedback models. We show that the feedback process in these models can be exploited to design an influence limiter heuristic that can be very effective against manipulation of such systems.

1. Introduction

Top-10 (or top-N) lists such as "most popular" or "most viewed" are prominently displayed on the front page of most media websites. Some serious limitations of these mechanisms have been recently highlighted. Prawesh and Padmanabhan [1] showed that these systems are susceptible to amplifying negligible differences in the initial counts of the N^{th} and $(N+1)^{th}$ article. In similar vein the New York Times [2] notes that widespread use of these lists is leading us to less choice on the Web. It has also been noted that with some initial effort a manipulator can leverage the self-reinforcing nature of these lists to gain more popularity [1]. This potential threat has also been demonstrated by the managing editor of Newsweek, when he intentionally used a team to place a relatively old science story in the New York Times 'most e-mailed' list [3] using crowdsourcing to manipulate entry into this list (in this case the number of emails needed to get into the most e-mailed list of New York Times was shown to be as small as 1300).

One possible solution to these problems is to use a probabilistic NRS [1]. Such systems generate recommendations by sampling probabilistically from a pool of articles. An article's probability of being recommended is proportional to its current popularity. Though this has been shown to address the above issues to an extent, it has limitations as well. Perhaps most important this approach may sometimes select articles that are not as popular thereby potentially sacrificing short-term clicks. Second, probabilistic sampling in [1] does not provide media owners with flexibility in implementing variants to bias the system in certain directions as needed [4].

These limitations of probabilistic NRS can be addressed through sampling based on a feedback model [4]. Feedback models are used in applications where the behavior of the system creates either positive or negative feedback that affects the future behavior of the system. In economics attaining the monopoly share by one company among few competing companies is often considered an example of positive feedback [5]. Other examples of feedback models in different applications are modeling the early stage of neuron growth, Metcalfe's Law and even Microsoft's operating systems monopoly [6, 7].

We recently showed that such feedback models also have advantages in recommender systems in terms of allowing an implementer flexibility to incorporate the desirable properties of both probabilistic and top-N NRS [8]. Using a feedback exponent (γ) to bias probabilistic selection we show that a continuum of selection models can be elegantly modeled. Controlling this feedback parameter can help toggle between top-N and probabilistic selection elegantly, as well as other interesting selection models as well. We term these systems FNRS (News Recommender Systems with Feedback). Building on this recent work we study the issue of manipulation in FNRS. Through simulation experiments we show that the issue of manipulation in FNRS can be addressed effectively following the *influence limiter* approach of Resnick and Sami [9] in conjunction with the feedback parameter (introduced later).

2. Formalism

We use the simulation setup described in [1] to study feedback in Top-N recommenders. In this setup we maintain a Comprehensive List (CL) of articles and their counts. From CL, we "select" N articles into a Display List (DL) as recommendations. The Top-N recommender selects the N articles with highest counts. In FNRS, the probability that an article will be selected into DL is given by $p_a(t) = \frac{c_a{}^{\gamma}(t)}{\sum_j c_j{}^{\gamma}(t)}$. Where $c_a(t)$ represents the count of an article 'a' at a given time t and $\sum_j c_j^{\gamma}(t)$ represents the sum of counts of articles (those are not yet selected for DL) at time t to the power γ. This sampling process is repeated N times (without replacement, since articles selected into DL are no longer considered in the denominator) to generate recommendations.

The γ exponent in this selection probability creates different types of feedback. For $\gamma = 1$ the selection mechanism corresponds to probabilistic selection proposed by Prawesh et al. [1]. The higher the value of γ, the higher the probability that articles with high counts will appear in DL. At the extreme, when $\gamma \to \infty$ the selection mechanism corresponds to traditional top-N selection since each step will pick the current most popular article into DL. Since articles in DL are no longer considered when the next selection is made, this process simply picks the most popular articles where articles are selected corresponding to highest counts. Clearly $\gamma > 1$ creates a system with positive feedback where an article's chance of being recommended is more than its proportional share. Likewise, other settings of γ can be used in such a system to create random recommendations ($\gamma = 0$) when the focus is on diversity of recommendations. As the reader will note, though atypical, this exponent can also be set to recommend even the least popular articles if need be (perhaps to ensure that all articles - pre-selected when there is editorial oversight such as in places like the New York Times – have adequate exposure to readers). The use of this feedback model therefore addresses the limitation in [1] where a choice is made between either using the traditional top-N recommender or a simple probabilistic one.

In the simulation setup we implement a simple reader model in which a user upon arrival selects an article from DL randomly with probability p and from RL $(= CL - DL)$ with probability $1 - p$. The count of the selected article is increased by 1. For ease of exposition we intentionally leave out other factors of news arrival and reader behavior.

Accuracy and distortion have been used as properties to study these top-N systems. Intuitively, when articles other than the most popular are recommended, it is assumed there is an "accuracy loss". Likewise, if the recommendations over time distort the initial proportions of counts among articles there is "distortion". Traditional top-N systems have no accuracy loss but suffer from high distortion. Probabilistic selection [1] on the other hand has low distortion but some accuracy loss.

We define accuracy loss assuming the counts of articles that appear as recommendations in top-N NRS (at the corresponding time when implemented in parallel) as the benchmark [8]:

$$avg\ accuracy\ loss\ (\bar{E}_t) = \frac{1}{t}\sum_{j=1}^{t}\frac{1}{N}\ln\left(\frac{\sum_{i=1}^{N} C_{ij}^{\hbar}}{\sum_{i=1}^{N} C_{ij}^{\wp}}\right) \qquad (1)$$

In the above equation $C_{ij}^{\hbar}$ represents the count of i^{th} article, appearing in the Top-N NRS (the index "h" represents "hard cutoffs" that top-N systems use) at the j^{th} time step. Whereas $C_{ij}^{\wp}$ represents the count of i^{th} article appearing in the FNRS, at the j^{th} time step. Hence, $\sum_{i=1}^{N} C_{ij}^{\hbar}$ and $\sum_{i=1}^{N} C_{ij}^{\wp}$ represent the sum of counts of all articles that appear in Top-N NRS and FNRS respectively, at the j^{th} time step. This metric has been averaged over a number of iterations.

For distortion we use the $Kullback - Leibler$ distortion measure [10], averaged over the number of iterations (t) in the simulation. We denote the probability distribution of articles in the system in the presence of FNRS at the time step j as $q_j(x_i)$. Then the mean KL distortion for the articles $\{x_1, x_2, \ldots, x_n\}$ is given by

$$avg\ distortion(\overline{KL}_t) = \frac{1}{t}\sum_{j=1}^{t}\sum_{i=1}^{n} \mathrm{p}(x_i) \ln\left(\frac{p(x_i)}{q_j(x_i)}\right) \qquad (2)$$

Where n is the total number of articles in the system. The above expression represents the inefficiency of the distribution q_j when the true distribution of articles is given by p (given initially). Metrics defined in equation 1 & 2 can be used in following way to define a new combined loss metric that linearly combines accuracy loss and distortion [8]:

$$f_t(\beta, \gamma, p) = \beta * \bar{E}_t + (1 - \beta) * \overline{KL}_t \qquad (3)$$

The parameter $\beta \in [0,1]$ represents the designer's tradeoff between accuracy and distortion.

3. Manipulation Resistance and Simulation Results

The parameters used to present findings are as follows: number of readers = 100,000, DL = 10, CL = 200, initial counts of articles range from 0 to 1000, p = 0.9 and $\beta = 0.5$. At the start of the simulation the $(N+1)^{th}$ article was deliberately assigned a count of exactly one less than that count of N^{th} article to study impact of manipulation for the marginal next article. In the above setup the metric defined in equation-3 was obtained at the end of simulation. The corresponding feedback exponent (γ^*) for which the metric was minimized for these parameters was obtained as $\gamma^* = 4$. For the rest of the discussion on manipulation we have used this value of the feedback exponent.

By design, FNRS selects articles in such a way that it assigns higher probability in the next time step for the articles receiving higher counts. This property can be exploited by manipulators to create higher feedback for their target articles by injecting fake counts. To deal with this we present a manipulation resistant algorithm for FNRS, adapting the approach of Resnick and Sami [9]. We term this as Adapted Influence Limiter algorithm (AIL), and it modifies the exponent of the feedback mechanism for each article k at time t_i as follows:

$$\gamma_k^* = \gamma^*. \beta_{ik} = \gamma^*. \min\left\{1, \left(\frac{t_i - t_0}{\alpha^h. C_{\gamma ik}}\right)\right\} \qquad (4)$$

AIL operates between a preselected time interval (t_0, t_n); this can be the time interval when manipulation activities are most observed. For every $t_0 \le t_i \le t_n$ an article $k's$ exponent is updated as given in equation-4. The intuition is to penalize articles by reducing the feedback exponent if the arrival rate for clicks in a period is substantially different than might be expected. Readers may note that while not possible in a simulation study, in practice exponents need not be automatically reduced. Such articles can be highlighted and embedded in workflow that will alert media editors who may then approve any such (exponent reducing) penalty if the editors suspect foul play. In the expression above, α^h is the average time that is reasonable between two successive clicks for the recommended article in Top-N recommender and $C_{\gamma ik}$ is the new clicks (after t_0) received by the k^{th} article at the time t_i. After t_n, each new count received by any of the articles is modified through its corresponding value of β_{nk} at the time t_n.

Though the effect of an article manipulation in FNRS can be studied in different ways, in the present research we study the impact of manipulation in different scenarios based on article ranking, duration of attacks and the number of fake counts injected for the target article. In particular, articles of interest are $(N+1)^{th}$, a marginal article (selected randomly among next 10 articles after top-N list) and an article with low ranking (selected randomly among bottom 10 articles). All selections mentioned above are determined based on the ranking determined through initial counts of the article.

To get maximum 'benefit' from fake clicks manipulators will try to inject fake counts in the early part of the life-span of an article. So, we focus on the effects of early manipulation in FNRS. In addition, we also discuss cases where manipulators may try to inject large numbers of fake counts over a long time interval. In all cases manipulated counts were uniformly distributed over a given interval. These numbers are represented as (-,-) in the third column of Table 1 – where the first entry indicates the number of fake counts and the second entry represents the time interval over which these fake counts were uniformly distributed.

Table 1: Different Cases of Manipulation

Target article for the manipulation	Cases	Operationalization
Manipulation of $(N+1)^{th}$ article	Early little	(100, 5000)
	Early heavy	(500, 5000)
Manipulation of marginal next article (18^{th})	Early little	(100, 5000)
	Early heavy	(500, 5000)
Manipulation of a low ranked article	Early heavy	(1000, 10000)
	Uniform heavy	(5000,50000)

We examine the success of manipulation through two measures (a) total clicks received by the manipulated article over the complete simulation (excluding manipulated clicks, presented in the left panels of figures) and (b) counting the number of times the manipulated article appears in the recommended list (presented in the right panels of figures). The paths followed by these two measures have been produced in the graphs below.

Manipulation of $(N+1)^{th}$ *article*. By using $(N+1)^{th}$ article as target, in case of FNRS we show that carefully crafted manipulation may go undetected even in presence of AIL, especially if an article has just missed the cutoff for Top-N and the manipulator is in a position to determine this. This case is possible when manipulator has knowledge of the popularity of articles (Figure 1). Also, it should be noted that popularity of articles is often displayed by many media sites.

Table 2: Abbreviations

AIL*n*	Clicks received by the n^{th} article upon manipulation in the presence of AIL
MN*n*	Clicks received by the n^{th} article upon manipulation and in absence of AIL
UM*n*	Clicks received by the n^{th} article without any manipulation
ILappear*n*	Number of times the manipulated n^{th} article appears in DL in the presence of AIL
MNappear*n*	Number of times the manipulated n^{th} article appears in DL in absence of AIL
Uappear*n*	Number of times the n^{th} article appears in DL without any manipulation

Figure 1 (next page), presents the case of manipulation with 100 manipulated clicks distributed over initial 5000 clicks. Clearly, it can be observed that the performance of AIL algorithm is indistinguishable from the manipulated counterpart in terms of number of clicks received by the article (left panel). Hence, it suggests that in this case even in presence of influence limiter mechanism a manipulator may take advantage of the feedback nature of FNRS (but substantially less than Top-N NRS). The possible explanation for this finding is that fake clicks are distributed over a relatively large interval. Hence, the effect of AIL is not observed on the exponent. Similar findings were observed for the appearance of article in the recommended list in FNRS (Figure 1, right panel) with significant benefit in case of manipulation.

Similar findings were also observed in the case of manipulation for the randomly chosen article from the articles after top-10 list (not reported in detail for lack of space). For the case of heavy manipulation in the initial part of the simulation, the simulation paths are produced in the figure-2 (next page). It is evident that the case of heavy manipulation of an article may not be beneficial for a manipulator in FNRS with AIL. Often very quickly its feedback exponent is reduced to the extent that the manipulated article receives even less new clicks (figure 2, left panel) and appearance (figure 2, right panel) than the non-manipulated article in the similar scenario.

Again, similar trajectories were observed in the case of manipulation for the randomly chosen article from the next 10 articles after the top-10 list. Hence, this suggests that manipulators will have very little motivation to attack FNRS heavily in the presence of AIL.

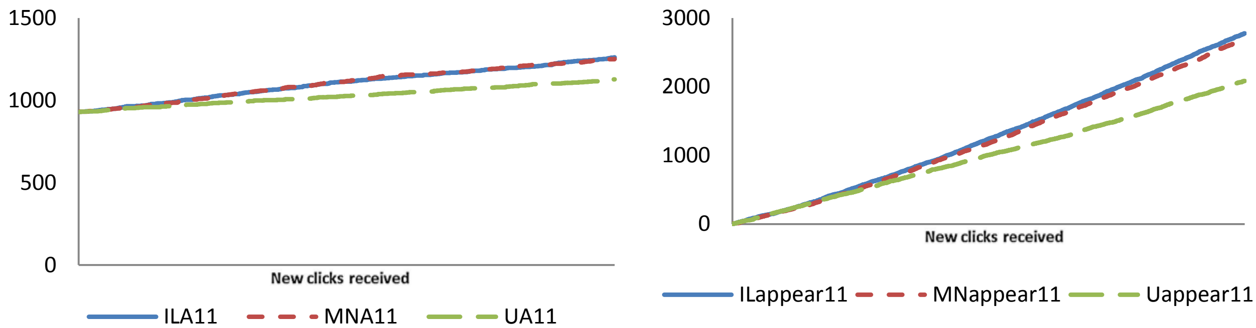


Figure 1: Small Manipulation Effects for the $(N+1)^{th}$article

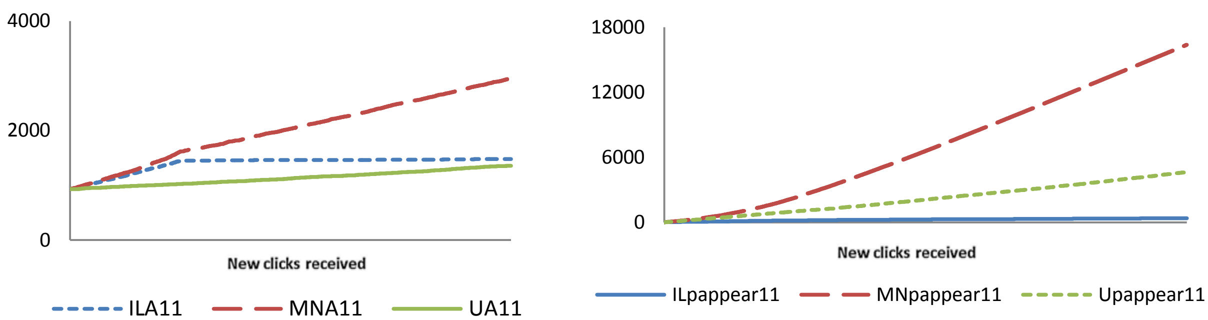


Figure 2: Heavy Manipulation Effects for the $(N+1)^{th}$article

Manipulation of a low rank article. Manipulation of a low ranked article needs substantially large number of clicks by manipulators. Hence we have considered two cases with 5000 and 1000 fake clicks randomly distributed over initial 50,000 and 10,000 iterations of the simulation respectively. As figure-3 suggests, through heavy manipulation in the presence of AIL for FNRS manipulators do not get much benefit. As, for both clicks (left panel) and appearance (right panel) we do not observe any significant increase in their count.

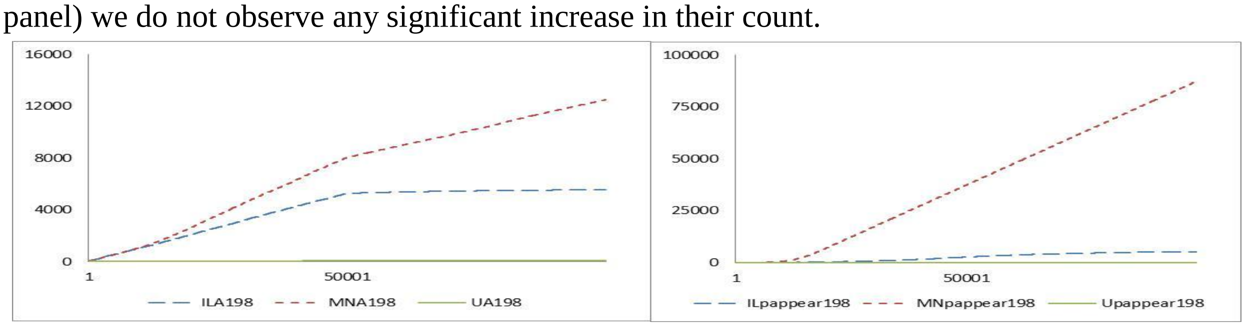


Figure 3: Heavy Manipulation for the Low Rank Article

These results suggest that FNRS with AIL can be effective to thwart manipulation efforts by heavily injecting fake counts through automated programs such as ‘bots’ and “zombie followers” [11, 12]. The case which may go undetected is when small numbers of fake clicks are distributed over long time intervals. These findings suggest that the feedback nature of FNRS can be exploited to make manipulation particularly difficult in top-N recommenders.

4. Conclusions

There is growing recognition of limitations of Top-N lists. Manipulation, in particular, still remains a largely unaddressed problem [13]. This research therefore makes an important contribution towards addressing these issues faced by firms who use top-N lists of any form. Building on recent work we show that the use of a feedback model in the probabilistic mechanism can be an elegant solution to prior limitations of probabilistic selection [1]. Further, the feedback mechanism here can be particularly valuable to counter heavy manipulation of such systems. The adapted influence limiter presented here can be extended in different ways to control the exponent continuously as new clicks (and information) are recorded by systems. In the case of controlling manipulation in the media we also suggest that this feedback modifier can be used in conjunction with workflow alerts to editors who may make final decisions before penalties kick in.

Though we have presented our findings in the context of news recommenders we believe our approach is easily implementable by users who face the issue of manipulation on app stores or for product recommendations. Indeed, uncontrolled manipulation can have significant human and innovation costs as well. As programmer Walter Kaman states [13], *“As an indie developer, this (manipulation) totally demoralizes my passion to continue making apps for this platform, knowing that at the end of the day, all that matters is $5,000 and a bunch of bots.”*

References

1. Prawesh, S. and B. Padmanabhan, *The "top N" news recommender: count distortion and manipulation resistance (RecSys' 11)*, in *Proceedings of the fifth ACM conference on Recommender systems*2011, ACM: Chicago, Illinois, USA. p. 237-244.
2. Bilton, N., *Disruptions: Top 10 Lists Lead to Less Choice on the Web. The New York Times.*, in *The New York TImes* April 1, 2012.
3. Weber, T.E., *Cracking the New York Times Popularity Code*, in *The Daily Beast* December 19, 2010.
4. Prawesh, S. and B. Padmanabhan, *Probabilistic news recommender systems with feedback*, in *Proceedings of the sixth ACM conference on Recommender systems*2012, ACM: Dublin, Ireland. p. 257-260.
5. Mitzenmacher, M., R. Oliveira, and J. Spencer, *A Scaling Result for Explosive Processes.* The electronic journal of combinatorics, 2004. 11(R31): p. 1.
6. Khanin, K. and R. Khanin, *A probabilistic model for the establishment of neuron polarity*. Journal of Mathematical Biology, 2001. 42(1): p. 26-40.
7. Adomavicius, G. and K. YoungOk, *Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques.* Knowledge and Data Engineering, IEEE Transactions on, 2012. 24(5): p. 896-911.
8. Prawesh, S. and B. Padmanabhan. *News Recommender Systems with Feedback*. in *ICIS 2012*. 2012. Orlando (To appear).
9. Resnick, P. and R. Sami, *The influence limiter: provably manipulation-resistant recommender systems*, in *Proceedings of the 2007 ACM conference on Recommender systems*2007, ACM: Minneapolis, MN, USA. p. 25-32.
10. Kullback, S. and R.A. Leibler, *On information and sufficiency*. The Annals of Mathematical Statistics, 1951. 22(1): p. 79-86.
11. *Zombie followers and fake re-tweets.* The Economist March 17th, 2012.
12. Rusli, E.M. and B.X. Chen, *For Apple, Pressure Builds over App Store Fraud*, in *The New York Times* March 15, 2012.
13. Chen, B.X., *Apple Fights App Makers Who Try to Game the System*, in *The New York Times* February 7, 2012.

Social Media-based Social TV Recommender System

Shawndra Hill, Adrian Benton, Jin Xu
University of Pennsylvania, 3730 Walnut Street, Suite 500
Philadelphia, PA 19104
shawndra@wharton.upenn.edu

Abstract

We demonstrate the effectiveness of a novel recommendation system built using user-generated content (UGC) found on social media. We have collected a unique and large data set using an approach we developed for data collection that enables us to make and evaluate product recommendations, in this case, television show recommendations. The two main contributions of this paper are: 1) A new methodology for collecting data from social media, including data on social networks, product networks, geographic location data and user contributed text, to validate social media-based recommendation systems; and 2) A new UGC-based recommendation system that relies on general publicly available free form text, as opposed to features extracted from UGC using a preselected set of keywords. We show that our approach performs remarkably well at predicting TV shows that users of the social networking site Twitter like.

1. Introduction

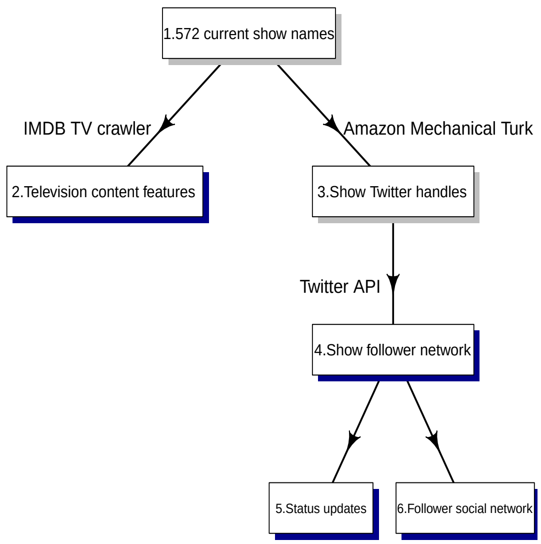
Recommendation systems (RS) that suggest products to consumers have been shown to increase sales and alert consumers to new products [1]. While RS have been extensively used by online firms, it is only recently that researchers have tried to expand them to include context (location, mood), social networks (what friends are doing and buying), and text (from consumer reviews and tweets) to extract meaningful information. In this paper, we look to build television RS that include all of these types of data, sourced from the social media site Twitter.

The micro-blogging platform Twitter, where people contribute tweets of 140 characters or less, provides a rich dataset of real-time commentary on almost every aspect of life, including consumer response to TV shows and TV advertisements [7, 6, 2]. Twitter has been used extensively as a research testbed. For example, it has been used in RS to recommend who a user should follow on Twitter [5]and websites or news stories that users might find interesting [3, 9]. Prior research has also focused on using the content of tweets to identify the characteristics of users, such as demographics and geo-location data. In addition, tweet content has been used to predict the future popularity of topics discussed on social media.

In this paper, we explore using the data generated when people interact online, in public about their daily lives. We show that the content of the text followers of TV shows generate on Twitter is extremely useful when calculating the similarity between TV shows. The similarity calculations are then used to predict whether users are likely to follow a particular show or not; although, preliminary, we think our work is the first to both compare different types of recommenders based on social media data in one context for many "products" and automate the use of user generated content in RS. User generated content has recently been shown to be useful in RS [4] and when calculating the similarity between brands [8]. However in prior work, researchers only consider a list of keywords associated with the products or brands in the user generated content for analysis. The purpose of this paper is to demonstrate that the user generated text, in general, has value without needing a predetermined set of keywords. In Section 2, we describe the novel dataset collection method used for this study.In Section 3, we describe our recommendation system methodology. In Section 4, we describe the results followed by conclusions and next steps in Section 5.

2. Testbed

We collected a large database of TV-related content for our recommendation system. A flowchart illustrating our data collection process can be found in Figure 1a. The approach is novel because it enables us to test recommendation systems using user revealed preferences found in publicly available user generated content. The process consisted of six phases.



(a) Data Collection Process

Algorithm 1 Recommendation Evaluation Process

```
1: Input: A recommendation engine e, 10 sets of users (with the shows
   list they followed) based on cross-validation {test[1], train[1],···,test[i],
   train[i],···,test[10], train[10]}.
2: for (i IN 1:10) do
3:    Prepare to list the results for each set:
      results[i] = [ ]
4:    Train a recommendation metric based on the training set:
      Metric[i] = TRAIN(e, train[i])
5:    Test on each user u_j in the test set
6:    for (u_j IN test[i]) do
7:       Randomly choose a show from u_j shows list:
         randshow(j) = GET_RANDOM_SHOW(u_j)
8:       Use the trained metric to recommend show for user u_j:
         recommended(j) = PREDICT(Metric[i], u_j, randshow(j))
9:       Evaluate the performance of recommendation:
         results_byuser[j]=EVALUATE(recommended(j), u_j, randshow(j))
10:   end for
11:   Get the average performance for each test set:
      results[i]=average(results_byuser)
12: end for
13: Output: (SUM(results)/10)
```

(b) Algorithm for Recommendation Evaluation Process

Figure 1: Overview of Processes

First, we selected a 572 current TV shows airing during the January - June 2012 timeframe. Second, we used the Internet Movie Data Base (IMDB.com) to collect metadata for all of the shows. The metadata included genre, year released, TV network, etc. The metadata collection process did not rely on Twitter data. Third, we collected the Twitter handles for each of the current shows on our list using Amazon Mechanical Turk.

The remaining three steps rely on data collected from the Twitter API. In the fourth step, we collected the lists of all users that follow the TV shows in our selected sample. We reduced the list to users that follow at least two of the selected TV shows‘ Twitter accounts. This is an important criteria that will enable us to evaluate our methods on these users. Additionally the users made at least 200 tweets in the past and did not have over 2000 followers (to rule out celebrities and atypical TV viewers). In total, we collected data on over 10 million users who followed more than 2 TV shows. We randomly sampled approximately 1000 show followers from each show follower list that met our criteria. In the fifth step, we collected up to 400 of the most recent posts from the sampled show followers. In the sixth and final step, we collected the social networks (the followers of the followers of the TV shows) of the TV show followers. So from Twitter, we were able to collect millions of TV show followers, their tweets and their follower friends. These data will be used to make and validate recommendations using our approach.

3. Method

In the section, we describe a set of RS that we built using user generated content. For each method, we assume one input TV show per user, picked at random from the shows the user follows, and we use that input show to make predictions for other shows the user might like by calculating the similarity between the input show and other shows using various metrics. For each approach, we calculate the similarity between shows using a training set of approximately 90,000 users and apply the similarity matrix to a set of test users. We perform 10-fold cross validation for all methods to report results on 10 training/test data pairs. Algorithm 1 in Figure 1b above describes our general approach for all metrics. With one exception, the social network-based recommender that we will describe below, the methods are not personalized to the user. In this paper, we evaluate our predictions using standard RS measures of precision and recall. Precision is the number of predictions we get right over the total number of predictions made and recall is the number of TV show predictions we get right over the number of shows the viewers actually follow.

The similarity score metrics used to build the RS are described in the subsections below. The list includes network-based, text-based and baseline approaches. We will discuss all of the methods we applied below. But, we will only present results for a select number of high performing text and network-based models due to space constraints.

3.1 Text-based Approaches To compute user-generated text-based similarity between all shows, we used the Tweets that we collected on followers of TV shows.

3.1.1 Text TFIDF Similarity Recall from Section 2 that we sampled approximately 1000 users per show. For each show, if a user was known to follow that show, then all of his/her tweets were added to that show's tweet corpus. Each show tweet corpus was then tokenized, hashtags, handles, URLs were excluded, and a "bag of words" was built for each show, along with counts for each token. The similarity between two shows was generated using the cosine similarity between their bags of words, with token counts transformed by using a Term Frequency-Inverse Document Frequency count (TF-IDF). This transformation was used to discount highly frequent words from overwhelming the *bag of words* vectors. The TF-IDF value for the tth token in a particular bag of words v_i was defined as follows: $v_{it}/|\{j|v_{jt} > 0\}|$.

3.1.2 Text-based TV Show Mention Tweets We also used an alternative approach where we considered only tweets that mentioned a show's Twitter handle to be included in its bag of words. This resulted in a significantly smaller corpus, just over 370,000 tweets. The similarity was computed as the cosine similarity between the log-transformed vectors of the two shows.

3.2 Network-based Approaches The network-based approaches took advantage of the link structure between users and shows. We considered two types of networks. The first is the show product network where shows are linked together through users. The second is the social network of users where recommendations are made to users based on what their friends follow.

3.2.1 Product (TV Show) Network In the TV show product network approach, we measured the association between pairs of shows using the association rule metrics, support and confidence. For a given show pair A and B, support is measured by the proportion of users in our data set that connect to the show pair. Directional confidence is measured by the number of users that connect to both A and B / the total number of users that connect to A. We took the average of support and confidence to generate directional scores between pairs of shows A and B.

3.2.2 Social Network In the social network based approach, we considered the TV show following patterns of a user's friends to make predictions for them. A user was recommended the most frequent show that user's friend follows. So if three of user A's friends follows American Idol, two follow the Voice, and one follows Duets, American idol would be recommended first, then the Voice, then Duets. Since there is no input show using this method, to make sure the denominator remained the same across all methods when calculating precision and recall, we randomly removed a show from the evaluation set for each user.

3.3 Baselines We considered a number of baselines. *Categorical Popularity-based Method:* When a user provides a past-liked show, the recommender engine will return the most popular shows for all training set users. *Geography-based Method:* By grabbing the available location information, at the State level, of users from the Twitter bio field, the system will returns the TV show with the largest number of followers in that area. We have applied our own sophisticated methods to infer the geographic location of users. However, space does not allow us to discuss this process in detail. *Gender-based Method:* Similar to the geography-based method, we recommend the most popular shows by gender. *Content-based approach:* We collected the metafeatures of recent TV shows from the IMDB and computed the similarity between all the shows with respect to each of these features separately. We then applied a linear weighting of these features from a reserved set of users to combine these features appropriately. We used ordinary linear regression, in R, to determine this weighting on a validation set of users.

4. Results

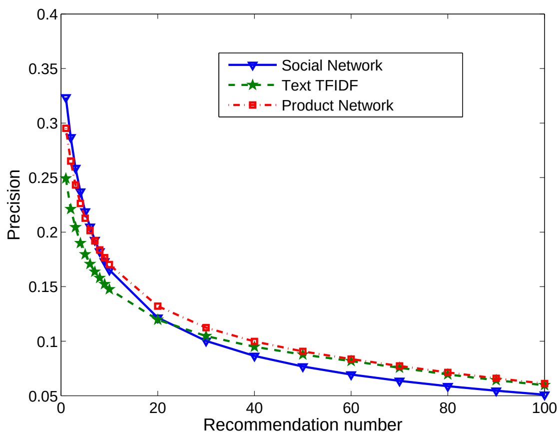


Figure 2: Precision

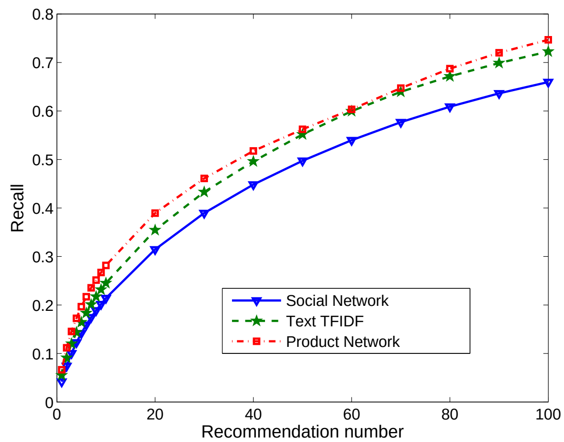


Figure 3: Recall

The primary goal of this paper is to demonstrate that our proposed general text-based TF-IDF approach, which is based on publicly available content, has significant value for TV show preference prediction. In this section we present the results of comparing a select set of the aforementioned recommendation strategies by precision and recall. In Figures 2 and 3, we show three things: 1) all highlighted approaches perform remarkably well in terms of precision when making recommendations (we perform with 29%, 32.5% and 25% accuracy for product network, social network and TF-IDF text respectively for making 1 recommendation compared to a random baseline of 3%) ; 2) the TF-IDF approach performs almost as well as the product network approach, which could be equated to a standard collaborative filtering approach; and 3) the social network-based approach performs exceptionally well when making a smaller number of recommendations but the performance of the social network-based approach deteriorates quickly.

The results are important for two reasons. First and foremost, the user generated content alone in the form of tweets are powerful predictors that can be used when a product or social network is not available or costly to acquire. Second, we find that the different types of social media content, although based on the same set of users, yield different types of predictions. We find that the TV show product network approach favors recommending high degree shows (or shows with a lot of followers) while the Text-based TF-IDF Tweets approach is more likely to recommend low degree shows adding to the diversity of recommendations. The combination of the two approaches allows us to increase the diversity of predictions as well as increase the accuracy of the recommendation system.

Our methods was performing so well that we further investigated the similarity networks that were derived from the user generated content. We investigated the words that were representative of the shows by ranking the words for each show by their TF-IDF scores. We restricted the words for inspection to words that were in the English dictionary since many terms on Twitter are not and are therefore hard to interpret. Table 1 lists top words that represent a number of shows in our TV network. The word lists are quite telling. They appear to represent both demographics of the viewers and features of the TV shows very clearly. Additionally, we wanted to see whether our similarity score matrix was enabling us to segment the shows by show features. In Figures 4 and 5, we present the results of clustering the shows using the network-based clustering algorithm

Table 1: Key words in shows

American Idol	Amsales Girls	Colbert Report	RuPaul's Drag Race	Thundercats Now	Beavis and Butthead
idol	bridal	petition	gay	samurai	f**k
birthday	wedding	bullying	lesbian	marvel	s**t
snugs	gown	newt	drag	barbarian	f**king
god	bride	republican	equality	cyborg	loco
recap	curvy	tax	marriage	batman	b**h
finale	meditation	president	maternal	comic	a**
bullying	fortune	f**k	cuckoo	wars	hate
love	coziness	debate	s**t	watchmen	damn
excited	respectable	freedom	b***h	spiderman	smoke
happy	hopefulness	unsigned	jewelry	extermination	stupid

METIS. For illustration purposes, the links are based on the Text-based TF-IDF similarity scores since we believe that approach is the most novel. When we calculate the diversity in the clusters measured by entropy over the different TV show attributes (in this example, we used the TV show networks, NBC, ABC, CBS, etc.), we find that there is a reduction in the amount of entropy in the clusters that are found based on the similarity score matrix compared to a network that is formed by placing random links between the shows. The plots indicate that the networks that we are constructing actually do have value at partitioning the set of shows, at least by TV show network.

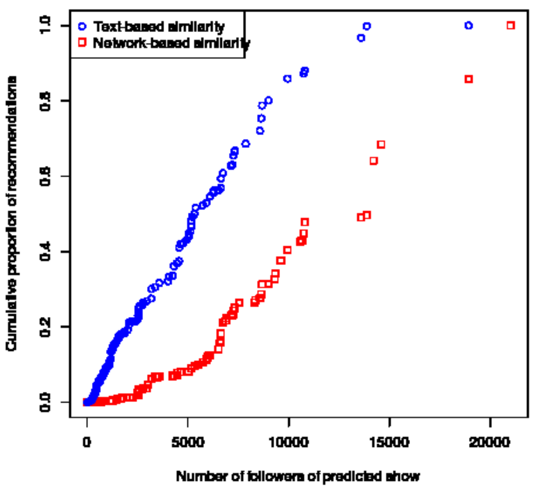


Figure 4: Followers of predicted show

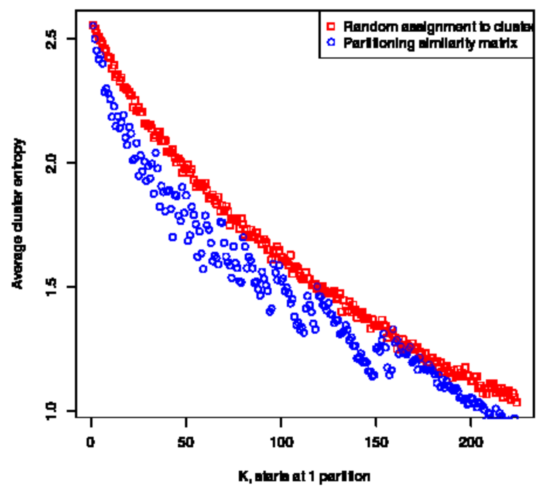


Figure 5: Average entropy of clusters

5. Conclusion

We have demonstrated how to collect a data set from publicly available Tweets that can be used to validate a wide range of recommendation systems using a number of evaluation measures. In addition, we have shown that user generated content has value. This work is a first step at constructing a *pyschographic* profile for products, in this case TV shows, using user generated content contributed by followers of products.

We find it is remarkable how good the recommendation engines based on user generated content perform compared to traditional methods like content-based and demographic based approaches. We find that not only are the user generated content-based approaches better at predicting the TV shows users follow, but they also enable the recommendation engine to make more predictions (as compared to say demographic based approaches where you can only make predictions for users you have demographics on). Using both the links and text of posters on Twitter, we are able to calculate the similarity between most pairs of shows. On the other hand for approaches based on location and gender, the recommendations are typically hard to make for users because demographic data is only available 1% of the time. We find that the same is true for the social network based approach. While the social networks enable better predictions for a small subset of users and for a small subset of shows, a social network approach can only be applied for a small number of users because of the sparsity of the tv-show following network. In both network-based approaches, the relationship between sparcity or support for links in the models used to calculate similarity and predictive power is extremely strong.

Although, our results are preliminary, there are many questions our data an analysis can answer. For example, under which conditions does network-based both TV show network and social network approaches work to make better predictions. In addition, we can begin to understand which features of the text are more useful for calculating similarity between shows and likewise, which types of users are the most helpful. In addition, we can ask how all of the answers change as a result of the network changing quickly as TV show popularity rises and falls in a seasonal way. One of our next steps is to build a real-time TV recommender. Another next step is to test out our approaches in a lab setting to see which recommenders users like and trust more. Finally, we intend to extend our results to include more personalized recommendations. We are very excited about these preliminary results and will certainly have a larger set of results to present at the WITS workshop.

6. Acknowledgements

We would like to thank Jie Qie for help with a much earlier prototype of our system. We would also like to thank the Office of Naval Research for research support. A visualization of the network of relationships between shows using our metrics can be found at www.thesocialtvlab.com.

An additional, extensive reference list available upon request.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] Adrian Benton and Shawndra Hill. The spoiler effect?: Designing social tv content that promotes ongoing wom. In *Conference on Information Systems and Technology*, Pheonix, AZ, 2012.

[3] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1185–1194, Atlanta, Georgia, USA, 2010. ACM.

[4] Anindya Ghose, Panagiotis G. Ipeirotis, and Beibei Li. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Science*, 31(3):493–520, 2012.

[5] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the Fourth ACM conference on Recommender systems*, pages 199–206, Barcelona, Spain, 2010. ACM. 1864746 199-206.

[6] Shawndra Hill and Adrian Benton. Social tv: Linking tv content to buzz and sales. In *International Conference on Information Systems*, Orlando, FL, 2012.

[7] Shawndra Hill, Aman Nalavade, and Adrian Benton. Social tv: Real-time social media response to tv advertising. In *ADKDD 2012: The Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, Beijing, China, 2012.

[8] Oded Netzer, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko. Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521 – 543, 2012.

[9] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM conference on Recommender systems*, pages 385–388, New York, New York, USA, 2009. ACM.

Product reputation manipulation: Exploring the linguistic characteristics of shill reviews

Toan Ong and Michael V. Mannino[1]

Abstract

This study explores the linguistic characteristics of shill reviews which are used to manipulate the reputation of products sold on websites. We collected shill reviews via an experiment. Using semi-automatic natural language processing techniques, we compared the shill reviews with normal reviews on informativeness, subjectivity and readability. The results show statistically significant differences between shill reviews and normal reviews in subjectivity and readability. Informativeness appears to be a weak separator of shill and normal reviews so additional studies may be necessary. Overall, the study provides improved understanding of shill reviews with an eventual goal to increase effectiveness of review filtering methods.

Introduction

As customer reviews have become an important information channel for online purchase decisions, there has been manipulation of product reviews to dishonestly change the consumers' perception toward the quality of targeted products. Although recent studies found evidence of product reputation manipulation on popular online marketplaces such as Amazon.com and TripAdvisor.com (Wu, Greene et al. 2010; Hu, Bose et al. 2011; Hu, Liu et al. 2011), the characteristics of the fake reviews were not fully explored. This gap is partially due to the difficulty to collect data on fake reviews (Jindal and Liu 2007).

In this study, we refer to fake reviews as *shill reviews*. We focus on the undisclosed relationship between review writer and product seller. A shill review misleads online shoppers because it is written by a reviewer with an undisclosed relationship to the seller. This shill review definition builds on definitions of spam reviews (Jindal and Liu 2007) and review management (Hu, Bose et al. 2011).

The popular press has numerous examples of review manipulation through shill reviews. In 2009, Belkin, a networking and peripheral manufacturer, was reportedly hiring people to write fake positive reviews for their products on Amazon.com[2]. Later, Belkin management issued an apology for this action[3]. In 2012, the Denver 7News channel discovered that a woman was hired to create more than 50 Google accounts to publish 5-star reviews for multiple local businesses[4]. In the music industry, marketers, disguised as consumers, promoted newly released CDs on discussion forums and fan sites (Mayzlin 2006).

In this study, we propose a method to reveal the characteristics of shill reviews by analyzing the free-text part of the review using natural language processing (NLP) techniques. Our shill review dataset was collected via an experiment in which reviewers were rewarded for deceptively submitting positive reviews even though reviewers never previously used the product. We compared the characteristics of the shill reviews and normal reviews on informativeness, subjectivity and readability.

---

[1] The Business School, University of Colorado Denver, Toan.Ong@ucdenver.edu and Michael.Mannino@ucdenver.edu

[2] http://www.thedailybackground.com/2009/01/16/exclusive-belkins-development-rep-is-hiring-people-to-write-fake-positive-amazon-reviews/

[3] http://news.cnet.com/8301-1001_3-10145399-92.html

[4] http://www.thedenverchannel.com/news/31087210/detail.html

This study is the first attempt to collect real shill reviews via an experiment setting. The comparison between the shill reviews and normal reviews reveals the shill reviewers' weakness, lack of knowledge and usage experience about the product. The findings of this study show the role of informativeness, subjectivity and readability as important separators between shill reviews and normal reviews. This result should help improve the effectiveness of current review filtering methods[5].

Shill review background and hypotheses

Because the relationship of reviewers to product seller is unknown, detection of shill reviews is difficult even if performed manually (Jindal and Liu 2007). Lack of labeled training data for shill reviews also hinders development of machine learning techniques to detect shill reviews (Hu, Liu et al. 2011). By applying the discretionary accrual-based earnings management framework, Hu, Bose et al. (2011) showed that review management occurred with several categories of products. However, the above study simply detects the occurrence of reputation manipulation without specifying shill reviews. Jindall and Liu (2008) detected shill reviews by looking for similar reviews but not all shill reviews are duplicated. Ott et al. (2011) attempts to detect spam reviews using psychological and linguistic characteristics of the review writing. The authors claim that they were able to accurately detect 90% of the hotel spam reviews. Other research (Lim, Nguyen et al. 2010; Wu, Greene et al. 2010) has analyzed rating behavior, an indirect indicator of shill reviews.

In this study, we use a direct approach to understanding the characteristics of shill reviews through examination of the main differences found in the free-text part of shill and normal reviews. We argue that main differences between shill reviewers and normal reviewers involve knowledge about the product and the product usage experience. We assume that shill reviewers have never used the target product, a reasonable assumption since it is too costly to send most products to shill reviewers.

A shill review can be either positive or negative. Positive shill reviews increase and negative shill reviews undermine the reputation of the target product. In the scope of this study, we only consider the case of positive shill reviews. We argue that positive shill reviews might be more popular than negative shill reviews because they have more direct impact on the target product. However, negative shill reviews are also interesting to study, especially for markets having only a small number of competitors..

Following the approach in (Liu, Cao et al. 2007), we define informativeness of a review as the amount of product information provided in the review. Informativeness of a review is measured by using the product features mentioned in the reviews. Both shill reviews and normal reviews are expected to include multiple product features. However, the characteristics of the product features included in shill and normal reviews can be different. Since normal reviewers have used the target product, they may have private information about features unknown to shill reviewers. We make another assumption that the shill reviewers do not read the normal reviews before posting the shill review. In contrast, the shill reviewers only use the product description containing the official features provided by the manufacturer to write their reviews. Therefore, shill reviewers are expected to include less unofficial product features and more official features in their reviews. We hypothesize that:

H1a: Shill reviews contain more official features per sentence than normal reviews.
H1b: Shill reviews contain less unofficial features per sentence than normal reviews.

---

[5] http://officialblog.yelp.com/2010/03/yelp-review-filter-explained.html

H1c: The percentage of sentences containing official features of shill reviews is more than that of normal reviews

H1d: The percentage of sentences containing unofficial features of shill reviews is less than that of normal reviews.

The product usage experience is measured by the subjectivity and objectivity of the sentences in the reviews. According to (Ghose and Ipeirotis 2004), a subjective sentence “gives a very personal description of the product”. An example of a subjective sentence can be “It’s really a great little player”. An objective sentence “lists the characteristics of the product”. An example of an objective sentence can be “It even includes a computer USB interface and built-in speaker”. Since normal reviewers have used the product, they know their feelings about the product. So, normal reviews are expected to include a lot of subjective sentences which express the specific assessment of the reviewer toward the product. In contrast, shill reviewers never used product. Therefore, it is expected that shill reviews contain more objective sentences that simply describe and repeat the product features. We hypothesize that:

H2: Shill reviews are less subjective than normal reviews.

Another measure to compare shill and normal reviews is the readability of the reviews. Readability is defined as the cognitive effort required for a person to understand and comprehend a piece of text (Zakaluk and Samuels 1988). Readability is usually measured by the length of the text, the complexity of the words and number of sentences. The length of the free-text comment part of the review is usually unlimited. However, normal reviewers can write without constraints about the product. Some normal reviews can contain just one sentence. In addition, normal reviewers may shorten their reviews to avoid repeating details in previous reviews. In contrast, shill reviewers are paid to write reviews feeling obliged to provide long and complicated reviews to make them appear real. We hypothesize that:

H3: Shill reviews are less readable than normal reviews

Research Methods

In this section, we discuss the NLP techniques and data collection process used to test the hypotheses. The three methods are product feature extraction, subjectivity analysis and readability analysis.

Product feature extraction and classification

The purpose of the product feature extraction method is to retrieve the product-related features mentioned in the reviews. With this method, we partly implement the approach proposed in (Hu and Liu 2004) by using a Part-of-Speech tagger to find noun and noun phrases possibly about product features. However, our approach tries to detect all product related features instead of just the features that the reviewer has opinion on. We used the OpenNLP[6]

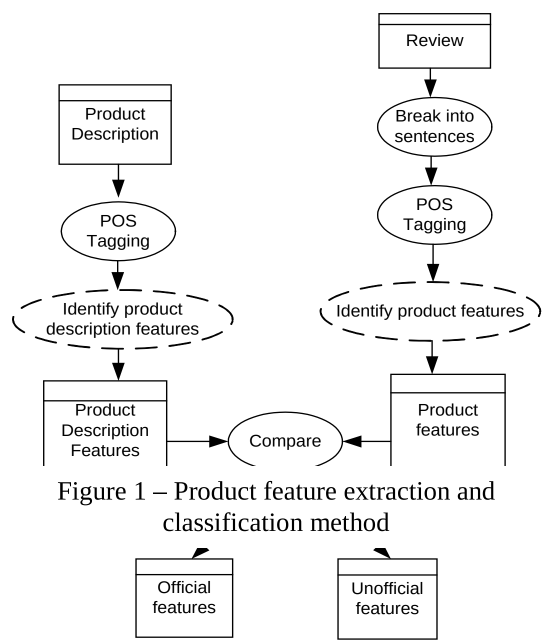


Figure 1 – Product feature extraction and classification method

[6] http://opennlp.apache.org/

toolkit to for the POS tagging job. Before the identification process, the product description and review were preprocessed using spellchecking, stemming and fuzzy matching. Figure 1 shows two separate product feature extraction tasks.

In the first task, product features from the product description were extracted. Then, similar product features were identified. For example, the words "sound" and "audio" were identified as similar features. This step is highly context-dependent so it was done manually.

In the second task, product features from reviews were extracted. After the reviews were parsed into sentences using the LingPipe[7] sentence model, the same process of POS tagging was used to retrieve noun and noun phrases. This step was performed manually to improve its accuracy and completeness.

In the final step, product features from reviews were classified. If a review product feature matched a product description feature, we classified it as an *official feature*. Otherwise, it was classified as an *unofficial feature*.

Subjectivity analysis

Subjectivity analysis determines if a sentence is subjective or objective. We applied the just the first step of sentiment analysis, subjectivity detection proposed in (Pang and Lee 2004). We used the subjectivity classifier called Dynamic Language Model with n-grams (n=8) from the LingPipe[8] toolkit. One of the most complicated problems of subjectivity analysis is the availability of labeled subjective and objective sentences used as training data. To obtain labeled training data, we used the solution in (Ghose and Ipeirotis 2004) to automatically obtain labeled sentences. Objective sentences were found in the product description page and subjective sentences were in the reviews. About 3800 objective sentences were retrieved from the product description page on Amazon.com. Although we had many more sentences from the reviews, we randomly selected about 3800 sentences from the reviews to build the subjective sentence dataset. 90 percent of each dataset were used in the training data set and the remaining 10 percent sentences were used in the testing data set. Table 1 shows the confusion matrix of the classifier.

		Response	
		Objective	Subjective
Reference	Objective	399	10
	Subjective	14	386

Table 1 - Subjectivity classifier confusion matrix

Readability analysis

We used two prominent readability measures in this study, the Coleman-Liau index and the SMOG index. The score range of both of these indexes (1 to 12) reflects the US grade level of education required to read a piece of text (O'Mahony and Smyth 2010; Korfiatis, GarcíA-Bariocanal et al. 2012).

Data collection

Our data collection process provided two sets of data: positive normal reviews and positive shill reviews. These reviews were submitted for the same product. We collected normal reviews on Amazon.com for a real MP3 player. To reduce the chance that the normal reviews were shill reviews, we only collected reviews posted by a reviewer with an Amazon.com verified purchase or disclosure of his real name with the review. A sample size of 93 positive normal reviews was obtained.

[7] http://alias-i.com/lingpipe/demos/tutorial/sentences/read-me.html
[8] http://alias-i.com/lingpipe/demos/tutorial/lm/read-me.html

The shill review data was collected via an experiment. The shill reviewers were undergraduate students who were asked to deliberately post positive shill reviews for the target product for a reward of course credit. No guidelines were provided to shill reviewers about the details of their reviews. We showed reviewers two product images and the product description of the product. To ensure that shill reviewers would not easily find other product information, we disguised the product. A sample size of 61 positive shill reviews was obtained.

Results

	Source	Mean	Std. Dev.	Std. Error	p-Value	Effect size
Coleman-Liau Index	Shill	8.824	1.591	0.204	0.000	0.630
	Normal	7.596	2.150	0.223		
SMOG Index	Shill	6.878	1.853	0.237	0.000	0.604
	Normal	5.553	2.393	0.248		
% Subjective sentence	Shill	0.680	0.259	0.033	0.000	-1.312
	Normal	0.934	0.135	0.014		
Official Feature Quantity per review	Shill	16.460	10.282	1.316	0.000	1.197
	Normal	6.470	6.775	0.706		
Unofficial Feature Quantity per review	Shill	3.200	2.719	0.348	0.446	0.126
	Normal	3.720	4.839	0.504		
% Sentence with an official feature	Shill	0.780	0.186	0.024	0.000	0.685
	Normal	0.611	0.279	0.029		
% Sentence with an unofficial feature	Shill	0.320	0.224	0.029	0.034	0.336
	Normal	0.408	0.285	0.030		

Table 2 – Analysis Results

Table 2 indicates that both readability measures showed statistically significant differences between shill and normal reviews. The effect size of the Coleman-Liau and SMOG was medium. This result provides evidence to support hypothesis H3. Shill reviews appear more difficult to read than normal reviews.

The percentage of subjective sentences in normal reviews was also statistically higher than that of a shill reviews and the effect size of the difference was large. This result strongly supports hypothesis H2. In most parts of the normal reviews, the reviewers expressed their personal opinions about the product. In contrast, shill reviewers used more of the review content to describe the features of the product instead of giving their opinions about it.

The results from Table 2 also show that shill reviewers concentrate their focus on the official features included in the product description page. The quantity of official features and the percentage of sentences containing official features of shill reviews were significantly larger than that of normal reviews. Therefore, hypotheses H1a and H1c are supported. While the quantity of unofficial feature per review of shill reviews was not significantly different from that of normal reviews, the percentage of sentences that contained unofficial features in of a normal review was significantly higher than that of a shill reviews. Thus, we conclude that hypothesis H1b is not supported and H1d is supported.

We strongly believe that informativeness about product features is an important variable to separate shill reviews from normal reviews. A review with details about many official features provides little information, especially when the reviewer is simply repeating the information in the product description page. This practice indicates that shill reviewers tried to convince consumers that they had extensive product knowledge. Most unofficial features were only mentioned in normal reviews. The unofficial features mentioned in shill reviews were general features that are not private information. Shill reviewers would not know most unofficial features without actually using the product.

Conclusion

It is difficult to distinguish a shill review from a normal review. To address this problem, this study explored linguistic differences between shill and normal reviews on informativeness, subjectivity and readability. We tried to separate shill and normal reviews using the fundamental factors, the knowledge about the product and the product usage experience. These factors were based on the assumption that shill reviewers never used the product. We strongly believe that these factors can increase the performance of the shill review filtering process. Our future work will be implementing automatic product feature detection techniques and using these variables to build a shill review detector model.

References

Ghose, A. and P. G. Ipeirotis (2004). Designing Ranking Systems for Consumer Reviews: The Impact of Review Subjectivity on Product Sales and Review Quality. Proceedings of the International Converence on Decision Support System.

Hu, M. and B. Liu (2004). Mining opinion features in customer reviews. AAAI'04 Proceedings of the 19th national conference on Artifical intelligence

Hu, N., I. Bose, et al. (2011). "Manipulation in digital word-of-mouth: A reality check for book reviews." Decision Support Systems **50**: 627-635.

Hu, N., L. Liu, et al. (2011). "Fraud detection in online consumer reviews." Decision Support Systems **50**: 614-626.

Jindal, N. and B. Liu (2007). "Review Spam Detection." WWW 2007 / Poster Paper.

Jindal, N. and B. Liu (2008). Opinion Spam and Analysis. WSDM'08. Palo Alto, California, USA.

Korfiatis, N., E. GarcíA-Bariocanal, et al. (2012). "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content." Electronic Commerce Research and Applications **11**(3): 205-217.

Lim, E.-P., V.-A. Nguyen, et al. (2010). Detecting product review spammers using rating behaviors. CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management

Liu, J., Y. Cao, et al. (2007). Low-Quality Product Review Detection in Opinion Summarization. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague.

Mayzlin, D. (2006). "Promotional Chat on the Internet." Marketing Science **25**(2): 155-163.

O'Mahony, M. P. and B. Smyth (2010). Using readability tests to predict helpful product reviews. RIAO 2010 the 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information. Paris, France.

Ott, M., et al. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011. Portland, Oregon.

Pang, B. and L. Lee (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts Proceedings of the ACL.

Wu, G., D. Greene, et al. (2010). Distortion as a Validation Criterion in the Identification of Suspicious Reviews. 1st Workshop on Social Media Analytics (SOMA '10). Washington, DC, USA.

Zakaluk, B. L. and S. J. Samuels (1988). Readability: its past, present, and future, International Reading Association.

Recommender Systems Position and Orientation Study in E-commerce Websites

Noura A. Alhakbani Abdulrahman A. Mirza
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
E-mail: nhakbani@ksu.edu.sa
E-mail: amirza@ksu.edu.sa

Abstract - The proliferation of e-commerce created the need for recommender systems which aid the customers in finding their way among the endless choices and helps reduce their search times. Recommender systems have been seen to affect the customer behavior. Most research has focused on the different proposed algorithms and the effectiveness of these algorithms. Very few research papers, however, have been found to study the effect of the position and layout of the recommended products on a customer's purchasing behavior. Existing research suggests usability design guidelines based on subjective findings such as questionnaires and surveys. This research aims to investigate the effect of different positions and/or layouts of recommender systems on the customers' behavior objectively through the use of an eye tracking system to examine visual attention of users during their interaction with recommender systems.
Keywords: E-commerce, HCI, Eye tracking, Recommender systems, Usability.

I. INTRODUCTION

E-commerce websites are increasingly offering more products online for consumers where they can easily get bewildered among the vast options. In brick-and-mortar stores, usually a sales person helps customers find their desired item(s) amongst the different products. In the virtual world, recommender algorithms have emerged trying to imitate this process.

Eye tracking systems help researchers determine where a user is looking while performing a task (Pernice & Nielsen, 2009). Eye tracking is becoming a more prominent tool for understanding user behaviors in many contexts ranging from studying web navigation techniques for dyslexic web users (Al-Wabil, et al., 2007), to studying a user behavior pattern towards a search result page (SERP) (Lorigo, et al., 2008). In recent years, a growing body of research has provided evidence of how eye tracking offers insights into which elements on the interface attract users' attention, hold their attention, and more importantly, what users fail to pay attention to or ignore.

Research has focused on the accuracy of the recommender systems' algorithms, suggesting that accuracy of algorithms leads to more user satisfaction. While accuracy is an important factor to user satisfaction, other factors such as presentation of recommended products and location within a webpage shouldn't be neglected due to their influence on user satisfaction and behavior.

Since very little research has focused on recommender system design and usability, and the research that studied the usability and design interface of recommender systems based their results and guidelines subjectively on user opinions (Ozok, et al., 2010), (Pu & Chen, 2010). So we got motivated to examine various recommender system designs with eye tracking to assess their effectiveness in persuasive computing contexts. As a result of this study we would like to be able to maximize the benefits from recommender systems. We plan to investigate how users/shoppers interact with the recommenders' lists by examining their visual attention (duration and spatial distribution on e-commerce interfaces) using an eye tracker. Of main interest in our study is the effect of the difference in layout and position of the recommended products. A comparative study will be conducted to find the differences in behavior towards different presentations of recommender systems.

II. RELATED WORK

Recommender systems have become an important research field. Park, et al. (2012) categorized 210 articles on recommender systems published from 2001 to 2010 according to their application field and data mining techniques. They showed the growing interest in recommender systems by the increasing number of research published in recent years.

Recommender systems for e-commerce sites take different forms and use different algorithms. Schafer, et al. (1999) created taxonomy of recommender systems and the main classifications were non-personalized recommendations which mostly depend on other customers' ratings; item to item correlation, which bases its recommendation on chosen items from the customer; and, people to people correlation, or collaborative filtering, which recommends products based on other people preferences with similar interests or taste. Some recommender systems take information from customers about their requirements and recommend products that most-closely fit their requests (Sarwar, et al., 2000). Some recommender systems will customize the online retail store webpage to customers' needs. Products that are believed to be of interest to a specific customer will be displayed. Recommender systems have been used successfully as a marketing tool. For example, Amazon has been using a hybrid of algorithms which uses three different approaches for its recommender systems: personalized recommendation, which recommends things based on the individual's past behavior; social recommendation, which recommends things based on the past behavior of similar users; and, item recommendation, where things are recommended based on the item itself, with the aim of getting customers to add more products into their shopping carts (MacManus, 2009).

While much research has studied recommender systems, little research has studied the usability and layout of recommender systems. Yet the style of presentation of recommended products affected user satisfaction. Most of the usability studies and design guidelines were based on subjective studies. Ozok, et al. (2010) recommended a set of 14 design guidelines for recommender systems interface and usability based on a subjective survey result. They found that shoppers preferred maximum of three recommendations on one page. Pu & Chen (2010) studied users' subjective evaluations of a recommender interface in terms of its interface label and layout adequacy and clarity. Chen & Pearl (2010) investigated the effect of different presentations of recommendation systems using a subjective user centric model. They found that quadrant layout attracted user attention most when compared between list and organization layouts.

III. METHODOLOGY

This study will examine the effectiveness of recommender systems in e-commerce websites using an eye tracker, complimented with retrospective think-aloud protocols, and interaction logs of key presses and selections on the sites' recommender systems. This triangulation approach offers us insights into what users see, do, and think, by capturing their cognitive thoughts during the website navigation. Participants will be observed by recording and analyzing their visual attention to the recommender systems while conducting unconstrained tasks. The purpose of the study was explained to the participants as assessing different interface designs of the different e-commerce sites, but without specifically telling them that we are interested in recommender elements of the interfaces.

A. ***Participants*** Twenty female participants were asked to browse 3 different e-commerce websites. Participants were recruited from the female campus of King Saud University. They were lecturers, teacher assistants, and students. Age of selected participants ranged between 20–30 years of age, with a mean age of around 23 years. All participants are bilingual in both Arabic and English. All participants had previous experience in online browsing and/or shopping, except for two that never had such experience. Almost three quarters of the participants had never visited any of the selected websites for this study.

B. ***Apparatus*** A Tobii x120 eye tracker was used to record eye movements of participants as they browsed different websites. The eye tracker has infrared corneal reflection to measure point of gaze. Data rate of 60 Hz was used. Tobii Studio version 3.0.2 was used for eye and mouse movement analysis. The screen resolution of the monitor was set to be 1011*835 pixels. Participants were seated approximately 65cm (i.e., 2’ 2”) away from the LCD monitor.

C. ***Stimuli*** Three e-commerce websites that sell apparel and shoes, among other things, were presented to the participants. The main language of the selected websites is English. All selected websites present pictures of the recommended products and a descriptive sentence. Table 1 presents the different placement of recommendations provided by websites used in the study. Most frequently used orientations found to be used on e-commerce websites are the ones exhibited by websites 1 and 2. These positions in particular were chosen to be tested because in English websites mostly recommended products appear to be on the right or bottom and rarely on the top or on the left side. The positioning of the recommender system utilized by Website 3 is less common, yet it is found to be used, and therefore was included in the study. The names of the actual websites in the study are being withheld because we feel that such information is irrelevant to the purpose of this research.

Table 1. Stimuli description of recommender systems used in the study

Website	Position	Orientation	Descriptive Sentence	Type of Recommended Product
Website 1	Right	Vertical	Customers also loved	Different
Website 2	Bottom	Horizontal	You may also like	Same
Website 3	Bottom	Vertical	People also viewed	Same

D. ***Task Scenarios*** The participants were asked to choose a pair of shoes that they liked from each website. The objective from this unconstrained browsing task was to control the external validity of the experiment by simulating a real experience of interacting with recommender systems, and to avoid having the demand characteristics exhibited by participants influencing the data when they have an understanding of what is expected of them in the session.

E. ***Procedure*** Consents were obtained prior to enrolling participants in the study. Sessions started with a calibration procedure with the eye tracker to develop a 2D model of the viewer's eyes. Following that, the stimuli were presented in random and the participants were asked to choose one preferred pair of shoes from each website.

IV. DATA ANALYSIS

A. Areas of Interest (AOI)

Three areas of interest (AOI) were selected; one on each website that surrounded the recommended products' area as shown in Figure1.

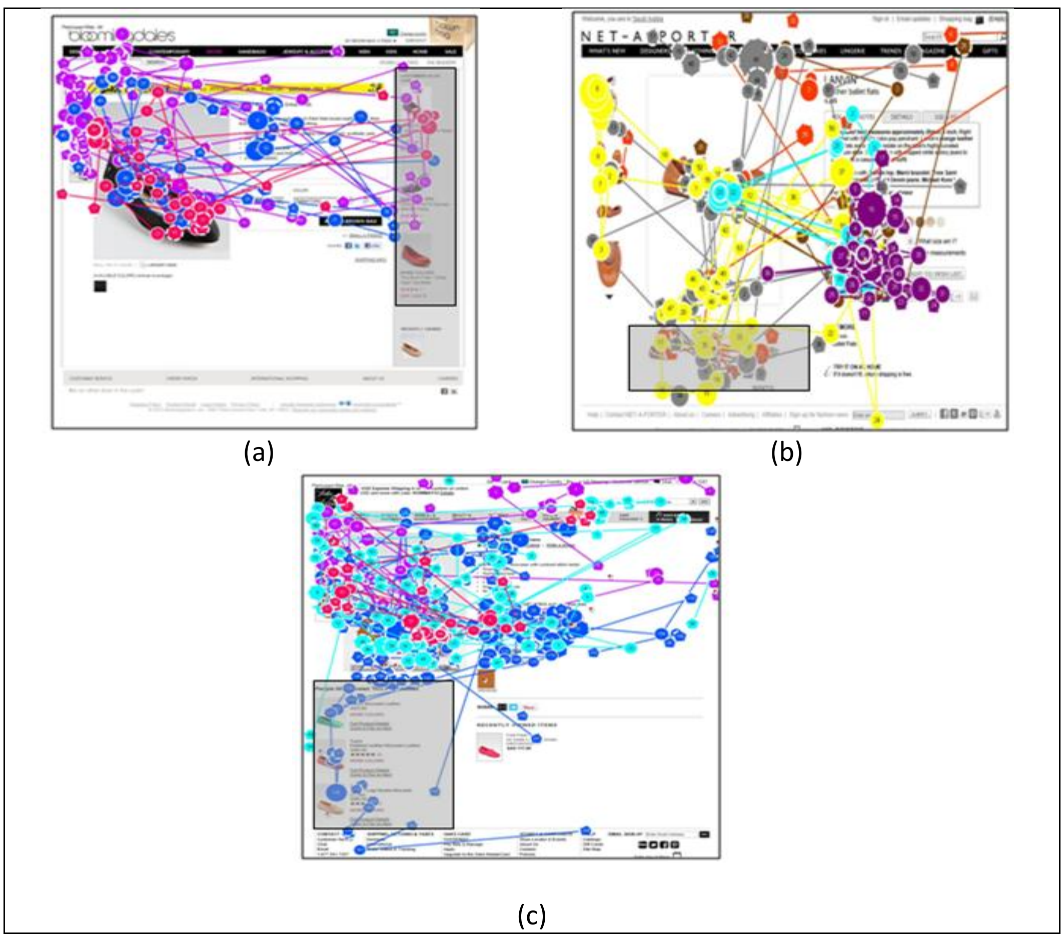


Figure 1: Areas of Interest (AOI)
(a) Website 1 (b) Website 2 (c) Website 3

Among all participants, we found that 26.31% have looked at the AOI or recommended products on website 1 and website 2, while only 10.52% of them looked at the AOI on website 3.

Measures of fixation count and fixation duration: the duration of each individual fixation within an AOI as shown in Table 2 and Table 3 respectively. On average, participants exhibited more fixations and higher mean fixation durations when using website 2. Mean fixation duration is an indication of the intensity of attention exhibited by the participant while shopping.

Table 2 Fixation Count

Fixation Count	**Count**	**Mean**	**Sum**
Website 1	5	6	30
Website 2	5	13.4	67
Website 3	2	15.5	31

Table 3 Fixation Duration

Fixation Duration	**Count**	**Mean**	**Sum**
Website 1	27	0.1933	5.22
Website 2	67	0.2217	14.86
Website 3	31	0.1370	4.25

B. Retrospective Think Aloud Protocol

After the participants finished their task from the selected websites they were asked about their preferred website design; 73.68% of the participants preferred the website 2 design, while 21.05% preferred the design of website 3. Only 5.26% of the participants preferred the design of website 1. They were then asked about their attention to the recommended products; 63% of them mentioned that when browsing at home or in a real life scenario where the environment is more relaxed; they'll pay more attention to recommended products than in an experiment and lab environment. This explains the reason behind the low numbers of participants that looked at the AOI.

V. CONCLUSION

A key goal of this study was to test whether location and orientation of recommended products affected attention paid to them by shoppers. All results were analyzed to find if differences in orientation and positions of the recommender system within the website layout affected the user decision and resulted in more attention to recommended products. In our study we found that most attention was paid to website 2 where the recommended products where displayed below the products and oriented horizontally. On the other hand the least attention was paid to recommended products displayed below the selected product but oriented vertically. And that might be due to the scrolling needed to view all recommended products. The findings of this paper would help e-commerce website designers to select the most viewed or prominent location and orientation to display the recommended product.

VI. ACKNOWLEDGMENT

We would like to thank the participants for their time and effort, and Ms. Abeer Alnafjan for her assistance with the experiment. And we would like to extend our sincere gratitude and appreciation to Dr.Areej Alwabil for her guidance and support throughout the study.

REFERENCES

Aciar, S., Zhang, D., Simoff, S. & Debenham, J., 2006. *Recommender System based on Consumer Products Reviews.* s.l., IEEE.

Al-Wabil, A., Zaphiris, P. & Wilson, S., 2007. *Web Navigation for Individuals with Dyslexia: An Exploratory Study.* Beijing, China, LNCS 4554, p. 593–602.

Chen, L. & Pearl, P., 2010. Eye Tracking Study of User Behaviour in Recommender Interfaces. *In: P.De Bra, A,Kobsa, and D.Chin (Eds.):UMAP 2010, LNCS 6075,* pp. 375--380.

Granka, L., Joachaims, T. & Gay, G., 2004. Eye Tracking Analysis of User Behavior in WWW Search. *SIGIR '04, ACM,* pp. 25-29.

Hu, R. & Pu, P., 2011. Enhancing Recommendation Diversity with Organization Interfaces. *IUI 2011*, pp. 13-16.

Lorigo, L. et al., 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology,* May, 59(7), pp. 1041-1052.

MacManus, R., 2009. *A Guide to Recommender Systems.* [Online]
Available at: http://www.readwriteweb.com/archives/recommender_systems.php
[Accessed 4 April 2012].

Ozok, A. A., Fan, Q. & Norcio, A. F., 2010. Design Guidelines for Effective Recommender System Interfaces Based on a Usability Criteria Conceptual Model: Results from a College Student Population. *Behaviour & Information Technology,* 29(1), pp. 57-83.

Park, D. H., Kim, H. K., Choi, I. Y. & Kim, J. K., 2012. A literature review and classification of recommender systems research. *Expert Systems with Applications,* 39(11), p. 10059–10072.

Pernice, K. & Nielsen, J., 2009. *Eyetracking Methodology How to Conductt and Evaluate Usability Studies Using Eyetracking,* Fremont, California: Nielsen Norman Group.

Pu, P. & Chen, L., 2010. *A User-Centric Evaluation Framework of Recommender Systems.* Barcelona, Spain, Sep 30, 2010, CEUR-WS.org, pp. 14-21.

Pu, P., Zhou, M. & Sylvain, C., 2009. Critiquing Recommenders for Public Taste Products. *RecSys'09, ACM,* p. 23–25.

Sarwar, B., Karypis, G., Konstan, J. & Riedl, J., 2000. *Analysis of Recommendation Algorithms for E-Commerce.* Minneapolis, Minnesota, ACM, pp. 158-167.

Schafer, J. B., Konstan, J. & Riedl, J., 1999. *Recommender Systems in E-Commerce.* New York, ACM, pp. 1588-166.

Reputation in an Open Source Software Community

Yuanfeng Cai
Iowa State University
Ames, Iowa, 50010
yfcai@iastate.edu

Dan Zhu
Iowa State University
Ames, Iowa, 50010
dzhu@iastate.edu

Abstract

Developer's reputation in the OSS community is defined as the positive evaluations from his peers. While a large body of studies focuses on studying the importance of developers' reputation in their participation motivations, there is still a lack of understanding for two issues. First, which factors can lead to a high developer's reputation? Second, how the reputation level in a project development group impacts its project performance? In this study, we develop a theoretical model and conduct an empirical analysis in a large online open source community. The results show that a developer's reputation level is based on 1) his previous work quality, 2) his contribution percentage in previous projects, 3) the time spent on previous projects, and 4) the number of previously participated projects. In addition, we find that the group with an overall higher level of reputation would achieve a better performance, while the individual reputation level deviation within the group would impair its performance. The implications of our findings and the future research direction are then discussed.

1. Introduction

The success of open source software (OSS) development highly relies on self-motivated behaviors among software developers, and the field of OSS could not have survived without their voluntary contributions (Roberts et al. 2006). Scholars have long been interested in the developer's motivation in the OSS community (Brabham 2008; Hu et al. 2012; Subramanyam and Xia 2008). They have found intrinsic and extrinsic motivations including personal attributes, attitudes, behavioral patterns, and job-related factors (Krishnamurthy, 2006). The extrinsic motivations are related to direct compensation while the intrinsic motivations are related to psychological factors (Hars and Qu 2002; Lerner and Tirole 2002). Among the widely recognized motivations, a key one is a developer's desire for community reputation (Lakhani and Wolf 2005; Markus et al. 2000). *Reputation* is defined as a distribution of opinions, estimation or evaluation about a person or other entity, such as group, organization, or activity, in the group (Bromley 2001). Adapted from this definition, we represent the developer's received evaluation generated from peers in the community as his/her *community reputation*. The higher value of the evaluation indicates that the developer has a higher reputation. In the OSS community, the desire to build and maintain reputation motivates new developers to join OSS community and complete tasks effectively (Sharma et al. 2002).

Due to the importance of community reputation to motivation, two issues are necessary to be explored. The first is the antecedents that influence a developer's community reputation level. In this study, we define antecedent as the factor that facilitates obtaining a higher reputation for developers. The second issue relates to the outcome of the community reputation. In this study, we suggest the outcome of the community reputation as its influence on project performance. Since the projects in the OSS community are conducted by a group instead of an individual, it should be the reputation level of the group rather than that of an individual affecting the project performance. In this study, we model the *group reputation* as the average of all its members' individual reputations as previous studies conducted under the Web context (Mui 2002; Josang et.al 2007). Though previous studies have found that the failures of OSS projects are mostly due to the lack of sustained voluntary developers' contribution (Crowston, et al. 2003; Krishnamurthy 2002), whether enough motivation of developers indicates a good project performance is still unknown. Understanding the influence of group reputation is important, since it relates to the quality of the software produced in the OSS community. Therefore, in this study, we raise two research questions: 1) What are antecedents that help improve the developer's reputation in the OSS community? 2) How does group reputation in OSS community affect their project performance?

A growing body of research has explored the reputation in the OSS community. But they mostly focus on the role of reputation as a motivation for joining the community. Although a few studies have explored the first issue, their findings are not applicable for the fresh developers who have already participated in projects but have not received any evaluations from peers. In addition, though several studies devoted to the software quality issue in the OSS community, to the best of our knowledge, there is no systemic analysis on the role of group reputation on the project performance.

2. Theoretical Background and Research hypotheses

2.1 Antecedents

The theoretical foundation for understanding the antecedents for community reputation is adapted from Homans's (1950) studies of small groups. Homans has suggested that there are three elements in a social system, *activities interactions and sentiments*. Tasks that people participate and accomplished are defined as activities; communications between members who participate in the tasks are defined as interactions; attitudes among those participated members are defined as sentiment (Homans 1950). The evaluation, which is the base for an individual's social reputation, is the released sentiment. Homans (1950) has also postulated in principle the impact of activities and interactions on social reputation: the closer the person's activities conform to the norm--the goal that a group wishes to reach, the higher his/her reputation tend to be; the wider the individual interact, the higher the individual's social reputation. Therefore, the feature of activities and interactions could be viewed as antecedents for the individual's social reputation.

We build our framework based on Homans' work. In the OSS context, the developers' activities are mostly coding bugs. The norm of every group in OSS project development may be distinct in content, yet most groups have an essentially same requirement that the project should be developed effectively and efficiently. Thus, a degree of a developer's activity comes to norm could be measured by his/her coding quality. Specifically, we suggest that two developer's attributes could be used to measure his/her coding quality: work quality and contribution percentage. Second, in the OSS context, while the main task of the developer is project development, the interaction is the behavior between developers in performing the project development. Although developers seldom have face to face interaction in the OSS community, they have rich opportunities for mutual participation in the same project. Hence, participation experience, which indicates the interaction experience, is suggested to be another potential factor. Likely, we suggest that other two developer's attributes could be used to measure his/her participation experience: time spent and project experience. We will next discuss the relationship between each factor and individual reputation respectively.

2.1.1 Coding Quality

Coding quality reflects the credibility of the developer's previous work in this OSS community. In the OSS community, it may be hard to review the codes of each developer so that there may not be a direct measurement of the developer's coding work. Despite that we could still measure the developer's coding quality from his/her previous project quality. If most of the previous projects that the developer has developed are high quality, other community members may have relative high confidence in the developer's coding capability. However, besides the previous work quality, the extent of a developer's involvement in the group is also a crucial indicator of his/her coding capability. As in the OSS project development group, each developer works voluntarily without any workload requirement. A developer may luckily join in several teams which have successful outcomes, while he/she does not have too much contribution. The developer's contribution could be measured by his/her workload percentage of the whole project. Hence, a high quality of previous work combined with a large contribution percentage may indicate a high level of coding quality for developers.

In the OSS community, developers seldom have face-to-face communication between each other. Individuals have a limited acquaintance about their peers (Daniel et al. 2009). The previous coding quality is the most convenient source that developers could learn about their peers. For the two developers who have worked on the same project, it is possible for one developer to judge another's capabilities. But for the developer without any prior collaboration experience, the previous coding quality can be a reliable basis. Though the evaluator may not know what the specific norm of the evaluatee's group is, an excellent record implies a less possibility of norm violation. In addition, the voluntary developers in OSS community are highly likely to have a strong interest in coding. A developer with a higher coding capability is more possible to be venerated. Thus, a better coding quality may indicate a larger possibility to attain a higher reputation. Sharma et al. (2002) suggests that reputation is developed through consistently quality contributions. The better quality of the developer's previous coding, the higher positive evaluation he/she may receive from others. Thereby, we propose the following hypothesis:

H1a: Work quality is positively related to developer's reputation.

H1b: Contribution percentage is positively related to developer's reputation.

2.1.2 Participation Experience

Participation experience here intends to measure the previous interaction experience of a developer. Homans (1950) defined interaction as long and frequent communication opportunity. Project collaboration provides communication opportunities between developers. The communication duration could be measured by a developer's time spent in coding. The more time spent in coding may indicate that developers may have known about each other for a longer duration. In addition, the communication frequency could be measured by the number of project participation. We define project experience as the number of a developer's participated projects. If a developer has participated in many projects, he/she will have a large number of collaborators as well as a high amount of frequent collaborators.

Both time spent and project experience are suggested to be the influential factors for the individual community reputation from two aspects. First, spending more time on the project can facilitate knowledge sharing with other group members so that developers can benefit them personally (Fleming and Waguespack 2007; Rafaeli and Ariel 2008). As mentioned in Section developers may ask help or they can be asked for help from other members. The more time a developer spent seeking from other peers' contributed ideas, the more knowledge he/she learnt from peers; the more time a developer spent replying to others' questions, the more information he/she provided for others (Faraj et al. 2011). Thus, time spent has a positive impact on knowledge sharing. An effective knowledge sharing process requires a smooth interaction, and further positively contributes to reputation attainment. Second, a rich participation experience provides the developer opportunities to learn from other developers. Monge et al (1985) have stated that undergoing various environments can provide multiple types of relevant knowledge to developers. By participating in projects, the developer could access to the peers' skill sets so that his/her knowledge is expanded (Kuk 2006). A developer with strong project participation experience is more sensitive to the features necessary to the software (Daniel et al. 2009). Hence, he/she is more likely to interact with other developers widely and effectively, either by contributing his/her knowledge in their own project, or by solving others' problems. Hereby, the project experience positively contributes to reputation attainment. Similarly, we propose the following hypothesis:

H2a: Time spent is positively related to developer's reputation.

H2b: Project experience is positively related to a developer's reputation.

2.2 Outcome

Regarding the outcome of reputation, we observe a relationship between the reputation of a group and its project performance. We do not explore the outcome of individual community reputation since our concern is the project performance and the project quality is always conducted by a group. We study two characteristics of the reputation for a group: group reputation and reputation variance. Group reputation is defined as the mean value of all its members' individual reputations (Mui 2002; Josang et.al 2007). And reputation variance refers to the variance of the individual reputation level in the same group.

2.2.1 Group Reputation

A relatively high level of group reputation is expected to positively contribute to its project performance. While a reputation is not totally corresponding to the developer's technical capability, a higher reputation at least suggests that their developers' capability is reliable. Furthermore, a significant barrier for the group collaboration in OSS community is that group members usually are geographically dispersed. Without a face-to-face communication, they are highly likely to misunderstand each other's purpose as reading others' codes is not an easy task for developers. In addition, insufficient long-term participant has already led a high percentage of OSS projects to fail (Colazo and Fang 2009). A high level reputation drives the developers' collectivistic spirit. It encourages them to persist in working and spend effort to solve problems (Kane et al. 2009). Thus, a relatively high level of reputation in a group should positively contribute to the project performance. Hereby we the following hypothesis:

H3: Group reputation has a positive impact on its project performance.

2.2.2 Reputation Variance

The variance of the reputation level in the same group is suggested to affect the project performance. When a group has a high deviation of individual reputation, it will have a mixture of developers of all reputation levels. Two possible cases are existed. First, the developer's lower reputation is due to lack interaction. The newcomers mostly have lower reputation. Since the experienced developers are mostly highly respected and motivated, the newcomers may be stressed out. Their suggested are less possible to be adopted and their enthusiasm may be attacked. Faraj et al. (2011) have suggested that team members with less enthusiasm have less discourse power and their arguments are less possibility to be taken into consideration. Thus, not all the developers could display their capability so that the overall quality may be affected. Second, the developers' lower reputation is due to coding quality. In this case, the developers with a lower reputation still contribute their knowledge to the OSS project. But when deciding the group norm, the developers with a higher reputation may accommodate themselves to the degree that the developers with a lower reputation feel acceptable (Phillips et al. 2009). Hence, the overall group performance will be impaired as well. Based on the above discussion, we could bring up the hypothesis:

H4: Reputation variance has a negative impact on its project performance.

3 Data

Out data-set was collected from a large online OSS community – Ohloh. This data-set provides OSS project participation information as well as the evaluation information among

Ohloh community members. Each Ohloh member can send a "Kudo" link to other members, which contains a simple attitude. A rating score will be generated based on the "Kudo" links. The rating is an integer from 1 to 10. "Kudo" rating represents the reputation of other members in the OSS community. A larger value of rating indicates a higher level of reputation. Ohloh also contains the rating for each project. The rating is a floating point value from 1.0 to 5.0. It is provided by the users of this project and is the average value of all user ratings. 1.0 is the worst possible rating and 5.0 is the highest possible one. In our study, the information for a total of 2159 developers was retrieved.

The individual community reputation is measured by the "Kudo" rating, which represents the reputation level of each developer. All community members in the Ohloh community could send an evaluation for another developer, regardless of prior coloration experience. The project performance is measured by the project rating. The value of this rating is an average rating value from the users of this project. Work quality is measured by the average project rating of all his/her participated project. We measure the developer's contribution as the total number of commits made by this developer divided by the total number of commits made by all contributing developers on a project. The developer's contribution percentage is the mean value of his/her contribution in every participated project. Similarly, we measure the time spent on each project as the total number of calendar months in which this contributor made at least one commit, while the time spent for a developer is the mean value of time in every participated project. Project experience is measured by the number of projects that this developer has participated. The group reputation is the mean value of individual reputation in the same group while the reputation variance is the variance of the individual reputation in that group.

4. Results and Discussion

Using stepwise multivariate regression model of analysis we arrived at a set of results, which we can use to interpret the relationship between dependent variables of coordination and independent variables. The inter-correlation between variables has also been examined. The data sizes of the first and second models are 2159 and 196. The regression models are shown below:

F1: reputation=a+b1 work quality +b2 * contribution percentage+ b3*time spent+ b4* project experience*

F2: project rating=c+ d1 group reputation+d2* reputation variance*

Since the group size is one criterion we used when selecting projects, we control the group size when analyzing the group performance. More developers may bring more resource to the project so that it is expected to have a positive impact on group performance (Butler 2001). Table 1 and Table 2 show the results for each model. In the first model, all four suggested independent variables are statistically significant. Thus, all hypotheses are supported. In the second model, both group reputation and reputation variance are statistically significant. The positive coefficient indicates that group reputation positively influence the project performance, while the negative coefficient indicates that reputation variance negatively impacts the project performance. Thus, both H3 and H4 are supported. We should notice that the value or R^2 is not high, which is around 16%. And both group reputation and reputation variance are not statistically significant when the alpha level is below 0.05. As discussed in the literature section, we know that there are a lot of known predictors for the project performance besides the control variables included in the current study. The group reputation may neither be the only determinant of project performance, nor the most important one. Thus, we do not expect a high explanatory power for this variable. There are other important factors which we do not explore in the current study but are planned to conduct in the future. Since few of the previous studies consider reputation as an explanatory variable for the project performance, we want to detect whether group reputation indeed affects project performance.

Table 1 Result for the First Model (N=2159, '***'p<0.001, '**' p<0.05, '*'p<0.1)	
Coefficient	**Estimated**
Work Quality	0.663***
Contribution Percentage	1.638***
Time Spent	0.0173***
Project Experience	0.105***
Multiple R-squared:	0.284
Adjusted R-squared:	0.283
F-statistic:	213.6***

Table 2 Result for the Second Model (N=196, '***'p<0.001, '**' p<0.05, '*'p<0.1)	
Coefficient	**Estimated**
Group Reputation	0.101*
Reputation Variance	-0.021*
Group Size	0.000
Multiple R-squared:	0.164
Adjusted R-squared:	0.151
F-statistic:	12.55***

In this study, we have explored the antecedents for a developer's community reputation and outcome of the group reputation in the OSS community. We contribute to the literature in OSS for discerning the concept of reputation. Though previous studies have shown the importance of developer's reputation, few of them discuss their antecedents and its influence on the project quality. In addition, the relationship between the group level reputation and the project performance has practical implications for virtual groups.

We end this discussion by noting some limitations and additional directions for future research. Several project related factors such as its market structure and complexity are not available in Ohloh data-set and thus not examined in our study. Thus an important extension of this paper would be studying the effects of all major project factors and other potential determinants using empirical data from other major OSS communities such as Sourceforge.net. To summarize, our future work mainly consists of three research directions, including 1) investigating reputation's influence in other OSS communities to examine other possible determinants, and 2) developing computational mechanisms or collaborative information systems to support reputation management in OSS communities based on our empirical findings.

References

Butler, B. S. 2001. "Membership Size, Communication Activity, Sustainability: A Resource-Based Model of Online Social Structure"s, Inform. Systems Res. (12:4), pp, 346-362.

Bromley D.B . 2001. "Relationships between personal and corporate reputation,"*European Journal of Marketing* (35: 3), pp.316 – 334

Brabham, D.C. 2008. "Moving the crowd at iStockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application,"*First Monday*(13: 6)

Colazo, J., and Fang, Y. 2009."The impact of license choice on open source software developmentactivities,"*Journal of the American Society for Information Science and Technology*

Crowston, K., Annabi, H., and Howison, J. 2003."Defining open source software project success," in *Proceedings of the 24th International Conference on Information Systems*, Seattle, WA

Daniel, S. L., Agarwal, R., and Stewart, K.J. "The Diverse Effects of Diversity in Global, Distributed Collectives: A Study of User Participation in Open Source Projects," *Information Systems Research*(Conditionally Accepted)

Faraj,Jarvenpaa, and Majchrzak. 2011."Knowledge Collaboration in Online Communities,"*Organization Science* (22: 5), pp. 1224–1239.

Fleming, L., and Waguespack. D. M. 2007. "Brokerage, boundary spanning, and leadership in open innovation communities," *Organ. Sci.*(18: 2),pp. 165–180.

Hars, A. and Qu, S. 2002. "Working for free - Motivations for participating in open-source projects," *International Journal of Electronic Commerce* (6), pp. 25-39.

Homans, G. C. 1950. *The Human Group*, New Yourk: Harcourt, Brace.

Hu. D, Zhao, J. L, and Chen, J. 2012. "Reputation Management in an Open Source Developer Social Network: An Empirical Study on Determinants of Positive Evaluations,"Decision Support Systems

Josang, A., Ismail, R., and Boyd, C. 2007. "A survey of rust and reputation systems for online service provision," *Decision Support Sysems* (43), pp.618-644.

Kane, G. C., Majchrzak, A., Johnson, J. and Chen, G. 2009. "A longitudinal model of perspective making and perspective taking within fluid online collectives,"in *Proc. Internat. Conf. Inform. Systems*, Phoenix, AZ, pp. 10.

Krishnamurthy, S. 2006. "On the intrinsic and extrinsic motivation of free/libre/open source (FLOSS) developers," *Knowledge,Technology,and Policy* (18), pp.17-39.

Lakhani, K.R., and Wolf, B. 2005. "Why hackers do what they do: Understanding motivation and effort in free/open source software projects," in *Perspectives on Free and Open Source Software*, J. Feller, B. Fitzgerals, S.A. Hissam, and K.R. Lakhani (eds.), Boston: MIT Press, pp. 3–23.

Kuk, G. 2006. "Strategic interaction and knowledge sharing in KDE developer mailing list,"*Management Sci*. (52: 7), pp. 1031-1042.

Monge, P., Rothman, L., Eisenberg, E., Miller, K., Kirste, K. 1985. "The dynamics of organizational proximity," *Management Sci.*(31: 9), pp. 1129-1141.

Moody, G. 2001.*Rebel Code: Linux and the Open Source Revolution*, Cambridge, MA: Perseus Press.

Mui, L. 2002. "Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks,"*Ph.D. thesis*, Massachusetts Institute of Technology (December).

Phillips, K. W., Rothbard, N. P., and Dumas, T. L. 2009."To disclose or not to disclose? Status distance andself disclosure in diverse environments," *Acad. Management Rev.*(34: 4), pp. 710-732.

Sharma, S., Sugumaran, V., and Rajagopalan, B. 2002."A Framework for Creating Hybrid-OSS Communities,"*Information Systems Journal*(12: 1), pp. 7-26.

Rafaeli, S., Y. Ariel. (2008). Online motivational factors: Incentives for participation and contribution in Wikipedia. A. Barak, ed. Psychological Aspects of Cyberspace: Theory, Research, Applications. Cambridge University Press, Cambridge, UK, 243–267.

Roberts, J., Hann, I.-H., and Slaughter, S (2006) "Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects", Management Science, 52(7), pp. 984-999.

Subramanyam, R. and Xia, M. 2008."Free/Libre Open Source Software development in developing and developed countries: A conceptual framework with an exploratory study,"*Decision Support Systems* (46: 1), pp. 173-186.

A Case for a Workflow Driven Workflow Execution Engine

Shubhangi Sharma, Kamalakar Karlapalem
International Institute of Information Technology,
Hyderabad, India
shubangi@students.iiit.ac.in; kamal@iiit.ac.in

P. Radha Krishna
Infosys Labs, Infosys Limited,
Hyderabad, India
radhakrishna_p@infosys.com

Abstract. Workflow Management Systems (WfMSs) execute workflows according to a fixed set of rules coded in the workflow engine. *Once deployed, the WfMS's execution engine cannot be changed easily to support new execution requirements.* Our approach is based on – *a workflow executing workflows* – so as to allow multiple control flows to be specified and supported for workflow execution. An *Execution Workflow* (EW) is a workflow modeling the procedure (execution process) to execute user workflow instances. The EW can be modified or multiple EWs can be specified to cater to different user workflow execution semantics. We present a *Task Dispatcher* which schedules and assigns the workflow and EW tasks to agents/applications for execution and returns events to the EW for executing workflow instances. We present the system built based on EW with experimental results.

1. Introduction

Workflows are specified, driven and executed by Workflow Management Systems (WfMSs). Typically a WfMS [4] consists of (i) *a specification unit*, (ii) *a workflow engine*, and (iii) *a worklist*. Current WfMSs (ex.[2],[5]) follow a fixed execution control flow or a fixed execution procedure (with multiple control flows) to execute workflow instances. While workflow definitions can be modified, the procedure to execute workflows is hard-coded as a workflow engine, and cannot be modified or enhanced easily. The challenge is *whether we can dynamically change the way a WfMS engine functions without re-coding the engine code.* Our solution addresses this challenge, and allows execution procedure to be modified, according to business and technological needs. In this paper, we present a *WfMS that supports workflows executing workflows.* We specified and implemented a WfMS engine procedure as *Execution Workflow (EW).* To support execution of the workflows, a Task Dispatcher is introduced which allocates the task instances to humans/applications and orchestrates the workflow execution.

2 Execution through EW based System

Many times it is required to change the way a workflow is executed. Consider workflow execution procedure, such as: (i) simple workflows that do not require resources, (ii) requiring resources, and (iii) further needing exception handling.

(i) *Get a task - execute - continue till no more tasks*
(ii) *Get a task - check resources - execute - continue till no more tasks*
(iii) *Get a task - check resources - execute - call exception manager if required – handle exceptions - continue till no more tasks*

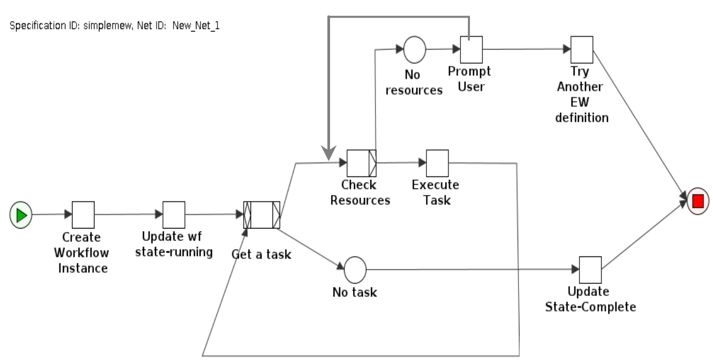


Fig. 1. A basic Execution Workflow

Thus, there may be a need to shuffle or add new execution steps, depending on how we want the workflow execution engine to execute workflows. Our system consists of two modules: the *Execution Workflow* and the *Task Dispatcher*.

2.1 Execution Workflow (EW)

In our framework, the procedure for running user workflows is expressed in the form of a workflow, called EW. For the purpose of uniformity, YAWL [4] is used to specify all user workflows and EW definitions in this work. Fig. 1 shows an example specification of EW. Workflow specifications consist of (atomic) tasks associated with some application or authorized human(s) for execution. Similar is the case with an EW. Separate routines (for every task) have been coded to execute each of the EW tasks. During specification, EW tasks need to be associated with either these routines or any other external applications that can perform the desired task. These set of EW enabling routines are collectively given the name **Execution Controller** of the system. EW undertakes various stages of workflow execution as depicted in Fig. 2. EW responsibilities include workflow instance creation, workflow instance state maintenance and completion of execution of workflow instances.

Table 1. Difference between a user workflow & EW

Workflow Type	Purpose of Workflow	What is Specified in Workflow
EW	Execute user workflows just like workflow execution engine	Engine procedure as workflow
User Workflow	Specify a user business process to automate it	User business process as workflow

Note that, EW in the present work can be viewed as a meta workflow. However, EW is not exactly a control workflow like a meta workflow in [1]. *The meta workflow in [1] has tasks such as Start, Wait, Suspend, Resume and Stop. These are the tasks that control other workflows. But in our execution workflow, the workflow execution procedure is specified as a workflow (see Table 1).* The principle of a workflow (i.e., modelling the process to be executed) remains same in case of EW. A user (process) workflow upon execution completes a user business process, while an EW completes a workflow engine process. Though, the purpose of tasks of an EW and a user workflow is different, their specification syntax remains the same. The specification requirements of an EW task are exactly similar to that of a user workflow. Thus, we assert that EW does not need a different specification language. Available workflow specification languages (ex., YAWL) can be used to define an EW.

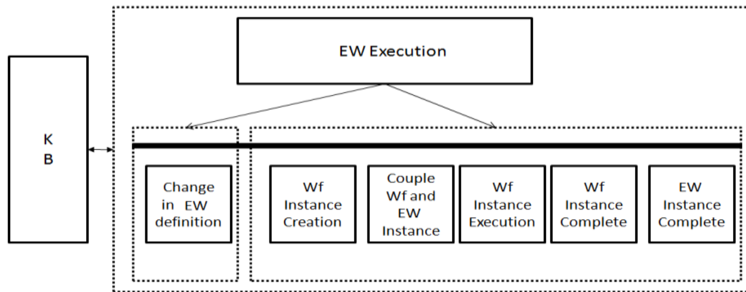


Fig. 2. EW Driven Execution of Workflows

Another difference between EW and a user workflow is that, EW is specified by a Workflow System Designer or Workflow System Manager, rather than a Workflow Application Designer. The EW execution is transparent to end users executing user workflows. The workflow system manager has the ability to modify and to update execution procedure by editing EW (when necessary), rather than making changes in the engine code. As and when a need comes to change the control flow of execution (such as, to involve new technologies or business execution needs), the workflow system manager shuffles, edits, adds or removes the tasks in EW definition, and a new EW definition is arrived at. The **Knowledge Base** (KB) in Fig. 2 stores various definitions of EW as and when they evolve over time. KB is a database which stores tuples of the form: *<EWSpecificationID, EWXMLfile, TimeOfDefinition, ParentEWID, ReasonForUpdate>.* The KB stores (as EWXML file) the new specification ID and ID of the EW it was modified from, along with other details as suggested by the field names in the tuple.

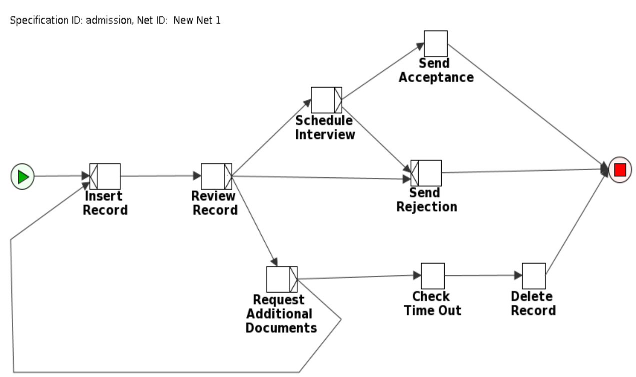


Fig. 3. Workflow for PG Admission in an Institution

2.2 Task Dispatcher (TD)

To facilitate execution by the EWs, a Task Dispatcher (TD) is deployed. From a pool of tasks ready to be executed, TD selects a task and assigns it to appropriate agents, applications or humans for execution. Both user workflows and EWs are made up of tasks that are atomic. Based on the post events of the task's execution, the course of the workflow or the EW is decided [3]. Thus, at any time there will be a pool of ready to be executed tasks, belonging to various active user workflow and EW instances. The TD selects tasks from this pool, giving priority to a user task, unless an EW task has to be executed to progress some long waiting user workflow instance. The TD then sends request to appropriate applications or humans to execute that task and monitors the events raised during the execution. Finally, it returns the post-events to the respective EW instances which had sent those tasks to the Task Dispatcher for execution. The post events involve the events happening as a result of execution of task, along with the data exchange related to the task. The TD is responsible for the data flow between applications and EW while a workflow instance executes. It receives data and events from the application/person executing a task, and returns the data to appropriate EW associated with it in order to decide the control flow of execution.

Table 2. Trace of tasks of a EW instance executing a User Workflow Instance

No.	EW Task	User Workflow Instance	User Task	Tasks in Dispatcher	Post-Event Data
1	Create Wf instance	Nil	Nil	Create Workflow Instance	Success; Instance id=4
2	Update State as Running	4	Nil	Update State as Running	Success;
3	Get a Task	4	Nil	Get a Task	Success; User Task = 'Insert Record'
4	Check Resources	4	Insert Record	Check Resources for Insert Record	Success;
5	Execute Task	4	Insert Record	Insert Record	Success; Admission Application Record No. 10
6	Get a Task	4	Nil	Get a Task	Success; Task = 'Review Record'
7	Check Resources	4	Review Record	Check Resources for Review a record	Success;

2.3 The Coordination and Procedural Flow in EW with an Example

We illustrate how the system components coordinate to build a complete workflow execution engine. The Post Graduate admission process of our institute (see Fig. 3) is taken as an example to demonstrate the execution of a user workflow instance through EW.

Step 1: Choosing an EW specification. A *daemon thread* instance that constantly runs in the system is waiting for user requests to create workflow instances. When such a request is received, the daemon thread creates an EW instance to execute the user workflow instance execution.

Initially, the daemon chooses the default EW definition from the Knowledge Base and creates an instance of it. The user instance and EW instance are coupled together (*one to one mapping*) until the execution of the user instance is complete (See Fig. 4).

The trace of execution for the admission workflow instance (Fig. 3) by the EW (Fig. 1) instance is shown in Table 2. For the first EW task, the user workflow and EW instances have not been

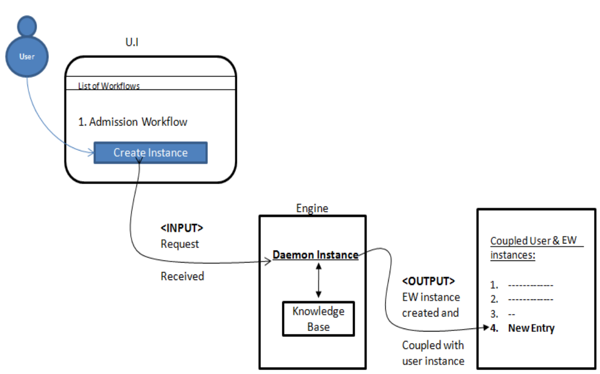


Fig. 4. Initiation step of execution: EW instance creation by Daemon instance

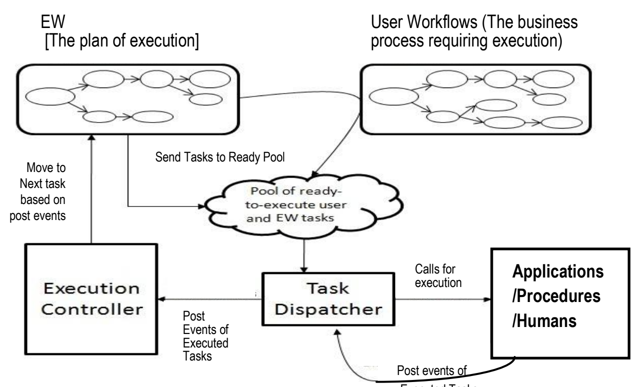


Fig. 5. Coordination of various components of the system

coupled and thus, there is no associated user workflow instance or a user task. The EW task goes to the dispatcher for execution and as a result, a `success' post event is returned along with it and an Instance Id for the new user workflow instance that got created. From this step on, the instances of EW and user workflow remain coupled. The states of both the instances are updated as running by the EW instance and the execution of user workflow instance begins.

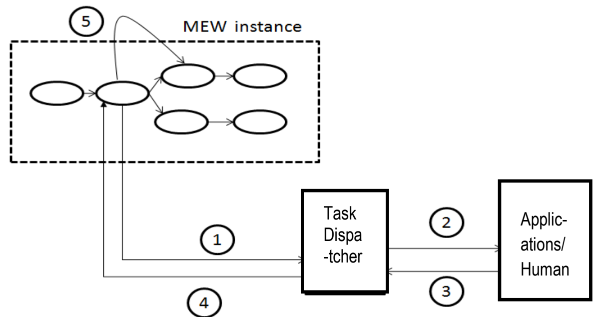


Fig. 6 Stepwise Data and Control flow

Step 2: Execution through EW. Referring to Fig. 1, the next task in the EW definition is to attempt to get a new user task and if successful, proceed with checking of the resources (such as data, actors, rules, roles of actors, procedure or execution logic) in order to execute the user task. The EW instance updates and maintains the *state* of the workflow throughout the process.

The interaction between the system components to perform execution is shown in Fig. 5. The EW instance puts EW and user workflow tasks for execution in the pool and in turn waits for the TD to handle task to execute and capture the post events of the task after execution. TD now has to identify whether the task is of type EW or user workflow and accordingly assign it to corresponding humans/applications that execute the task. The TD collects the post events and output data for the respective tasks from the humans/applications and posts it back to the Controller. The process continues till the user instance is completely executed. In Fig. 5, only one coupled user workflow and EW pair is depicted, whereas, at any time, there are many such pairs of user workflow and EW instances executing and interacting with the TD. Note that the user workflow instance does not directly interact with the TD; it is only when the EW decides that a user task is ready for execution, it puts that user workflow task in the pool. The process of moving from one task to next and dispatching it for execution continues till the user workflow instance completes its execution. The EW stops when the execution of `Get a Task' returns `End' as the post-event (i.e., the EW has no more user tasks to work on). The state of the workflow instance is updated to `complete'. Following this procedure, execution of the admission workflow instance is completed by EW and is shown (partially) in the Table 2.

The EW tasks may return `fail' as a post event in case of an unanticipated exception. In that case, the control is returned to the daemon instance, which then selects another EW definition from the Knowledge Base and the execution process repeats from that state. If execution is not successful with all the EW definitions, the current EW definition can be modified or a new definition can be evolved by the Workflow System Manager or human intervention is sought to handle the exception.

2.4 Data and Control Flow

Consider the admission workflow instance running to illustrate data and control flow. A single workflow instance is taken for simplicity:

User Workflow Instance ID :*13*
Coupled EW Instance ID: *10*
Coupled EW Instance Current Task*: Check Resources*
User Workflow Instance Current Task: *Schedule Interview*

At this point, the control is with the EW instance.

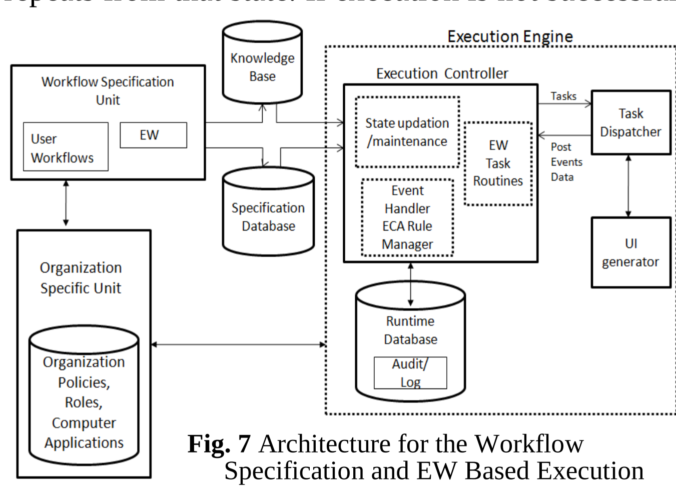


Fig. 7 Architecture for the Workflow Specification and EW Based Execution

The shift of control and data is described in the following steps and depicted in Fig. 6.

1. The EW instance 10 passes the control to the TD, for the execution of its task “Check Resources" for “Schedule Interview”.

2. TD sends the task for execution to an application/human.
3. The task is executed; events are captured by the TD and the control if returned to the dispatcher.
4. The TD returns a post event data tuple to the EW instance ID 10. With this step, the control and data tuple is returned back to the EW instance task – "Check Resources" (where it started from).
5. Based on the post event (suppose "success" in this case), the current EW task passes on the control and the post event data to the new EW task – "Execute".

So on, the control and data flows between these entities (Fig. 6) till user workflow instance completes its execution. Data flow is constantly happening between Applications, TD and the EW instance's tasks, which decides the control flow of execution.

3 Architecture & Implementation of EW System

The architecture of the EW based framework (enhancement to existing WFMSs architectures) is shown in Fig. 7. The components - Specification Unit (the process modeler), Organization Specific Unit (domain related knowledge required for workflow execution) and UI generator are the standard components of any WfMS [4, 5]. Whereas, the components Execution Controller and TD are the new additions in the architecture and constitute the execution engine in our framework. A process modeler can be used for the workflow specification. The workflow system designer specifies EWs, the workflow application designer specifies the user workflows. Once specified, EWs and user workflow specifications are stored in the Specification Database.

The KB stays connected with the Specification Unit to store various definitions of EW when they are updated. During specification, the user workflow tasks also need to be associated with user roles, organizational policies or applications for workflow tasks. Thus, the specification unit is connected to the Organization Specific Unit which contains the user workflow domain specific data (environment). Execution Controller and TD are the major components interacting with each other to execute user workflows supported by a runtime database and a UI generator. The KB and specification database interact only with the Controller. Controller needs the workflow definitions to move from one task instance to another and complete the instance execution. Execution Controller consists of a set of programming routines that execute EW tasks. So, when a new task is added to the EW, the new task defined needs to be added to the controller section along with its execution details. Note that, as part of the EW task definitions, the controller also incorporates the event handler and the Event Condition Action (ECA) rule manager. For instance, the EW task - `Find a User task' initiates the ECA rule manager and finds out the next task based on the execution of previous task and occurrence of events. TD receives requests from the Controller to execute task instances. Dispatcher extracts the corresponding applications/humans from the organization unit that are authorized to execute the tasks. The dispatcher then updates worklists of appropriate humans or runs the computer applications to execute the requested tasks. It gets the user task instances executed and returns post event data to the Controller. Dispatcher also sends to the Controller, any process data exchanged with the user through the forms generated by the UI unit on a request from the Dispatcher. The runtime information of all instances (running/ paused/ complete) such as user workflow and EW instance coupling information, current running tasks and states are stored in the runtime database regularly populated and accessed by the Execution Controller. This database also includes the Audit/Log section which helps in recovery in case of a system crash.

3.1 Implementation details

We implemented a prototype based on this architecture with the default EW (Fig. 1). The user workflows and EWs are specified in YAWL process modeler and exported to XML for use. User workflows and EWs are given workflow Ids and stored in user Specification Database and the KB respectively. The user workflow ID is prefixed with a *`u'* and an EW with an *`m'*. It helps the TD at a later stage in distinguishing user and EW tasks. The KB, Specification unit and the runtime database are MySQL databases accessed and populated by the modules as shown in the architecture (Fig. 7). The EW task routines that constitute the Execution Controller are coded in PHP, which include the initiation algorithm, algorithm for the task scheduling by the dispatcher, execution control flow algorithms and algorithms for exception handling. TD receives requests of the form Execute(Instance Id, Task id) based on the Instance Id, the dispatcher identifies whether it is a EW or a user task instance. It maintains two queues one for the user task requests and one for the EW tasks. The dispatcher schedules tasks and sends them for execution. Scheduling is done on FCFS basis after giving preference to the user tasks over EW tasks. There are N running threads, where 0^{th} and 1^{st} are daemon threads for initiation and thread 2 to N-1 are for execution. Multiple TD instances exist for dispatching and sending tasks for execution. The number `N' or the number of TD instances will depend on the load on the system and the number of task execution requests coming in. When execution requests per minute per dispatcher exceed a certain number, a new TD is initiated for the system to stay efficient. We start with two dispatchers and increase the number, with the increase in load, while keeping an upper limit check on N. After every iteration the flow of data continues for the EW to update the runtime database for consistency and to coordinate to successfully execute workflow instances.

4 Experimental Results

A WfMS based on the concept of workflows executing workflows has been built. It was successfully used to run the admission workflows of our institution. For every new application that arrives, a new workflow instance is created and complete execution of the instance results in completion of admission process for that application. Table 3 shows that our system can add more TDs to handle increased workload. With increase in number of tasks per dispatcher per minute, new instances of TDs are created, and the system is balanced to efficiently execute the tasks. We compared execution times of the admission workflow instances, executed by two systems: (i) A traditional CapBasED WfMS [3] (ii) our WfMS. We specified the workflows such that all the tasks are automated and executed by computer applications (to avoid any human related time lag). We compared the execution times for both of these systems, executing 50 and 100 instances of the workflow. In both the cases, the execution times were found better for EW driven WfMS (see Table 4). This is mainly due to allocation of tasks by TD to multiple workflow instances (including EW instances) in parallel and availability of pre-specified EWs. With changing needs of the admission workflows, we changed the execution procedure and tested our system with different definitions of EWs, as they kept evolving. Through the added knowledge of EW and user workflow tasks while execution, we observed that our system provides better support for exception handling.

Table 3. System Stats as Number of Applications Increase

Days	Average Number of Applications Received Per Day	Tasks/ Minute	No. of Task Dispatcher Instances	Tasks per Dispatcher Per minute
1-10	5	20	2	10
11-20	12	50	4	12
21-30	30	100	9	11

Table 4. Comparison of workflow instance execution time

No. of Instances	Workflow System	Execution Time (minutes)
50	CapBasED WfMS	25
	Our WfMS	20
100	CapBasED WfMS	50
	Our WfMS	35

5 Conclusion

Current WfMSs need to be rewritten/extended to support complex business processes by facilitating modifications in the workflow execution procedure. In this paper, we present a framework that facilitates workflows executing workflows. We introduce EW and TD. We presented architecture and implementation details of a prototype system, which facilitates an alternate way to execute workflows and provides flexibility for workflow execution engine in executing workflows. The results showed that workflows executing workflows is a practical idea and there is no lag in efficiency in execution through EW.

References

[1] Kumar, A. and Wainer, J., Meta workflows as a control and coordination mechanism for exception handling in workflow systems. Decision Support Systems, 40(1):89-105, 2005.

[2] Reichert, M., Rinderle, S., and Dadam, P., Adept workflow management system: flexible support for enterprise-wide business processes. In Proc. BPM'03, pp. 370-379, 2003.

[3] Hung, Patrick C. K. and Yeung, Helen P. and Karlapalem, Kamalakar, CapBasED AMS: a capability-based and event driven activity management system. SIGMOD, 1996.

[4] van der Aalst W. and ter Hofstede, A., Yawl: yet another workflow language. Information Systems, 30(4):245-275, 2005.

[5] van der Aalst, W. and Pesic, M. and Schonenberg, H., Declarative workflows: Balancing between flexibility and support. Computer Science – Research and Development, 23:99-113, 2009.

Creating a Repository for the Design and Delivery of Web Services

John D. Delano
School of Business Administration
Cedarville University
Cedarville, OH

Atish P. Sinha and **Hemant Jain**
Lubar School of Business
University of Wisconsin-Milwaukee
Milwaukee, WI

Abstract
Existing web service repositories are not well suited to handle the multi-layered representation of web services, nor are they able to support multiple development methodologies. We describe the design and development of a repository called Web Service Crawler that supports both a traditional development methodology for the initial design of services, and an agile approach for the design of composite applications. Web Service Crawler is developed based on a set of theory-based design characteristics, and includes novel facets to represent multi-layered web services, such as workflow, composition, and layer. The positive evaluation results indicate that Web Service Crawler can be used to guide web service designers, as well as enable consumers to more easily find and use services.

Introduction
As the adoption of service-oriented architecture grows, reuse plays an increasingly important role in the development of modern applications. Repositories that store and provide access to software components are key enablers of reuse. Existing component repositories have demonstrated that they can manage the task of storing and retrieving components with high levels of accuracy, precision, and recall (Vitharana et al. 2003). With a shift toward service orientation, however, existing component repositories and existing web service technologies, such as UDDI, may not be able to satisfy the demands of the new paradigm. Two factors primarily contribute to this inadequacy. First, existing technologies are not well suited to handle the multi-layered representation of web services. Second, service orientation introduces the challenge to support both the selection and composition of services to form new systems and the initial design of services to encourage future reuse.

These two factors suggest the need for a repository that can support both a traditional development methodology with a focus on long-term planning for the initial design of services, and an agile approach for the design of composite applications (Haines and Rothenberger 2010). However, it is not clear from the extant literature how to design a repository that supports the two different methodologies. Our paper addresses this issue by adopting the design-science research paradigm (Hevner et al. 2004) to create and evaluate a prototype repository based on a set of theory-guided design characteristics.

Web Service Crawler Design Specification
Designing a repository to fulfill the dual role of supporting design and delivery of web services requires a solid foundation in conceptual modeling as well as in human information processing. Although there is some overlap, the conceptual modeling literature primarily impacts the repository with respect to the initial design of services, while the human information processing literature impacts the repository with respect to the creation of composite applications. We therefore began the development of our prototype repository, *Web Service Crawler*, with a thorough review of those bodies of knowledge. Based on the review, we identified ten design characteristics that we use to inform the development of Web Service Crawler. Because of space constraints, we briefly describe the design characteristics below.

Simulation Capable. Experienced designers often use mental simulations to confirm the validity of a design. The use of concept maps can be used to allow the designer to visually specify new service compositions or select existing compositions, while guiding the designer through a simulation of his or her design.

Usability. Repository users should be able to view their interaction with the repository as a seamless interface (Matook and Indulska 2009). The repository should enable more experienced designers to quickly interact with the repository, relying on visual cues to trigger activation of prior knowledge, but for service selection, the user interface should be simpler in nature by presenting intuitive visual cues to enable easier exploration of the repository.

Queryability. Users can query the repository based on values stored in facets. The repository can accommodate both keyword-based and facet-based search strategies. Locating web services in the repository should not rely solely on a simple string-matching algorithm, but it should introduce some method for matching semantically similar services.

Orientation. Because the repository has to support the initial design of web services as well as the selection of existing services for composition, it must be designed to support two distinct audiences: service designers and service consumers.

Feedback. The web service repository should provide feedback relating to the cost of web service selection and the complexity of web service composition.

Understandability. A faceted approach to service representation can influence the understandability of each service stored in the repository by prompting the designer to consider specific attributes of each service that focus not only on current requirements, but also on future reuse potential. A faceted approach should also improve understandability of web services during search and retrieval (Matook and Indulska 2009).

Flexibility. The repository should use an extensible, faceted classification system that can evolve with the web services technology (Matook and Indulska 2009).

Completeness. The repository should contain an appropriate number of facets within its schema, such that web services are represented as completely as possible, while minimizing the cognitive burden of the analyst (Matook and Indulska 2009).

Correctness. Uses some kind of mechanism to encourage the entry of correct facet values, using either formal grammar checking or peer review (Lindland et al. 1994).

Generality. Represents multiple layers of web services as steps along an abstraction continuum between business logic and application logic (Matook and Indulska 2009).

Design and Development of Web Service Crawler

We created the repository using a 3-tier distributed architecture. The client tier was developed using Microsoft Silverlight, the middle tier using Windows Communication Foundation (WCF), and the database tier using Microsoft SQL Server 2008. The conceptual development of the prototype is guided by the design characteristics discussed above. From Semantic Network Theory, we understand that human memory is organized similar to a graph data structure, where nodes that represent memory engrams are connected via arcs that represent a single activation path (Ashcraft 2002). We pattern the design of Web Service Crawler after this structure to enable the repository to fill the role of external memory. In this way, the repository is able to support and extend user simulations, fulfilling the first design characteristic. Specifically, Web Service Crawler views the basic unit of memory as a single web service, and represents each web service as a node within a graph data structure.

The repository implements two types of arcs to act as activation paths when retrieving services from the repository. One is the composition relationship. Whenever a service initiates or makes use of the services at the same or lower layer in the service hierarchy, the repository creates an arc between the parent and child service to represent the composition relationship between them. The second type of arc is for semantic relationships, which represent dynamic connections between services based on their similarity. The repository calculates semantic similarity based on the ranking score provided by SQL Server's Full Text Search (FTS).

As suggested by multiple design characteristics, a faceted classification approach is used to represent the meta-data about each web service in the repository. The facets identified by Vitharana et al. (2003) are modified and extended to capture the features of web services more precisely (see Table 1). In addition to modifying the workflow facet to include a visual diagram of the business process automated by the web service, we introduce two completely new facets in the repository – *layer* and *composition* – to support the representation of web service hierarchies.

Facet	Description
Name	Used as a means to identify a web service.
Layer	One of four possible layers: utility, entity, task, or orchestration (Erl 2005)
Description	Descriptive information to facilitate retrieval
Keywords	Elicit thoughts about the important characteristics of the service.
Operations	Defines the capabilities that are offered by each service.
Role	Defines what part the service might play in a composite application.
Rule	Defines high-level business rules that are enforced by the service.
User	Identifies the end-user of the service.
Workflow	Contains a workflow diagram for web services that are categorized at the task or orchestration level to provides an efficient means by which the analyst can achieve a more detailed understanding of the web service's functionality.
Composition	Represents each web service as an object node within a directed graph. Each edge that connects two nodes represents a relational link between the two objects: either child (i.e. composition relationship) or friend (i.e., conceptual similarity between the two services).

Table 1 – Facets Used in Web Service Crawler Repository

The user interface of the repository is created to guide the user through the process of designing web service meta-data, as well as to identify and select existing services for later composition. Web Service Crawler accomplishes this through the use of two separate modes: design and exploration. When in design mode, the repository enables a web service designer to manage a list of web services that have been created in the repository by the designer as well as other designers who belong to the same group. This group sharing mechanism allows a team of designers to collaborate on the creation of a web service composition hierarchy. A designer can create a new web service by navigating to a wizard-like series of screens that guide the user by prompting him/her for the facets of information described earlier. When the user hovers the mouse pointer over help icons that are next to each data entry field, a tooltip window appears containing example data for the given field. This helps to encourage completeness and correctness of the data stored within each facet.

On navigating to the next page, the user completes the workflow facet by creating a graphical process flowchart for the new web service (see Figure1a). The interface provides a navigation option to proceed to the final page, where the value for the composition facet can be entered. The composition facet records the relationship between two services when one of the services invokes the functionality of the other service (see Figure 1b). The service that initiates the invocation is referred to as the parent, while the recipient of the invocation is the child. For each service composition facet, only immediate child services are represented in the context of their relationship with the service being edited.

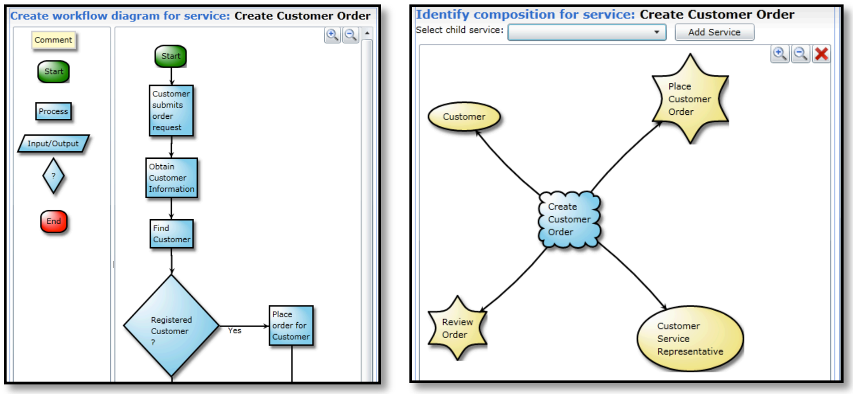


Figure 1a (left) – Web Service Crawler Workflow Screen; 2b (right) – Composition Screen

Web Service Crawler also provides an exploration mode for service consumers interested in building a composite application. It provides the ability to search for services in the repository using two search techniques: simple text search and facet-based search. Both techniques rely on SQL Server FTS to produce a relevance score for each search result, providing the user with valuable feedback. Each service listed in the search results provides a hyperlink to an exploration screen, where the user can visually examine the composition hierarchy for the selected service, along with the semantically related services in the repository (see Figure 2).

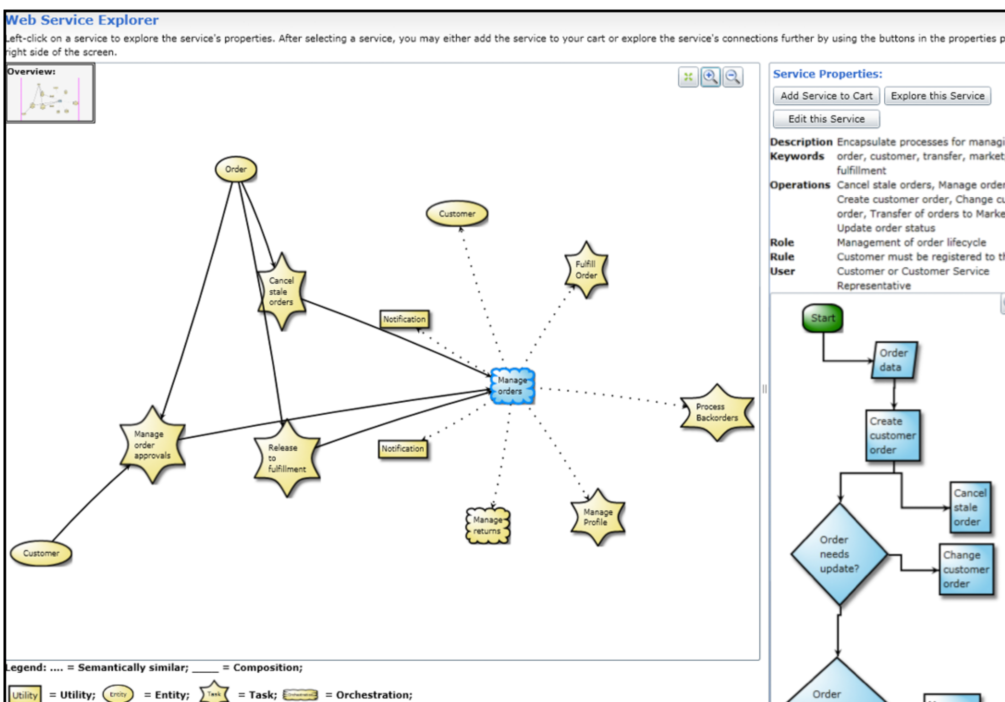


Figure 2 – Web Service Crawler Exploration Screen

The user can view the meta-data of a selected service on the right side of the screen. The exploration screen also provides a means for the user to “crawl” through the service linkages to view similar services or other services in the same composition hierarchy. The user can also add the selected service to a shopping cart. Once a service is inside the shopping cart, an indicator at the border of the screen displays information regarding the number of services in the cart, along with the total cost of the services, meeting the requirement of the feedback design characteristic.

Web Service Crawler Evaluation

We evaluated Web Service Crawler with respect to its role as a web service design tool and its ability to deliver web services. The validation process involved both the service designers who designed and stored the services in the repository and an independent panel of IT reviewers, serving as a proxy for service consumers, who were given a live demonstration of the system.

Two separate classes of students, enrolled in a Systems Analysis and Design course in two successive semesters, worked in teams to produce a design for a reusable web service hierarchy that consisted of at least 20 services. Students were asked to base their service designs on business rules that were publicly available at IBM's Infocenter website. The site contains high-level, hierarchical descriptions of e-commerce business processes.

The prototype was evaluated based on the "quality in use" HCI standard, which measures the degree of excellence to which a product meets the needs of its users. Its components include functionality, productivity, and usability. We developed an evaluation instrument from the standpoint of service design that consisted of 25 statements on a 5-point Likert scale. Sixty-five subjects participated in the evaluation (36 in the first semester and 29 in the second semester).

Exploratory factor analysis with Varimax rotation was utilized to demonstrate the discriminant validity of the factors represented by the questionnaire. The reliabilities for the usability, productivity, and functionality scales were 0.812, 0.778, and 0.779, respectively. We follow the convention of reporting usability on a 0 to 100 scale by normalizing the score for each component. The component scores for each of the two classes of students appear in Table 2.

	Service Designers			Service Consumers
Factor	**Class 1**	**Class 2**	**Difference**	
Usability	66.78	81.17	14.39 (t=4.435, p=0.000)	74.77
Productivity	71.18	72.43	1.25 (t=0.284, p=0.777)	73.24
Functionality	74.58	82.24	7.66 (t=1.999, p=0.050)	68.37

Table 2 – Quality in Use Scores by Factor

The scores for all the components are relatively high, especially given that human respondents tend not to rate any product at extreme ends of a scale. The scores provide strong evidence that the designers who used the repository found it to be of high quality in terms of functionality, productivity, and usability. These scores look particularly impressive when considering the complexity of the business processes for which the web services were designed.

An independent panel of reviewers validated the system from a service consumer perspective. Fourteen experts working for a major, global IT consulting firm and 50 IT students participated in the validation process. The Web Service Crawler system was demonstrated to both groups, and a questionnaire was then administered, similar to the one for designers, but from a consumer's perspective. Table 2 shows the average scores for the consumer group. The reviewer scores are quite high, providing support for the external validity of the repository.

When asked if a service workflow in the repository is helpful for understanding that service, the reviewers provided a high rating (74.206). When asked if composition of a service helps in understanding, the reviewers provided another relatively high rating (70.23). There is strong evidence, therefore, that the modified workflow facet and the newly introduced composition facet facilitate understandability of web services.

In addition to the objective scores, subjective feedback was also solicited from the reviewers on the strengths and weaknesses of the repository. With respect to the strengths, the recurring themes in the feedback were reusability, development speed, ease of use, and search capabilities. One reviewer stated, "Easy to use, nice GUI, good search capabilities, great idea – many companies suffer from reinventing the wheel every project despite efforts of developers to make code repositories that aren't supported by management." As for the weaknesses of the system, most of the comments were related to implementation issues with the common theme being security; however, this was not the focus of the study.

Conclusion

This study contributes to the extant research in the area of storing, classifying, and locating web services. The number of web services that will be available to developers in the future will be extraordinarily high, and the need for an efficient classification, storage, and retrieval infrastructure will be essential for developers to assemble high-quality applications. The findings from this study indicate that the solution to that need lies in a lightweight, facet-based, web service repository that provides facets such as workflow, composition, and layer to support the design and delivery of web service hierarchies. From the standpoint of practice, the development of a repository that can be used to guide web service designers, as well as enable consumers to more easily find web services, should enhance the distribution of services within and across organizational boundaries. Unlike existing tools, such as IBM's Service Registry and Repository, Web Service Crawler is lightweight so that it can better support the agile exploration and composition of web services.

Finally, it is important to address two limitations of this study. First, Web Service Crawler does not yet support the design and retrieval of web services based on non-functional characteristics, such as quality of service metrics. Future iterations of the repository will include this capability (Thurow and Delano 2010). Second, technical challenges prevented a hands-on evaluation of the repository by the service consumers. Future evaluations of the repository will include a more in-depth, hands-on evaluation of the system.

References

Ashcraft, M. H. 2002. *Cognition*, (3rd ed.) Prentice Hall: Upper Saddle River, NJ.

Erl, T. 2005. *Service-Oriented Architecture: Concepts, Technology, and Design*, Prentice Hall: Upper Saddle River, NJ.

Haines, M. N., and Rothenberger, M. A. 2010. "How a Service-Oriented Architecture May Change the Software Development Process," *Commun ACM* (53:8), pp 135-140.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quart* (28:1), pp 75-105.

Lindland, O. I., Sindre, G., and Solvberg, A. 1994. "Understanding Quality in Conceptual Modeling," *IEEE Software* (11:2), pp 42-49.

Matook, S., and Indulska, M. 2009. "Improving the Quality of Process Reference Models: A Quality Function Deployment-Based Approach," *Decis Support Syst* (47), pp 60-71.

Thurow, N. A., and Delano, J. D. 2010. "Selection of Web Services Based on Opinion Mining of Free-Text User Reviews," in *ICIS 2010 Proceedings*: St. Louis, MO.

Vitharana, P., Zahedi, F. M., and Jain, H. 2003. "Knowledge-Based Repository Scheme for Storing and Retrieving Business Components: A Theoretical Design and an Empirical Analysis," *IEEE T Software Eng* (29:7), pp 649-667.

Online Review Sentiment Classification Using a Dictionary-Enriched Text Mining Model

Thomas L. Ngo-Ye
School of Business
Dalton State College
tngoye@daltonstate.edu

Atish P. Sinha
Lubar School of Business
University of Wisconsin-Milwaukee
sinha@uwm.edu

Abstract

We propose a conceptual framework to organize different review sentiment classification models based on the underlying knowledge type. While pure bag-of-words (BOW) models represent specific knowledge, pure dictionary-based models represent general knowledge. We operationalize a hybrid model by pooling together a subset of review words and dictionary categories as predictors. We empirically compare these models with reviews from Yelp.com. An interesting finding is that, besides the positive and negative dimensions, around a dozen dictionary dimensions are also useful for review sentiment classification. This result reveals the complexity of review text language for sentiment analysis. The hybrid model performs the best among the three types of models. We find that both specific knowledge and general knowledge are important for review sentiment classification.

Introduction

Online customer reviews contain valuable consumer information for product manufacturers, service providers, and e-Commerce retailers. The sentiment carried in a review text may affect the readers' perceptions and attitudes toward the product or brand, and may influence the purchase decision. For businesses, review sentiment is an invaluable barometer of consumer attitude. Detecting the sentiment of online customer reviews helps businesses better understand the customer perspective and adjust operations accordingly.

In this study, we investigate the efficacy of specific knowledge model (BOW-based) and general knowledge model (dictionary-based) for predicting review sentiment (positive or negative). We propose and examine a hybrid model that combines the two types of knowledge.

Sentiment Analysis

In the past few years, there has been a growing research interest in sentiment analysis (Abbasi et al. 2008) and its applications to online reviews (thumbs up or thumbs down). Sentiment analysis is of real practical need and is recognized as a technically challenging task compared to the traditional topic-based text classification (Liu 2010; Pang and Lee 2008). In the context of text processing, sentiment analysis refers to the identification and assessment of written expressions of subjective mental states (Wright 2009). An important dimension of sentiment analysis is the polarity (positive or negative). Four dominant research topics have been identified in the area: subjectivity classification, word sentiment classification, document level sentiment classification, and opinion extraction (Tang et al. 2009).

Dictionary Approach and General Inquirer

Instead of using the actual words that appear in the text as independent variables, dictionary categories (tags) are used as predictors in text analysis. A computer program assigns tags to the words according to the conception of meaning, which is facilitated by a dictionary with word families. For example, many words express a positive sentiment and they are grouped into a category with the label "positive". This represents the abstraction of meaning from the actual word to its word family (Krippendorff 2004).

General Inquirer (GI) (http://www.wjh.harvard.edu/~inquirer/) is one of the earliest and most widely used dictionaries in content analysis. As a psycho-sociological dictionary, GI contains dimensions such as positive and negative, and David McClelland's theories regarding the need for achievement, power, and affiliation. The domain-independent positive and negative adjective sub-dictionaries have been used in various studies of text analysis (Higashinaka et al. 2007; Kim et al. 2006).

Conceptual Framework of Models

In this research, we explore the efficacy of three broad kinds of models for classifying review sentiment. They are pure bag-of-words (BOW) models, pure dictionary-based models, and hybrid models combining words from the BOW model and dictionary dimensions. These three kinds of models have the same target variable – review sentiment. The unit of study is an individual review and sentiment classification is at the document level. The datasets for all the models are derived from the same review data. The main difference among the models is in the choice of predictors.

In the pure BOW models, the independent variables are the words that appear in the review text. The review words represent the specific information unique to the particular review collection. Therefore, we can argue that a BOW model focuses on the specific knowledge in modeling review sentiment. In contrast, in a pure dictionary-based model, the independent variables are the categories or dimensions of a dictionary (e.g. positive emotion, negative emotion, cognitive processes, perceptual processes, etc.) The dictionary categories are pre-defined without specific consideration of the review domain examined. The dictionary dimensions represent general concepts of interest to theorists in the disciplines of psychology, sociology, and linguistics. Therefore, we argue that a dictionary-based model embodies general knowledge. In this research, we propose a hybrid model, pooling together predictors from both the BOW model and the dictionary-based model. Conceptually, the hybrid model represents both specific knowledge and general knowledge. We argue that both types of knowledge are useful for predicting review sentiment, and that a hybrid model combing the two is an attractive option.

Another important difference among the models arises from whether a dimension reduction technique is applied or not. Dimension reduction methods, such as Correlation-based Feature Selection (CFS), can help reduce the distractive influence of irrelevant, redundant, and noisy features. Dimension reduction may improve model accuracy, increase computational efficiency, and generate smaller models. Table 1 presents a conceptual framework of the models.

Table 1. Conceptual Framework of Review Sentiment Classification Models

	BOW	General Inquirer (GI)	Hybrid
Without Dimension Reduction	BOWFull	GIFull	
With Dimension Reduction	BOW-CfsBF	GI-PosNeg, GI-CfsBF	BOW-CfsBF-GI-CfsBF

The BOWFull model is a pure BOW model using all the unique words that appear in a review text collection as independent variables. CfsBF (Hall 2000) is applied to the BOWFull model to select a subset of review words as predictors to generate the BOW-CfsBF model. The chosen variables in the subset have high correlation with the dependent variable and low correlation among themselves. Therefore, BOW-CfsBF focuses on only the more important and less redundant review words. BOW-CfsBF represents the filtered salient specific knowledge found in the review collection.

The GIFull model is a pure dictionary-based model using all the categories of GI. The GI-PosNeg model only uses two categories – positive words and negative words in a text – as independent variables. These two dimensions directly measure the amount of positive and negative sentiment expressed in a text. Therefore, GI-PosNeg can be interpreted as focusing on only these two most seemingly relevant dictionary dimensions – salient general knowledge. CfsBF is applied to GI to select a subset of categories as independent variables to produce the GI-CfsBF model. GI-CfsBF can be interpreted as focusing on only those dictionary dimensions believed to be the most essential for predicting review sentiment – filtered salient general knowledge.

The BOW-CfsBF-GI-CfsBF model is a hybrid one. Its independent variables include both CfsBF-selected review words and CfsBF-selected categories of GI. The hybrid model represents both the filtered salient specific knowledge found in the review text collection and the filtered salient general knowledge found in the GI dictionary.

Empirical Study and Results

We collected real-world restaurant review data from Yelp.com. Consistent with the existing literature, we derive review sentiment based on the star rating assigned by the review author. We treat 1 and 2 stars as negative, 4 and 5 stars as positive, and 3 stars as neutral. We remove the neutral sentiment group and reduce the problem to a binary classification problem. In Table 2, we report the number of reviews for each sentiment category.

Table 2. Number of Reviews for Sentiment Analysis

Dataset	Total	Positive	Neutral	Negative	Pos+Neg
Yelp Reviews for 2 Restaurants	540	278	160	102	380

For pure BOW models, we experiment with four index weighting schemes (Binary Occurrence, Term Occurrence, TF, and TF/IDF) to assign a term weight to each feature in the review text. In dictionary-based models, each column (independent variable) corresponds to a particular category (dimension) in the GI dictionary. The realized value of each entry in dictionary-based datasets is an aggregate measure of a dictionary dimension for a review text. All these values can be calculated by matching a review text with the GI dictionary dimensions and then normalizing the count frequency. For example, if a review text with a total of 100 words contains five words that belong to the positive category in the GI dictionary, the realized value for its positive dimension is 0.05. In Table 3, we present the number of variables of the six models we examined. In Table 4, we report the 14 GI categories selected by CfsBF, which captures the general knowledge relevant for review sentiment analysis.

Table 3. Number of Variables for the Six Models (including target class)

BOWFull	BOW-CfsBF	GIFull	GI-PosNeg	GI-CfsBF	BOW-CfsBF-GI-CfsBF
3059	89	187	3	15	103

Classification algorithm and experiment configuration

We apply radial basis function network (RBFNetwork) as a classification algorithm to all datasets to generate the models. To obtain reliable results, we design a text mining experiment with 10 runs and, within each run, we use 10-fold cross-validation for performance estimation.

The overall average across 10 runs and 10-fold cross-validation (100 observations) is reported as the classification performance (see Tables 5 and 6). For each pair of models of interest, we conduct paired *t*-tests and report if the differences are statistically significant.

Table 4. Description of 14 GI Categories Selected by CfsBF for Classifying Review Sentiment

Category	Description	Category	Description
Positiv	words of positive outlook	Negativ	words of negative outlook
Ngtv	negative words, an earlier version of Negativ	Weak	words implying weakness
Pleasur	words indicating the enjoyment of a feeling, including words indicating confidence, interest and commitment	Vice	words indicating an assessment of moral disapproval or misfortune
POLIT	words having a clear political character	PLACE	References to places, locations and routes between them
Exert	a movement category	Name	only contains 86 names identified in the Harvard IV dictionary
Negate	words that refer to reversal or negation and generally signal a downside view	AffOth	affect words not in other categories
AffTot	words in the affect domain	NegAff	words of negative affect denoting negative feelings and emotional rejection

Six classification performance measures

Besides Accuracy, we report five additional classification performance measures. They are true positive rate (TPR), false positive rate (FPR), Precision, F-Measure, and area under ROC curve (AUC). To incorporate information of classification performance into both the positive and the negative class, we take the macro average, weighted by the relative class frequency, for the above measures except Accuracy. Among the six measures, only for FPR, a smaller realized value means better performance. For the remaining five measures, a larger realized value indicates a better result.

Results of Text Mining Experiments

In Table 5, we present the results of comparing BOWFull and BOW-CfsBF models across four index weighting schemes with respect to the six classification performance measures. We highlight the better performing model in the pair in bold.

Table 5. Comparing BOWFull and BOW-CfsBF for Classifying Review Sentiment

	BinaOccu		TermOccu		TermFreq		TFIDF	
	BOWFull	BOW-CfsBF	BOWFull	BOW-CfsBF	BOWFull	BOW-CfsBF	BOWFull	BOW-CfsBF
Accuracy	71.947%	**89.711%**	71.84211%	**89.500%**	69.86842%	**87.263%**	70.42105%	**87.132%**
TPR	0.71947	**0.89711**	0.71842	**0.89500**	0.69868	**0.87263**	0.70421	**0.87132**
FPR	0.51161	**0.26676**	0.51456	**0.27248**	0.60000	**0.29115**	0.56191	**0.29912**
Precision	0.70352	**0.90780**	0.70211	**0.90600**	0.65926	**0.87644**	0.67667	**0.87539**
F-Measure	0.70337	**0.88756**	0.70231	**0.88506**	0.66613	**0.86289**	0.68054	**0.86070**
AUC	0.60182	**0.90484**	0.60322	**0.87462**	0.55107	**0.83606**	0.57238	**0.82992**

Table 5 shows that BOW-CfsBF always outperforms BOWFull. Based on paired *t*-tests, we find that in all the 24 scenarios, the performances are significantly different ($p < 0.001$). Therefore, the feature selection method CfsBF helps to improve review sentiment classification performance of the BOWFull model. Moreover, we find that BinaOccu always outperforms the other three indexing schemes. Therefore, we will use BinaOccu-based BOW-CfsBF to represent specific knowledge in the hybrid model.

Table 6. Comparing GIFull, GI-PosNeg, GI-CfsBF & BOW-CfsBF-GI-CfsBF for Classifying Sentiment

	GIFull	GI-PosNeg	GI-CfsBF	BOW-CfsBF-GI-CfsBF
Accuracy	66.842%	75.684%	76.026%	**91.026%**
TPR	0.66842	0.75684	0.76026	**0.91026**
FPR	0.54201	0.58471	0.46947	**0.22335**
Precision	0.66060	0.73548	0.74678	**0.91742**
F-Measure	0.66050	0.70652	0.74118	**0.90374**
AUC	0.59804	0.71211	0.76978	**0.90934**

From Table 6, we find that across the six classification performance measures, GI-CfsBF always outperforms both GIFull and GI-PosNeg. For all the six measures, GI-CfsBF is significantly better than GIFull ($p < 0.001$). For FPR, F-Measure, and AUC, GI-CfsBF is significantly better than GI-PosNeg ($p < 0.001$). For Accuracy, TPR, and Precision, GI-CfsBF and GI-PosNeg are not significantly different. When comparing GIFull and GI-PosNeg, for five out of the six measures, GI-PosNeg performs significantly better ($p < 0.001$). Only for FPR, GIFull significantly outperforms GI-PosNeg ($p < 0.01$). Since GI-CfsBF is the best among these three models, we will use GI-CfsBF to represent general knowledge in the hybrid model.

From Table 5 and Table 6, the results show that for all the six measures, the hybrid model – BOW-CfsBF-GI-CfsBF – always outperforms both BOW-CfsBF (BinaOccu-based) and GI-CfsBF. For five of the six measures, the hybrid model performs significantly better than BOW-CfsBF ($p < 0.001$). Only for AUC, the hybrid model and BOW-CfsBF are not significantly different. For all the six measures, the hybrid model is significantly better than GI-CfsBF ($p < 0.001$). When comparing BOW-CfsBF and GI-CfsBF, we find that BOW-CfsBF significantly outperforms GI-CfsBF under all the six measures ($p < 0.001$).

Conclusion and Future Directions

In this study, we proposed and empirically compared different models. Categorizing different models according to the underlying knowledge type is a contribution of this paper. This framework helps us to develop a deeper understanding of review sentiment.

Comparing all the six models presented in this study, for all the six classification performance measures, we find that the hybrid model (BOW-CfsBF-GI-CfsBF) always performs best (see Tables 5 and 6). The finding suggests that both specific knowledge and general knowledge are important for review sentiment analysis. Combining both types of knowledge helps form a more complete understanding of review sentiment. The finding that the fusion of these two types of knowledge is even more powerful is a major contribution to the sentiment classification literature.

The fact that BOW-CfsBF outperforms GI-CfsBF suggests that, compared to general knowledge, filtered specific knowledge is more powerful for review sentiment classification. The implication is that specific keywords are instrumental in determining review sentiment. General knowledge by itself is not a strong predictor as specific knowledge. The sentiment of a review phrase can be context and domain dependent. The richness and complexity of review text makes the specific knowledge model more effective than the general knowledge model. Empirically comparing the specific knowledge model and the general knowledge model, and offering a plausible explanation is a contribution of this study.

In previous sentiment analysis studies (Higashinaka et al. 2007; Kim et al. 2006), when the dictionary approach was used, most of the attention was devoted to the positive and negative word dimensions. Our empirical study has revealed an interesting result – the GI-CfsBF model outperforms the GI-PosNeg model (see Table 6.) This finding suggests that the model with only positive and negative word dimensions might be over-simplistic and other seemingly non-relevant dimensions have turned out to be subtly expressive with respect to sentiment. Therefore, it indicates the complexity of language used in review text. This study contributes to the sentiment analysis literature with the new insight that we should broaden our focus to include dictionary dimensions other than positive and negative words (see Table 4).

We investigated review sentiment classification with Yelp restaurant review data. In the future, we can extend this work by examining other review domains. We used RBFNetwork as the classification algorithm and CfsBF as the dimension reduction method. We can experiment with other algorithms and dimension reduction methods to generalize the results. Also, the framework proposed in this paper can be applied to perform sentiment analysis for other emerging social media, such as blog posts, Facebook updates and comments, and tweets. We believe that this line of research will generate interesting theoretical and practical insights.

References

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems,* 26(3), 12:1-12:34.

Hall, M. A. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, 359–366.

Higashinaka, R., Walker, M. A., and Prasad, R. (2007). An Unsupervised Method for Learning Generation Dictionaries for Spoken Dialogue Systems by Mining User Reviews. *ACM Transactions on Speech and Language Processing*, 4(4), 1-25.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically Assessing Review Helpfulness. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 423-430.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, California: Sage Publications, Inc.

Liu, B. (2010). Sentiment Anlaysis and Subjectivity. In N. Indurkhya, and F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed.), 1-38.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval* (Vol. 2), 1-135.

Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications,* 36, 10760-10773.

Wright, A. (2009). Our Sentiments, Exactly. *Communications of the ACM,* 52(4), 14-15.

Improving Sentiment Analysis by Global Feature Dimension and Parameter Optimization in Support Vector Machines

Xinmiao Li[1], Yilu Zhou[2], Jing Li[1], Pengzhu Zhang[3]
[1] Shanghai University of Finance and Economics, xinmiaoli@shufe.edu.cn
[2]The George Washington University, yzhou@gwu.edu
[3]Shanghai Jiaotong University, pzzhang@sjtu.edu.cn

Abstract. *Sentiment analysis has become an important component of business intelligence. While many researches rely on Support Vector Machine to perform text sentiment classification, few have studied global optimization of feature dimension and kernel function. We propose a holistic model that finds the optimized combination of feature dimension and SVM kernel. Our experiment shows that feature reduction alone does not always improve SVM-based sentiment analysis. However, SVM performance can be improved by optimizing feature dimension as well as optimizing SVM kernel function. Furthermore, finding a global optimization can generate the best performance in sentiment analysis.*

1. Background

Sentiment analysis is the process of identifying sentiment and opinion (e.g. negative or positive) of a given piece of text (Pang et al. 2002). It is also referred to as emotional polarity computation, opinion extraction or semantic classification (Pang et al. 2002). Sentiment analysis has draw significant attention with the abundance of unstructured user generated textual data in online forums, group discussions and user reviews. Identification of sentimental orientation can provide important information to support business decision making.

Sentiment analysis is often treated as a text classification task where texts are classified into negative, positive and neutral sentiments. It has developed in recent years from a simple dictionary-based approach to a rule-based approach and a more advanced machine learning approach. Machine learning methods provide effective solutions to deal with large data set. Correct classification relies on finding the right set of features for training purpose and an effective learning algorithm. Many previous studies use bag of words approach where each word or phrase is treated as a feature. A full bag of words approach often suffers from high dimensionality and noise from irrelevant keywords. A sentiment dictionary can reduce the noise of keywords. Semantic analysis can be performed using language model theories such as Latent Semantic Analysis (LSA). Such analysis can reduce the dimension by finding the most representative keyword vectors. However, previous research used a generic LSA model only to reduce the dimension and failed tested the optimal feature dimension during such semantic analysis. Maas et al. (2011) demonstrated that using LSA does not improve performance over a bag of words approach.

Machine learning techniques such as Support Vector Machines (SVM), Maximum Entropy and the Naïve Bayes classifier have been used in sentiment polarity classification. Among them, SVM is most widely used in sentiment analysis (Prabowo and Thelwall 2009). For example, Pang and Lee (2002) applied and compared several machine learning methods, Naïve Bayes, Maximum Entropy and SVM, on movie review data from Internet Movie Database (IMDb) to determine the polarity. Tan and Zhang (2008) conducted a study of sentiment categorization on Chinese documents and find that SVM exhibited the best performance in sentiment classification. Mullen and Collier (2004) also studied movie reviews using SVM as the learning algorithm and investigated performance of various features including word unigrams and lemmatized unigrams. While SVM demonstrated superior performance over other learning algorithms in these studies, they are based on a generic SVM model. Performance of SVM classification depends on the selection of kernel function and kernel parameters (Bernd et al., 2003). While this is a well-known problem in other fields such as bioinformatics, according to our knowledge, it has not raised attention in sentiment analysis community.

Most important, optimal feature vector and optimal SVM kernel parameter estimation are not two isolated problems. The local optimal value of either feature vector or SVM kernel may not achieve the global optimal in sentiment analysis. We provide a summary of selected works in sentiment analysis and text classification using SVM as the learning machine in Table 1. Most previous research used some feature reduction method to deal with high dimension in textual data. However, feature reduction was often performed randomly without dimension optimization. SVM optimization was considered in some of them. However, there was no research study the global optimization of both feature dimension and SVM optimization. We have noticed other gaps identified in previous sentiment analysis research. For example, current studies are dominant by language corpus in English (Tan and Zhang 2008) and there is a need to study other popular languages such as Chinese. Majority of research also studied 2 classes of sentiment (positive and negative, where there is a need to study the strength of polarity with multi-class classification techniques (Kim and Hovy 2005).

Table 1. Summary of Previous Research

Research	SVM *optimization*	Feature dimension *reduction*	Dimension *optimization*	Global *optimization*	# of classes	Corpus language
Pang and Lee (2002)	No	No	No	No	2	English
Mullen and Collier (2004)	No	Dictionary-based.	No	No	2	English
Tan and Zhang (2008)	No	No	No	No	2	Chinese
Cao et al. (2010)	No	Using generic LSA with 50, 100, 150 and 200 dimensions	No	No	2	English
Li and Wu (2010)	No	Dictionary-based	No	No	5	Chinese
Maas et al. (2011)	No	Latent Dirichlet Allocation (LDA) and LSA	No	No	2	English
Lane et al. (2012)	Yes	Randomly projected 250 dimensions	No	No	5	English

3. Research Challenges and Research Question

To address the above problems, we propose a holistic model for effective multi-level polarity sentiment analysis that uses Support Vector Machines. We pose the following research questions:

1. How to achieve global optimization in SVM kernel function and feature dimensions in sentiment analysis?
2. Does feature reduction method improve sentiment analysis performance?
3. Does feature optimization method improve sentiment analysis performance?
4. Does SVM kernel optimization improve sentiment analysis performance?
5. Does global optimization further improve the performance of SVM?

4. System Design

To address these research questions, we design a conceptual model of automatic identification of multi-level polarity using LSA and SVM as shown in Figure 1. The goal is to build a model that classifies target text into multi-level polarities.

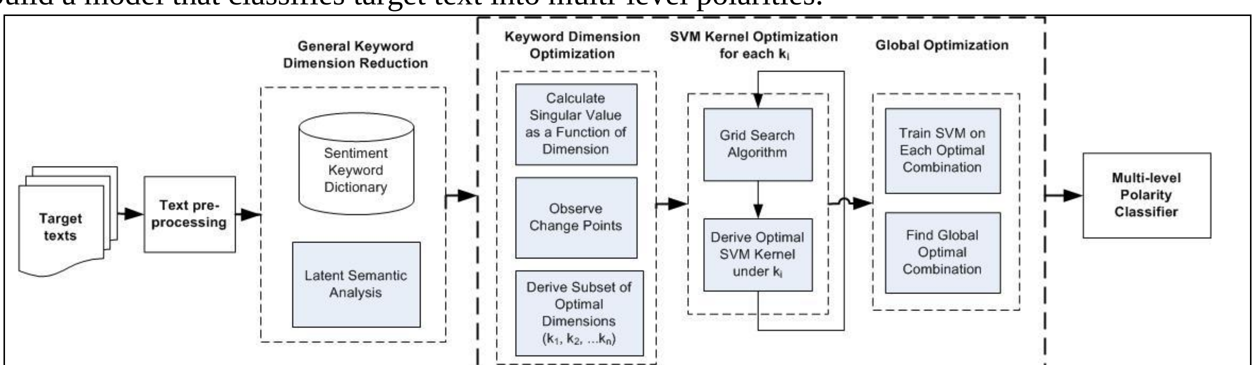


Figure 1. A Global Optimization Approach for Multi-level Polarity Sentiment Analysis

Text Pre-processing. Text pre-processing component is designed to index and extract terms that are used in later classification step. Words that do not carry a useful meaning are considered stop words and are removed to optimize the performance. Using Part-Of-Speech could potentially improve the performance of sentiment analysis. Furthermore, when a dependency parse tree is used, one can find out the importance of each word in the sentence structure and make weight adjustment in later classification.

General Keyword Dimension Reduction. In textual analysis, it is normal to have thousand term features in target texts. If the entire term space is used as feature set for SVM, terms that do not convey useful meanings will likely to create noise during training. There are several methods to reduce the keyword dimension. Using a sentiment dictionary can largely eliminate the irrelevant keywords and is widely adopted in sentiment analysis. Another effective method is Latent Semantic Analysis (LSA). The underlying idea of LSA is that the aggregation of all the word contexts provides a set of mutual constraints that determine similarity of meaning (Landauer and Dumais, 1997). LSA employs truncated singular value decomposition (TSVD) to reflect terms and text in latent semantic space. Suppose there are *m* terms *and n* documents. The initial matrix $A_{m\times n}$ can be represented as

$$A_{m\times n} = \lfloor a_{ij} \rfloor = (doc_1, doc_2, ..., \ doc_n) = (term_1, term_2, ..., \ term_m)^T \qquad (1)$$

Matrix A is the relation between feature terms and polarity of text. a_{ij} indicates the frequency a term *i* in document *j*. The initial matrix A_{mxn}, can be presented as the product of two orthogonal matrices and one diagonal matrix.

$$A = TSD^T \qquad (2)$$

T is the $m \times m$ orthogonal matrix with each column being a left singular vector in *A*. D is the $n \times n$ orthogonal matrix with each column being a right singular vector in *A*. S is a $m \times n$ rectangle diagonal matrix, with diagonal elements $s_1 \geq s_2 \geq s_3 \geq \ldots \geq s_r > 0$. s_1, s_2, s_3, …, s_r are singular values of A, sorted descending in matrix S in the diagonal. With this transformation, the *k* largest values in the matrix are preserved and the rest values are set to zero. The new matrix is a corresponding *k*-order diagonal matrix from S, a *k*-column matrix form T, and a *k*-row matrix from D^T. Consequently, a similar matrix of A is computed by using formula (3).

$$A \approx T_k S_k D_k^{\ T} \qquad (3)$$

To our knowledge, most previous research in sentiment analysis selected an arbitrary *k* at this step and assumed such dimension reduction improves SVM performance. However, Maas et al. (2011) showed that the original feature set can perform better than a truncated feature set from LSA. Thus, it is necessary to find the optimal dimension *k*.

Keyword Dimension Optimization. The optimal keyword dimension *k* by definition is the one that achieves best SVM outcome. A straight forward approach is to test all possible *k*s in building SVM learning engines. However, with the original keyword dimension being couple hundred sometimes couple thousand, this approach is inefficient. Furthermore, the selection of *k*s is not isolated from SVM kernel function and parameters. To reduce the computational cost, we propose that a subset of typical *k* can be identified by observing the varying of singular values of the term-document matrix. After the initial term matrix is transformed by TSVD, non-zero singular values are sorted with descending order in the diagonal in matrix *S*. The singular value of each matrix can be calculated quickly for each *k* value and can be treated as a function of dimension *k*. When there is a dramatic change between the singular values of two adjacent *k* values, we call this a *change point*. The *k* values at these change points are selected to truncate the matrix *D* separately shown in formula (2). Consequently, each *k* value represents a new term matrix produced with dimension of *k*.

SVM Kernel Optimization. Although SVM overcomes difficulties such as dimensionality curse, local minimization, and over-fitting problems, its performance relies on kernel function. A number of kernel functions have been used in previous research. Among them, Radial Basis Function (RBF) is a most widely used kernel. Here, we use RBF kernel as an example to illustrate the process of kernel optimization. RBF has two parameters, punishment parameter c and core parameter γ. There is a close relation between parameter estimation of RBF and the performance of SVM classifier. Therefore, SVM parameter estimation is an important component in SVM classification. Previous machine learning research on SVM has used Grid

Search algorithm to perform parameter optimization. Grid Search algorithm is a non-linear parameter estimation method. It searches the optimal solution in two steps. A general search is performed first followed by a fine search. We illustrate the algorithm in Figure 2. More importantly, our Grid Search is performed on the entire subset of candidate *k*s. We estimate optimal punishment parameter c and core parameter γ with each *k* in the subset. Each combination of (c, γ, *k*) can be viewed as a local optimal solution to SVM.

(1) Define the domain of parameters.
(2) Define the fixed step length for searching. Form a two-dimensional net on the coordinates of core parameter and punishment parameter.
(3) For further accuracy, double the filter to select parameters for SVM with general selection and fine selection.
(3.1) During parameter general selection, select a large parameters range and a long step. Run the object function.
(3.2) During parameter fine selection, we continuously examine the results to narrow the parameter value range and step. Then, put the adjusted parameter area and step into the function and run again until the satisfied results are achieved.
(4) The average accuracy of each pair of parameters on the grid is calculated by cross-validation. The goal is to identify a set of (C; γ) so that the classifier can accurately predict the polarity of testing texts

Figure 2. Grid search algorithm

Global Optimization. During global optimization, we train SVM models for each local optimal (c, γ, *k*) combination. Cross-validation can be used to identify the performance of each (c, γ, *k*) combination. By comparing the SVM performance results, including precision, recall of each (c, γ, *k*), the best performed SVM is the optimal solution and can be applied to classify multiple-level polarities.

5. Experiments

5.1 Testbed and Dataset

We chose to test our sentiment analysis model on a group discussion dataset from an insurance company in Chinese. The discussion was about ways to improve various aspects in insurance underwriting and claim assessment process. Being able to detect the sentiment polarity and the strength of polarity in group discussion will assist business' decision making capability. We randomly selected 373 texts from this group discussion. Three experts read all 373 texts and labeled them into 5 categories: strongly positive, positive, neutral, negative and strongly negative. We used a Chinese sentiment lexicon provided by HOWNET. The lexicon contains 3,116 negative sentiment terms, 1,254 negative feeling terms, 3,730 positive sentiment terms, 836 positive feeling terms and 219 terms that express a degree of sentiment. While all these terms can potentially all be keyword features, we only observed the appearance of 198 sentiment expressions in our 373 text dataset. We grouped the 373 texts into 2 parts. 187 texts were randomly picked among them to determine global optimal parameters of SVM in experiment 1. The remaining 186 texts were used to test the validity of the model in experiment 2.

5.2 Experiment 1: Finding Global Optimal Feature Dimensions and Parameters for SVM

In experiment 1, our focus is to test if our global optimal approach combining feature dimension optimization and SVM kernel optimization generates satisfactory performance. Using our proposed method, we initially formed the eigenvector of argumentation text for automatic identification. The initial term matrix in our instance study is a 198*373 matrix. Each element in the matrix indicates the frequency that the row term appears in the column text. Based on formula (2), the SVD of the initial term matrix was performed to achieve dimension reduction.

Table 1. Selected *k*, the corresponding singular value and term matrix

K value	Singular Value s_k	term matrix *D*
132	4.13e-229	$D_{373\times 132}$
125	0.6180	$D_{373\times 125}$
95	1.0000	$D_{373\times 95}$
66	1.5827	$D_{373\times 66}$
35	2.0000	$D_{373\times 35}$
13	3.2434	$D_{373\times 13}$

After the decomposition, we obtained $S_{198\times 373}$ which is the diagonal matrix and corresponding $D_{373\times 373}$ which is the orthogonal matrix of text. We plotted Singular Values for all 198 *k*s and select six *k*s with a large jump in Singular Value from the diagonal matrix. The six *k*s is used to truncate the matrix *D*. The new term matrix with the six different *k* is produced. Table 1 illustrates the six selected *k* values, its singular values in diagonal matrix, and its term matrix space D.

The term matrix from LSA dimension reduction is input into SVM. We used radial basis function (RBF) for the SVM kernel function, and Grid Search algorithm to determine punishment parameter c and core parameter γ. To improve accuracy, we double the filter to select parameters for SVM with general selection and fine selection.

During each round of SVM kernel parameter optimization, a comparatively broad domain for parameters is set for this case. We started with $2^{-10}<c<2^{10}$, $2^{-10}<\gamma<2^{10}$, cstep=0.8, γstep=0.8. Cstep is the changing step length for c, and γstep is the changing step length for γ. Using *k*=66 as an illustrating example, General selection is accomplished with c=21.1121, γ=0.43528, average Cross Validation (CV) Accuracy=74.3316%. The corresponding 3D graph is Figure 3.

After general selection, the range of c is confined to 2^{-2}-2^{8}, and γ is limited within 2^{-4}-2^{4}. We set cstep =0.5 and γstep = 0.5. Thus, selection is realized, with optimal parameters c=11.3137, γ=0.70711, corresponding average Accuracy=81.8182%, when k=66. The contour figure and the 3D graph of fine selection are illustrated in Figure 4 with k=66.

Figure 5 further illustrates the relation of dimension *k* and accuracy of classification. We observe that SVM achieved best performance when *k* is around 66. The performance of original *k* without any LSA is slightly below the performance of optimal *k.* However, if one randomly chooses a *k* value of 120 during LSA feature dimension, the performance can be lower than original feature dimension which is consistent with the study by Mass et al. (2011).

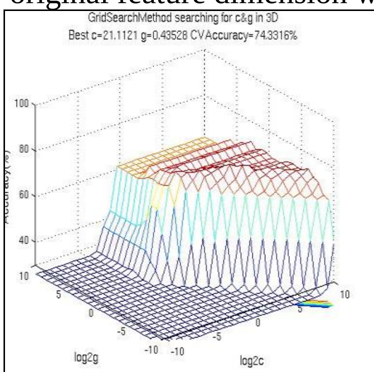


Figure 3. Contour map on general selection for c and γ roughly when k=66

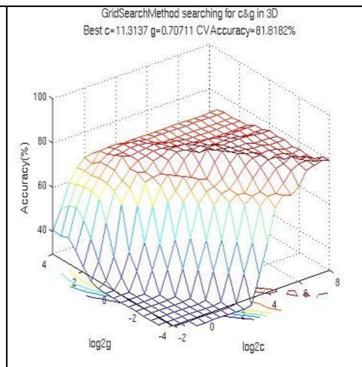


Figure 4. Contour 3D map on fine selection for c& γ accurately when k=66

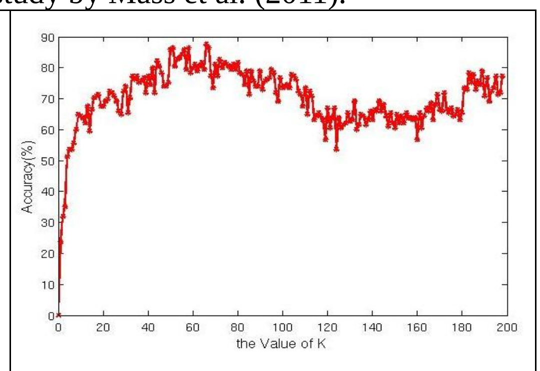


Figure 5. The relationship between the value of k and accuracy

5.2 Experiment 2: Validate Performance of Optimal Multi-level SVM Classifier

In experiment 2, we validate the performance of our optimal SVM using the remaining 186 texts. In this experiment, the model is tested with different dimensions again to see if our previous chosen optimal dimension still holds. For unification and comparison concerns, we also process the test data with five-fold Cross Validation method, and evaluate performance under different dimensions (*k* values) by comparing their recall precision and F measure. Due to the limit of pages, we only present the average recall, precision and F measure in Table 2. The original *k* = 198 is used as our benchmark, as well as the default setting of SVM without parameter optimization. We also calculated other *k* values for comparison that are not presented in the table. The result was consistent with experiment 1. The best performance was achieved when k=66, c=11.3137, γ=0.70711. We conclude that the optimized parameters significantly improved the performance of a default SVM with either default or optimized result

Table 2. Validate Model

	Average Precision	Average Recall	F measure
Global Optimization	**91.43%**	**88.89%**	**0.9014**
Default SVM + Optimal *k*	85.83%	83.90%	0.8485
Optimal SVM + Default *k*	81.86%	81.10%	0.8148

of LSA. Meanwhile, we also observed that optimizing dimension *k* alone also improve the performance over original *k* value from LSA in sentient analysis.

6. Conclusions and Future Directions

Sentiment analysis has many useful business applications. However, dealing with textual data is a challenging task because of high dimension of data. In this study, we propose a global optimization approach for SVM-based automatic sentiment analysis. Our study has the following contributions. *First*, high dimension of textual data is a common problem affecting performance. Selection of the number of dimensions is a challenge problem. Our experiment showed that dimension reduction using LSA does not always improve performance especially when the dimension is randomly chosen. Finding the optimal dimension of keyword features is important. *Second*, feature dimension optimization is not an isolated problem. It can achieve the best performance with optimization of SVM kernel function. To address the high computational cost in this process, we proposed a way to find candidate dimension values. The method is shown feasible and effective. *Third*, most existing sentiment analysis used general SVM package without tap into the kernel parameter optimization. Using RBF kernel as an example, our experiment showed that an optimized kernel can improve the performance over a generic kernel. *Fourth*, compared to existing simple classifier, our method provides a holistic way to optimize classification model. Our global optimization approach has shown to be effective in sentiment analysis in terms of recall, precision and accuracy. In the future, we will apply our sentiment analysis model in other domains such as user review and news article. The method of multi-level polarity sentiment analysis will be applied to identify more levels polarity with higher dimensions. We would like to compare the parameter selection with different test dataset. We also want to study the parameter optimization process using dynamic programming. This will reduce the computational complexity and allow real time training and classification.

7. Acknowledgement

This research is supported by National Natural Science Foundation of China (NSFC:71001059)：
"Research on task-oriented open team innovation collaboration theories and methods"

References

Cao, Q., Duan, W., & Gan, Q. (2011). “Exploring determinants of voting for the ‘helpfulness’ of online user reviews: A text mining approach.” *Decision Support Systems*, 50(2), 511-521.

Kim, S. M. and E. Hovy (2005). "Automatic detection of o bearing words and sentences." *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*: 61-66.

Lane, P. C., Clarke, D., & Hender, P. (2012). “On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data.” *Decision Support Systems*.

Li, N. and D. D. Wu (2010). "Using text mining and sentiment analysis for online forums hotspot detection and forecast." *Decision Support Systems* 48(2): 354-368.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). “Learning word vectors for sentiment analysis.” In *Proceedings of the 49th annual meeting of the association for computational Linguistics* (*ACL-2011*).

Mullen, T. and N. Collier (2004). "Sentiment analysis using support vector machines with diverse information sources." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 412–418.

Pang, B., L. Lee, et al. (2002). "Thumbs up? Sentiment classification using machine learning techniques." *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, Association for Computational Linguistics: 79-86.

Prabowo, R. and M. Thelwall (2009). "Sentiment analysis: A combined approach." *Journal of Informetrics* 3(2): 143-157.

Rui, H. and Whinston, A. (2011). Designing a Social-Broadcasting-Based Business Intelligence System. ACM Transactions on Management Information Systems. 2(4): 1-19.

Tan, S. and J. Zhang (2008). "An empirical study of sentiment analysis for chinese documents." *Expert Systems with Applications* 34(4): 2622-2629.

A Statistical Model for Hierarchical Topic Modeling and User Interest Discovery

Xuning Tang, Christopher C. Yang, Mi Zhang

College of Information Science and Technology

Drexel University, Philadelphia, PA, 19104, USA

Abstract

In light of the rapid development of online social media sites, web users are shifting from data consumers to data producers. It is important to understand the hot topics being discussed on the web in order to catch the pulse of our rapid changing world and make quick response, especially for business decision-making system. This paper presents a statistical model that can detect hierarchical topics from a document corpus. At the same time, our proposed model is able to discover web users' interests to those detected topics and stories, which brings the opportunities to recommend either breaking news or popular products to users who might be interested to. Experiments on the MySpace dataset indicate that our proposed model can capture meaningful topics and stories, outperform baseline approach in terms of perplexity on held-out dataset and improve the prediction result of user's participation by knowing web users' interests on different stories.

1. Introduction and Background

The evolution from Web 1.0 to Web 2.0 is accompanied by the generation of a mass of web data, particularly by web users, which brings great challenges for the existing knowledge management systems to organize and interpret them. Therefore, there is a desire to develop new techniques to automatically extract and understand hot topics and ongoing stories from web contents. Since a decade ago, many research works have been devoted to topic modeling. Hofmann proposed the PLSI model given an assumption that each document contains different topics and each word is generated by one of these topics (Hofmann 1999). Blei et al. proposed the LDA model which used a Dirichlet prior to solve the over-fitting problem and also increase the flexibility of the model (Blei et al. 2003). However, none of these approaches considered the potential relationships between topics. From a statistical point of view, a topic is merely a cluster of words that frequently co-occur using the bags of words approach of topic modeling. Therefore, it can be a set of words describing a facet of a story or an event of a series of continuous events, which decreases the interpretability of the detected topics. In addition, the topics identified by the previous techniques are isolated. However, different storylines or events may share common topics so that the connections between topics should not be ignored. As a result, there is a clear research need to model the hierarchical relationships between topics and identify topics with different granularity, where a **general topic**, namely as **story** in this paper, may consist of **sub-topics**. More recently, Ahmed et al. and Kawamae proposed to employ a latent variable, which can be considered as either a story or a trend, to depict the relationships between topics (Ahmed et al. 2011; Kawamae 2011). In their works, each document belongs to a story/trend and each story/trend consists of a collection of topics. Different story/trend can share common topics. However, both of these two models neglected the information of the contributing users which limits their applications to model user interest.

Inspiring by both Ahmed's and Kawamae's research (Ahmed et al. 2011; Kawamae 2011), in this paper we proposed a new statistical model, aiming to capture topics with hierarchical structure and model users' interests toward these topics. Specifically, our model can extract two layers of topics where the upper layer of topics are considered as stories and the lower layer of topics are considered as components of the stories. For instance, bankruptcy of Lehman Brothers can be a story and this story may contain financial topic, political topic and more. In the experiment section, we demonstrated the capability of our proposed model to discover stories and topics and compared with the state-of-the-art approach.

2. Methodology

2.1 Probabilistic Model

In this work, given a document corpus, our goal is to develop a new statistical model which can simulate the generation process of user interest and content, detect stories and topics in the document corpus, and illustrate the relationships between story and topic. Our model, namely as User Interest and Topic Detection Model (UTD), is designed based on a real generation process of content terms and users. The major difference between our proposed model and Kawamae's model lies in the modeling of user interest. Take online forum as an example, we consider each thread as a document. Content terms of a document are composed of the terms from both the initial post of a thread and all its following comments. Similarly, the initiator of a thread and all commenters are considered as the users of this document. The generation process of each document d in our UTD model is as follows: first of all, UTD chooses a story label c_d for document d. Once c_d is determined, users' interests play an important role to determine whom this new document will attract. UTD model will then sample $|U_d|$ users, each from a multinomial distribution with a Dirichlet prior depending on c_d. At third, UTD generates content terms in a document one by one following a LDA-like process. Concretely, we assume that the generation of each term of d is influenced by one of three factors: 1) a general background topic; 2) a story-background topic; 3) topics belonging to the whole corpus. We adopt a switch variable to control the influence of these three factors on content term generation.

The story label c_d of document d is sampled by a multinomial distribution φ with a Dirichlet prior α. Once c_d is determined, let $|U_d|$ denotes the number of users participated in document d, the user list of d is generated by repeating the sampling process $|U_d|$ times based on the story-user distribution Ω_{c_d}. Each word w_{d_i} of d is drawn from either the general background topic or the story-background topic or one of K general topics. First of all, a switch variable x_{d_i} is drawn from a multinomial distribution for word w_{d_i} to control its generation process. If $x_{d_i} = 0$, w_{d_i} is drawn from the general background topic, ϕ_{bg}. If $x_{d_i} = 1$, w_{d_i} is drawn from the story-background topic, ϕ_{c_d}, which consists of the signature terms of a story. If $x_{d_i} = 2$, a topic z_{d_i} is first sampled from the story-topic distribution θ_{c_d}, and then the word w_{d_i} is drawn conditioning on the sampled topic. Overall, the generation process of users and words in the UTD model can be described as follows:

1. Draw 1+C+Z multinomial distributions ϕ_z from prior γ, one for each topic (1 general background topic, C story-background topics and Z general topics)
2. Draw D multinomial distribution μ_d from prior ϵ, one for each document d;
3. Draw C multinomial distributions Ω_c and θ_c from prior λ and β respectively, one for each story c;

For each document d:

A. Draw an story label c_d from φ

B. Draw $|U_d|$ users from multinomial distribution Ω_{c_d}

For each content term w_{d_i} in document d:

Draw a switch variable x_{d_i} from multinomial μ_d

if $x_{d_i} = 0$: Draw word w_{d_i} from multinomial ϕ_{bg}

else if $x_{d_i} = 1$: Draw word w_{d_i} from multinomial ϕ_{c_d}

else if $x_{d_i} = 2$: i. Draw topic z_{d_i} from multinomial θ_{c_d}

ii. Draw word w_{d_i} from multinomial $\phi_{z_{d_i}}$

Table 1. Notations

C, c_d	Story labels of all documents, the story label of document d	Ω_c	the multinomial distribution of users specific to story c
P, P_u	All users , user u	θ_c	the multinomial distribution of topics specific to story c
Z, z_{d_i}	All topics , the topic of the i^{th} word in document d	μ_d	the multinomial distribution of switch variables specific to document d
W, w_{d_i}	All words , the i^{th} word of document d	$X_{d,x=0,1,2}$	number of word tokens assigned to switch x = 0/1/2
X, x_{d_i}	All switch variables , the switch variable of the i^{th} word in document d	$M_{z,w}$	number of times that word w is assigned to topic z
$\alpha, \gamma, \beta, \varepsilon, \lambda$	parameters of symmetric Dirichlet priors	K_c	number of times that a document is assigned to story c
φ	the multinomial distribution of story	$N_{c,z}$	number of times that topic z is assigned to story c
ϕ_z	the multinomial distribution of words specific to topic z	$L_{c,u}$	number of times that user u is assigned to story c

2.2 Inference and Learning

In our model, nodes P_u and w_{d_i} are observed variables; $\alpha, \gamma, \beta, \varepsilon$ and λ are predefined Dirichlet priors; $\varphi, \phi, \Omega, \theta$ and μ can be eventually integrated out and c_d, z_{d_i} and x_{d_i} are the only hidden variables need to estimate based on the observations. Gibbs sampling is an effective way to estimate these hidden variables (Ahmed et al. 2011; Kawamae 2011; Griffiths et al. 2004). In Gibbs sampling, Markov chains are constructed to approximate the conditional probabilities of the hidden variables. The key of a Gibbs sampling algorithm for our UTD model is to approximate the conditional probabilities of story c_d , topic z_{d_i} and switch variable x_{d_i} . Therefore, the sampling scheme consists of two steps. The first step is to sample the story c_d for document d, $\Pr(c_d = c| \ldots)$. The second step is to sample the switch variables and the topics for each individual word of a document: $\Pr(x_{di} = 0| \ldots)$, $\Pr(x_{di} = 1| \ldots)$ and $\Pr(z_{di} = k, x_{di} = 2| \ldots)$.

Story Label Sampling

In the sampling schema, for each document, we use the chain rule to derive the conditional distribution of observing document d belongs to story c, $\Pr(c_d = c|\ldots)$, as: $\Pr(c_d = c|\ldots) = \frac{K_c+\alpha_c-1}{\sum_c^C(K_c+\alpha_c)-1} \cdot \left[\frac{\prod_u^U \Gamma(L_{c,u}+\lambda_u)}{\prod_u^U \Gamma(L_{c,u\backslash d}+\lambda_u)} \cdot \frac{\Gamma(\sum_u^U(L_{c,u\backslash d}+\lambda_u))}{\Gamma(\sum_u^U(L_{c,u}+\lambda_u))}\right] \cdot \left[\frac{\prod_z^Z \Gamma(N_{c,z}+\beta_z)}{\prod_z^Z \Gamma(N_{c,z\backslash d}+\beta_z)} \cdot \frac{\Gamma(\sum_z^Z(N_{c,z\backslash d}+\beta_z))}{\Gamma(\sum_z^Z(N_{c,z}+\beta_z))}\right]$

where Γ means a gamma function, $L_{c,u\backslash d}$ represents the number of times that user u has been assigned to story c, except for document d. $N_{c,z\backslash d}$ represents the number of times that topic z is assigned to story c, except for document d.

Switch Variable and Topic Sampling

For each word token, the posterior probability of adding word w_{d_i} in document d to background topic is:

$$\Pr(x_{-di} = 0|\ldots) \propto \frac{X_{d,0}+\varepsilon_0-1}{\sum_x^X(X_{d,x}+\varepsilon_x)-1} \cdot \frac{M_{z=bg,w}+\gamma_w-1}{\sum_w^W(M_{z,w}+\gamma_w)-1}$$

Similarly, the posterior probability of adding word w_{d_i} in document d to the story-background topic c is:

$$\Pr(x_{-di} = 1|\ldots) \propto \frac{X_{d,1}+\varepsilon_1-1}{\sum_x^X(X_{d,x}+\varepsilon_x)-1} \cdot \frac{M_{z=c,w}+\gamma_w-1}{\sum_w^W(M_{z,w}+\gamma_w)-1}$$

Similarly, the posterior probability of adding word w_{d_i} in document d to topic k is defined as:

$$\Pr(z_{-di} = k, x_{-di} = 2|\ldots) \propto \frac{X_{d,2}+\varepsilon_2-1}{\sum_x^X(X_{d,x}+\varepsilon_x)-1} \cdot \frac{N_{c,k}+\beta_k-1}{\sum_z^Z(N_{c,z}+\beta_z)-1} \cdot \frac{M_{z=k,w}+\gamma_w-1}{\sum_w^W(M_{z,w}+\gamma_w)-1}$$

2.3 Parameter Estimation

Once the sampling processes converged based on the posterior distributions calculated in the previous section, we estimate the five parameters by:

$$\theta_{c,k} = \frac{N_{c,k}+\beta_k}{\sum_z^Z(N_{c,z}+\beta_z)}, \ \phi_{z,w} = \frac{M_{z,w}+\gamma_w}{\sum_w^W(M_{z,w}+\gamma_w)}, \ \varphi_c = \frac{K_c+\alpha_c}{\sum_c^C(K_c+\alpha_c)}, \ \mu_{d,x} = \frac{X_{d,x}+\varepsilon_x}{\sum_x^X(X_{d,x}+\varepsilon_x)}, \ \Omega_{c,u} = \frac{L_{c,u}+\lambda_u}{\sum_u^U(L_{c,u}+\lambda_u)}$$

3. Experiment

In this section, we tested our proposed model in a public dataset which was made available in workshop CAW 2.0 (http://caw2.barcelonamedia.org/). The original dataset consists of threads from three different sub-forums: campus life, news & politics and movies. In our experiment, we only used the threads of news & politics sub-forum and movies sub-forum because campus life sub-forum is about students' daily life and is very noisy. We preprocessed and transformed the dataset into the input format of our model. Specifically, each thread corresponds to a document in our experiment. All words from a thread's initial post and its following comments were extracted to be the content of a document. Similarly, the thread initiator and all commenters were extracted to be the users of this document. Accordingly, we obtained a set of 3,886 documents and 1,373 users. We randomly split the whole dataset into a training set with 90% of the documents and a test set with 10% of the documents.

3.1 Qualitative Analysis

In the first experiment, we applied our proposed model to the training dataset and investigated the discovered stories and topics. A case study was conducted and the result is showed in Fig. 1.

In Fig. 1, each rectangle represents a topic identified by our model. Each topic consists of a list of keywords and their probabilities of belonging to this topic. Each oval denotes a story and a link from a story to a topic indicates that this topic is an important element of the story. We manually labeled the oval based on its important topics and keywords. Blue rectangle means that a topic is shared by different stories.

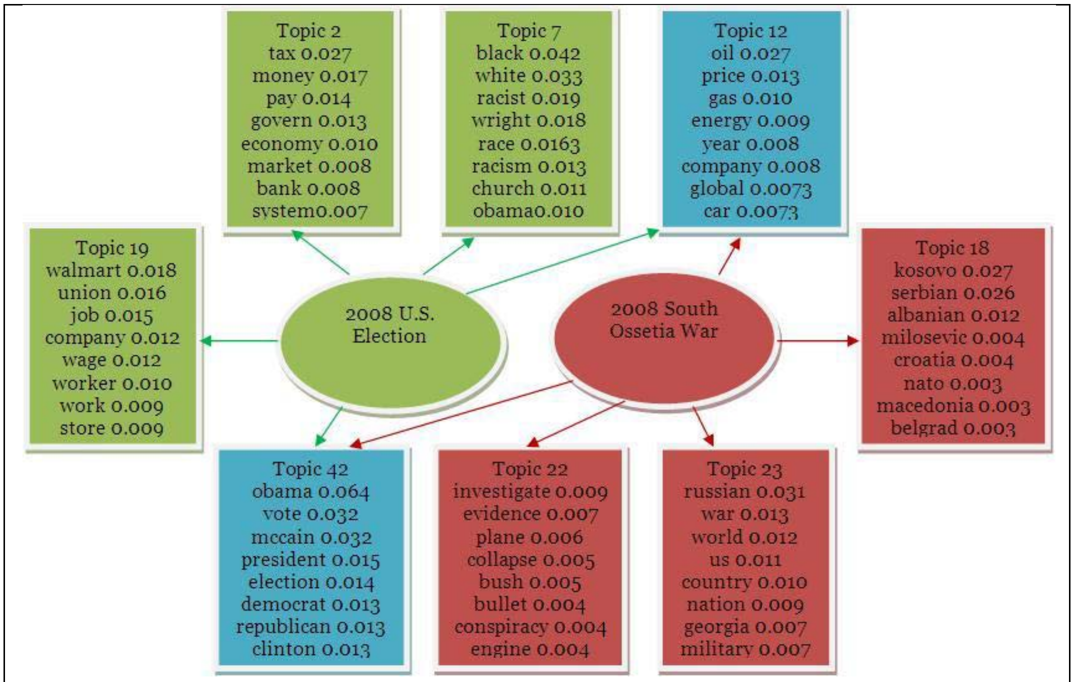


Figure 1. A Case Study

The two stories that we picked and showed in Fig. 1 are about 2008 U.S. Election and 2008 South Ossetia War respectively. Threads belonging to U.S. Election exhibited topics including "taxation"(topic 2), "racism"(topic 7), "energy problem"(topic 12), "employment"(topic 19) and "election"(topic 42). On the other hand, threads belonging to South Ossetia War involved topics such as "Kosovo War" (topic 18), "Russia-Georgia War"(topic 23), "plane crash"(topic 22), "energy problem"(topic 12) and "election" (topic 42). Among them, topic 12 and 42 are shared by both stories. Since the South Ossetia War happened around the same time of 2008 U.S. Election, at that moment, presidential candidates' attitudes toward this war was critical to show their capability of dealing with foreign affair, which explains why these two stories shared topic 42. In addition, energy problem or energy security is always highly correlated to a war, which might explain why topic 12 was also shared by these two stories.

3.2 Quantitative Analysis

In this section, we conducted a quantitative analysis to study the performance of our proposed model. We compared the performance of our model to the TAM model proposed recently by Kawamae (Kawamae 2011).

We used test-set perplexity as the criterion for model evaluation. The value of perplexity reflects the ability of a model to generalize to unseen data. The perplexity is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity indicates better generalization performance. We follow Kawamae (Kawamae 2011) and define the perplexity score for a model in test set D_{test} as $PPX(D_{test}) = \exp\left(-\frac{1}{W}\sum_{d=1}^{|D_{test}|}\sum_{w\in d}^{|d|}\log\left(\Pr\left(w_d\right)\right)\right)$, where

$$\Pr(w_d) = x_{d,0}\phi_{bg,w_d} + x_{d,1}\phi_{c_d,w_d} + \sum_z^Z x_{d,2}\theta_{c_d,z}\phi_{z,w_d}$$

probabilities ϕ_{bg,w_d}, ϕ_{z,w_d}, ϕ_{c_d,w_d} and $\theta_{c_d,z}$ are learned by Gibbs sampling process in the training dataset. $x_{d,0}$, $x_{d,1}$ and $x_{d,2}$, are estimated by a Gibbs sampling process on the test set.

3.3 Result

Due to limited space, we arbitrarily selected the number of stories to be 20 for both UTD and TAM (the concept "trend" in TAM is equivalent to the concept "story" in UTD). The number of topics was set to 50 and 100 respectively to run two tests. As shown in Fig. 3, UTD model performed slightly better than TAM under these experiment settings, which means that by incorporating user information, our model has better generalization performance on the test set.

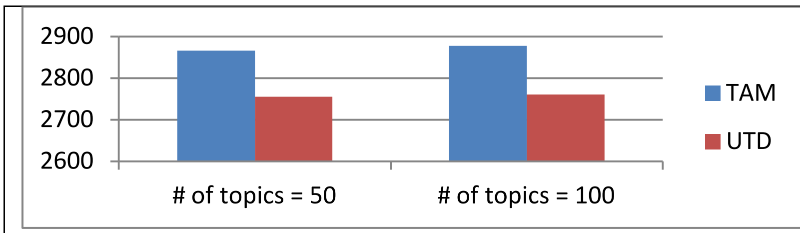


Figure 3. Perplexity Scores of TAM and UTD on Test Set

4. Conclusion

In this paper, we proposed a novel statistical model to detect topic and story in a document corpus, and model user interest. Experiment on a public dataset demonstrated that our model can extract clear and meaningful stories and topics, and also generalize to unseen documents with better performance than TAM.

References

Ahmed, A., Ho, O., Eisenstein, J., Xing, E., Smola, A.J., and Teo, C.H. 2011. "Unified Analysis of Streaming News," in *Proceedings of the 20th International Conference on World Wide Web*, S. Sadagopan, K. Ramamritham, A. Kumar, M.P. Ravindra, E. Bertino, and R. Kumar (eds.), Hyderabad, India, pp. 267-276.

Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3: 4-5),pp. 993-1022.

Griffiths, T.L. and Steyvers, M. 2004. "Finding Scientific Topics," *PNAS* (101:1),pp. 5228-5235.

Hofmann, T. 1999. "Probabilistic Latent Semantic Indexing," in *Proceedings of 22nd Annual International ACM Conference on Research and Development in Information Retrieval*, F. Gey, M. Hearst, and R. Tong (eds.), Berkeley, CA, pp. 50-57.

Kawamae, N. 2011. "Trend Analysis Model: Trend Consists of Temporal Words, Topics and Timestamps," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, I. King, W. Nejdl, and H. Li (eds.), Kowloon, Hong Kong, pp. 317-326.

Network Influences Comparison in Large Social Network – An Empirical Investigation

Bin Zhang†, Ramayya Krishnan‡, David Krackhardt‡
†: Fox School of Business, Temple University ‡: Heinz College, Carnegie Mellon University

Abstract

Cohesion and structural equivalence are two competing social network models to explain diffusion of innovation. While considerable work has been done on these models, the question of which network model explains diffusion has not been resolved. This paper examines adoption of Caller Ring Back Tones (CRBT) in a cellular telephone conversation network. Since these societal scale networks are very large, the study of diffusion in these settings require the development of methods to extract connected subpopulations from the network. We use a novel technique to detect densely connected and self-contained components of the network. Using a new network auto-probit model, we study the competing influences of cohesion and structural equivalence on each of the subpopulation detected using our technique. The results show CRBT adoption is affected by both cohesion and structural equivalence. Cohesion have consistent pattern across different sizes of subpopulation, while structural equivalence changes with subpopulation size.

1 Introduction

The debate about which network influence, cohesion or structural equivalence, plays a more influential role in technology diffusion is still inconclusive. When forming opinions or making decisions, people usually use someone they know or someone in their social network as their frame of reference, taking their opinions into account. This progress, in which an actor adapts his behaviors to those of alters in his social network, is known as contagion or social influence (Duncan et al., 1968; Leenders, 1997). There are two social network models, cohesion and structural equivalence, to explain contagion. Cohesion is made through communication, which is direct contact between actor and alter; while structural equivalence is created through comparison, which occurs when an actor competes with other alters who he considers in a similar social position to him in the network. Both models have been used to explain the progress of contagion. Coleman et al. (1966) studied diffusion of medical innovation and found medical doctors adopted new technology at the early stage because of cohesion. Burt (1987) reanalyzed Coleman et al's data and concluded that contagion does not happen through cohesion but rather through structural equivalence. Since then, both camps, cohesion (Rogers and Kincaid, 1981; Harkola and Greve, 1995) and structural equivalence (Strang and Tuma, 1993; Van den Bulte and Lilien, 2001), have found quantitative evidence to support their claims. However, most of these analyses only used network autocorrelation models that only include one network effect at a time, and compare the effect sizes of these two models in two independent models. The assumption made by such method is that only one network term would be significant. Since then, some quantitative methods for social networks have been developed to work in the situation where both terms are significant and direct comparison of these two influence sizes can be made. For example, Doreian's (1989) two regimes of network effect autocorrelation model. However, it does not support dichotomous response variable. Actually, quite often in social sciences, the behaviors need to be analyzed are binary, for example, whether adopt a new technology. But no model is available if we want to compare multiple network terms at the same time and the response variable is dichotomous. Furthermore, the only data being used for the comparison is Coleman et al.'s Medical Innovation data, which only has a node size of 125. Whether network influence shows similar mechanism in large social networks has not been studied yet.

In this paper, we apply a newly designed statistical model, multi-network auto-probit (mNAP) Zhang et al. (Forthcoming), to investigate the effects of both direct influence (cohesion) and competition (structural equivalence) on Caller Ring Back Tone (CRBT) adoption, which is represented by binary variable, within cellular phone communication social networks. CRBT is becoming one of the most attractive mobile content with a projected revenue of $4.7 Billion in United States by the end of 2012[1]. The penetration rate of CRBT is even larger in Asia and Africa - 95% of the market for digital music in Indonesia comes from CRBT[2]. CRBT replaces plain ring-back tones with music for a caller to hear as he/she waits for the receiver to answer. Cellular phone is the major medium of communications nowadays. Unlike other instances of large social networks, which are often extracted from online networking sites. The interaction between two individuals entails a stronger notion of intent-to-communicate, thus celluar phone call network is a better approximation of individual's real social network. It is important to understand how the dynamics of adoption are likely to unfold within the underlying phone call social network: whether individual adopts a new technology because of direct contact with adopters, or because of competition with them. However, how new products spread within such networks has not been well studied.

2 Theory and Hypothesis

In the cohesion model, a focal person's adoption is influenced by his neighbor who he directly connects to. found that medical doctors prescribe a new drug because of directed ties with other doctors. The greater a doctor's connection to his colleagues, the earlier he prescribes that drug for the first time. Coleman et al. (1966) and Rogers and Kincaid (1981) both found cohesion had effect on innovation diffusion. In the first hypothesis below we want to test whether the average probability of CRBT adoption by people whom a focal person calls influences the probability of adoption by that person.

> **Hypothesis 1: Cohesion and CRBT Diffusion.** *The probability of CRBT purchased by a focal person is positively related to the ratio of CRBT adopters among the focal person's neighbors.*

Structural equivalence is a positional model (Burkhardt, 1994). An actor is structurally equivalent to an alter if they connect to the same others. Structural equivalence model describes the competition between actor and alter. Individuals in the same position in a social network will "use each other as a frame of reference for subjective judgments and so make similar judgments even if they have no direct communication with each other" (Harkola and Greve, 1995).

The hypothesis about the structural equivalence effect on the diffusion is described as below:

[1] Broadcast Music Inc, 2011

[2] Indonesia Finance Today, 03/15/2011

Hypothesis 2: structural equivalence and CRBTs Diffusion. *The probability of CRBT purchased by a focal person is positively related to the extent to which role equivalent alters have purchased CRBTs.*

Comparing Cohesion and structural equivalence is important because they represent two different target scheme. Suppose we want to trigger more adoptions in the network. If the adoption is influenced by cohesion, then we need to influence individuals with many ties, while if adoption is mainly influenced by structural equivalence, we need to influence individuals with many alters in the same social position.

3 Model and Analysis

A statistical model that accommodates multiple regimes of network effects for the same group of actors, multiple network auto-probit model (m-NAP), is adopted for this study. It provides a solution for researchers who wish to compare multiple network effects on actors' behaviors and attitudes in social network contexts. The model is shown below:

$$\mathbf{y} = \mathbb{I}(\mathbf{z} > 0)$$

$$\mathbf{z} = \mathbf{X}\beta + \theta + \varepsilon,\ \varepsilon \sim \text{Normal}_n(0, I_n)$$

$$\theta = \rho_1 \mathbf{W}_1 \theta + \rho_2 \mathbf{W}_2 \theta + \mathbf{u},\ \mathbf{u} \sim \text{Normal}_n(0, \sigma^2 I_n)$$

$\mathbf{y}$ is the vector of observed binary choices, whether adopt CRBT or not. IT is an indicator function of the latent preference of consumers $\mathbf{z}$. If $\mathbf{z}$ is larger than a threshold 0, consumers would adopt; if $\mathbf{z}$ is smaller than 0, then consumers would not adopt. $\mathbf{z}$ could be represented as a function of both exogenous covariates $\mathbf{X}$ and autocorrelation term θ. $\mathbf{X}$ is an matrix includes all observed attributes of consumers in the network, such as age and gender. θ is the autocorrelation term and can be described as the aggregation of two network structures $\mathbf{W}_i$ and coefficient ρ_i where $i = 1, 2$. $\mathbf{W}_1$ represents structures of cohesion, and $\mathbf{W}_2$ represents structural equivalence. Detailed description of mNAP is in Zhang et al. (Forthcoming).

Our dataset is from one of the largest cellular phone services in an Asian country. There are over one million subscribers and over one billion phone call records in our data after preprocessing. Given the size of the data, it is impossible to analyze the whole data size using the current computing power available. The heterogeneity across different groups of subscriber also increase with the population size. Hence in order to investigate CRBT diffusion, we use subpopulations to decrease heterogeneity and make analysis tractable. The subpopulations extracted using T-CLAP (Zhang et al., 2011) are relatively isolated and densely connected subnetworks so they can show stronger network influences on CRBT diffusion.

Before addressing the question of network influence over the diffusion of CRBT technology, we face another challenging problem: finding an ideal subset of the network. Our network is large, a network with millions of users. With this wealth of interconnectedness and access to data, have come two major challenges. The first and more important is with statistics. Heterogeneity across subsets of the population increases with the size of the population. Assuming that we have enough computing power, analysis of the whole population would confound across these distinct effects existing in different subsets. The second challenge is computational cost. Most of the social network analysis packages do not scale well. That is, for some class of questions and analytic routines, our standardized desktop systems are not able to finish analyzing large networks in a realistic time frame. Memory-wise systems do not have enough resources to accommodate a structure of a large network. One solution to this problem is using subpopulations of a smaller size. Since there is no extraction algorithm existing to provide an ideal balance of extracted subpopulation quality and running speed, we developed T-CLAP algorithm to identify dense and relatively independent subpopulations.

The information included in our dataset is cellular phone call records and CRBT purchase records over a three-month period, and phone account holders' demographic information such as age and gender. We use symmetric phone call as the connections of social because it implies equal and stable connections (Hanneman and Riddle, 2005). One could argue that the size of subpopulation is likely to have an impact on network effects. So we extract subpopulations at different sizes from the population. If network effects are universal in the global network, when we extract subpopulation from the global network, these network effect should be consistent. Four subpopulations in total were extracted by using T-CLAP. The first two were at the size of about 200, The other two are at the size of about 600.

The results of network autocorrelation model for all the subpopulations are provided at below. Table 1 shows the results for subpopulations ranging from 202 to 597. We find that cohesion effect is consistent across all subpopulations. The result shows that cohesion effects from all subpopulations are significant at 0.05 level, the size ranges from 0.06 to 0.075. Such result confirms our Hypothesis 1. It shows that callers receive strong influence through direct connections to alters in the same group who have already adopted. The explanation could be: if a caller calls more CRBT subscribers, he gets exposure to more ring-back tones, and is more likely to hear ring tones interest him, thus he is more likely to buy ring tones too.

We also observed significant effect of structural equivalence. The results show that the adoption of CRBT is impacted by both cohesion and structural equivalence. For the structural equivalence model, caller evaluate alters who are in the same position of a phone call network as him. Same position in a social network means people are in the same kinds of relations, with the same kinds of people. Interestingly the effect size of structural equivalence varies with the size of subpopulation. When the subpopulations are at the size of about 200, the effect of structural equivalence is significant and negative, meaning individuals with more adopters having the same social position are less likely to adopt. The result seems absurd at the first glance, but actually is reasonable. One explanation is that in a smaller group, people weight individuality more than being fashion. When the subpopulations are at the size of about 500, the structural equivalence effect is significant at positive value. One explanation is that in a cellular phone call social network, parties who call each other are likely to be friends or belong to same group under some relationship. In this case the enthusiasm of showing others about his adoption of frontier fashion and individuality is higher. The satisfaction of letting friends appreciate his fashion taste or simply an interesting tone is also higher. Motivated by this thought, a perceived competition is created among these subscribers. Actor will know about an alter he does not necessary call to has adopted ring-back tone through common friends they both call. The more ring tones those alters bought, the more likely the actor will adopt CRBT.

In general cohesion is more statistically significant than structural equivalence, which suggests cohesion effect is more influential on CRBT adoption than structural equivalence. Cohesion has a significant effect on a caller's decision on adoption of CRBT – adoption is influence by adopters among neighbors. Structural equivalence also has a significant effect on a caller's decision on adoption of CRBT. When the group size is large, people tend to imitate competitors. When the group size is small, people tend to differentiate themselves from others. Based on the results, we should use different trigger schemes for groups at different sizes. If we want to trigger more adoptions For smaller groups, we only need to concentrate on individuals with many connections. For larger groups, we should not only concentrate on users who are popular, but also these who are in the similar positions in the group. Finally gender and age do not seem to have any relationship with CRBT adoption.

Table 1: Results of cohesion vs structural equivalence

Parameter	Subpop. 1	Subpop. 2	Subpop. 3	Subpop. 4
n	202	213	563	597
Gender	-0.20	-0.034	0.094	0.90
	(3.3)	(0.33)	(0.46)	(0.87)
Age	-0.026	0.053^{**}	0.065	-0.0080
	(0.017)	(0.014)	(0.040)	(0.017)
Degree	0.040^{**}	0.030^{**}	0.043^{**}	0.024^{**}
	(0.0098)	(0.0095)	(0.044)	(0.0070)
Cohesion (ρ_1)	0.06^{**}	0.065^{*}	0.075^{*}	0.065^{**}
	(0.022)	(0.026)	(0.062)	(0.040)
Role	-0.0098^{*}	-0.01^{*}	0.034^{*}	0.0049^{*}
equivalence (ρ_2)	(0.0041)	(0.0052)	(0.015)	(0.0024)

**: $p < 0.01$, *: $p < 0.05$

4 Conclusion

The debate among researchers about two classes of network models, cohesion and structural equivalence, and the impact on diffusion in social networks still persists. However, other than Coleman's classical *Medical Innovation* data, few new data sets have been used to address this research question. Reconciling these findings is very important because the social network is a key medium of diffusion, and figuring out which network effect drives social influence can help us understand the mechanism of diffusion. In our study, we attempt to readdress this important but unresolved question. One large challenge from this research is to make our analysis tractable. This challenge comes from the size of our data – both the number of actors and connections are in millions. we used an innovative algorithm T-CLAP to extract subpopulations from large scale network. Given the fact that more and more social networks data are large scale, this method provides a solution for fast and efficiently extract subpopulations with high quality. Using the subpopulations extracted, we analyze the effects of cohesion and structural equivalence's on adoption of CRBT by using mNAP model. My study is one of the very few to investigate multiple network effects on diffusion. Our results show that the adoption of CRBT is consistently correlated with cohesion. When the size of subpopulation is small (at the level of 200), CRBT adoption is negatively correlated with structural equivalence; when the size of subpopulation is large, at the level of 500, adoption is positively correlated with structural equivalence. Between the two network effects, cohesion has a more significant impact. So the strength of communication or connection still has stronger influence on adoption.

References

Burkhardt, M. E. (1994). Social interaction effects following a technological change: A longitudinal investigation. *Academy of Management Journal*, 37(4):869–898.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92(6):1287.

Coleman, J. S., Katz, E., and Menzel, H. (1966). *Medical innovation: A diffusion study*. Bobbs-Merrill Co.

Doreian's, P. (1989). *Two Regimes of Network Effects Autocorrelation*. The Small World. Ablex Publishing.

Duncan, O. D., Haller, A. O., and Portes, A. (1968). Peer influences on aspirations: A reinterpretation. *The American Journal of Sociology*, 74(2):119–137.

Hanneman, R. and Riddle, M. (2005). *Introduction to social network methods*. Online.

Harkola, J. and Greve, A. (1995). Diffusion of technology: Cohesion or structural equivalence? In *Academy of Management Best Papers Proceedings*, pages 422–426, Vancouver, Canada.

Leenders, R. T. (1997). *Longitudinal behavior of network structure and actor atributes: modeling interdependence of contagion and selection.*, chapter Evolution of Social Networks. Gordon and Breach, New York.

Rogers, E. M. and Kincaid, L. D. (1981). *Communication Networks: Toward a New Paradigm for Research*. Free Press.

Strang, D. and Tuma, N. B. (1993). Spatial and temporal heterogeneity in diffusion. *The American Journal of Sociology*, 99(3):614–639.

Van den Bulte, C. and Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 106(5):1409–1435.

Zhang, B., Krackhardt, D., Krishnan, R., and Doreian, P. (2011). An effective and efficient subpopulation extraction method in large social networks. *Proceedings of International Conference on Information Systems*.

Zhang, B., Thomas, A. C., Doreian, P., Krackhardt, D., and Krishnan, R. (Forthcoming). Contrasting multiple social network autocorrelations for binary outcomes, with applications to technology adoption. *ACM Transactions on Management Information Systems*.

Quality of Service Tiering: Implications for Content Innovation and Broadband Coverage

Hong Guo (hguo@nd.edu), Robert Easley (reasley@nd.edu)
University of Notre Dame

Abstract

The proposed introduction of quality of service tiers, which would allow network providers to charge content providers for preferential handling of their content, has been one of the most controversial topics to be considered in recent years. This paper addresses research questions paralleling the key concerns facing policy makers – innovation of digital content and broadband market coverage. The main contribution of this research is the development and solution of a two-sided market model that simultaneously considers: both the content providers' and consumers' decisions concerning market entry and participation, consumer demand for content dependent on congestion, heterogeneous consumes in terms of content valuation, and heterogeneous CPs in terms of sensitivity to delay. Our most interesting result concerns content innovation, where we illustrate that what the network provider perceives to be a free rider problem is actually a critical feature of net neutrality that under certain market conditions enables superior content innovation. Findings of this paper provide new policy insights to the net neutrality debate.

Introduction

The proposed introduction of Quality of Service (QoS) tiers, which would allow Internet Service Providers (ISPs) to charge Content Providers (CPs) for preferential handling of their content, has been one of the most controversial topics to be considered in recent years. This proposal has been characterized by opponents as a fundamental attack on the Network Neutrality (NN) principles that they see as critical to fostering the continued innovation of digital content. In December 2010, the Federal Communications Commission (FCC) issued new rules governing network providers (NPs), which require transparency of network management practices and prohibit blocking legal content and offering QoS tiers to CPs (FCC 2010).

Research questions in this paper parallel the key concerns facing policy makers in the NN debate. The main argument for strong NN regulation is that QoS tiering (QT) would undermine innovation of digital content, which has become an important driver of economic growth. We address this concern by comparing the effects of QT and NN regimes on content innovation. Another concern of policy makers is that the US is lagging in broadband penetration, and NPs are quick to point out that QT could readily provide funding for network expansion. We address this concern by examining the relative effects of QT and NN regimes on broadband coverage.

Relative to the vast media coverage and policy discussions about NN, formal economic analysis of NN is limited (Hermalin and Katz 2007, Choi and Kim 2010, Cheng et al. 2011, Economides and Hermalin 2012, Economides and Tåg 2012, Krämer and Wiewiorra 2012). This paper's unique contribution to the literature is to allow both CP entry and variable broadband market coverage in the context of managing congestion-sensitive content, leading to new insights.

Model

We model the internet data transmission market as a two-sided market, with the NP serving both CPs and consumers.

We consider a monopoly NP with a fixed capacity μ. We model the NP's network management options based on the concept of effective bandwidth, which is a widely adopted measure for resource usage and can be used as the basis for usage charges (Kelly 1996, Courcoubetis et al. 2000). Under NN, the NP charges a fixed fee F to consumers for internet access and does not charge CPs. The effective bandwidth α is the same for all traffic. Under QT, there are two service levels – premium (H) with α_{H} and standard (L) with α_{L}, where $\alpha_{\mathrm{H}} \geq \alpha_{\mathrm{L}}$. The NP would charge a usage-based fee p to CPs for premium service but none for standard service.

The CPs are uniformly distributed on $\left[0, S\right]$ based on their minimum bandwidth requirement s, which measures congestion sensitivity. Congestion only occurs when allocated bandwidth does not match the minimum requirement, leading to reduced consumer demand. CPs generate revenue from advertisements with revenue rate r and incur a fixed entry cost E. Under NN the CP's profit is $\varphi(s) = r\alpha d(s,\alpha) - E$, where $d(s,\alpha)$ is the consumer demand for CP s. A CP will enter the market if and only if $\varphi(s) \geq 0$. Under QT the CP's profit is $\varphi_{\mathrm{H}}(s) = (r-p)\alpha_{\mathrm{H}} d(s,\alpha_{\mathrm{H}}) - E$ with premium service, or $\varphi_{\mathrm{L}}(s) = r\alpha_{\mathrm{L}} d(s,\alpha_{\mathrm{L}}) - E$ with standard service. Let $\bar{s}$ denote the most congestion-sensitive CP indifferent to entering the market or not. Then all CPs located in $\left[0, \bar{s}\right]$ will enter the market. $\bar{s}$ measures the amount of available content on the internet, which serves as a proxy for content innovation.

We consider a unit mass of heterogeneous consumers uniformly distributed on $\left[0,1\right]$ in terms of their content valuation v. Consumer v gets a gross utility $u(v,s,\alpha) = \begin{cases} v, & \text{if } s \leq \alpha \\ v - b(s-\alpha), & \text{if } s > \alpha \end{cases}$ from content s, where b measures consumers' congestion sensitivity. Consumer v will subscribe to s if and only if $u(v,s,\alpha) \geq 0$. The net utility of internet access for v is $U(v,\alpha,F) = \int_0^{\bar{s}} u(v,s,\alpha) I\left\{u(v,s,\alpha) \geq 0\right\} ds - F$, where $I\left\{u(v,s,\alpha) \geq 0\right\}$ is an indicator function. Consumer v will subscribe to internet access if and only if $U(v,\alpha,F) \geq 0$. Let $\underline{v}$ be the consumer indifferent between subscribing to internet access or not. Then $\underline{v}$ solves $U(v,\alpha,F) = 0$, and the resulting broadband market coverage is $1 - \underline{v}$. Let $\hat{v}(s,\alpha)$ be the marginal consumer who is either $\underline{v}$ or indifferent between subscribing to content s or not. Then $\hat{v}(s,\alpha) = \begin{cases} \underline{v}, & \text{if } 0 \leq s \leq \alpha + \underline{v}/b \\ b(s-\alpha), & \text{if } \alpha + \underline{v}/b < s \leq \bar{s} \end{cases}$, and all consumers with $v \geq \hat{v}(s,\alpha)$ will subscribe to content s. Therefore consumer demand for content s is $d(s,\alpha) = 1 - \hat{v}(s,\alpha)$. Let $\hat{s}(v,\alpha)$ be the marginal content, which is also the amount of subscribed content, for consumer v. Then $\hat{s}(v,\alpha)$ solves $u(v,s,\alpha) = 0$, i.e., $\hat{s}(v,\alpha) = \alpha + v/b$.

Net Neutrality (NN)

Under NN, the fixed internet access fee from the consumers is the sole revenue source for the NP. As shown in Figure 1, the size of consumer market is $1 - \underline{v}_{\mathrm{NN}}$ and consumer v only subscribes to content located on $\lfloor 0, \hat{s}(v,\alpha) \rfloor$. As a result, the shaded area in Figure 1 corresponds to the total demand for content from all consumers with internet access. The NP's profit maximization problem under NN can be formulated as: $\max_{F_{NN},\alpha} \; \pi_{NN} = F_{NN}\left(1-\underline{v}_{NN}\right)$

s.t. $\alpha\int_0^{\bar{s}_{NN}} d(s,\alpha)\,ds \le \mu$ (NP's Capacity Constraint)

$U\left(\underline{v}_{NN}, \alpha, F_{NN}\right) = 0$ (Participation Constraint for Consumer $\underline{v}_{NN}$)

$\varphi\left(\bar{s}_{NN}\right) = 0$ (Participation Constraint for CP $\bar{s}_{NN}$)

$\bar{s}_{NN} \ge \alpha + \underline{v}_{NN}/b$ (Most Congestion-Sensitive Available Content Constraint)

$F_{NN} \ge 0$, $\alpha \ge 0$, $0 \le \underline{v}_{NN} \le 1$, $\bar{s}_{NN} \ge 0$ (Non-negativity Constraint)

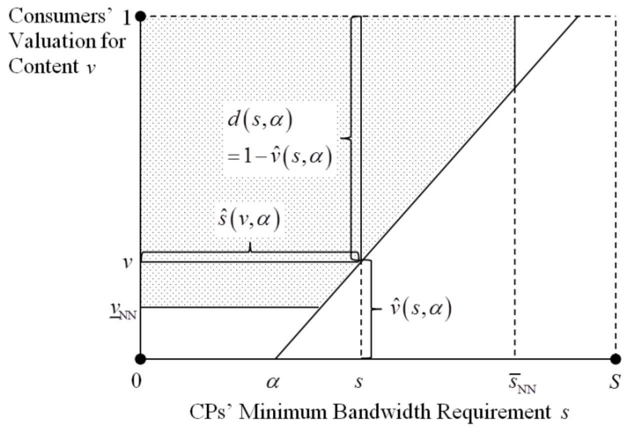


Figure 1: Net Neutrality

The NP maximizes its profit subject to the NP's capacity constraint and the participation constraints for consumers and CPs. The most congestion-sensitive available content constraint ensures that CP $\bar{s}_{NN}$ is the most congestion-sensitive content available on the internet, especially when there are multiple CPs indifferent to entering the market or not. The resulting NN equilibria are of the two types shown in Figure 2: the rectangular "no free-riders" type (bordered by $\underline{v}_{NN}$ below and $\bar{s}_{NN}$ to the right) and the "with free-riders" type (bordered by $\underline{v}_{NN}^{'}$ and $\bar{s}_{NN}^{'}$).

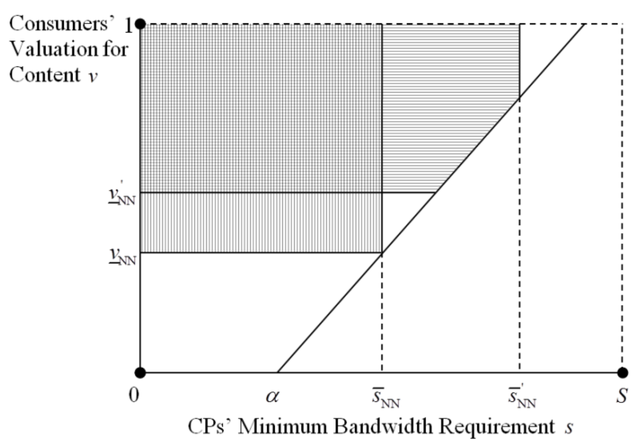


Figure 2: Equilibrium of Net Neutrality

Under NN, the NP can only control the demand side of the market through setting internet access fee F_{NN} and allocating effective bandwidth α. When the revenue rate for CPs, r, is relatively low, the NP selects F_{NN} and α such that only CPs that are of interest to the marginal consumer enter the market. This is the "no free-rider" equilibrium, where all consumers subscribe to the same content and thus the NP is able to indirectly extract surplus from all available content via F_{NN}. When r is sufficiently high, there will be CPs that will enter the market even though their content is not of interest to all the consumers in $\lfloor \underline{v}_{\mathrm{NN}}^{'}, 1 \rfloor$. We refer to these CPs as free-riders because they make use of the NP's bandwidth but do not contribute even indirectly to the NP's revenue, since they are not of interest to the marginal consumer at $\underline{v}_{\mathrm{NN}}^{'}$ and thus cannot contribute to F_{NN}. In these conditions the NP would like to block the free-riding CPs if possible, which is a strong argument for the need for the FCC's existing "no-blocking" rule.

QoS Tiering (QT)

Under QT, there are two QoS levels: premium (H) and standard (L). CPs have the choice to either pay a fee p to get the premium service or stay with the standard service for free. We use s_0 to denote the CP who is indifferent between H and L, i.e., $r\alpha_{\mathrm{L}} d(s_0, \alpha_{\mathrm{L}}) - E = (r-p)\alpha_{\mathrm{H}} d(s_0, \alpha_{\mathrm{H}}) - E$. We use $\bar{s}_{\mathrm{QT}}$ to denote the most congestion-sensitive CP who is indifferent between entering the market or not, i.e., $\max\{r\alpha_{\mathrm{L}} d(\bar{s}_{\mathrm{QT}}, \alpha_{\mathrm{L}}) - E, (r-p)\alpha_{\mathrm{H}} d(\bar{s}_{\mathrm{QT}}, \alpha_{\mathrm{H}}) - E\} = 0$. There are three segments of CPs based on their service choices. The CPs located in $[0, s_0]$ choose standard service, the CPs located in $\lfloor s_0, \bar{s}_{\mathrm{QT}} \rfloor$ choose premium service, and the CPs located in $\lfloor \bar{s}_{\mathrm{QT}}, S \rfloor$ stay out of the market. The NP's decision problem in the case of QT can be formulated as follows:

$$\max_{F_{\mathrm{QT}}, p, \alpha_{\mathrm{H}}, \alpha_{\mathrm{L}}} \pi_{\mathrm{QT}} = F_{\mathrm{QT}}(1 - \underline{v}_{\mathrm{QT}}) + p\alpha_{\mathrm{H}} \int_{s_0}^{\bar{s}_{\mathrm{QT}}} d(s, \alpha_{\mathrm{H}})\, ds$$

s.t. $\alpha_{\mathrm{L}} \int_0^{s_0} d(s, \alpha_{\mathrm{L}})\, ds + \alpha_{\mathrm{H}} \int_{s_0}^{\bar{s}_{\mathrm{QT}}} d(s, \alpha_{\mathrm{H}})\, ds \le \mu$ (NP's Capacity Constraint)

$U(\underline{v}_{\mathrm{QT}}, \alpha_{\mathrm{L}}, \alpha_{\mathrm{H}}, F_{\mathrm{QT}}) = 0$ (Participation Constraint for Consumer $\underline{v}_{\mathrm{QT}}$)

$\varphi_{\mathrm{L}}(s_0) = \varphi_{\mathrm{H}}(s_0)$ (Service Choice Constraint for CP s_0)

$\varphi_{\mathrm{H}}(\bar{s}_{\mathrm{QT}}) = 0$ (Participation Constraint for CP $\bar{s}_{\mathrm{QT}}$)

$\bar{s}_{\mathrm{QT}} \ge \alpha_{\mathrm{H}} + \underline{v}_{\mathrm{QT}}/b$ (Most Congestion-Sensitive Available Content Constraint)

$F_{\mathrm{QT}} \ge 0$, $p \ge 0$, $\alpha_{\mathrm{H}} \ge \alpha_{\mathrm{L}} \ge 0$, $0 \le \underline{v}_{\mathrm{QT}} \le 1$, $\bar{s}_{\mathrm{QT}} \ge s_0 \ge 0$ (Non-negativity Constraint)

Under QT, the NP's ability to charge a usage-based fee p to CPs for premium service allows it to exclude any potential "free-riding" CPs from the market. Figure 3 shows the resulting QT equilibrium in which the NP sets α_{H}, α_{L}, and p such that only CPs with minimum bandwidth requirements lower than or equal to the bandwidth level offered for standard service will subscribe to standard service, i.e., $s_0 = \alpha_{\mathrm{L}}$.

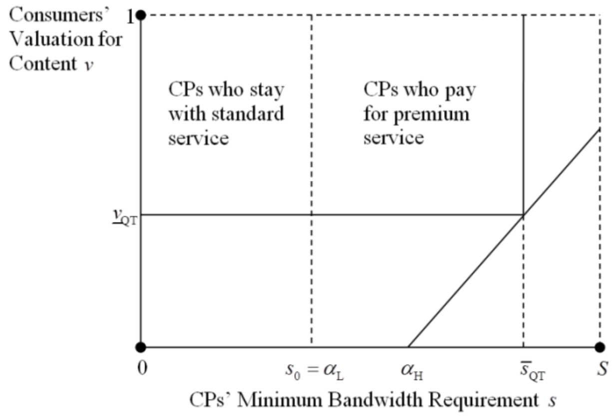


Figure 3: Equilibrium of QoS Tiering

Findings

Three parameters are important in comparing the NN and QT regimes. Two of these, the entry cost E for CPs and their revenue rate r are closely related in that revenues must be sufficient to overcome entry costs for CPs to participate. We can therefore look just at different levels of r for a fixed level of E. The other parameter of interest is consumers' sensitivity to congestion b. In Figure 4, we show a general case of what emerges for different combinations of b and r with a fixed level of E. Based on these three identified scenarios, we discuss the results of content innovation for the CPs' market and the results of broadband coverage for the consumer market.

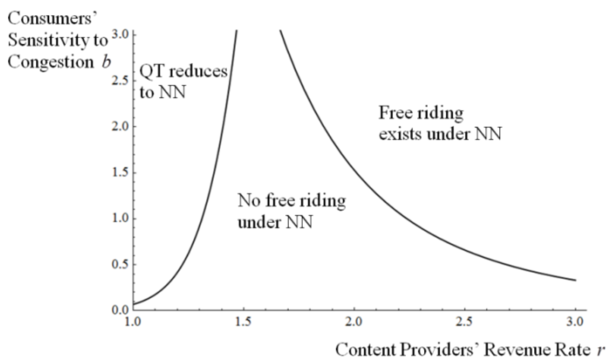


Figure 4: Comparison of NN and QT under Changing Market Conditions

To determine how content innovation varies between the NN and QT regimes, we compare the values of $\bar{s}_{NN}$ and $\bar{s}_{QT}$. A higher value indicates greater entry of CPs and thus higher level of content innovation. Under NN, the NP has no market mechanism available to control entry of CPs and will under some conditions thus have to tolerate free-riding CPs. The existence of these free-riders is detrimental to NP profits, since they use bandwidth capacity to serve just a portion of consumers while not enhancing the utility of the marginal consumer, and thus not contributing to the fee income of the NP. On the other hand, the ability of these CPs to enter the market at the high end of bandwidth requirements and serve consumers with high valuations, is exactly what leads to higher content innovation levels under NN. This result contradicts the findings of Krämer and Wiewiorra (2012), and we believe this is primarily due to their assumption of full market coverage for consumers. Our model relaxes this assumption, and allows both sides of the market to be responsive to the network management and pricing decisions made by the NP.

Comparing broadband market coverage for NN and QT involves comparing $\underline{v}_{\mathrm{NN}}$ and $\underline{v}_{\mathrm{QT}}$. Lower values of $\underline{v}_{\mathrm{NN}}$ and $\underline{v}_{\mathrm{QT}}$ indicate more consumers with internet access, and the value is 0 when full market coverage occurs. Under QT, the NP is able to extract CPs' advertising revenues, which in turn grow with the number of consumers in the market. In some market conditions this leads to equilibria where the NP fully subsidizes the consumer market by setting F_{QT} to 0 and depends solely on p for income. Since the NP cannot obtain profits from p under NN, F_{NN} is always positive, and market coverage under NN never exceeds that achieved under QT. This result is consistent with prior literature (Economides and Tåg 2012).

Concluding Remarks

The main contribution of this research is the development and solution of a two-sided market model that simultaneously considers: both the CPs' and consumers' decisions concerning market entry and participation, consumer demand for content dependent on congestion, heterogeneous consumes in terms of content valuation, and heterogeneous CPs in terms of sensitivity to delay. Our most interesting result concerns content innovation, where we illustrate that what the NP perceives to be a free rider problem is actually a critical feature of NN that under certain market conditions enables superior content innovation. This analysis makes it very clear why NPs would be motivated to block some legal content, and thus underscores the importance of the FCC's no-blocking rule to protecting content innovation. The government's current position defending NN with the no-blocking rule for hard-wired service provision seems reasonable given that at present broadband access concerns can be addressed to a degree through other means such as subsidies and public access points, whereas it is not as clear what alternative mechanisms may exist to preserve incentives to innovate in digital content provision.

References

Cheng, H. K., S. Bandyopadhyay, H. Guo. 2011. The debate on net neutrality: A policy perspective. *Information Systems Research*, 22 (1), 1-27.

Choi, J. P., B. Kim. 2010. Net neutrality and investment incentives. *RAND Journal of Economics*, 41 (3), 446-471.

Courcoubetis, C., F. P. Kelly, R. Weber. 2000. Measurement-based usage charges in communication networks. *Operations Research*, 48 (4), 535-548.

Economides, N., B. Hermalin. 2012. The economics of network neutrality. *RAND Journal of Economics*, forthcoming.

Economides, N., J. Tåg. 2012. Net neutrality on the internet: A two-sided market analysis. *Information Economics and Policy*, 24 (2), 91-172.

FCC. 2010. *Report and Order: Preserving the Free and Open Internet*.

Hermalin, B. E., M. L. Katz. 2007. The economics of product-line restrictions with an application to the network neutrality debate. *Information Economics and Policy*, 19, 215-248.

Kelly, F. P. 1996. *Notes on effective bandwidths*, 141-168.

Krämer, J., L. Wiewiorra. 2012. Network neutrality and congestion sensitive content providers: Implications for content variety, broadband investment and regulation. *Information Systems Research*, forthcoming.

Service Consolidation and Interconnection among Internet Service Providers

I. Robert Chiang
Gabelli School of Business
Fordham University
ichiang@fordham.edu

Jhih-Hua Jhang-Li
Department of Information Management
Hsing Wu University, Taiwan
jhangli@gmail.com

Abstract

The modern-day Internet is managed by two classes of providers: internet service providers (ISPs) catering to retail customers; and internet backbone providers (IBPs) serving large organizations and ISPs on long-haul routes. IBPs, facing excessive capacity and pricing pressure, have started offering content delivery network (CDN) services for content providers. We model the market competition on the CDN services as sequential games, and the aim is to study the game equilibriums by considering factors such as market share, cost structure, product pricing, and consumer preference. We prescribe conditions under which the content provider should choose an IBP over a traditional CDN operator; we also show how the IBP's venture into CDN could affect its interconnection relationship with downstream ISPs.

Introduction

The modern-day Internet is a complex coalition of IP networks managed by two classes of providers: internet service providers (ISPs) and internet backbone providers (IBPs). While there is not a strict delineation, the former generally operate regional networks for retail customers whereas the latter (including the wholesale side of major telecomms) maintain long-haul routes serving institutional clients and ISPs. As the Internet's global reach requires that individual networks properly connect with each other, it is of significant interest to investigate the incentive and economic issues surrounding the interconnection among providers.

An interconnection generally falls into one of the two types: transit and peering. Transit occurs when one provider (likely an ISP) pays the other (likely an IBP) as a customer to gain access to the rest of the Internet. In contrast, peering is de facto among IBPs and gaining popularity among ISPs. Under *regular* (or *settlement-free*) peering, both parties reciprocally terminate (i.e., packets not relayed further) each other's traffic (Weiss and Shin, 2004) without financial exchanges. Free peering is thus more likely when the traffic is relatively symmetric (Shrimali and Kumar, 2008). A *paid* peering differs from its settlement-free counterpart in that one side incurs transport and/or facility costs. Two networks can also negotiate to treat part of the traffic exchanged as transit, with the rest as peered.

During the nascent era of the commercial Internet, overhyped traffic projection fueled a frenzied build-out of national and intercontinental backbones. The ensued reality, however, had led to glutted debts, unlit fibers, and diminished role for IBPs[1]. The demand for the IBPs' network is further diminished by the ISPs' "doughnut peering," essentially using private exchanges to bypass the backbone. Consequently, IBPs reportedly have faced pricing pressure (Besen et al., 2001) at an annualized decline of 61%[2]. Seeking better return from their network infrastructure and data centers, IBPs have branched into the realm of content delivery. A content delivery network (CDN) caches and replicates digital contents on the edges of the Internet on behalf of content providers and eCommerce sites to improve responsiveness and customer

---

[1] http://www.economist.com/node/2098913
[2] http://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php

experience. Traditional CDNs lease capacity from ISPs and IBPs and in turn charge their clients a premium over cost. IBPs, with extensive infrastructure capacity and reach in place, are well positioned to compete in the CDN space.

This research is motivated by an ongoing and well-publicized dispute[3] between an ISP and an IBP. The contention was due to the IBP's new CDN role for a leading streaming media content provider (CP), after supplanting the incumbent. When a major IBP[4] consolidates the content streaming chain, not only does the ISP stand to lose the transit fee from the incumbent CDN, it also receives much higher traffic volume from the IBP. Another issue from the ISP's perspective is that the CP's traffic used to be more evenly distributed from the CDN's edge servers but now has to be hauled from the exchange points with IBP, which in turn may require substantial infrastructure upgrade[5]. This development has created a very dynamic landscape, as its outcome could have significant economic and regulatory implications on provider interconnection. We next show how the CP chooses its CDN and how the ISP responds accordingly.

Modeling the Service Consolidation

We consider the market for a media streaming service, with a number of IBPs and ISPs, a CP, and the CP's incumbent CDN. While a CP may deliver contents to the customers of every ISP through its CDN, an ISP (say, ISP-1) can only make its streaming content available to its own subscribers[6]. We assume that the CP's incumbent CDN leases capacity from the ISPs to minimize streaming delays and that ISP-1 transits with its upstream IBP-1 but also has peering agreements on certain traffic types.

To capture the decision dynamics on provider selection and service pricing, we use a two-stage decision process and show its structure and the three scenarios under analysis in Figure 1. The first stage reflects the ongoing trend of service consolidation, with backbone providers also offering content caching, duplication, and delivery. The decision during this stage is whether CP should stay with the incumbent CDN or switch to an IBP. The decision in Stage 2 will be for ISP-1 to choose between settlement-free peering (Case B) or paid peering (Case C) with its IBP for media streams. We construct and analyze each subgame to find equilibrium solutions.

We derive the demand for the two streaming services using the Hotelling model to capture consumer preference. Without loss of generality, the total number of network subscribers is normalized to one. The network subscribers' preference toward streaming media services is uniformly distributed on the Hotelling line segment [0,1], with ISP-1 (CP) at the 0 (1) end of the Hotelling line. A consumer's net utility from using either service depends on the value attributes (as measured by the number and variety of titles, ease of search, types of device supported, etc.) of the service, the distance between the subscriber' ideal service to those offered by ISP-1 and CP, and the cost of subscription. We use parameter $\theta \in [0,1]$ to capture individual preference and to measure the "fit cost" (Hotelling, 1929) in the Hotelling framework when a service deviates from a consumer's perceived ideal. Let the value of the media streaming service from ISP-1 (CP) as $V_{ISP,1}$ (V_{CP}). Additionally, ISP-1 and the CP charge $t_{ISP,1}$ and t_{CP} of subscription for each period, respectively. The utilities of service from ISP-1 and CP to a consumer located at point θ are $U_{ISP,1} = V_{ISP,1} - \theta - t_{ISP,1}$ and $U_{CP} = V_{CP} - (1-\theta) - t_{CP}$, respectively.

---

3 http://www.usatoday.com/tech/news/2010-12-01-comcast01_ST_N.htm reported the Comcast vs. Level3 dispute

4 http://www.level3.com/en/products-and-services/video/cdn/

5 http://arstechnica.com/tech-policy/news/2010/12/comcastlevel3.ars

6 http://online.wsj.com/article/SB10001424052970204909104577237321153043092.html

Solving $U_{ISP,1}\left(\hat{\theta}\right)=U_{CP}\left(\hat{\theta}\right)$ and $U_{CP}\left(\underline{\theta}\right)=0$ yields the indifference points of misfit at $\hat{\theta}$ and maximum level of misfit of $\underline{\theta}$. If α is ISP-1's market share, then the service demands are:

$$D_{ISP,1}=\alpha\cdot\hat{\theta} \text{ and } D_{CP}=\left(1-\hat{\theta}\right)+\left(1-\alpha\right)\left(\hat{\theta}-\underline{\theta}\right)$$

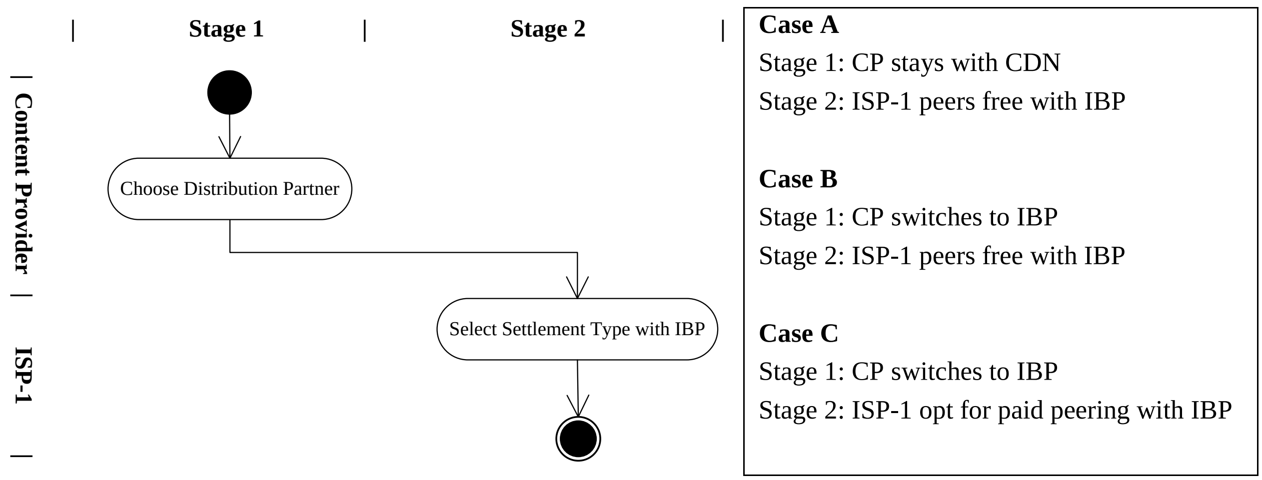


Figure 1. The game stages and scenarios

ISP-1 charges the CDN a fee F for each unit of traffic under the required level of transmission quality. The traffic within ISPs and IBPs can be classified into three types: local traffic, outbound traffic, and inbound traffic (Weiss and Shin, 2004). Prior economic literature assumes that a network incurs different carrying costs between each unit of local traffic (c_t) and each unit of inbound/outbound traffic (c_o) (Armstrong, 1998; Laffont et al., 1998; Laffont et al., 2001; Gans and King, 2001; Hau et al., 2011). To identify the equilibrium solution, we derive profit functions for all three cases and identify conditions/regions under which one set of decisions is preferred over the others.

For Case A, the decision sequence is as follows. First, the CDN charges CP a service fee a_{CDN} for content delivery. Then, both ISP-1 and CP make their pricing decisions simultaneously. If ISP-1 and CP charge $t_{ISP,1}$ and t_{CP} for the streaming services, then the providers' profits are:

$$\pi_{ISP,1}=\left(t_{ISP,1}-c_t\right)\alpha\cdot\hat{\theta}+\left(F-c_t\right)\alpha\cdot\left(1-\hat{\theta}\right) \tag{1}$$

$$\pi_{CDN}=\left(a_{CDN}-F\right)\alpha\cdot\left(1-\hat{\theta}\right)+\left(a_{CDN}-F\right)\left(1-\alpha\right)\cdot\left(1-\underline{\theta}\right) \tag{2}$$

$$\pi_{CP}=\left(t_{CP}-a_{CDN}\right)\cdot\alpha\cdot\left(1-\hat{\theta}\right)+\left(t_{CP}-a_{CDN}\right)\cdot\left(1-\alpha\right)\left(1-\underline{\theta}\right) \tag{3}$$

Lemma 1. (The profits derived for Case A)

ISP-1's and CP's profits are given by $\pi_{ISP,1}=2\alpha\cdot\hat{\theta}^2+\left(F-c_t\right)\cdot\alpha$ and $\pi_{CP}=\dfrac{2}{2-\alpha}\cdot D_{CP}^2$, where $D_{CP}=\left\lfloor\left(\alpha^2-2\alpha\right)V_{ISP,1}+\left(3\alpha^2-10\alpha+8\right)V_{CP}-\left(4\alpha^2-12\alpha+8\right)F+\left(6\alpha-3\alpha^2\right)\right\rfloor/4\left(8-5\alpha\right)$ and

$$\hat{\theta}=\frac{\left(32-50\alpha+19\alpha^2\right)V_{ISP,1}-\left(8-10\alpha+3\alpha^2\right)V_{CP}-\left(24-40\alpha+16\alpha^2\right)F+\left(32+3\alpha^2-26\alpha\right)}{4\left(4-3\alpha\right)\left(8-5\alpha\right)}.$$

For Case B, the decision sequence is as follows. First, IBP-1 charges CP a service fee $a_{IBP,1}$ to stream the media while incurring a margin cost of c_o to carry each unit of traffic. Then, both ISP-1 and CP make their pricing decisions simultaneously. The profit functions are:

$$\pi_{ISP,1}=\left(t_{ISP,1}-c_t\right)\alpha\cdot\hat{\theta}-c_t\cdot\alpha\cdot\left(1-\hat{\theta}\right) \tag{4}$$

$$\pi_{IBP,1}=\left(a_{IBP,1}-c_o\right)\alpha\cdot\left(1-\hat{\theta}\right)+\left(a_{IBP,1}-c_o\right)\left(1-\alpha\right)\cdot\left(1-\underline{\theta}\right) \tag{5}$$

$$\pi_{CP}=\left(t_{CP}-a_{IBP,1}\right)\cdot\alpha\cdot\left(1-\hat{\theta}\right)+\left(t_{CP}-a_{IBP,1}\right)\cdot\left(1-\alpha\right)\left(1-\underline{\theta}\right) \tag{6}$$

Lemma 2. (The profits derived for Case B)

ISP-1's and CP's profits are given by $\pi_{ISP,1}=2\alpha\cdot\hat{\theta}^2-c_t\cdot\alpha$ and $\pi_{CP}=\frac{2}{2-\alpha}\cdot D_{CP}^2$, where

$$\hat{\theta}=\frac{\left(32-50\alpha+19\alpha^2\right)V_{ISP,1}-\left(8-10\alpha+3\alpha^2\right)V_{CP}+\left(8-10\alpha+3\alpha^2\right)c_o+\left(32+3\alpha^2-26\alpha\right)}{4\left(4-3\alpha\right)\left(8-5\alpha\right)}$$ and

$$D_{CP}=\left\lfloor\left(\alpha^2-2\alpha\right)V_{ISP,1}+\left(3\alpha^2-10\alpha+8\right)V_{CP}-\left(3\alpha^2-10\alpha+8\right)c_o+\left(6\alpha-3\alpha^2\right)\right\rfloor/4\left(8-5\alpha\right).$$

For Case C, the stages of the subgame are the same as those in case B. We assume that the transit fee p is determined exogenously by the bargaining power. When monitoring traffic between ISP-1 and IBP-1 incurs a fixed cost M, the profits for ISP-1, IBP-1, and CP are given by:

$$\pi_{ISP,1}=\left(t_{ISP,1}-c_t\right)\alpha\cdot\hat{\theta}+\left(p-c_t\right)\cdot\alpha\cdot\left(1-\hat{\theta}\right)-M \tag{7}$$

$$\pi_{IBP,1}=\left(a_{IBP,1}-c_o-p\right)\alpha\cdot\left(1-\hat{\theta}\right)+\left(a_{IBP,1}-c_o\right)\left(1-\alpha\right)\cdot\left(1-\underline{\theta}\right)-M \tag{8}$$

$$\pi_{CP}=\left(t_{CP}-a_{IBP,1}\right)\cdot\alpha\cdot\left(1-\hat{\theta}\right)+\left(t_{CP}-a_{IBP,1}\right)\cdot\left(1-\alpha\right)\left(1-\underline{\theta}\right) \tag{9}$$

Lemma 3. (The profits derived for Case C)

ISP-1's and CP's profits are given by $\pi_{ISP,1}=2\alpha\hat{\theta}^2+\left(p-c_t\right)\cdot\alpha-M$ and $\pi_{CP}=\frac{2}{2-\alpha}\cdot D_{CP}^2$, where

$$\hat{\theta}=\frac{\left(32-50\alpha+19\alpha^2\right)V_{ISP,1}-\left(8-10\alpha+3\alpha^2\right)V_{CP}+\left(32+3\alpha^2-26\alpha\right)+\left(8-10\alpha+3\alpha^2\right)c_o-\left(32+20\alpha^2-52\alpha\right)p}{4\left(4-3\alpha\right)\left(8-5\alpha\right)}$$

and $D_{CP}=\left\lfloor\left(\alpha^2-2\alpha\right)V_{ISP,1}+\left(3\alpha^2-10\alpha+8\right)V_{CP}-\left(3\alpha^2-10\alpha+8\right)c_o+\left(6\alpha-3\alpha^2\right)\right\rfloor/4\left(8-5\alpha\right).$

We can now analyze ISP-1's best decision if CP decides to contract with IBP-1. When interconnecting with IBP-1 under paid peering, ISP-1 receives a transit fee (to compensate for its loss revenue from CDN) which indirectly influences CP's pricing decision having to cover IBP-1's higher CDN cost. Under free peering, IBP-1 can operate at lower cost than CDN and can afford to charge CP less; CP in turn can pass the saving down to its subscribers and in turn affects the ISP-1's profit from its own media service. By comparing $\pi_{ISP,1}^{(B)}$ with $\pi_{ISP,1}^{(C)}$, we have the following proposition.

Proposition 1.

If CP contracts with IBP-1, then ISP-1's best response is to adopt paid peering contract when

$$\left\{(1-\alpha)\left[\frac{-\Delta+\left(16+10\alpha^2-26\alpha\right)p}{(4-3\alpha)^2(8-5\alpha)}\right]+1\right\}p\cdot\alpha > M \text{ , where}$$

$$\Delta \equiv (32-50\alpha+19\alpha^2)V_{ISP,1}-(8-10\alpha+3\alpha^2)V_{CP}+(8-10\alpha+3\alpha^2)c_o+(32+3\alpha^2-26\alpha)$$

Otherwise, the ISP-1 should peer settlement-free with its IBP.

We use Figure 2 to compare $\pi_{ISP,1}^{(B)}$ and $\pi_{ISP,1}^{(C)}$. The paid peering contract leads to a higher profit when the market share of ISP-1 is sufficiently high. As the market share of ISP-1 increases, more ISP-1' subscribers purchase CP's media streaming service; the traffic revenue from IBP-1 to ISP-1 also increases, so ISP-1's profit will rise with the paid peering. However, if the revenue from paid peering is not high enough to offset ISP-1's monitoring overhead, then keeping the peering settlement-free is preferred. When comparing $\pi_{CP}^{(B)}$ with $\pi_{CP}^{(C)}$, we have the following finding.

Lemma 4.

If CP contracts with IBP-1, CP's profit will not be affected by ISP-1's interconnection type with IBP-1. Formally, $\pi_{CP}^{(B)}=\pi_{CP}^{(C)}$.

Lemma 4 shows that CP is not concerned if IBP-1 charges a higher fee for content delivery when under a paid peering with ISP-1. A higher fee for content delivery service will make CP raise its service fee; however, doing so will also raise that of ISP-1. Consequently, CP's sales won't decline if ISP-1 chooses paid peering. Built on Lemma 4, we can compare $\pi_{CP}^{(A)}$ with $\pi_{CP}^{(B)}$ to identify conditions that CP will choose IBP-1 over the CDN.

Proposition 2.

CP stays with the incumbent CDN when $\dfrac{F}{c_o}\le\dfrac{3\alpha^2-10\alpha+8}{4\alpha^2-12\alpha+8}$; otherwise, it will switch to IBP-1.

Based on Proposition 1 and 2, given $\chi_1 \equiv (3\alpha^2-10\alpha+8)/(4\alpha^2-12\alpha+8)$ and

$$\chi_2 \equiv \left\{(1-\alpha)\left[\frac{-\Delta+(16+10\alpha^2-26\alpha)p}{(4-3\alpha)^2(8-5\alpha)}\right]+1\right\}p\cdot\alpha .$$ We plot the equilibrium result in Figure 3.

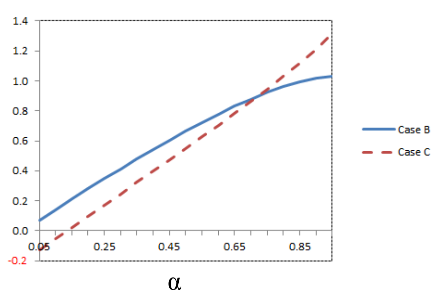


Figure 2. $\pi_{ISP,1}^{*}$ in case B and C

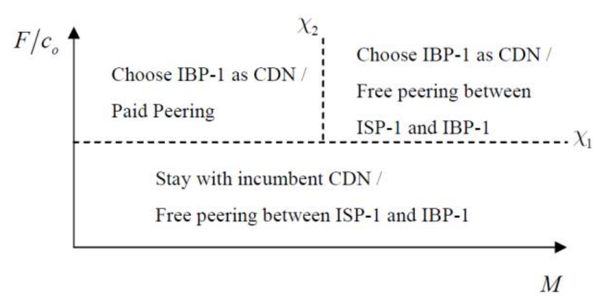


Figure 3. Equilibrium of the game

Corollary 1.

- When $\alpha \approx 1$, the equilibrium is that CP contracts with the CDN.
- When $\alpha \approx 0$, the equilibrium is that ISP-1 agrees with the peer contract.

Conclusions

As IBPs entering the CDN business, it is expected that the interconnection among Internet providers will be rapidly changing in the near future. By not incurring the cost of monitoring traffic, the free peering used to be a reasonable approach to settling the finance of data exchange among Internet providers. However, a revision to such practice appears to be inevitable as the industry is going through consolidation, especially along the line of managing content delivery. Given that many ISPs (including cable companies and the retail side of telecomm establishments) already or are vying to become the content providers for digital media streaming, delivery efficiency and provider relationship discussed here likely won't be the sole drivers for success in this market space. Rather, ownership to media rights will also play a crucial role. Partnerships, rather than direct competition among ISP's and CPs may also become commonplace as the market matures.

Acknowledgement

National Science Council of Taiwan supported this research (grant NSC 101-2410-H-266-002).

References

Armstrong, M., 1998. "Network Interconnection in Telecommunications," *The Economic Journal* (108:448), pp. 545-564.

Besen, S., Milgrom, P., Mitchell, B., and Srinagesh, P. 2001. "Advances in Routing Technologies and Internet Peering Agreements," *The American Economic Review* (91:2), pp.292-296.

Gans, J. S., and King, S. P. 2001. "Using 'bill and keep' interconnect arrangements to soften network competition," *Economics Letters* (71:3), pp. 413-420.

Hau, T., Burghardt, D., and Brenner, W. 2011. "Multihoming, content delivery networks, and the market for Internet connectivity," *Telecommunications Policy* (35:6), pp. 532-542.

Hotelling, H. 1929. "Stability in competition," The Economic Journal (39:153), pp. 41-57.

Laffont, J.-J., Marcus, S., Rey, P., and Tirole, J. 2001. "Internet Peering," *The American Economic Review* (91:2), pp. 287-291

Laffont, J.-J., Rey P., and Tirole, J. 1998. "Network Competition: I. Overview and Nondiscriminatory Pricing," *The RAND Journal of Economics* (29:1), pp. 1-37.

Shrimali, G., and Kumar, S. 2008. "Bill-and-Keep peering," *Telecommunications Policy* (32:1), pp. 19-32.

Weiss, M. B., and Shin, S. J. 2004. "Internet Interconnection Economic Model and its Analysis: Peering and Settlement," *Netnomics* (6:1), pp. 43-57.

Proof of Proposition 1 and 2

$\pi_{ISP,1}^{(B)} = 2\alpha \cdot \hat{\theta}^2 - c_t \cdot \alpha$, where $\hat{\theta} = \dfrac{\Delta}{4(4-3\alpha)(8-5\alpha)}$ and $\pi_{ISP,1}^{(C)} = 2\alpha \cdot \hat{\theta}^2 + (p - c_t) \cdot \alpha - M$, where $\hat{\theta} = \dfrac{\Delta - (32 + 20\alpha^2 - 52\alpha)p}{4(4-3\alpha)(8-5\alpha)}$. We complete the proof of P1 by examining $\pi_{ISP,1}^{(C)} - \pi_{ISP,1}^{(B)}$. Likewise, we complete the proof of P2 by examining $\pi_{CP}^{(A)} - \pi_{CP}^{(B)}$ because of $\pi_{CP}^{(B)} = \pi_{CP}^{(C)}$.

Surviving Hyper-Competitive, Unforgiving Platform Ecosystems: Examining Developer Strategies in iOS and Android Marketplaces

Narayan Ramasubbu
University of Pittsburgh
naayanr@pitt.edu

Kajanan Sangaralingam, Nargis Pervin, Kaushik Dutta, Anindya Datta
National University of Singapore
{skajanan, nargisp, duttak, datta} @nus.edu.sg

Abstract

With the rapid proliferation of mobile applications ("apps") running on smart handheld devices, individual app developers are greatly empowered, having unprecedented market reach and access to powerful tools created by platform sponsors such as Apple and Google. At the same time, achieving sustained success in these mobile platform ecosystems is challenging for developers because of the heavily crowded marketplaces and easy substitutability of apps. In this research, using a large panel dataset of 183,285 observations pertaining to 12,222 apps released in the Apple iOS and Google Android marketplaces, we examine the supply-side dynamics of mobile platform ecosystems. We systematically investigate how developers position, price, and maintain their apps along with the resulting success of the apps in the marketplaces to draw insights on winning strategies in hyper-competitive mobile platform ecosystems.

Introduction

Mobile platforms (Apple iOS, Google Android, Microsoft Phone 7, Blackberry, etc.) and the applications ("apps") created to run on them have disrupted the handheld smartdevice industry, and the app ecosystems have been classified as the fastest growing consumer product segment of this decade (Gigaom 2011; MarketGrowth 2011). To get a sense of the remarkable evolutionary speed and scale of the mobile app ecosystem, consider the following:

- Each week about 100 new movies and 250 new books are released worldwide. In contrast, over 15,000 new apps, on average, are launched weekly; currently there are over 1.2 million apps on the Apple, Android, Blackberry and Microsoft native app markets (Datta et al. 2011; Murphy 2011).
- On smart devices, whose annual sales have now eclipsed that of fixed-line and larger screen devices (e.g., PCs, laptops), apps account for 70% of usage by time, while users spend the remaining 30% on the web (Nielsenwire 2012).
- The dollars associated with apps are enormous as well. According to market estimates, the total global mobile applications market was worth $6.8 billion in 2010 and is expected to grow to about $25 billion by 2015 (Marketsandmarkets2011).

The success of new mobile platform ecosystems and the rapid growth of apps have greatly empowered individual publishers and developers (we will use the term "developers" to refer to both groups) by allowing them to achieve unprecedented success with comparatively minimal investment. Simultaneously, the rapid growth of apps has also created a number of hurdles for them (BusinessInsider 2011). There are simply too many apps and far too much fragmentation in certain mobile operating systems (like Android), making app discovery, assessment, and continuous usage difficult and often frustrating for consumers. There are almost 1 million mobile apps between the Apple and Android platforms alone, but most smartphone owners can only recall and identify a handful (Freierman 2012). It is extremely difficult for an app to gain user's mindshare and get noticed in the crowded app stores. For example, Apple ranks, prominently displays, and promotes only 1000 popular apps in a category in its apps web store, making the discovery of thousands of other apps in the category extremely challenging (similarly, Android promotes only 600 apps in a category). Moreover, it is typical for an app to reach the top 10 status in its category one day, only to slip into oblivion the next day. The high substitutability of apps and the ease with which consumers can install and delete apps in their new generation smart mobile devices, often with the flick of a finger, make the mobile app marketplace a hypercompetitive turf for developers.

Motivated by conceptualizations of competitive strategy as ecology (see Iansiti and Levien 2004a, 2004b), in this research, we take an important first step in examining the supply side dynamics of the two-sided mobile platform ecosystems. We systematically investigate developer strategies in the Apple iOS and Android marketplaces to explicate the factors that contribute to success in these hypercompetitive but potentially enormously lucrative platform ecosystems. Using a large panel dataset of 183,285 observations pertaining to 12,222 apps released in the Apple iOS and Google Android marketplaces, we empirically investigate how developers position, price, and maintain their apps along with the resulting performance of the apps in the two marketplaces.

The rest of the paper is structured as follows. In the immediately following section, we describe the data collection and our regressions model used to analyze developer actions. After presenting our empirical results, we discuss our findings and conclude with a recommendation for future work.

Empirical Context and Data Collection

Key segments of the empirical data for this study were collected from the Apple iOS and Google Android marketplaces ("app stores"). There are two key types of data available in these app stores (i) *Static data*, such as names and descriptions of apps and developer, versions, price etc., and (ii) *Dynamic data*, such as user rating, ranks, and reviews that change continuously (daily, even hourly). For this particular study the dynamically changing data, whose collection poses many technical challenges, was obtained from an independent app rating agency called Mobilewalla (Datta et al. 2011). The data obtained from Mobilewalla was audited and verified for data accuracy multiple times by independent assessors not associated with this research project, and has also been utilized for independent reporting by reputed publications (e.g., Freierman 2011; Murphy 2011). Thus, the reliability of the dataset we utilized is very high.

Mobilewalla crawls app stores and gathers the dynamic changing attributes of an app on a daily basis and stores it in a proprietary database. We integrated a separate research database with the Mobilewalla database and downloaded data on a weekly basis from 1st May 2011 until 30th September 2011. During this time period, we traced 12,222 new and visible apps (i.e., apps that reached the visible ranking zones in the marketplaces) throughout their lifecycle. There is a need for extensive post-processing on the crawled data from the app stores before it is ready for statistical analysis. We performed three key post-processing steps beyond the basic auditing and quality assurance of the data. First, we normalized the data from the two marketplaces to scale the variables to a common measurement scheme. For example, the visible rank zone in the Apple app store (at the time of data collection) was 1-240 whereas in the Android store it was 1-500; we normalized the rank score to a comparable percentage measure.

Second, we derived sentiment scores for user posts using text mining and classified every user comment on an app as positive, negative, or neutral to calculate the overall dissatisfaction score associated with the app. Third, we identified the sets of apps that are similar to each other. Typically, app developers release app variants such as “pro” and “lite” versions with the same root name. For example, the app called “Angry Birds” has several variants in the iTunes store such as “Angry Birds Free” and “Angry Birds HD Free”. To identify these app pairs, we calculated a semantic similarity score using TF-IDF distance (Salton and McGill1986; Spärck J.K. 1972) between a pair of app names. The basic idea is that two strings are more similar if they contain many of the same tokens with the same relative number of occurrences. We identify all the app-pairs that cross a threshold score value, and once these app pairs are identified, we determined if the apps are from the same developer or different developers. If the app is from the same developer, we classify the app as either a functional variant or price variant depending on whether there exists any price difference between the two apps. If the apps are from different developers, we classify the apps as competition apps.

Fourth, we applied the above-mentioned semantic similarity detection procedure on the two app stores (Android and iOS) across app-name, description and developer’s name to determine if an app exists in the two different platforms. Finally, we detected change events such as an app release event or price variation of an app in the two app stores, and coded them appropriately. Eventually, our data set had 183,285 observations for 12,222 apps. The list of variables we collated for our analysis, along with their measurement and post-processing calculations, along with their summary statistics are listed in Table 1.

Table 1: Data derived from iOS and Android App Stores

Variable	Calculation	Mean	Std. Dev.	Min	Max
App Rank	iOS App Rank = [(241-Rank)/240]*100 Android App Rank = [(501-Rank)/500]*100	51.606	22.367	0.2	100
Installed Base	Unique number of user review inputs per app version	1222.301	11310.48	0	347013
Review Score	Average user review score for current version	1.486	1.928	0	5
Popular Category	Number of apps released in the category	806.282	172.298	241	11365
Price	Market price in USD	1.10589	4.173	0	249.99
Price Change	Price reduction or increase in USD	0.004	0.438	-109	90.79
Age	Days since release	35.790	27.387	1	119
Size	Application footprint in Mega Bytes (MB)	14.433	17.404	0.099	99
Multi-Home	Is there an app version for the other platforms? Yes=1	0.067	0.249	0	1
Competition	Number of similar apps in the platform	0.286	0.633	0	10
Developer Experience	Number of apps released prior to this app	8.206	18.841	1	479
Platform	1=Apple, 0=Android	0.712	0.453	0	1
Frequent Updates	Number of versions released	1.307	1.301	0	27
Functional Variant	Is the released app a functional variant of an existing app? 1=Yes	0.105	0.306	0	1
Price Variant	Is the released app a price variant of an existing app? 1=Yes	0.122	0.327	0	1
Dissatisfaction	% of negative posts in reviews, assessed using sentiments of the user reviews	8.718	15.299	0	100

Data Analysis

We measured success of an app using its rank achieved in the Apple and Android app stores. Both the Apple and Android platforms utilize the number of downloads associated with an app as an important factor in determining the rank of an app. Moreover, app rank is an important measure of success in the mobile platforms because it impacts the visibility of an app in the app store. Because the rank visibility zones varied in the Apple and Android app stores (Apple ranks

240 apps and Android lists top 500 apps at the time of data collection), we normalized the app rank measure to a visibility-adjusted percentage measure as detailed in Table 1.

We analyzed the effect of three key developer strategies—price variation, functionality variation, and frequent updates—on the success of app, controlling for other factors such as competition, user sentiments, original price, app size, and developer experience. Since we hypothesized that these strategies would be contingent on platform charactersitcs and user sentiments, we examined the interaction effects between these variables in our analysis. The regression equation specification of the variables of interest is shown in Eq.1. We estimated Eq.1 using random effects generalized least squares (GLS) regression; the results from the regression are presented in Table 2. We performed a variety of diganostics and robustness tests including the Hausman specification test, Breusch-Pagan Lagrange Multiplier test for random effects, checks for autocorrelation, heteroskedsticity, and the effect of outliers, and verified that the results are robust and unbiased.

$$\text{App Rank}_{i,t} = \mu + \beta_1{*}\text{Installed base}_{i,t} + \beta_2{*}\text{Review Score}_{i,t} + \beta_3{*}\text{Popular Category}_{i,t} + \beta_4{*}\text{Price}_{i,t} + \beta_5{*}\text{Price Change}_{i,t} + \beta_6{*}\text{Age}_{i,t} + \beta_7{*}\text{Size}_{i,t} + \beta_8{*}\text{Multi-Home}_{i,t} + \beta_9{*}\text{Competition}_{i,t} + \beta_{10}{*}\text{Developer Experience}_{i,t} + \beta_{11}{*}\text{Platform}_{i} + \beta_{12}{*}\text{Frequent Updates}_{i,t} + \beta_{13}{*}\text{Functional Variant}_{i,t} + \beta_{14}{*}\text{Price Variant}_{i,t} + \beta_{15}{*}\text{Dissatisfaction}_{i,t} + \beta_{16}{*}(\text{Platform}_{i} * \text{Frequent Updates}_{i,t}) + \beta_{17}{*}(\text{Platform}_{i} * \text{Functional Variant}_{i,t}) + \beta_{18}{*}(\text{Platform}_{i} * \text{Price Variant}_{i,t}) + \beta_{19}{*}(\text{Dissatisfaction}_{i,t} * \text{Frequent Updates}_{i,t}) + \beta_{20}{*}(\text{Dissatisfaction}_{i,t} * \text{Functional Variant}_{i,t}) + \beta_{21}{*}(\text{Dissatisfaction}_{i,t} * \text{Price Variant}_{i,t}) + u_i + e_{i,t} \quad \textit{\textbf{Eq.1}}$$

Table 2. Estimation Results

Variable	Coefficient		z	p>z
Installed Base (1000s)	β_1	0.035	2.63	0.008
Review Score	β_2	0.188	5.81	0.000
Popular Category	β_3	0.006	26.47	0.000
Price	β_4	-0.015	-0.48	0.632
Price Change	β_5	-0.369	-5.07	0.000
Age	β_6	-0.060	-33.87	0.000
Size	β_7	-0.022	-2.23	0.026
Multi-Home	β_8	8.6759	18.98	0.000
Competition	β_9	-1.067	-3	0.003
Developer Experience	β_{10}	0.1381	21.63	0.000
Platform	β_{11}	32.782	83.59	0.000
Frequent Updates	β_{12}	1.704	28.85	0.000
Functional Variant	β_{13}	2.476	3.23	0.001
Price Variant	β_{14}	3.985	4.99	0.000
Dissatisfaction	β_{15}	-0.035	-4.64	0.000
Platform X Frequent Updates	β_{16}	-0.794	-6.27	0.000
Platform X Functional Variant	β_{17}	-3.031	-3.8	0.000
Platform X Price Variant	β_{18}	-5.529	-6.93	0.000
Dissatisfaction X Frequent Updates	β_{19}	-0.011	-2.74	0.006
Dissatisfaction X Functional Variant	β_{20}	-0.076	-6.33	0.000
Dissatisfaction X Price Variant	β_{21}	-0.089	-5.33	0.000
Number of observations = 183285; Number of groups = 12222; Wald Chi2(21) = 11018.76, Prob > Chi2 = 0.000				

Discussion

We can glean several important insights from the empirical results presented in Table 2. Starting with the control factors, as expected, we found a bigger installed base (β_1) to be positively associated with app rank, corroborating the view that the proprietary Apple and Android ranking schemes use download measure as an important factor in ranking apps. We also noticed that apps with higher review scores (β_2), apps from experienced developers (β_{10}), and those that face lower competition (β_9) achieve higher ranks. Also, among the apps we observed in this study, those listed in the Apple app store had higher ranks than their Android counterparts (β_{11}). Moreover, apps that were multi-homing (i.e., present in both Apple and Android stores) were ranked higher on average (β_8).

Effect of perceived quality and affordability: The negative coefficient on price change (β_5) indicates that the iOS and Anrdoid marketplaces are price sensitive and holding all else constant, the rank of an app is likely to go down if a developer increases the price of the app. Similarly, consumer dissatisfaction with an app (β_{15}) is negatively associated with app rank. These results indicate that an initial user perception of quality and affordability of apps has a big impact on their success.

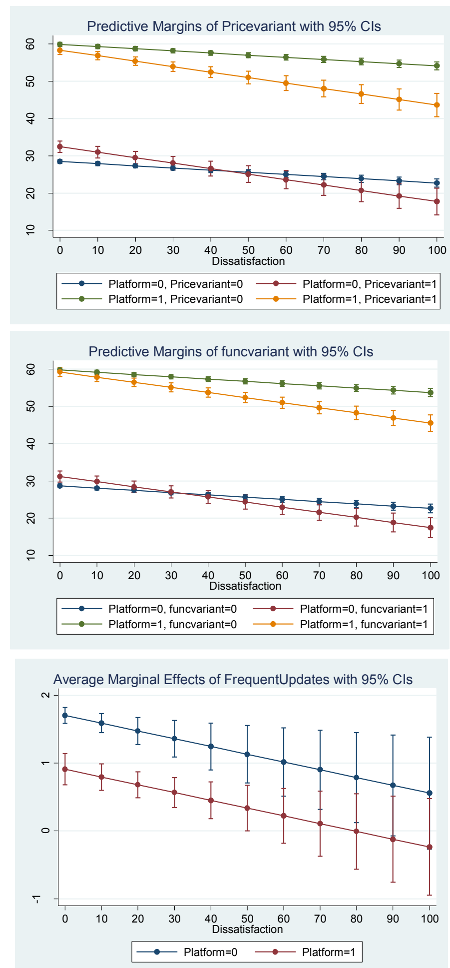


Figure 1. Interaction Plots

Users are unforgiving: Beyond the initial release of the app, the effectiveness of strategies of developers—price and functional variation, and frequent updates—are different in Apple and Android platforms, and are heavily contingent on consumer sentiments. While frequent updates (β_{12}), price variation (β_{14}), and functional variation (β_{13}) of an app are generally beneficial on average, these strategies are not effective if the app received negative sentiments on its initial release (see interaction terms β_{19}, β_{20}, and β_{21} , comparing the coefficient sign change with β_{12}, β_{13}, and β_{14} respectively; refer decreasing marginal effects in Figure 1). Thus, a 'quick release and fix later' developer strategy could be ineffective in positively impacting the rank of the app.

Effect of Platform Properties: We noticed important differences in the effectiveness of developer actions in the Android and Apple ecosystems:

(1)***Freemium Success varies in iOS and Android:*** Releasing an app as a functional or price variant ("HD", "pro", "lite", etc), which is typical in a freemium business model, is met with lower success in Apple iOS as compared with the Android ecosystems (refer to negative coefficients β_{17}, and β_{18}, comparing with β_{13} and β_{14} respectively).

(2) Frequent updating of an app was less effective in positively impacting its rank in the Apple app store than in the Android market (compare β_{12} and β_{16}).

(3) Dissatisfaction scores also had slightly different net impacts on the price and functional variants in the two platform ecosystems, as can be seen from the differing slopes in the marginal effect plots (Figure 1).

These differences could be attributed to the different policies of the platform sponsors (Apple and Google). Apple enforces stringent proprietary quality controls and a time-consuming review process, but provides a more homogeneous experience for consumers across different smart devices (iPhone, iPod touch, and iPad). Thus, it could be tedious for developers to operationalize timely frequent updates. Apple consumers are also less tolerant of functional and price variation across apps, indicating that these app variants are likely perceived as disrupting the otherwise uniform experience across the Apple devices. In contrast, Google has a less-stringent quality control process, but the Android hardware landscape is diverse and fragmented across different manufacturers. Thus, frequent updating of Android apps could be more efficient and timely for developers, perhaps facilitating them to address any incompatibility bottlenecks faced by consumers. Also, since the Android device landscape is heterogeneous and controlled by several different manufacturers, consumers seem to be welcoming functional and price variants of apps that could be potentially tailored specifically for the different Android devices.

Overall, the key insights from this study are:

- iOS and Anrdoid marketplaces are extremely price sensitive and consumers are typically unforgiving. That is, the initial user perception of quality and affordability of an app has a significant impact on its success on both platforms. Following this key initial impression, the effectiveness of adaptation strategies of developers—price, functional variation, and frequent updating of apps—are distinctly different in Apple and Android platforms, and are heavily contingent on consumer sentiments.
- While frequent updates and variations of an app are generally beneficial, these strategies are not effective if the app received negative sentiments on its initial release. Thus, a developer strategy to quickly release an app and fix bugs later is ineffective in positively impacting the rank of the app.
- Releasing an app simultaneously in both Apple and Android helps in achieving higher app rank and hence an improved visibility in the marketplace, but sustaining the higher visibility requires different strategies in the iOS and Android platforms. Functional and price variations of apps with frequent updates were associated with better success in the Android marketplace, whereas developers were better off with less frequent updates, and avoiding price and functional app variants in the Apple appstore.

References

Businessinsider 2011. "Why Native App Stores Like iTunes And Android Marketplace Are Bad For Mobile Developers," http://articles.businessinsider.com/2011-05-12/tech/30060154_1_blackberry-app-world-android-marketplace-native (last accessed August 1, 2012).

Datta, A., Dutta, K., Kajanan, S., and Pervin, N. 2011. "Mobilewalla: A Mobile Application Search Engine," *Proceedings of 3rd International Conference on Mobile Computing, Applications and Services*.

Freierman, S. 2011. "One Million Mobile Apps, and Counting at a Fast Pace," http://www.nytimes.com/2011/12/12/technology/one-million-apps-and-counting.html?_r=1,URL retrieved on 02 May 2012

Gigaom. 2011. "Appsfire scores $3.6m as app discovery demands grow," http://gigaom.com/2011/05/30/ appsfire-scores-3-6m-as-app-discovery-demands-grow/ (last accessed April 13th, 2012).

Iansiti, M., and Levien, R. 2004. "Strategy as ecology", Harvard Business Review, 82(3), pp 68-81.

Iansiti and Levien, 2004. "The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability", Harvard Business Review Press.

MarketGrowth 2011. "Android Market grows a staggering 861.5 per cent," http://www.knowyourmobile.com/smartphones/smartphoneapps/News/781397/android_market_grows_a_staggering_8615_per_cent.html (last accessed April 11th, 2012).

Marketsandmarkets 2011. "World Mobile Applications Market - Advanced Technologies, Global Forecast (2010 - 2015)," http://www.marketsandmarkets.com/Market-Reports/mobile-applications-228.html (last accessed April 11th, 2012)

Murphy, D. 2011. "Mobilewalla: The Market is Inches Away from One Million Mobile Apps," http://www.pcmag.com/article2/0,2817,2397200,00.asp, URL retrieved on 02 May 2012.

Nielsenwire. 2012. "State of the Appnation," retrieved from http://blog.nielsen.com/nielsenwire/online_mobile/state-of-the-appnation-%E2%80%93-a-year-of-change-and-growth-in-u-s-smartphones/ May 16, 2012.

Salton, G., and McGill, M. J. 1986. "Introduction to Modern Information Retrieval," McGraw-Hill, Inc., New York, NY, USA.

Spärck J.K. 1972. "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation* (28:1), pp. 11–21.

The Economics of Shared Data Plans

Soumya Sen, Carlee Joe-Wong, Sangtae Ha
Princeton University, NJ, USA
{soumyas, cjoe, sangtaeh}@princeton.edu

Abstract

The demand for mobile broadband is doubling every year, forcing ISPs to use pricing as a congestion control mechanism. The largest US ISPs, AT&T and Verizon, have already terminated unlimited offerings in favor of "cap and meter" pricing plans with $10/GB overages. More recently, both ISPs have introduced shared data plans in which multiple devices share a data cap. However, such measures to curb demand can have adverse implications for mobile commerce and online content consumption. Hence, a study of the role of overage fees, usage caps, and shared data plans on the consumer's utility is needed. In this work, we introduce an analytical framework for studying the economics of such shared data plans and model the consumer utility of choosing between shared and separate (individual) data plans for their devices. We utilize usage data from a trial of 34 iPhone and iPad users to explore this tradeoff, and show that the choice between individual and shared data plans depends heavily on consumer's willingness to reduce usage upon exceeding the data cap. Our work creates a framework to study this dependency and indicates the importance of analyzing the impact of such data plans on consumer choice.

1. Introduction

The demand for mobile data is rising every year; the Cisco Visual Networking Index for 2012 projects an 18-fold increase in global mobile data traffic between 2011-2016. Mobile video is predicted to be the fastest-growing consumer mobile service, increasing from 271 million users in 2011 to 1.6 billion users in 2016 [Cisco VNI 2012]. Much of this demand growth is driven by the popularity of smart devices, bandwidth-hungry applications, cloud-based services, and media-rich web content. This adoption of smartphones and portable devices by consumers has also fuelled the popularity of mobile commerce [Kerschberg 2012], which is now projected to account for almost 25% of total e-commerce revenues by 2017 [ABI 2012]. In 2011, the mobile commerce market doubled in size to $65.6 Billion, partly due to increases in adoption and consumption by smartphone users in both mature and developing markets [ABI 2012].

However, the sustainability of this growth in the consumption of online content and mobile commerce is increasingly under threat from Internet Service Providers' (ISP) inability to manage the capacity of their wireless networks to accommodate the increasing traffic volume. The major US ISPs, like AT&T, Verizon, and Comcast, have abandoned their traditional unlimited "flat rate" data plans in favor of throttling, tiered usage-based (metered) pricing, and data caps [Sen 2012]. More recently, AT&T and Verizon introduced shared data plans (AT&T's Mobile Share and Verizon's Share Everything) for families and users with multiple devices to share their limited quota across devices. Verizon has also replaced individual plans by shared data plans [Chen 2012]. Hence, understanding how shared data plans can help or hurt consumers is crucial in sustaining demand for mobile data consumption and m-commerce activities. We therefore consider the issue from a consumer's perspective, although similar economic modeling from an ISP's perspective is also equally important, and will be addressed in future works.

In this paper, we introduce a simple analytical framework to model the tradeoffs between choosing a shared data plan and individual data plans for different devices from a consumer perspective. In particular, we show that even in this simplified setting, this choice is non-trivial for a consumer, with a complex dependency on users' psychology of self-censoring their usage upon exceeding their data caps. Restricting the consumer's choice by eliminating individual device data plans can therefore hurt certain types of consumers. Although many shared data plans currently provide free texting and talk time, VoIP has effectively made these services free. Hence, we focus only on data usage caps, which directly impact the Internet ecosystem.

2. Model

We develop a framework to understand the consumer utility and the trade-offs in choosing between individual data plans and shared data plans for his/her mobile devices. Specifically, our model considers two devices with different marginal values for bandwidth consumed, v_1 and v_2, and different amounts of maximum usage, C_1^{max} and C_2^{max}, which can be either put on a shared plan or individual plans by their owner. The nature of the devices under consideration can determine the relative values of these parameters, as we discuss next.

In the first example scenario, consider two devices in a family with two different owners who have different priorities (*e.g.,* a parent and a child). Typically, the value for each unit of bandwidth that a smart device consumes is lower for a child than a parent ($i.e., v_1 < v_2$), and so is the data cap on a child's device ($C_1^{max} < C_2^{max}$). In the second setting, consider a single user with two devices (*e.g.,* iPhone and iPad): one is a personal productivity device for wireless data access (*e.g.,* iPhone) and the other is for entertainment (*e.g.,* iPad). The value of each unit of bandwidth consumed on the productivity device, v_1, is higher than that of the entertainment device, v_2 ($i.e., v_1 > v_2$), but the latter has a higher maximum data demand ($C_1^{max} < C_2^{max}$)[1].

In both these scenarios, a monopolist ISP is assumed to offer a continuum of usage-based pricing plans with different bandwidth caps from which consumers can choose[2]. If a consumer exceeds the chosen bandwidth cap, overage fees are charged. Mathematically, an ISP's individual base plans are priced at p_1 and p_2 per unit of bandwidth consumed (measured, *e.g.*, in MB) for the two types of devices, respectively, and a shared data plan of p_s per unit of bandwidth consumed on either device ($p_s > p_1, p_2$). Additionally, the shared plan also has a fixed monthly fee g_1 and g_2 for each of the two devices (*e.g.,* iPads and iPhones have different flat fees for being on the shared plan). For instance, AT&T charges $85 for 1GB of data per month for one smartphone. Adding a smartphone to this shared plan costs $45, with an overage charge of $15/GB [Molen 2012]. For the case of individual data plans for the two devices, a consumer chooses monthly bandwidth caps of B_1 and B_2 and can incur overage fees of o_1 and o_2 for each unit of bandwidth consumed above these caps. Under the shared data plan, a consumer chooses a bandwidth cap of B_s across the two devices and incurs an overage cost of o_s per unit of extra bandwidth consumed on either device ($o_s > o_1, o_2$).

Because of the overage fees involved, we assume that for individual device data plans, when the device with a lower value (*e.g.,* a child's smartphone or the iPad, in the two examples above) exceeds the consumer's chosen cap for that device, then the user continues to consume data but cuts back on his demand and consumes only a fraction $(1 - \alpha)$ of the demand over the cap. This typically happens because overage warnings and higher fees trigger the user psychology of self-restraint. Similarly, for the shared case, when the cap is exceeded, the user curtails the usage on the lower-value device by a factor α. Due to space constraints we only perform the analysis for the case $v_1 < v_2$ and $C_1^{max} < C_2^{max}$ (*i.e.,* a child's and parent's devices); the analysis for other scenarios is similar.

---

[1]Typically smartphones are actively used by consumers for their work-related data access and consume about 450 MB/month on an average, whereas users consume roughly 3.6 GB on iPads by streaming music and mobile videos.

[2]We assume usage-based pricing for simplicity, but tiered data plans can be considered using a similar framework. AT&T's tiered plans of $30 for 3GB, $50 for 5GB, already approximates *usage-based* plans of $10/GB. We consider a continuum of data cap choices for users (like a continuum of tiered plans over a range of data cap options), but the model can be easily discretized.

The consumer's objective is to choose between shared or individual plans and an optimal data cap. This decision involves three stages: first a consumer needs to choose the type of plan (shared or individual) and the data cap, then the demand on the two devices is realized, and third, if the demand exceeds the cap, then the consumer react by reducing some usage from the device with lesser value due to the higher overage fees. This sequential decision process is shown in Fig. 1. The problem needs to be solved backward by computing the expected utility for a given choice of data cap, and then choosing the cap and plan that maximizes the consumer's utility.

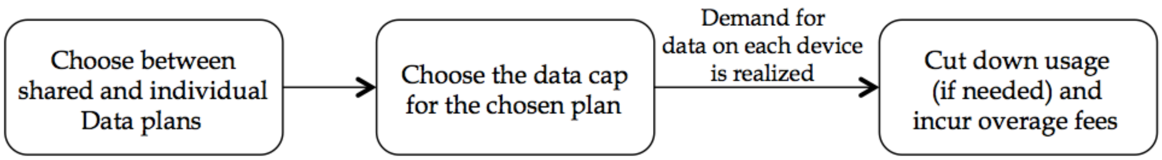


Figure 1: The stages of sequential decision process for a data plan consumer.

3. Analysis

We suppose that the consumer's realized demand for device i is c_i, which follows a distribution $f_i(c_i)$ that varies from *0* to C_i^{max}. The utility derived by a consumer from consuming some amount c_i is assumed to be a concave function of data consumed, $v_i \log(1+c_i)$. For the individual data plan of device 1, the consumer's utility for choosing a data plan of cap B_1 is:

$$U_1 = \begin{Bmatrix} v_1 \log(1+c_1) - p_1 B_1, & c_1 < B_1 \\ v_1 \log(1+c_1-\alpha(c_1-B_1)) - p_1 B_1 - o_1(1-\alpha)(c_1-B_1), & c_1 \geq B_1 \end{Bmatrix}$$

Note that upon exceeding the cap on device 1, a fraction of excess demand α is lost.

For device 2's individual plan, the utility for the consumer for choosing a data plan of cap B_2 is:

$$U_2 = \begin{Bmatrix} v_2 \log(1+c_2) - p_2 B_2, & c_2 < B_2 \\ v_2 \log(1+c_2) - p_2 B_2 - o_2(c_2-B_2), & c_2 \geq B_2 \end{Bmatrix}$$

Under individual plans, a consumer's utility from choosing a data plan with cap B_1 for device 1 is

$$E(U_1|B_1) = \int_0^{B_1} [v_1 \log(1+c_1) - p_1 B_1] f_1(c_1) dc_1 + \int_{B_1}^{C_1^{max}} [v_1 \log(1+c_1-\alpha(c_1-B_1)) - p_1 B_1 - o_1(1-\alpha)(c_1-B_1)] f_1(c_1) dc_1 \quad (1)$$

Setting the derivative $dE(U_1|B_1)/dB_1 = 0$, the optimal data plan cap for consumer's device 1, B_1^*, and the resulting utility, $U_1^* = E(U_1|B_1^*)$, can be estimated. A similar calculation for $U_2^* = E(U_2|B_2^*)$ gives the individual data plan with the cap that is optimal for device 2.

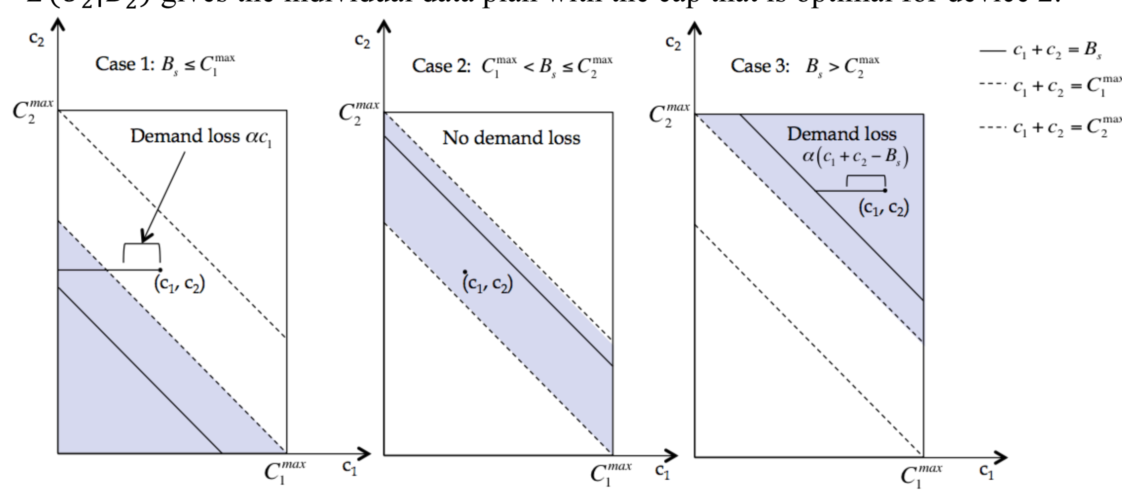


Figure 2: The three regions for which the optimal B_s need to be calculated. The one with the highest $E(U|B_s)$ is chosen as the optimal data cap for a consumer.

For the shared data plan, if the demand across the devices exceeds the shared cap B_s, then the lower priority device 1 curbs its realized usage c_1 by a fraction α of the amount by which it exceeds the shared cap B_s. As in the case of individual data plans, the utility of a user for a chosen value of data cap B_s needs to be computed by integrating over the demand distribution. But the limits of the integration and the utility derived are different depending on the value of the chosen B_s compared to C_1^{max} and C_2^{max}. As shown in Figure 2, there are 3 such regions in the (C_1^{max}, C_2^{max})-plane for which the expected utility needs to be calculated and the value of B_s^* that maximizes it should be chosen as the data cap for the user's shared plan.

If $c_1 + c_2 \leq B_s$, the shared data cap is not exceeded and the user experiences no loss in demand. If $c_2 \leq B_s$ but $c_1 + c_2 \geq B_s$, device 1 experiences a loss in demand proportional to the amount by which the cap B_s is exceeded, i.e., a loss of $\alpha(c_1 + c_2 - B_s)$; if, however, $c_2 > B_s$ and $c_1 + c_2 \geq B_s$, device 1 experiences a demand loss proportional only to the amount by which usage of device 1 exceeds the cap, i.e., αc_1. The expected utility for $B_s < C_1^{max}$ is thus

$$\begin{aligned} E\left(U_s | B_s < C_1^{\max}\right) = & \int_0^{B_s} \int_0^{B_s - c_2} \left[v_1 \log\left(1 + c_1\right) + v_2 \log\left(1 + c_2\right) - p_s B_s\right] f_1\left(c_1\right) f_2\left(c_2\right)\, dc_1\, dc_2 \\ & + \int_0^{B_s} \int_{B_s - c_2}^{C_1^{\max}} \Big[v_1 \log\left(1 + (1-\alpha) c_1 - \alpha\left(c_2 - B_s\right)\right) + v_2 \log\left(1 + c_2\right) - p_s B_s \\ & \qquad - o_s (1-\alpha)\left(c_1 + c_2 - B_s\right)\Big] f_1\left(c_1\right) f_2\left(c_2\right)\, dc_1\, dc_2 \\ & + \int_{B_s}^{C_2^{\max}} \int_0^{C_1^{\max}} \Big[v_1 \log\left(1 + (1-\alpha) c_1\right) + v_2 \log\left(1 + c_2\right) - p_s B_s \\ & \qquad - o_s\left((1-\alpha) c_1 + c_2 - B_s\right)\Big] f_1\left(c_1\right) f_2\left(c_2\right)\, dc_1\, dc_2 \end{aligned}$$

Similarly, we compute

$$\begin{aligned} E\left(U_s | C_1^{\max} < B_s < C_2^{\max}\right) = & \int_0^{B_s} \int_0^{B_s - c_2} \left[v_1 \log\left(1 + c_1\right) + v_2 \log(1 + c_2) - p_s B_s\right] f_1(c_1) f_2(c_2)\, dc_1\, dc_2 \\ & + \int_0^{B_s - C_1^{\max}} \int_0^{C_1^{\max}} \left[v_1 \log(1 + c_1) + v_2 \log\left(1 + c_2\right) - p_s B_s\right] f_1(c_1) f_2(c_2)\, dc_1\, dc_2 \\ & + \int_{B_s - C_1^{\max}}^{C_1^{\max}} \int_{B_s - c_2}^{C_1^{\max}} \Big[v_1 \log\left(1 + (1-\alpha) c_1 - \alpha\left(c_2 - B_s\right)\right) + v_2 \log\left(1 + c_2\right) - p_s B_s \\ & \qquad - o_s (1-\alpha)\left(c_1 + c_2 - B_s\right)\Big] f_1(c_1) f_2(c_2)\, dc_1\, dc_2 \\ & + \int_{B_s}^{C_2^{\max}} \int_0^{C_1^{\max}} \Big[v_1 \log\left(1 + (1-\alpha) c_1\right) + v_2 \log\left(1 + c_2\right) - p_s B_s \\ & \qquad - o_s\left((1-\alpha) c_1 + c_2 - B_s\right)\Big] f_1(c_1) f_2(c_2)\, dc_1\, dc_2 \end{aligned}$$

$$\begin{aligned} E\left(U_s | C_2^{\max} < B_s < C_1^{\max} + C_2^{\max}\right) = & \int_0^{B_s - C_1^{\max}} \int_0^{C_1^{\max}} \left[v_1 \log\left(1 + c_1\right) + v_2 \log\left(1 + c_2\right) - p_s B_s\right] f_1\left(c_1\right) f_2\left(c_2\right)\, dc_1\, dc_2 \\ & + \int_{B_s - C_1^{\max}}^{C_2^{\max}} \int_0^{B_s - c_2} \left[v_1 \log\left(1 + c_1\right) + v_2 \log\left(1 + c_2\right) - p_s B_s\right] f_1\left(c_1\right) f_2\left(c_2\right)\, dc_1\, dc_2 \\ & + \int_{B_s - C_1^{\max}}^{C_2^{\max}} \int_{B_s - c_2}^{C_1^{\max}} \Big[v_1 \log\left(1 + (1-\alpha) c_1 - \alpha\left(c_2 - B_s\right)\right) + v_2 \log\left(1 + c_2\right) - p_s B_s \\ & \qquad - o_s (1-\alpha)\left(c_1 + c_2 - B_s\right)\Big] f_1\left(c_1\right) f_2\left(c_2\right)\, dc_1\, dc_2. \end{aligned}$$

These equations may then be individually solved for the optimal budgets B_s^{*i}, $i = 1,2,3$ corresponding to the three cases above. The expected utility values at these optimal budgets can then be compared to find $E(U_s | B_s^*)$ for the shared data plan. Similar methods can be applied if $v_1 > v_2$ (e.g., iPhones and iPads).

4. Data and Simulations

We now apply the analytical framework introduced above to empirical data. We first estimate the distributions of the c_1 and c_2 variables, i.e., the distribution of the monthly usage volume for devices 1 and 2. As shared mobile data plans have come onto the market only recently, we use data from devices on individual data plans. Our data comes from 19 iPhone (i.e., device 1) and 15 iPad (i.e., device 2) users. We recruited trial participants from fourteen different academic and administrative divisions of our university, as well as their family members, and measured their 3G usage at an hourly granularity for three months. Based on this usage data, our trial participants fell into two groups: high usage (over 900MB for iPad users, and over 400MB for iPhone users) and low usage. We used maximum likelihood estimation to fit a β distribution to the monthly usage data for users of each device, in each group.[3] We obtain similar β distribution parameter values for the different groups of users: $(a, b) = (0.482, 0.549)$ for low-usage iPhone users, (0.344, 0.224) for high-usage iPhone users, (0.384, 0.408) for low-usage iPad users, and (0.515, 0.191) for high-usage iPad users. In the simulations below, we utilize the parameter values for the low usage groups of iPhone (device 1) and iPad (device 2) users; similar results may be obtained with the parameters for high-usage users.

We suppose that the user values iPhone over iPad usage and that iPad demand is lost over the data cap. Figure 3 shows the optimal budgets for the shared and individual data plans with their corresponding expected utility values, for a given set of marginal valuations and prices. If the user is not willing to cut back on her over-the-cap usage ($\alpha < 0.1$), then she prefers an individual data plan: since the shared plan's marginal price $p_s = 14$ is more expensive than the individual plans' ($p_1 = 8, p_2 = 10$), her optimal cap on a shared data plan does not account for large usage amounts on both devices. Similarly, if the user reduces her iPad's usage over the cap ($\alpha > 0.6$), she prefers individual data plans; most over-the-cap usage will take place on device 1, which can be accounted for with device 1's data cap. The expected utility of both individual and shared data plans is thus non-monotonic, reflecting the tradeoff between the utility of consuming more and disutility of exceeding the data cap.

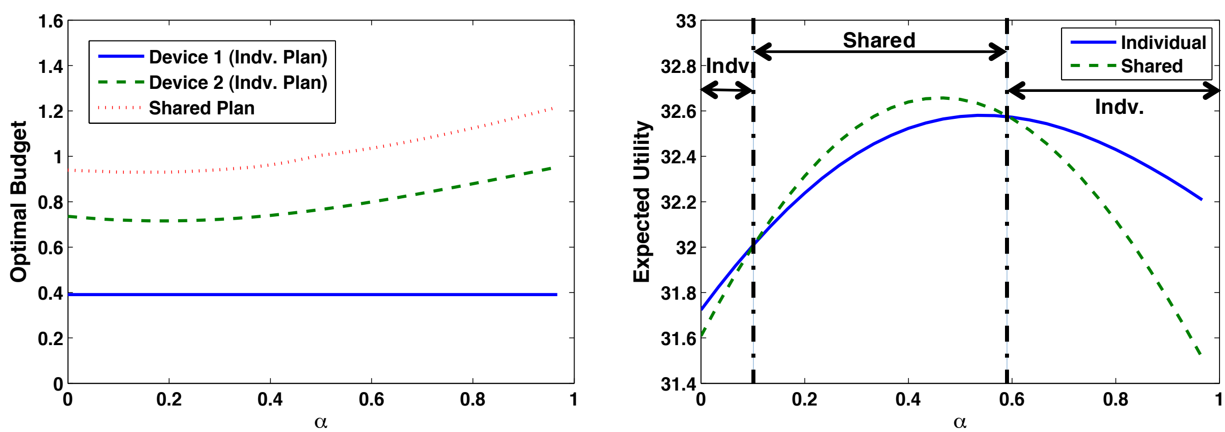


Figure 3: Optimal budgets and expected utility with β distribution parameters from real data. Prices and valuations (\$/GB) are $v_1 = 60, v_2 = 40, p_1 = 8, p_2 = 10, o_1 = 15, o_2 = 18, p_s = 14, o_s = 20$; maximum demand is $C_1^{max} = 1$ GB, $C_2^{max} = 2$ GB. Fixed fees for the shared data plans are (in \$) $g_1 + g_2 = 2$.

[3] Though our model assumes a demand distribution for one user of each device, we aggregate data from a group of users in order to obtain a sufficient number of monthly usage data points to fit a distribution.

Qualitatively similar behavior, with the user switching between preferring shared and individual data plans, can be obtained for other demand distributions (e.g., a uniform distribution) and for the case in which demand is lost from device 1.

Related Work

The question of choosing between shared and individual data plans has some parallels to the decision in manufacturing systems in provisioning between flexible and dedicated resources in presence of demand uncertainty [Fine 1990, Van Mieghem 1998]. There are two key differences between this paper and these models. First, rather than exploring the benefits of hedging against uncertainty with shared resources; the focus of this work is to understand the impact of the fee structure on the data plan choices users make. Secondly, in manufacturing systems, excess demand is lost due to time lag of ramping up capacity; but in case of shared data plans, users can reduce their usage in response to exceeding the data caps. As we show in this work, the level of willingness of a user to curb down on their excess usage can not only affect the data caps they choose a priori but also the very choice (shared or individual) of data plans for their devices.

Conclusions

In this work, we introduce an analytical framework to study user choice between individual and shared data plans for their mobile devices. In particular, we study the impact of overage fees on users' decisions of whether to adopt individual or shared data plans. We use empirical usage data from iPhone and iPad users to demonstrate that user choice of shared or individual data plans depends non-trivially on their psychological response in reducing usage upon exceeding their monthly data cap. Users who are either very willing or very unwilling to reduce usage prefer individual plans, while those between these extremes prefer shared plans. Thus, the economic impact of shared data plans deserves future study: their effect on users' mobile data consumption and the Internet ecosystem, can be non-intuitive. Our work provides an initial framework and results in this direction. Future directions for investigation include considering the issue from an ISP's perspective of how to choose between plans with different usage and cap structures, how to price them, and the social welfare realized under the profit maximizing decisions of the ISP.

References

[Cisco VNI 2012] Cisco Visual Networking Index White Paper. February, 2012. Available online at http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

[Kerschberg 2012] "How Instant Mobile Commerce Will Disrupt The Way You Do Business," B. Kerschberg, Forbes Tech, Online, July 2012.

[ABI 2012] "M-Commerce Growing from 24% of Total E-Commerce Market Value on Back of Smartphone Adoption, Claims ABI Research," ABI Research Press Release, Online, July 30, 2012.

[Chen 2012] "Shared Mobile Data Plans: Who Benefits?" B. X. Chen, The New York Times, July 19, 2012.

[Fine 1990] "Optimal investment in product-flexible manufacturing capacity," C. H. Fine and R. M. Freund, Management Science, 36(4), 1990, pp. 449-465.

[Molen 2012] "AT&T and Verizon shared data plans compared," B. Molen, Online, EnGadget, July 2012.

[Sen 2012] "Incentivizing Time-Shifting of Data: A Survey of Time-Dependent Pricing for Internet Access", S. Sen, C. Joe-Wong, S. Ha and M. Chiang, IEEE Communications Magazine, November, 2012.

[Van Mieghem 1998] "Investment strategies for flexible resources," J. A. Van Mieghem, Management Science, 44(8), 1998, pp. 1071-1078.

Assessing the contribution of EHR systems to medical decision-making

Ofir Ben-Assuli (Ono Academic College, ofir@ono.ac.il), Moshe Leshno (Tel-Aviv University), Itamar Shabtai (College of Management, itamar@colman.ac.il), Shawndra Hill (The Wharton School of the University of Pennsylvania, shawndra@wharton.upenn.edu)

Abstract

This paper evaluates the contribution of an electronic health records (EHR) systems to decision-makers (physicians) by investigating whether these systems contribute to improved medical outcomes in emergency departments (ED). Log-files were retrieved from an integrative EHR system that serves seven main hospitals owned by a big health maintenance organization (HMO). We found that using an EHR system in the EDs correlates with a decreased number of readmissions within seven days as well as with a reduced number of single-day admissions. Our results provide evidence that viewing medical history via EHR system can assist ED physicians in better decision-making. Likewise, the lack of use of EHR system in EDs might lead physicians to make medical decisions that are less effective, and hence lead to unnecessary expenses for the HMO. The study focused on single-day admissions and readmissions within seven days, problems that concern hospitals around the world.

Introduction

The healthcare sector has invested heavily in IT in recent years to enhance medical decision-making and increase its efficiency through improved medical processes (Goldschmidt 2005), streamlining costs, and the use of integrative and interoperable electronic health record (EHR) information systems (IS). EHR IS compiles data from multiple health sources such as laboratories, other hospitals, specialized clinics, etc. Despite the advantages of such IT, physicians do not always consult the vital medical information required for critical decision-making, and the lack of information may result in a decreased level of quality of care and unnecessary costs (Lawson 2011).

One of the most important decisions a physician has to make based on medical information in an emergency department (ED) is whether to admit a patient or not. Admissions and readmissions are a key quality measure in the healthcare system. If a patient is readmitted shortly after a hospital stay, this might indicate that the hospital discharged the patient without proper care or the right diagnosis. In addition, existing scales have shown that unnecessary short-term admissions can also be reduced and even eliminated if physicians had access to proper medical history (Ben-Assuli et al. 2012; Shabtai et al. 2007).

Readmissions are of particular concern in the United States because Medicare, the main provider of healthcare services, will no longer reimburse hospitals for readmissions within 60 days. A report prepared for Congress details the rationale for the new policy that hopes to curb hospital costs and improve overall quality (Stone and Hoffman 2010). The report discusses the comprehensive health care reform legislation, the Patient Protection and Affordable Care Act (PPACA; P.L. 111-148), that was signed into law on March 23, 2010 by President Obama.

This paper seeks to understand the role that EHR can play in improving the decision making process, thus leading to a reduced number of hospital readmissions and single day admissions.

Background

Theoretical Studies of Medical Information

The effects of medical IS at the point of care have been studied from a variety of perspectives, yet little research actually documented the effectiveness of their use (Basu and Meltzer 2007).

Walker et al. (2005) argued that there is a relationship between the viewing medical history and improved medical care performance including admission decisions. Shabtai et al. (2007) evaluated the contribution of IT to medical decision-making and concluded that accessing the medical history of patients by using IT can improve decision-making and its outcomes. Goldman et al. (2006) showed that children with abdominal pain receive more effective medical care (including better admission decisions) when physicians review their medical history.

Although clinicians expressed a high level of confidence after receiving the information (Westbrook et al. 2005), the introduction of additional information and care options has been shown to increase decision complexity (Redelmeier and Shafir 1995). In practice, physicians cannot wait for the results of time-consuming diagnostic procedures (Walter et al. 2011), and even if information is available to physicians, time constraints can restrict their availability to pursue it (Tierney 2001). In fact, physicians retrieve only a limited amount of relevant information even without time constraints (Hersh and Hickam 1998) and use EHR systems for far fewer tasks than the systems support (Laerum et al. 2003). One study reported that even though many ED physicians believe a majority of their patients would benefit from patient health information, they attempted to obtain such data less than 10% of the time (Hripcsak et al. 2007).

The Main Health Maintenance Organization and the Focal EHR IS

This study focused on one of the world’s largest HMOs, a non-profit organization that serves over 3.5 million customers. All seven general hospitals owned by the HMO are surveyed in this research. In this study we analyze an EHR system implemented by the HMO in year 2004. This system provides and shares an integrated real-time virtual patient record to all points of care of the HMO (including hospitals and clinics). The log-file used in this study covered patients’ data from 2005 to 2007. The data included demographic information, previous encounters data, past diagnoses, permanent medications, adverse reactions, detailed lab and blood tests, imaging results, a list of past surgeries, etc. This EHR IS, provided full integrative medical information only on patients of the main HMO. On all other HMO patients, only partial information was collected and thus available in the system.

Research Hypothesis

Sox et al. (2007) explained how, through the processes of decision analysis (such as admissions' decisions) a physician can reach valid, reasonable conclusions regarding medical care, emphasizing the importance of medical history as a means to attain those conclusions. These are our hypotheses followed by explanation and appropriate background:

H1: There is a negative relationship between viewing medical history and a decision to admit a patient to a hospital, which is later followed by a readmission within seven days.

Explanation and background: This readmission measurement is widely used as a means for monitoring the efficacy of critical care pathways (Ramachandran et al. 2007). Decreasing incidences of readmission have increasingly become a goal of caregivers, hospital administrators, and policy makers (Schneider et al. 2012). An accepted notion is that the shorter the period between discharge and readmission, the more likely that the patient was discharged prematurely (Ather et al. 2004). Alternatively, readmission rates are also used as a proxy for quality of care rendered during the hospitalization (Ather et al. 2004; Welch et al. 1992).

H2: There is a negative relationship between viewing medical history and a decision to admit a patient to a hospital, resulting in a single-day admission.

Explanation and background: Existing scales have shown that such short-term admissions can be reduced using medical information (Ben-Assuli et al. 2012; Shabtai et al. 2007). Single-day admission is equivalent to a 24 hour ED stay unit. However, similar to many EDs around the world, the researched hospitals maintain observation wards in which patients are supervised for periods of 12–24 hours (this observation period was not calculated).

Material and Methods

The research method selected for this study was track log-file analysis using statistical tools such as multivariate logistic regression. The log-files were retrieved from the main HMO databases for ED referrals from 2005 to 2007. The log-file consists of about 2.4 million referrals (restricted to internal medicine, surgical, obstetrics, orthopedics, gynecology, ENT, primary, and dermatology). In this paper, we study the percentage of readmission and single day admission when physicians used the EHR compared to when physicians did not use the EHR. Figure 1 shows that for all hospitals, there were reductions in the percentages of readmissions within seven days when the EHR was viewed, and, for almost all hospitals there were reductions in the percentages of single day admissions.

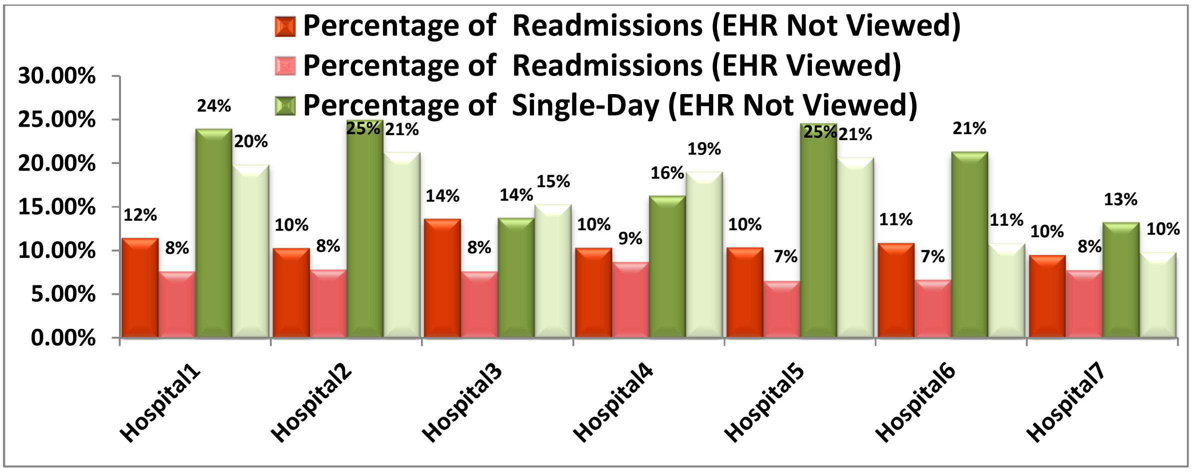


Figure 1. Readmissions and Single-day Admissions when EHR was viewed vs. not viewed.

Our main independent variable is **viewed medical history**. The "viewed medical history" refers to access to at least one of several medical history components in the EHR IS. This was measured as a dichotomous variable (1=history viewed; 0 if not). There were other independent variables including: HMO (1= member of the main HMO; 0 = other HMO), **ED Department**, **hospital**, **patient age** and **patient gender** (coded 1= male; 0 if female).

Our main dependent variables were: readmission within seven days and single-day admissions. **Readmission within seven days**, Quantified whether a patient was readmitted to a hospital within seven days since previous discharge from the hospital (coded 1) or otherwise (coded 0). **Single-day admissions,** Quantified whether a patient, as a result of the decision to admit, was admitted for a single day (coded 1) or for a longer period of time (coded 0).

Research Findings

The first step in our analysis was to calculate the percentages of histories viewed in terms of patient characteristics. Tables 1 and 2 reveal several differences between our examined variables. According to Table 1, in only 19.53% of all referrals to hospitals were the patients' medical histories viewed, a large portion of which resulted in a decision to admit. According to Table 2, medical history was used in the medical cases that involved older patients and probably in cases that tended to be more severe or urgent and ended in an "admit" decision (25.55% admissions rate when the EHR IS was not viewed vs. 41.93% when the EHR IS was viewed). There were significant correlations between viewing medical history and both the number of readmissions within seven days and single-day admissions. Following, we analyze these effects.

Table 1. Analysis of Patient Characteristics

Characteristics	Total Sample n = 2,397,878	EHR Was Not Viewed n = 1,929,474 (80.47%)	EHR Was Viewed n = 468,404 (19.53%)
Age (years) ± s.d.	39.28±24.92	36.36±23.91	51.29±25.38
Male (%)	1,107,537 (46.2%)	897,303 (46.5%)	210,234 (44.9%)
Insurance (% main HMO)	1,756,126 (73.2%)	1,384,450 (71.8%)	371,676 (79.4%)
Admission days ± s.d.	3.10±4.49	2.86±4.11	3.68±5.27

Data is mean or proportion of subjects; all univariate comparisons were significant at 0.001.

Table 2. Comparison of Patients whose Histories were Viewed vs. Patients whose Histories were not Viewed

Characteristics (%)	Total Sample	EHR Was Not Viewed	EHR Was Viewed
Admissions	689,317 (28.75%)	492,892 (25.55%)	196,425 (41.93%)
Readmission within Seven Days	250,369 (36.32%)	214,444 (43.51%)	35,925 (18.29%)
Single-day Admissions	139,012 (20.17%)	105,013 (21.31%)	33,999 (17.31%)

Multivariate logistic regressions were calculated on the two main dependent variables. Three blocks of variables were included: treatment variables (history viewing, age, HMO, and gender), control variables for type of department (for example internal medicine and surgical), and control variables for different hospitals (due to various differences such as policies). This regression reflected the pure contribution of viewing EHR to the rate of readmissions and to the rate of single-day admissions. Tables 3 and 4 present the regression outcomes (Block 2-control for type of department and Block 3-control for type of hospital are not shown here).

Table 3. A Logistic Regression on Readmission within Seven Days (H1)

Variables in the Equation	β	Standard Error	Odd Ratio	95% C.I. for Odd Ratio	
				Lower	Upper
History Viewed**	-.309	.007	.734	.725	.744
Age***	.002	.000	1.002	1.001	1.002
Insurance**	.265	.005	1.304	1.290	1.317
Gender**	.018	.005	1.019	1.008	1.030
Constant	-2.664	.016	.070	-	-

Note: *** p<0.001, ** p<0.01, *p<0.05, + p<0.1 (the table below use the same conventions).

Table 3 shows that when the history was viewed, the odds of readmissions to the ED within seven days decreased significantly (**H1 approved**) by 26.6% (p<0.01, adjusted OR=0.734). Insurance type also had a substantial effect. When the insured patient was a member of the main HMO, the odds of readmission within seven days increased 30.4% (p<0.01, adjusted OR=1.304). This result is reasonable, since we expect non-HMO insured patients to visit their own HMO hospital and not necessarily revisit the main HMO's hospitals where they were not insured. Both age and gender had negligible effects on readmission within seven days.

As shown in Table 4, when history was viewed the odds of single-day admission to the ED decreased (**H2 approved**) by 17.5% ($p<0.001$, adjusted OR=0.825). Age, insurance type, and gender also played significant though considerably smaller roles. First, when the age of the patients increased by an additional year, the odds of single-day admissions decreased by 2.1% ($p<0.001$, adjusted OR=0.979); this is apparently due to longer admission periods that correlate with older age. Second, when the insured patients were members of the main HMO, the odds of single-day admission decreased by 5.8% ($p<0.001$, adjusted OR=0.942), and finally, for male patients, the odds of single-day admission increased by 5.7% ($p<0.001$, adjusted OR=1.057).

Table 4. Logistic Regression on Single-day Admissions (H2)

Variables in the Equation	β	Standard Error	Odd Ratio	95% C.I. for Odd Ratio	
				Lower	Upper
History Viewed**	-.193	.008	.825	.812	.838
Age***	-.021	.000	.979	.979	.980
Insurance**	-.059	.007	.942	.929	.956
Gender**	.055	.007	1.057	1.042	1.071
Constant	-.603	.028	.547	–	–

Discussion

The main purpose of this study was to provide additional insights into the field medical informatics. In this study we shed light on the positive relationship between using medical history and improved admission decisions. Both of our hypotheses were approved as our results showed a reduction in the volume of readmissions within seven days as well as of single-day admissions, some of which are likely to be unnecessary and preventable; the results are in accordance to previous findings (Ben-Assuli et al. 2012; Cooke, Higgins and Kidd 2003; Shabtai et al. 2007). Nonetheless, alternative interpretations of our results are possible and should be examined in future research. For instance, it could be argued that viewing a patient's detailed information may lead some physicians to prolonging unnecessary hospitalization, instead of a single-day hospitalization.

Limitations and Future Research

It should be made clear that our results do not prove causality. While causality is a possibility, as implied in this paper, other possibilities should be considered for examination.

Future work thus should involve examining why lower levels of readmissions and single-day admissions were observed after using the EHR in the ED context. Possibly moderation as well as mediation effect examination may be conducted. An inter-hospital examination may also yield value, by learning the differences between different medical facilities.

References

Ather, S., Chung, K.D., Gregory, P., and Demissie, K. 2004. "The Association between Hospital Readmission and Insurance Provider among Adults with Asthma," *Journal of Asthma* 41, 709–713.

Basu, A., and Meltzer, D. 2007. "Value of Information on Preference Heterogeneity and Individualized Care," *MDM* 27, 112-127.

Ben-Assuli, O., Leshno, M., and Shabtai, I. 2012. "Using Electronic Medical Record Systems for Admission Decisions in Emergency Departments: Examining the Crowdedness Effect," *Journal of Medical Systems*. doi: http://dx.doi.org/10.1007/s10916-012-9852-0.

Cooke, W., Higgins, J., and Kidd, P. 2003. "Use of Emergency Observation and Assessment Wards: A Systematic Literature Review," *BMJ* 20, 138-142.

Goldman, R.D., Crum, D., Bromberg, R., Rogovik, A., and Langer, J.C. 2006. "Analgesia Administration for Acute Abdominal Pain in the Pediatric Emergency Department," *Pediatric Emergency Care* 22, 18-21.

Goldschmidt, P.G. 2005. "HIT and MIS: Implications of Health Information Technology and Medical Information Systems," in *Communications of the ACM 2005* 48, 68-74.

Hersh, R., and Hickam, H. 1998. "How Well Do Physicians Use Electronic Information Retrieval Systems? A Framework for Investigation and Systematic Review," *JAMIA* 280, 1347-1352.

Hripcsak, G., Sengupta, S., Wilcox, A., and Green, R. 2007. "Emergency Department Access to a Longitudinal Medical Record," *JAMA* 14, 235-238.

Laerum, H., Karlsen, H., and Faxvaag, A. 2003. "Effects of Scanning and Eliminating Paper-based Medical Records on Hospital Physicians' Clinical Work Practice," *JAMIA* 10, 588-595.

Lawson, A.E., and Daniel, E.S. 2011. "Inferences of Clinical Diagnostic Reasoning and Diagnostic Error," *Journal of Biomedical Informatics* 44, 402-412.

Ramachandran, S., Erraguntla, M., Mayer, R., and Benjamin, P. 2007, "Data Mining in Military Health Systems - Clinical and Administrative Applications," in *IEEE International Conference on Automation Science and Engineering 2007*, 158-163.

Redelmeier, A., and Shafir, E. 1995. "Medical Decision Making in Situations that Offer Multiple Alternatives," *Journal of the American Medical Association* 273, 302-305.

Schneider, E.B., Hyder, O., Brooke, B.S., Efron, J., Cameron, J.L., Edil, B.H., Schulick, R.D., Choti, M.A., Wolfgang, C.L., and Pawlik, T.M. 2012. "Patient Readmission and Mortality after Colorectal Surgery for Colon Cancer: Impact of Length of Stay Relative to Other Clinical Factors," *Journal of the American College of Surgeons* 214, 390-398.

Shabtai, I., Leshno, M., Blondheimc, O., and Kornbluth, J. 2007. "The Value of Information for Decision-Making in the Healthcare Environment," *Medical and Care Compunetics* 4, 91-97.

Stone, J., and Hoffman, G.J. 2010. Research Report "Medicare Hospital Readmissions - Issues, Policy Options and Patient Protection and Affordable Care Act," CRS Report for Congress PPACA; P.L. 111-148.

Tierney, M.W. 2001. "Improving Clinical Decisions and Outcomes with Information: A Review," *International Journal of Medical Informatics* 62, 1-9.

Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D.W., and Middleton, B. 2005. "The Value of Health Care Information Exchange and Interoperability," *Health Affairs* 24, 10-18.

Walter, S., Kostopoulos, P., Haass, A., Lesmeister, M., Grasu, M., Grunwals, I., Keller, I., Helwig, S., Becker, C., Geisel, J., Bertsch, T., Kaffine, S., Leingartner, A., Papanagiotou, P., Roth, C., Liu, Y., Erith W., and Fassbender, K. 2011. "Point-of-care Laboratory Halves Door-to-therapy Decision Time in Acute Stroke," *Annals of Neurology* 69, 581-586.

Welch, H.G., Larson, E.H., Hart, L.G., and Rosenblatt, R. A. 1992. "Readmission after Surgery in Washington State Rural Hospitals," *American Journal of Public Health* 82, 407-411.

Westbrook, I., Coiera, W., and Gosling, S. 2005. "Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions?," *JAMIA* 12, 315-321.

Mining Patient Orders to Select Point of Care Tests in Hospital Emergency Departments

Thomas Y. Lee
Institute for Business Innovation, UC Berkeley
thomasyl@haas.berkeley.edu

Esther H. Chen
Department of Emergency Medicine, UC San Francisco
esther.chen@emergency.ucsf.edu

Abstract

Crowding in Emergency Departments (EDs) leads to long waiting times that reduce service quality and can even result in death. Point of Care Testing is a strategy for reducing wait times through a service redesign that shifts the analysis of selected diagnostic tests from a central hospital laboratory to the patient's bedside. We describe a process for mining and analyzing patient data to help hospital administrators assess the suitability of a diagnostic test for conversion to POCT based upon the test's potential for improving operational performance. In this paper, we describe the approach using data from a major, urban tertiary care ED.

1. Introduction

Emergency Department (ED) overcrowding is a significant problem reportedly affecting more than 90% of all hospitals, more than 40% on a daily basis [1]. Overcrowding is known to result in long delays that result in decreased care quality and even death [2]. Point of Care Testing (POCT) is an innovation with roots in process management for decreasing service time and ED crowding [3]. In POCT, diagnostic patient samples are analyzed (at the Point of Care) in the ED rather than in a central hospital laboratory. Because the service time impacts of a diagnostic test depend upon many variables including patient population and test ordering patterns [4], the managerial challenge is to determine what test(s) to convert from lab to POCT.

In this paper, we describe a data-driven process for comparing and contrasting the suitability of laboratory tests for conversion to POCT on a hospital-specific basis. This is the first attempt that we know of to impose a more formal approach to POCT test selection. We apply text and process mining techniques to distil operational variables from patient order histories. Administrators can then rank tests by comparing and contrasting variables that influence the waiting time impacts of converting a test to POCT. We demonstrate the process on 6000 patient observations drawn from the ED of a major, urban tertiary care facility.

The remainder of this paper is organized as follows: after reviewing related work, we describe the process and illustrate it using data from one urban hospital's real POCT adoption. After discussing the managerial context and limitations, we conclude with future work.

2. Related work

POCT is not new to Emergency Medicine. Hospitals in various settings with a range of patient populations have empirically tested POCT in numerous formats (randomized control trial, case control, etc.) with inconsistent results [4]. The same test, converted to POCT in different hospitals, might decrease, leave unchanged, or even increase a patient's service time. In the same vein, a single hospital, converting different tests, produces similarly inconsistent results [3-8]. In all prior studies, the tests converted to POCT were selected on an ad hoc basis. Lack of a principled selection strategy and inconsistent empirical outcomes reinforces the need for a data driven approach to test selection.

This paper follows one theme in the literature applying well-understood mining techniques to clean and analyze real-world data [9]. Rather than ranking manholes for possible failures, we rank medical tests for possible conversion to POCT. Entity resolution and constraint-based categorical clustering techniques are used in preprocessing [10]. Sequence mining techniques are used to analyze patterns of physician test-ordering behavior that might benefit from POCT [11]. Specifically, we adapt two interestingness measures: statistical correlations [12] and information gain [13] to screen for (un)interesting sequences.

3. Approach

The process is summarized in Figure 1 and detailed below. Our goal is to provide mangers with decision support: a set of variables to rank candidate tests for possible POCT conversion. Beginning with patient records, a series of automated and manual pre-processing steps normalizes the raw data into a set of comparable entries. The normalized data is modeled as a set of sequences in a vertical data store. Data mining and knowledge discovery steps rooted in queuing theory are used to construct variables and a candidate ranking.

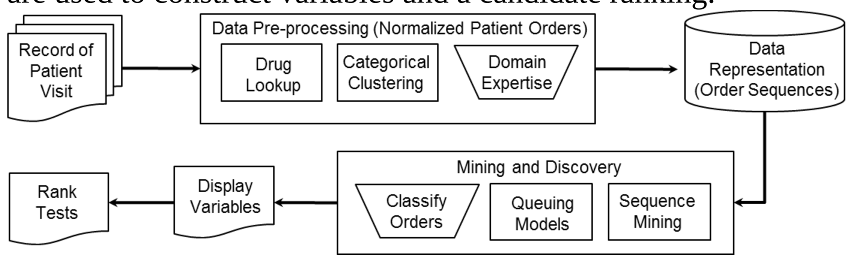


Figure 1. Process Overview

3.1. Raw data. Every patient visit is recorded in the Electronic Medical Record (EMR) system as a list of entries where each entry is a diagnostic test or medical procedure manually entered in text by the physician as an "Order Description." EMR entries are sorted first by a unique patient visit ID (Encounter No.) and second by when the order was placed (Start Time) (See Table 1).

Row	Encounter No	Order Description	Start Time
[1]	637535	knee 3 views	5/9 0:29:14
[2]	637535	Ibuprofen 600 mg po x1	5/9 3:20:24
[3]	637542	pcn vk 500 mg po Per Mouth	5/9 6:04:40
[4]	637542	percocet 5/325 mg 2 tablets po Per Mouth	5/9 6:05:12
[5]	637544	SMA7 Metabolic Panel	5/9 4:32:15

Table 1. Patient visits summarized as a history of physician orders

Row	SID	EID	Order Desc
[1]	1	1	knee 3 views
[2]	1	2	Ibuprofen …
[3]	2	1	pcn vk 500 …
[4]	2	1	percocet …
[5]	3	1	SMA7 …

Table 3. Order sequences

3.2. Data pre-processing. As illustrated in Table 1, "Order Descriptions" are entered manually and contain abbreviations, jargon, misspellings, etc. Data cleaning and entity resolution are a long-studied problem in data integration [10]. We apply a constraint-based categorical clustering used in entity resolution [14] to cluster similar orders. For example, the order "SMA7" a test for analyzing blood chemistry, is synonymous with "BMP," "Chem7," and "SMAC." As illustrated in Table 2, entity resolution begins by tokenizing Order Descriptions. We then cluster similar orders (Table 2 rows (1) – (4)). An online pharmaceutical dictionary is used to check spellings

Row	Normalized	Encounter No	Tokenized Order Description					
[1]	KAYEXALATE	637134	Kaexylate	30	g	PO	Per	Mouth
[2]	KAYEXALATE	637325	kayexalate	Per	Mouth			
[3]	KAYEXALATE	638474	Kayexalate	30	grams	PO		
[4]	KAYEXALATE	639965	30	cc	keyaxalate	po	Per	Mouth
[5]	LACTULOSE	620442	lactulose	30	mg	po	Per	Mouth

Table 2. Pre-processing raw order descriptions

(Table 2 rows (1) and (4) show misspellings of Kayexalate) and map between trade names and generic names (e.g. Sodium polystyrene is the generic equivalent of Kayexalate). A final check is performed by domain experts to validate transformations.

3.3. Data representation. The raw EMR data (§3.1) is modeled hierarchically as a set of Order Sequences. All tuples sharing the same Encounter No. are mapped to a single sequence and uniquely identified by a Sequence ID (SID). Every sequence therefore corresponds to a distinct patient visit. Each Order Description, Start Time tuple of a patient visit is mapped to its own sequence element (EID) and distinguished by an (SID,EID) pair. Temporally adjacent EMR entries of one SID are merged into a common EID when their corresponding Start Time(s) are within a parameterized window (e.g. two minutes) of one another. The intuition for the window is of a physician entering multiple orders "at the same time" while allowing for some interruptions during order entry. Table 3 shows the order sequences for the EMR entries in Table 1. Table 1 rows (3) and (4) are mapped to a single (SID, EID) in Table 2 rows (3) and (4) because they were entered within two minutes of one another.

3.3. Mining and discovery. We draw upon project management and basic queuing theory to identify a set of variables for evaluating POCT test candidates. For a single server system supporting one type of order (x) in steady state, if p is the average service time for order (x) and a is the average interarrival time of order (x), the waiting time $T(x)$ is calculated as:

$$T(x) = p \times \frac{p}{a-p} \times \frac{1}{2}\left(\left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_p}{p}\right)^2\right) \text{ [15]}$$

where the second term $p/(a\text{-}p)$ represents utilization and the third term represents variability (the coefficient of variation in arrival time and service time respectively. Although not directly applicable to the ED setting, this basic model motivates our first two variables:

Service time. POCT is intended to reduce service times by relocating the point of analysis. Orders that already experience short absolute service times are likely to promise less in overall patient experience. For each normalized order description, we calculate p and σ_p based upon the Start Time and End Time (End Time records when lab results are returned to the ED; End Time is not shown in Tables 1 and 2 for parsimony).

Interarrival time. The impact of POCT on ED service performance is attenuated by order frequency. A test that is seldom ordered will have limited impact on patient waiting times. We calculate the interarrival time between normalized order descriptions based upon Start Time.

While a helpful heuristic, the basic queuing model assumes that service time is defined by process(es) on the critical path [15]; yet this may not hold in the ED. In this context, an order is on the critical path if it has discrimination value (it results in a disposition decision such as "admit" or "discharge."). Discrimination value is challenging to predict because the same test (e.g. SMA7) may only discriminate in certain cases. Instead we use SPADE, an algorithm which uses a lattice theoretic approach to search for frequent item sets and sequences [16]. We use item and sequence frequency to identify orders that are less likely to have discrimination value:

Within element discrimination. For normalized order descriptions with the same EID, we down rank the discrimination value of an order in a frequent item set because it is (always) ordered with something else. The traditional metric of item set significance is *support*. However, our data reflects a highly skewed frequency distribution (figure omitted for space); support as a measure is confounded by highly skewed frequency distributions [13]. Instead, we use the Jaccard Index as a measure of correlation within a frequent 2-set [12, 17]. For an order x and all frequent 2-sets (x,y), we define:

$$Within(x) = 1 - \max\left(\forall_{frequent(x,y)}\left(Jaccard(x,y)\right)\right).$$

High correlation suggests low discrimination potential so we use the *max* of all frequent 2-sets as a bound on orders unlikely to discriminate.

Between element discrimination. Between orders of different EIDs within the same SID, we might infer discriminatory power (causality) between the preceding order and the order that follows. However, the discriminatory power of a preceding order is diminished if that preceding order is instead part of a frequent, preceding pair. For orders *x*,*y*, and *z* where <*x*, *z*> is a frequent sequence such that *x* and *z* are in different EIDs of the same SID, and (*x*,*y*) is a frequent 2-set, we define:

$$Between(x) = 1 - \frac{\max\left(\forall_{frequent\langle(x,y),z\rangle}\left(Support(\langle(x,y),z\rangle)\right)\right)}{Support(\langle x,z\rangle)}.$$

The measure is related to *confidence* in an association rule where high confidence suggests low discrimination potential so we again use the *max* of the support as a bound on discrimination.

Order classification. The order sequence includes every billable activity. This includes the administration of drugs, consultations from sub-specialists (e.g. orthopedists, cardiologists), and tests that cannot be converted to POCT either because they are already performed at the bedside (e.g. electrocardiograms) or have no bedside equivalent (computerized tomography). With manually classify orders as POCT candidates or non-candidates.

PPS (POCT Propensity Score). Absent an underlying queuing model to empirically validate and training data to learn a ranking function, we suggest a weighted linear combination as a preliminary summary measure from which administrators can use sensitivity analysis to test the robustness of a resulting ranking. Higher values indicate greater potential for waiting time impact from conversion to POCT. To scale all inputs, we sum the Z-scores of each variable (e.g. *Arrival*(*x*) as the normalized measure of arrival rate).

$$PPS(x) = \propto_1 Service(x) + \propto_2 Arrival(x) + \propto_3 Within(x) + \propto_4 Between(x) \text{ s.t. } \sum_{\forall i} \propto_i = 1$$

4. Evaluation.

To evaluate the approach, we apply the technique to actual patient data from a major, academic urban hospital with an annual census of about 55,000 ED patients per year. Our data set, taken directly from the ED's Electronic Medical Record (EMR), consists of a patient ID, a visit ID, a free text order description, start time and end time. The 6,403 visits between 5/9/2008 and 7/8/2008 resulted in 60,330 entries with 18,838 unique order descriptions. After pre-processing, the orders are normalized to 603 distinct order labels. Table 4 reports results for the 20 most frequent orders. Order classification eliminated non-POCT candidate orders like CT scans. For this hospital, Urinalysis, Troponin, Beta-HCG, and Type screen are likely POCT candidates.

To validate our sequence mining, we compared our mining results to the self-reported order practices of an expert panel and find that we correctly identified the most frequent pairing (*Jaccard*(*x*) > .25) with accuracy of 0.91. Although the *PPS*(*x*) ranking had face validity to the same panel of experts, more interesting is the comparison to actual managerial outcomes in both our test setting and in other published studies. In our test setting, Troponin was converted from the central lab to POCT (though not as a result of this study). We also know that subsequent reductions in patient service time were attributed to the test conversion and that all patients, regardless of whether they received the Troponin test or not, experienced reductions in overall ED length of stay of 30 minutes or more [18].

Morever, while every ED's patient population is different, we know that findings from published studies on the operational impact of POCT are consistent with our test site rankings. For example, studies documenting the POCT conversion of the Chem7 test either reported no

ROW	ORDER	SERVICE(X)	ARRIVAL(X)	SUPPORT	JACCARD	WITHIN(X)	BETWEEN(X)	PPS(X)
[1]	AMYLASE	-0.68	-1.01	0.11	0.87	-1.08	1.00	-0.545
[2]	BETA-HCG	0.51	0.06	0.20	0.44	0.44	1.00	0.400
[3]	CBC	0.14	1.74	0.48	0.85	-1.00	0.16	-0.144
[4]	CHEM7	-0.55	1.41	0.48	0.85	-1.00	0.09	-0.441
[5]	LFT	-0.35	-0.66	0.14	0.25	1.13	1.00	0.176
[6]	LIPASE	-0.69	-0.94	0.12	0.67	-0.36	1.00	-0.351
[7]	MAG PHOS LEVEL	-1.00	-0.77	0.15	0.98	-1.45	1.00	-0.656
[8]	PT & PTT	-0.70	-0.14	0.22	0.38	0.66	0.10	-0.444
[9]	TROPONIN	0.36	-0.31	0.15	0.29	0.99	1.00	0.406
[10]	TYPE SCREEN	0.40	-0.64	0.12	0.21	1.24	1.00	0.394
[11]	URINALYSIS	2.56	1.24	0.20	0.44	0.44	1.00	1.206

Table 4. Results at support .05

change in overall ED length of stay [5, 6] or a one hour decrease in ED length of stay for discharged patients but no change for admitted patients [7]. By contrast, [8] finds that converting the Beta-HCG test resulted in a 40 minute decrease in overall ED length of stay for test patients. Our scoring likewise suggests that Beta-HCG is a promising POCT candidate.

5. Discussion.

The proposed approach to decision support for managerial selection of POCT test candidates has a number of strengths and limitations. First and foremost is a data driven approach that hospital administrators can tailor to their own, unique order distributions, patient populations, seasonality, staffing decisions, etc. Traditional queuing models offer limited support [18], perhaps in part because service time assumptions overlook critical path issues for which we sequence mine. Second, as a hallmark of good decision support, it is easy to explain the rationale underlying the rankings. In Table 3, CBC and Chem7 tests have the highest frequency explaining why many empirical studies focused on them. However, *Jaccard*(x) and *Within*(x) reveal that these tests are rarely ordered alone, reducing their discriminative value and explaining the limited impact from their POCT conversion [3-7]. Mag Phos Level shows low support but very high Jaccard, highlighting the limits of using support as a metric with skewed distributions. Third, the process is modular. The Affordable Healthcare Act promotes EMR adoption intended to eliminate pre-processing concerns like ours. Yet it is worth noting that the EMR is no panacea; our test site was an EMR early adopter.

At the same time, the approach does have a number of limitations. First, the ranking function omits cost. POCT investments to improve ED waiting/service time are prima facie justified by virtue of improved patient care. In our test setting, the medical director intentionally omitted cost in the performance analysis of POCT [18]. Because of Little's Law, patient volume is unlikely to increase due to POCT alone. However, decreased bouncebacks (repeat visits for the same complaint) and shorter wait times should improve patient satisfactory scores governing Medicare reimbursement. Moreover, prior studies suggest the potential for cost parity due to scale economies between POCT and the central lab [3]. Second is staffing vs. bed count. Many operations models of the ED assume that bed counts, not staffing, are the bottleneck. In hospitals where staffing is a bottleneck, POCT conversion will only increase the burden on a constrained variable with no slack. Third is the lack of guidance on calibration. The approach relies on two sets of parameters: [a] we set a minimum support threshold of .05 to prune frequent item sets and frequent sequences [b] we equally weight each variable in the linear combination. Absent a theoretical model, at best, sensitivity analysis can suggest tolerances surrounding our variable parameters. Fourth, the model assumes that managers select only one test to convert. Subject to cost and staffing, managers might convert more than one test. Our approach generalizes to mining for discriminating test *sets* allowing for potential gains from converting *multiple* tests to POCT. The danger is the slippery slope of converting too many tests, thereby diluting service time benefits by overloading staff at the point-of-care.

6. Conclusions and future work.

In this paper, we introduce a data-driven process to display and summarize statistics to support the managerial selection of a POCT trial. We combine several previously studied techniques in a novel way and introduce a straw man ranking function that provides transparency into the selection process (rebalance alpha) while supporting extensions such as cost considerations. As noted above, additional research is needed to develop a more principled summary statistic while testing the current model's sensitivity to linear weighting and queuing model assumptions like staffing. Finally, we are working to acquire new hospital data to test the generalization of our pre-processing and data/text mining for within and between-element discrimination.

7. References.

[1] G. Warden, G. et al. *Hospital-Based Emergency Care: At the Breaking Point*, 2006. 1-314.

[2] S. Bernstein, et al.,"The Effect of Emergency Department Crowding on Clinically Oriented Outcomes." *Academic Emergency Medicine*, 2009, 16(1): 1-10.

[3] W. Tsai et al. "Point-of-care versus central laboratory testing: an economic analysis in an academic medical center." *Clin Ther*. 1994, Sept-Oct, 16(5): 898-910.

[4] R. Taylor, "Point-of-care testing in the Emergency Department," in *Point of Care Testing, 2nd ed* C. Price, A. St John and J. Hicks, ed, Am Assoc for Clinical Chemistry, 2004.

[5] C. Parvin, "Impact of point-of-care testing on patients' length of stay in a large emergency department." *Clin Chem*. 1996 May, 42(5):711-7.

[6] J. Kendall, B. Reeves, and M. Clancy, "Point of care testing: randomized controlled trial of clinical outcome." *BMJ*. 1998 Apr 4, 316(7137): 1052-7.

[7] R. Murray et al., "Effect of point of care testing on length of stay in adult emergency department." *J Emerg Med*. 1999 Sep-Oct; 17(5):811-4.

[8] W. Pierhoples, F. Zwemer, and J. Bazarian, "Point of care pregnancy testing provides staff satisfaction but does not change ED length of stay." *Am J Merg Med*. 2004 Oct; 22(6):460-4.

[9] C. Rudin, R. et al. "A process for predicting manhole events in Manhattan," *Machine Learning*, 80(1) 2010.

[10] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Information systems*, 25(5) 2000.

[11] R. Srikant, "Fast algorithms for mining association rules and sequential patterns," U Wisconsin, 1996.

[12] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in *ACM SIGMOD Record*, 1997, vol. 26, pp. 265–276.

[13] G. Berger and A. Tuzhilin, "Discovering unexpected patterns in temporal data using temporal logic," *Temporal Databases: research and practice*, pp. 281–309, 1998.

[14] T. Lee and E. Bradlow, Automated Marketing Research Using Online Customer Reviews, *JMR*. 2011 48(5).

[15] G. Cachon and C. Terwiesch, *Matching Supply with Demand 2nd ed*. Boston: McGraw Hill Irwin, 2009.

[16] M. J. Zaki, "SPADE-An efficient algorithm for mining frequent sequences.pdf," *Machine Learning*, vol. 42, pp. 31–60, 2001.

[17] W. Jones and G. Furnas, "Pictures of Relevance: A Geometric Analysis of Similarity Measures," *JASIS*, 1987.

[18] N. DeHoratius, T. Lee, T. Olsen, L. Song, and E. Chen, "Point-of-Care Testing: Improving Emergency Department Performance Through Process Redesign," *SSRN* http://ssrn.com/abstract=2160044, 10/11/2012.

EVOLVING DECISION STRATEGIES FOR DYNAMIC ENVIRONMENTS: A GENETIC PROGRAMMING APPROACH

Georg Meyer[1], Paul E. Johnson[1], Gediminas Adomavicius[1], Patrick O'Connor[2], JoAnn Sperl-Hillen[2], William Rush[3], Mohamed Elidrisi[4], Sunayan Bandyopadhyay[4]

[1]Information and Decision Sciences, University of Minnesota, [2]HealthPartners Medical Group, [3]HealthPartners Research Foundation, [4]Computer Science, University of Minnesota

{meye1131, johns021, gedas, elidr002, band0064}@umn.edu,
{Patrick.J.Oconnor,JoAnn.M.SperlHillen,William.A.Rush}@HealthPartners.com

Abstract: *A decision strategy is systematic way of choosing among alternatives or eliminating options in order to arrive at a goal. Individuals apply decision strategies in dynamic environments that require repeated decision making where decisions are path-dependent, time-constrained, and the environment changes not only in response to the actions taken by the decision maker but also autonomously. In addition to being used by individual agents, decision strategies are found in organizations in the form of policies, guidelines, and algorithms. In this paper, we consider decision strategies (specifically in the context of chronic disease care) as the product of an evolutionary process and outline an approach to develop strategies by means of a computational process based on genetic programming. By evolving strategies in a simulated environment, we find that the evolutionary process is capable of developing strategies comparable to those found in practice. Moreover, the proposed approach can serve as a useful tool for analyzing various incentive structures and identifying their potential unintended consequences.*

Keywords: dynamic decision making, strategy development, genetic programming, simulation.

1. Introduction and Motivation

An environment is *dynamic* if it requires a series of decisions to be made, the decisions are path-dependent, and the environment changes in response to actions implemented based on previous decisions as well as autonomously [1]. In such environments (e.g., healthcare, emergency management, air traffic control), individuals and organizations adopt decision strategies, i.e., systematic ways of eliminating or choosing among alternative options [2], [3]. In practice, these strategies are encoded in policies, guidelines, and algorithms. Performance of decision strategies in dynamic environments can be further improved using strategy modification [4]. In this paper, we propose an approach to generating and modifying decision strategies, and demonstrate it in a healthcare context. We build on the genetic programming paradigm from computer science [5], the objective of which is to enable computers to assemble programs (for solving a given problem) by means of an *evolutionary* process rather than coding the solutions manually.

In particular, we consider decision strategies as the subject of an evolutionary process. The fitness of a strategy is defined in terms of outcomes. The representation of a decision strategy as an object to evolve will be discussed in the next section. Throughout the paper, for the purpose of illustration, we use the example of treatment strategies for chronically ill patients. The objective of treatment is to reduce patients' risk of *cardiovascular events*, such as heart attacks and strokes. Three chronic conditions, hyperglycemia (elevated blood glucose associated with diabetes), hypertension (elevated blood pressure), and dyslipidemia (unhealthy cholesterol levels), are major contributors to cardiovascular risk. Patients typically suffer from multiple of these conditions at a time, which increases the challenge of treatment because a physician can often not treat more than one of these conditions at a given visit (for reasons of medical safety, patient concerns, competing demands [6]). Therefore, it is desirable to construct strategies that prioritize the treatment of these conditions in an advantageous way for each individual patient.

2. An Evolutionary Approach to Developing Decision Strategies

For the purpose of this study, we evolve decision strategies that decide what condition to treat and when to schedule the next visit (number of days). Once a condition is selected, the specific treatment move (e.g., what medication to use) is taken from clinical guidelines [7]. A strategy takes a patient state as input and produces actions as outputs. This can be represented graphically as a decision tree, as shown in Figure 1, with rectangles and arrows representing conditions and hexagons representing actions.

2.1. Genetic Representation of Decision Strategies

A decision tree as shown in Figure 1 can be transformed into a set of condition-action rules [8]. We define a *decision strategy* as a set of rules that adheres to the constraints summarized in extended Backus-Naur form (EBNF) [9] in Figure 2. Each rule contains a predicate of the form $(X > Y)$? where X and Y represent expression-to-evaluate elements. If the predicate evaluates to true, the first action of the rule (branch true) will be executed, otherwise the second action (branch false) will be executed.

The specific nature of the decision strategy is defined by its primitives, which are inputs, outputs, and functions. E.g., a treatment strategy takes a patient state as input, and returns the condition to treat and when to schedule the next visit. The input is represented as a vector of state variables, including demographics (age, gender, ethnicity), health states (weight, blood glucose level, blood pressure, cholesterol, creatinine), and treatments (number and doses of medications). A state variable primitive primitive-state-var represents one of these variables. The output is presented as a tuple (condition-to-treat, next-visit) where condition-to-treat can be **B**=blood pressure, **G**=glucose, **L**=lipids (cholesterol), or **N**=nothing (i.e., take no action at this visit). Next-visit is the number of days until the next visit is scheduled. (Due to space constraints, the primitives are not fully expanded in EBNF but briefly discussed here conceptually.) Functions are used to modify inputs in order to compare them in predicates. Table 1 illustrates the primitive functions primitive-state-function, primitive-state-action-function, and primitive-action-function used in a treatment strategy. Examples of predicates that use these functions are given in Table 2.

The choice of functions impacts what type of decision strategies can be evolved. For example, Cur() is necessary for information gathering purposes (e.g., to assess the current state of a patient), and Exp() is useful for including expectations (e.g., of what an action will do) into the decision making process.

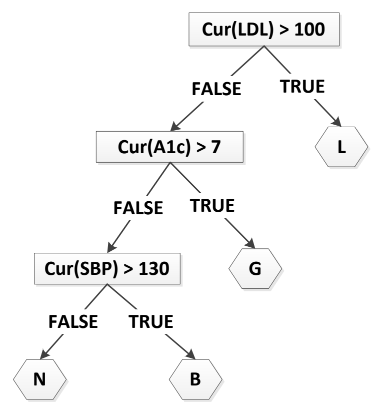


Figure 1. Example representation of a serial treatment strategy (L = treat lipids, G = treat glucose, B = treat blood pressure, N = do nothing).

strategy := rule
rule := predicate action action
action := primitive-action | action-func | rule
predicate := expression-to-evaluate expression-to-evaluate
expression-to-evaluate := state-function | state-action-function | literal
state-function := primitive-state-function (expression-to-evaluate | primitive-state-var) (expression-to-evaluate | primitive-state-var)?
state-action-function := primitive-state-action-function (state-action-function | primitive-state-var) (action-function | primitive-action)
action-function := primitive-action-function action-function?

Figure 2. Extended Backus-Naur Form (EBNF) representation of key elements of a decision strategy.

Table 1. Conceptual description of the primitive functions available to a decision strategy.

Function	Type	**Description**
Cur(X)	state-function	Current value of a state variable (age, A1c, SBP, creatinine, etc.)
Exp(X, a)	state-action-function	Expected value based on an action a
Goal(X)	state-function	Goal value for a state variable
Prev(X)	state-function	Value of a state variable at time t-1
PrevAct()	action-function	Last action (taken at time t-1)
Add(X,Y)	state-function	Add X and Y
Sub(X,Y)	state-function	Subtract Y from X
Mul(X,Y)	state-function	Multiply X and Y
Div(X,Y)	state-function	Divide X by Y, replace Y with ε if Y = 0

Table 2. Examples of predicates.

Condition	Functional Expression
Is the patient currently above goal for A1c?	Cur(A1c) > Goal(A1c)
Was the patient above goal at the last encounter?	Prev(A1c) > Goal(A1c)
Is the patient expected to be above the SBP goal if a BP move is made?	Exp(SBP, B) > Goal(SBP)
Is the patient expected to have an A1c above 9 if a BP move is made?	Exp(A1c, B) > 9
Is the current A1c below what is expected based on the last move?	Exp(Prev(A1c), PrevAct()) > Cur(A1c)

2.2. Fitness and Selection

Consider Figure 3a below, which represents a simplified illustration of the decision-making context. The horizontal plane represents a patient's health state – x-axis represents a patient's blood glucose, the y-axis represents cholesterol. The z-axis is a function of the patient state and provides the patient's risk of experiencing an adverse cardiovascular event, e.g., a heart attack or stroke, in the next 10 years. Patients with high levels of blood glucose and cholesterol have significantly higher risk of experiencing such an event than patients with low levels. The objective of treatment is to reduce a patient's risk by lowering blood glucose and cholesterol (and blood pressure, not displayed here to keep the figure simpler). Treatments can be plotted as a pathway along the risk surface, and different strategies result in different treatment paths.

For a given amount of time, each path produced by a treatment strategy results in a risk level for the patient and has associated treatment costs. For example, a strategy that treats very little will likely result in higher risk and lower cost compared to a strategy that treats aggressively. Figure 3b represents the cost/benefit trade-off, which, in the evolutionary approach defines the *fitness* of a treatment strategy. The dashed lines represent "isobars" of cost-effectiveness, i.e., two strategies on the same bar are equally cost-effective when trading off the cost of treatment compared to quality of life valuations and expected cost of complications [10]. More cost-sensitive environments (e.g., health systems) may have steeper slopes of the dashed lines. In contrast, if the only objective is to reduce risk regardless of cost, the dashed lines would be parallel to the x-axis in Figure 3b.

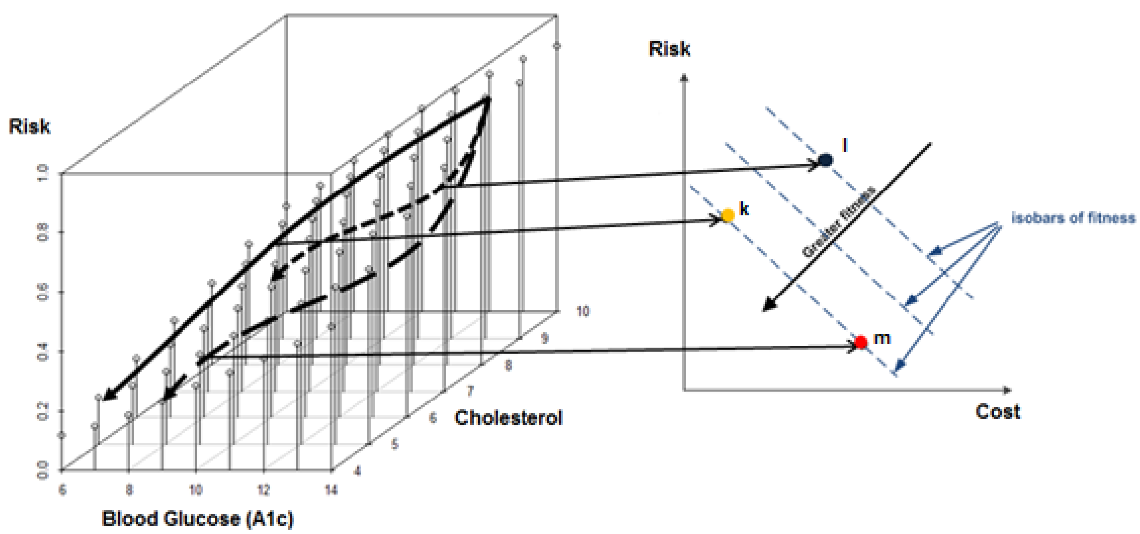


Figure 3a. Example of three treatment paths. **Figure 3b. The resulting risk and costs of treatment paths.**

2.3. Evolutionary Process

Step 0. Generate initial population. An initial population of decision strategies is generated by randomly assembling predicates and actions and combining them in a tree that satisfies the outlined grammar. Predicates $(X > Y)$? are generated by randomly selecting state variables, random constants, or functions for both X and Y. Actions are generated by randomly choosing a treatment type and a scheduling interval (next-action-time). A predicate or action is chosen as the root node for a tree. If the node is a predicate, a predicate or an action is chosen for the true branch, and another for the false branch. This process is repeated recursively for all nodes which are predicates. To keep the evolved strategies parsimonious as the decision trees grow larger, the choice whether an action or predicate is selected for a given node is stochastically biased to favor actions, which limits the growth of the trees. The generation of a strategy ends once all leaf nodes are actions.

Step 1. Selection: Fitness Evaluation. Each strategy treats a random sample of 1,000 simulated patients for one year. After treatment, the risk of cardiovascular events and the cost of treatment are computed for each patient. Risk and cost determine a strategy's fitness. Strategies are then selected for survival or reproduction according to their fitness by tournament selection, a process that selects strategies with likelihood proportionate to fitness [11].

Step 2a. Survival. A selected strategy survives into the next generation in unaltered form.

Step 2b. Recombination and Mutation. Two selected strategies beget a new strategy by recombining their genetic material. With a small probability, the resulting strategy will be subject to a mutation, a random change in its genetic code (for example, switching two predicates, changing the left hand and right hand side, adding a new rule or randomly deleting a rule). Examples of recombination and mutation are illustrated in Figures 4a and 4b.

Each strategy in the new generation (resulting from Step 2) now treats 1,000 randomly sampled simulated patients for one year and is evaluated at the end of the treatment period. The process returns to Step 1.

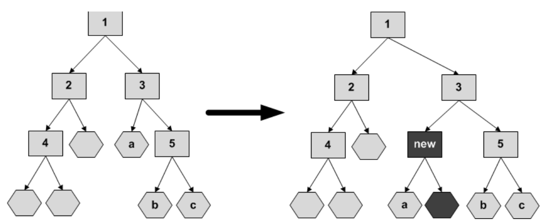


Figure 4a. Example of a mutation that adds a new rule.

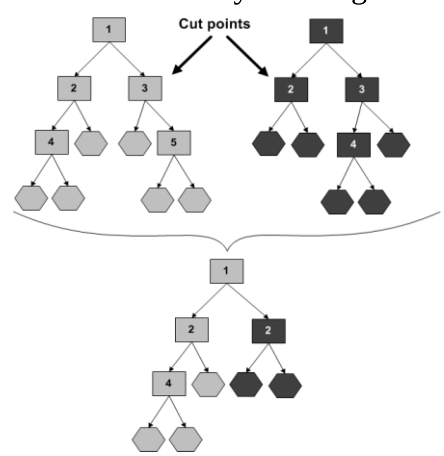


Figure 4b. Example of recombination.

3. Experiments

To demonstrate the proposed approach, we apply it to construct treatment strategies for chronic disease patients. To test the evolved strategies, we use a patient simulator, Simcare, that models patients with type 2 diabetes mellitus, chronic hypertension, dyslipidemia, and the effects of various treatments on patients [12], [13]. Simcare has been validated against randomized clinical trials and individual clinical patient cases [14]. Using Simcare, a population of simulated patients is created based on a snapshot of health characteristics of 18,356 patients under care at a US healthcare system. These patients are at various stages of treatment (no treatment, diet and exercise only, oral medications, insulin).

Using the evolutionary process described above, we generate 100 initial strategies, randomly select 1,000 patients and treat them for one year with each strategy S, and assess the fitness of S by calculating number of events prevented (based on risk estimated by the UKPDS Risk Engine and Monte Carlo simulation [15]) and additional treatment costs resulting from S ("additional" meaning the cost resulting from treatments prescribed by the strategy, ignoring cost of treatments the patient is already on). We consider two scenarios: (A) a purely risk-based fitness function that aims to maximize the number of events prevented, and (B) a cost-sensitive fitness function that aims to minimize cost of additional treatment per prevented event [15], [16].

Figures 5a and 5b show the results as the number of events prevented (solid line) and cost of treatment (dashed line) for the *best strategy* (i.e., with greatest fitness) in each generation. In scenario A, in the first 40 generations, there is little change in the prevention rate but actions that tend not to result in significant risk reductions, such as multi-medication blood pressure treatments, are eliminated from the genetic code of candidate strategies due to selection. Once wasteful actions have been removed, recombination yields strategies that make more effective moves, and the performance of the best strategy increases considerably in later generations. Occasional dips in performance are explained by the fact that a *new* random sample of patients is treated in every generation, which is important for properly representing variation in the environment and preventing over-fitting. In Scenario B, shown in Figure 5b, costly treatments are removed over time from the genetic code with little effect on event rate. This includes specifically later-stage blood pressure and glucose treatments (using multiple medications).

Analysis of the evolved decision trees reveals that the most successful risk-reducing strategy is a heuristic that focuses on lipids and glucose (in that order) and schedules more treatment opportunities for patients with very high levels of glucose and blood pressure. In several generations, through mutation and recombination, more complex strategies (with larger combinations of rules) evolved that performed better than these heuristics in a single generation, but failed to repeat their performance in following generations (when they were exposed to a different sample of patients) and as a result were selected against.

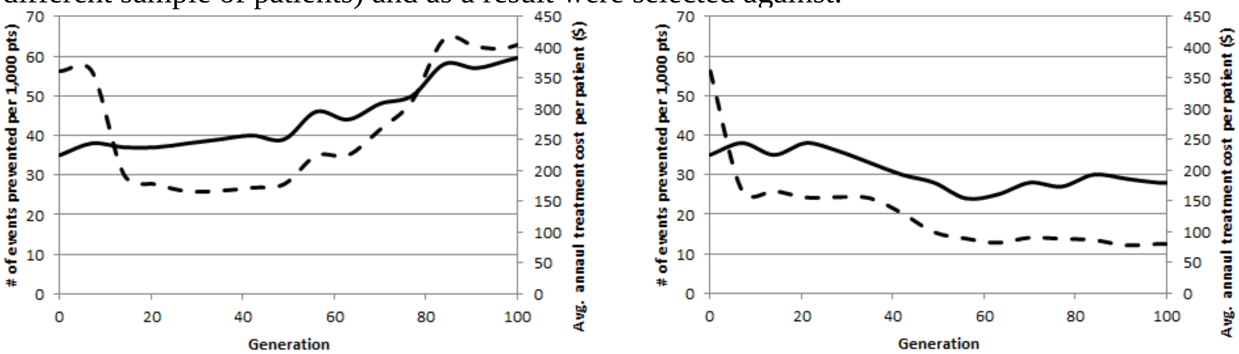


Figure 5a (left) and 5b (right). Prevented events (solid) and additional cost of treatment (dashed) for the best strategy in each generation in Scenario A (left) and Scenario B (right).

When using a cost-effectiveness fitness function, age and gender, which are both major risk factors, became dominant factors in guiding decision making in later generations. The best strategy focused primarily on giving lipids treatments (prescribing statins, which are generic and inexpensive) to older males. Remarkably, the resulting strategy is similar to a finding in medical literature about when it is cost-effective to start statin treatments [17]. However, we would like to point out that this strategy makes considerable savings in treatment costs by failing to treat groups of individuals that are above evidence-based goals [18]. This is demonstrated in Figure 6 which shows the proportion of patients that received some treatment adjustments from the best strategy in every generation. All patients in the population are in need of treatment, i.e., are above evidence-based goals. In the risk-based scenario, the best strategy treats every patient in later generations. In contrast, in the cost-effectiveness scenario, only 58% of the population receives treatment adjustments in the later generations. Especially younger patients are treated less because their baseline risk is lower due to their age. For example, a 45 year old male's heart attack risk can be lowered from 9.5% to 5.6% with statin treatment, the

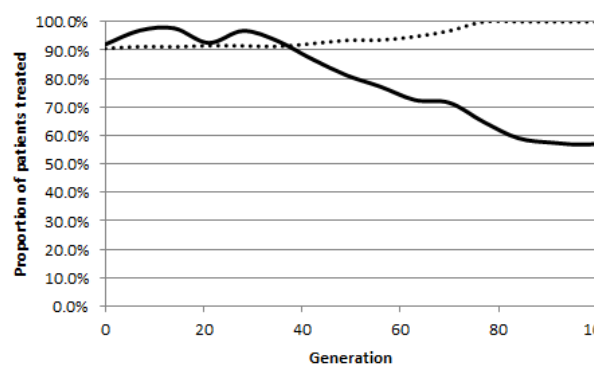


Figure 6. Patients (%) receiving treatment adjustments from best strategy in Scenario A (dotted) vs. B (solid).

same treatment will lower risk from 34.2% to 21.5% for a patient of age 70, all else being equal [15]. In the next section, we briefly discuss the implications of these results and, more generally, the contributions of the proposed approach.

4. Conclusions

In this research, we demonstrate an approach to developing decision strategies by means of an evolutionary process. We have outlined a grammar for describing decision strategies in general, and provided specific primitives to describe an example, i.e., treatment strategies for chronically ill patients. In this context, the approach evolved a successful heuristic strategy for prioritization and scheduling to reduce risk for patients. Moreover, it was found that a strategy evolved a form of discrimination against lower-risk patients (i.e., failed to adjust treatment even when patients were not at evidence-based goals) in cost-sensitive environments. If acting in an environment is costly, and selection pressure opposes costly strategies, strategies develop branches for which they take no action. In other words, strategies which take insufficient or excessive action may be an unintended consequence of the incentive structure of their environments. Therefore, in addition to generating well-performing decision strategies, the approach developed in this paper (combined with simulation) can serve as a useful tool for analyzing the driving forces behind the best performing decision strategies (i.e., the incentive structures modeled via fitness functions), potentially identifying their unintended consequences, and exploring their various modifications.

In summary, the evolutionary process developed in this paper makes promising contributions both to the explanation of why we observe certain decision making behaviors in practice as well as to the discovery of decision strategies with greater fitness for a given environment.

References

[1] B. Brehmer, “Strategies in Real-Time, Dynamic Decision Making,” in Insights in Decision Making, Chicago, IL: The University of Chicago Press, 1990.

[2] H. A. Simon, The Sciences of the Artificial. Cambridge, MA: The MIT Press, 1969.

[3] L. R. Beach and T. R. Mitchell, “A Contingency Model for the Selection of Decision Strategies,” *Academy of Management Review* 3(3), pp. 439–449, 1978.

[4] G. Meyer et al., “A Machine Learning Approach to Improving Process Control,” WITS 2009.

[5] J. R. Koza, Genetic Programming. Cambridge, MA: The MIT Press, 1992.

[6] M. L. Parchman, J. A. Pugh, R. L. Romero, and K. W. Bowers, “Competing Demands or Clinical Inertia: The Case of Elevated Glycosylated Hemoglobin,” Annals of Family Medicine, vol. 5, no. 3, pp. 196–201, 2007.

[7] R. S. Mazze et al., *Staged Diabetes Management*. Wiley-Blackwell, 2012.

[8] J. R. Quinlan, “Generating production rules from decision trees,” in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 1987, pp. 304–307.

[9] International Organization for Standardization, “ISO/IEC 14977:1996 Information technology -- Syntactic metalanguage -- Extended BNF.” Geneva, Switzerland, 1996.

[10] P. Clarke et al., “A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model,” *Diabetologia* 47(10), pp. 1747–1759, 2004.

[11] B. L. Miller and D. E. Goldberg, “Genetic Algorithms, Tournament Selection, and the Effects of Noise,” *Complex Systems* 9, pp. 193–212, 1995.

[12] P. Dutta et al., “SimCare: A Model for Studying Physician Decisionmaking Activity,” in *Advances in Patient Safety: From Research to Implementation*, AHRQ, 2005, pp. 179–192.

[13] R. McCabe et al., “Validation of the SimCare Model: A Computational Model of Individual Patients with Type 2 Diabetes,” *Diabetes* 59(Suppl 1), p. A353, 2010.

[14] R. McCabe, “Coherence and correspondence validation of the SimCare patient model,” U. of Minnesota, 2012.

[15] R. Stevens, V. Kothari, A. Adler, I. Stratton, R. Holman, “The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56),” *Clinical Science* 101, pp. 671–679, 2001.

[16] R. Chapman, P. Stone, E. Sandberg, C. Bell, P. Neumann, “Comprehensive League Table of Cost-Utility Ratios and a Sub-table of ``Panel-worthy'' Studies,” *Medical Decision Making* 20, pp. 451-458, 2000.

[17] B. T. Denton, M. Kurt, N. D. Shah, S. C. Bryant, and S. a Smith, “Optimizing the start time of statin therapy for patients with diabetes.,” *Medical Decision Making* 29, pp. 351–67, 2009.

[18] G. Ramsey, “Success and Failure in Dynamic Decision Environments: Understanding Treatment Strategies for Patients with a Chronic Disease,” University of Minnesota, 2010.

Managing Agile Software Development: A Control-Theoretic Approach

Subodha Kumar
Texas A&M University
subodha@tamu.edu

Yonghua Ji
University of Alberta
yji@ualberta.ca

Vijay S. Mookerjee
University of Texas at Dallas
vijaym@utdallas.edu

Abstract

Agile software development practices, such as *extreme programming*, call for intense participation by users to ensure that the systems are developed to meet current needs, yet be flexible enough to adapt to future needs. Hence, in such projects, the developers need to allocate their efforts optimally between understanding the requirements from users and developing the modules, throughout the development period. We propose an optimal control model to solve this problem with two different objectives: (a) maximizing the system value, which is the objective of user group, and (b) minimizing the development cost/time, which is the objective of development group.

1. Introduction

One of the most well documented reasons for IT project failure is the lack of understanding of user requirements (Field, 1997), with as much as 28 percent of large projects reportedly failing to meet the needs of the end users (Standish Group, 1995). Results of empirical studies show that the requirements errors are the most costly in software development (Boehm, 1981). Hence, eliciting user inputs and ensuring user involvement has long been considered a critical component in the success of software projects. Participation by users in articulating requirements reduces the likelihood that unimportant or unacceptable features get included in the system; thus increasing the value that users derive from the system (Baroudi et al. 1986).

The recent studies have empirically shown that the agile development methodologies, such as extreme programming (XP), improve software quality (Maruping et al. 2009). Such results have contributed to the increasing popularity of these methodologies (Dingsøyr et al. 2012). However, these methodologies introduce a new dimension to user participation in software development. Unlike earlier development methodologies where users are queried about requirements at the start of the project, users participate throughout the development process in XP. In these projects, continuous user involvement is encouraged so as to improve the ability of both users and developers to better conceptualize the system (Astels et al. 2002) and thus reduce the chances that faulty or incomplete user requirements are communicated to developers. Since developers spend less time on the initial planning, they rely on rapid feedback from the users to permit changes in code as requirements emerge (Beck 2000).

In this paper, we consider the problem of allocating developer effort between developing the system and understanding the requirements, with the objective of maximizing value obtained from the system as well as minimizing cost incurred in developing the system. Past studies show that the process of understanding user requirements becomes more efficient for the development group as it develops more modules because of (i) better understanding of the business domain (Davis 1992), and (ii) knowledge dissemination among developers (Dawande et al. 2008). A novel aspect of our paper is to incorporate this feedback term explicitly in our model. Furthermore, we analyze this feedback term in detail by studying the impact of its effectiveness on the development strategy. In agile software development practices, features are released to the customers incrementally in order to create value while the system is being developed (Wells 2009). Another novel aspect of our paper is to model this phenomenon by considering the value of the system before all the required modules are developed. By modeling the value of functionality that is created as the project unfolds, we can track the growth of value during system development. In our model, we also consider the impact of system complexity on the efficiency of development. To the best of our knowledge, it is first such study in this setting.

2. Problem Description and Solutions

In the proposed model, the user group gains value from the modules as the system develops and after it is released. On the other hand, the developer group incurs cost in understanding the requirements of users and developing the system. From the perspective of user group, the value needs to be maximized, whereas the developer group strives to minimize its cost. We first present and solve the generalized model with both the value and the cost. Then, we analyze user and developer objectives separately by setting appropriate parameter values. In this section, we present the expressions for module growth rate and the rate at which requirements are understood. Subsequently, we formulate the problem of optimally allocating effort between developing the modules and understanding the requirements. Finally, we present the solution.

2.1 Notations

We begin with presenting parameters and variables of the model in Table 1.

TABLE 1. MODEL PARAMETERS AND DECISION VARIABLES

SYMBOL	DEFINITION	REMARKS
$u(t)$	Proportion of developer effort allocated for developing modules at time t	*Control Variable*
$1 - u(t)$	Proportion of developer effort allocated for understanding requirements at time t	
$n(t)$	Total number of modules developed till time t	*State Variable*
$m(t)$	Number of modules understood but not developed (by developers) at time t	*State Variable*
N	Number of modules to be developed	*Exogenous Constant*
T	Total time required to develop N modules	*To be determined*
$r(n)$	Value rate (to the user group) as a function of developed modules	*Exogenous Function*
$R(N)$	Value rate of the released system as a function of the total number of modules	*Exogenous Function*
C	Cost per unit time incurred by the development group	*Exogenous Constant*
M	Useful life-time of the system	*Exogenous Constant*
$f(n,m)$	Effect of modules understood and/or developed on development efficiency	*Exogenous Function*
$g(n)$	Impact of the number of modules developed on understanding requirements	*Exogenous Function*

2.2 The Model

The rate of module growth is represented by $\dot{n}(t) = dn(t)/dt$, which is modeled as follows:

$$\dot{n} = f(n,m) \cdot u,\ n(0) = 0,\ n(T) = N \qquad (1)$$

For notational convenience, we suppress the time notation (t) when no confusion arises. As can be seen from Equation (1), the higher the proportion of effort allocated for developing modules (i.e., u), the faster new modules can be developed. The function $f(n,m)$ captures two effects:

First, as the system size n grows, the system becomes more complex and the effectiveness of development effort decreases (Kafura and Reddy 1987). Second, the growth rate is higher when more requirements are understood. Finally, the total number of modules that needs to be developed is a given constant N.

The growth in the number of modules understood but not developed is represented by

$$\dot{m} = g(n)(1-u) - \dot{n} = g(n)(1-u) - f(n,m)\cdot u\,,\ m(0) = m_0 \tag{2}$$

Clearly, as more effort is allocated for understanding the requirements, $\dot{m}$ increases. As discussed earlier, one novel aspect of our model is to consider the impact of the number of modules developed on the efficiency of understanding the requirements. It is incorporated in the function $g(n)$. Finally, by definition, as $\dot{n}$ increases, $\dot{m}$ decreases. The initial value of m (i.e., m_0) represents the initial understanding about the requirements before the project starts.

The objective function consists of the following three terms: (a) the value to the users when the software is being developed is represented by $\int_0^T r(n)dt$, (b) the value to the users from completed system (from T to the system lifetime M) is given by $R(N)\cdot(M-T)$, and (c) the cost of time incurred by developers in developing the software and understanding the requirements is written as $c\cdot T$. Now the optimal control problem is given by state equations (1) and (2) and the following objective function:

$$\max_{T,u} \int_0^T r(n)dt + R(N)\cdot(M-T) - c\cdot T \tag{3}$$

2.3 The Solution

The Hamiltonian function of the optimal control problem can now be written as (Sethi and Thompson 2000): $H = r(n) + \lambda_1 f(n,m)u + \lambda_2\left[g(n)(1-u) - f(n,m)\cdot u\right]$, where $\lambda_1(t)$ and $\lambda_2(t)$ are the adjoint variables corresponding to state variables n and m, respectively. Since the Hamiltonian H is linear in the control variable u, we have the following solution form (Sethi and Thompson 2000):

$$u = \begin{cases} 1 & H_u > 0, \\ \text{To be determined} & H_u = 0, \\ 0 & H_u < 0. \end{cases}$$

where $H_u = (\lambda_1 - \lambda_2)f - \lambda_2 g$. Now we can obtain the optimal solution using the above equations. The result is summarized in the following theorem (Proofs are omitted due to space limitation):

Theorem 1: *The general form of the optimal solution depends on* m_0 *as follows*:

(1). *When* m_0 *is smaller than a given threshold value* M_0*, then:*

(a) $u = 0$ *for* $0 \le t < T_1$, (b) $u = C(m,n)/B(m,n)$ *for* $T_1 \le t < T_2$, *and* (c) $u = 1$. *for* $T_2 \le t \le T$

(2). *When* m_0 *is larger than a given threshold value* M_0*, then:*

(a) $u = 1$ *for* $0 \le t < T_1'$, (b) $u = C(m,n)/B(m,n)$ *for* $T_1' \le t < T_2'$, *and* (c) $u = 1$ *for* $T_2' \le t \le T'$

When m_0 is small, the first time period (i.e., $0 \le t < T_1$) represents the pure requirement gathering period, where no module is developed and all the developer effort is allocated for understanding the requirements. The second time period (i.e., $T_1 \le t \le T_2$) represents the most interesting period, where developers concurrently understands the requirements and develop modules. Such concurrent approach is becoming prevalent in agile software development methodologies. Therefore, we will analyze the concurrent period in detail in our subsequent subsections. The last time period (i.e., $T_2 < t \le T$) represents a pure development period where developers don't spend any effort in understanding requirements. Results for high m_0 are omitted for the sake of brevity.

3. Discussions and Managerial Implications

In this section, we use some specific forms for $f(n,m)$, $g(n)$, $r(n)$, and $R(N)$ to illustrate the solution process and gain useful insights from numerical results. Let us first present these functional forms and their justification. First, as discussed earlier, the effectiveness of development effort decreases with an increase in the system size *n* (Kafura and Reddy 1987) and the module growth rate is higher when more requirements are understood. Hence, $f = (a_1 m - a_2 n)$, where a_1 and a_2 are positive constants. Here, a_1 and a_2 capture the impacts of *m* and *n*, respectively, on *f*. Second, the process of understanding user requirements becomes more efficient for the developers as they develop more modules (Davis 1992). Hence, $g = a_3(1 + a_4 n)$, where the positive constant a_4 represents the effectiveness of the feedback and a_3 is the multiplier. The term a_4 will be the focus of our discussion in next section. Finally, we have $r = b_1 n$ and $R = b_2 N$ to represent linear value functions. Baseline parameter values are as follows: $a_1 = 0.05$, $a_2 = 0.01$, $a_3 = 5$, $b_1 = 0.8$, $b_2 = 1$, $N = 50$, $m_0 = 0$, and $M = 50$.

3.1 Impact of the Effectiveness of Feedback in Value Maximization Problem

Note that the objective function given in Equation (3) reduces to the value maximization problem when the cost of time *c* is set to zero. Hence, in this subsection, we let $c = 0$ and study the impact of the feedback effectiveness term a_4. The objective of this study to analyze how the effectiveness of feedback impacts the optimal strategy and study when it is more beneficial to adopt the practice of concurrent development and requirement gathering.

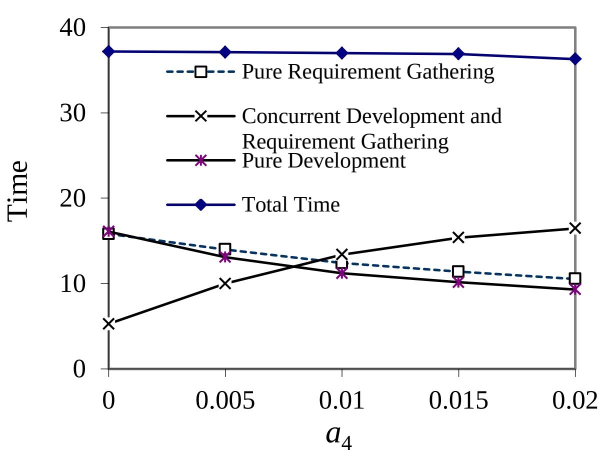


Figure 1 Impact of feedback effectiveness (a_4) on optimal time periods: Value Maximization

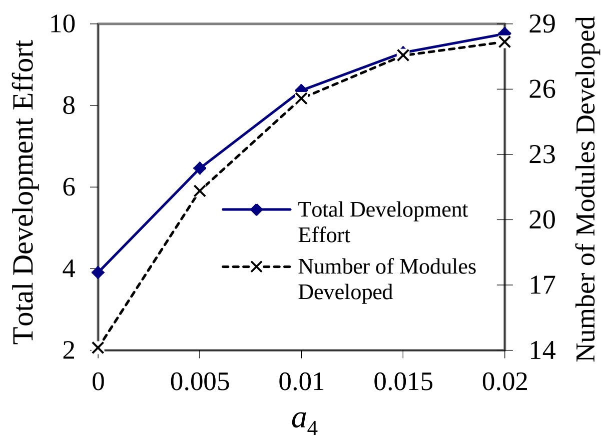


Figure 2 Impact of feedback effectiveness (a_4) on concurrent period: Value Maximization

We begin with Figure 1, where we study the impact of a_4 on the three time periods presented in Proposition 1. This figure shows that as the effectiveness of feedback increases, the duration of concurrent period increases in the optimal solution at the expense of pure development and pure requirement gathering periods. This result implies that the practice of concurrent development, adopted in agile development methodologies, becomes more effective when the effectiveness of feedback increases. Hence, the managers who chose to adopt concurrent practice should focus on improving the effectiveness of feedback in order to obtain the most value from the system. On the other hand, as the feedback becomes more effective, it is optimal for the managers to adopt more concurrent development and vice-versa.

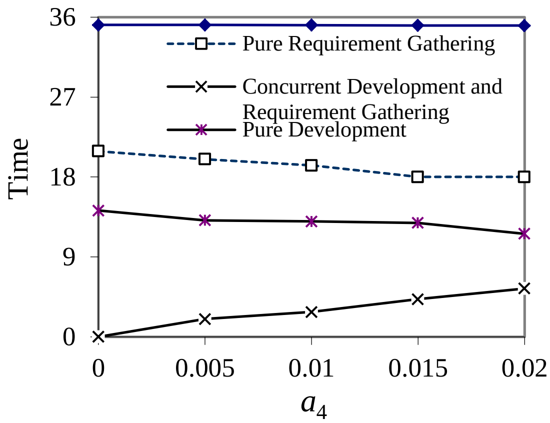


Figure 3 Impact of feedback effectiveness on optimal time periods: Cost/Time Minimization

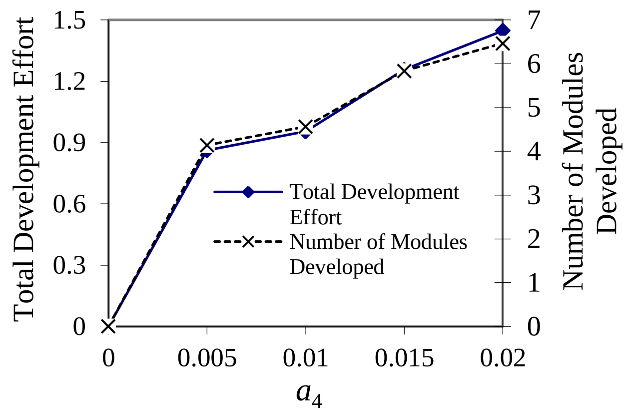


Figure 4 Impact of feedback effectiveness (a_4) on concurrent period: Cost/Time Minimization

We further analyze the impact of feedback effectiveness on the concurrent period in Figure 2. It shows that both the total development effort and the number of modules developed in the concurrent period increases with an increase in the effectiveness of feedback. This result further enforces the insights obtained from Figure 1.

3.2 Impact of the Effectiveness of Feedback in Cost/Time Minimization Problem

We solve the cost minimization problem (and the time minimization problem) by setting $b_1 = b_2 = 0$ in the objective function given by Equation (3). Similar to the previous subsection, we study the impact of a_4 in Figures 3 and 4. The following two results are similar to those in the value maximization problem: (a) as the effectiveness of feedback increases, the duration of concurrent period increases in the optimal solution at the expense of pure development and pure requirement gathering periods, and (b) both the total development effort and the number of modules developed in the concurrent period increases with an increase in the effectiveness of feedback. Hence, the similar recommendation follows.

In spite of the similarities between the results of two problems, there are some notable differences. First, there is no concurrent period in the cost/time minimization problem when the feedback is totally ineffective (i.e., $a_4 = 0$). Moreover, both Figures 3 and 4 shows that the increase in duration of concurrent period with an increase in a_4 is much slower in the cost/time minimization problem as compared to that in the value maximization problem. Hence, from the perspective of cost/time minimization, it is not beneficial for the managers to adopt the concurrent development practice when the feedback is not sufficiently effective.

3.3 Comparing the Solution of Two Problems

Clearly, the solution of value maximization problem provides higher value compared to than in the cost/time minimization problem, and the cost/time minimization problem provides lower cost and time as compared to those in the value maximization problem. The objective of this subsection is to study the difference in total time in the solutions of these two problems for different levels of the feedback effectiveness. We find that, this difference decreases with an increase in the effectiveness of feedback. It implies that the objectives of these two problems converge as the feedback effectiveness increases. It is an important result for managers, because they strive to find ways to align the objectives of user group and developer group. In this regard, a possible direction is to spend effort in improving the effectiveness of feedback.

4. Conclusions and Future Research Directions

We study the problem of allocating developer effort in a software project between understanding the requirements and developing the modules. The objective is to maximize value obtained from the system and minimize the cost required to develop the system. We are able to obtain a closed-form solution and gain useful insights from our analytical model. A novel aspect of this study is the consideration of incremental value that gets created as the project unfolds. It has been shown in earlier studies that the effectiveness of requirement gathering effort improves as more modules are developed. Another novel aspect of our paper is to incorporate this feedback explicitly in our model. We also analyze this feedback term in detail by studying its impact on the development strategy. We first present the optimal solution and then provide useful managerial insights. Since the agile development methodologies call for *concurrent* requirement gathering and development, the main focus of our numerical study is to analyze the impact of feedback effectiveness on the concurrent approach. Our results indicate that the concurrent approach becomes more effective as the effectiveness of feedback increases. Hence, the managers who chose to adopt concurrent practice should focus on improving the effectiveness of feedback in order to obtain the most value from the system. On the other hand, as the feedback becomes more effective, it is optimal for the managers to lean towards concurrent approach and vice-versa. We also find that the objectives of user group and developer group converge as the feedback effectiveness increases. Therefore, the managers can align the objectives of these two groups by spending effort in improving the effectiveness of feedback. An extension we are currently pursuing is to devise more mechanisms that align the objectives of these two groups. As another extension, we are further analyzing the joint optimization of value and cost. We plan to present these results during presentation.

References

Astels, D., G. Miller, M. Novak. 2002. *A practical guide to extreme programming*, Prentice Hall.

Baroudi, J.B., M. Olson, B. Ives. 1986. An empirical study of the impact of user involvement on system usage and information satisfaction. *Communications of the ACM* 29(3) 232-238.

Beck, K. 2000. *Extreme programming explained: Embrace change*, Addison-Wesley.

Boehm, B. 1981. *Software engineering economics*, Prentice Hall, New Jersey.

Davis, A. 1992. Operational prototyping: A new development approach. *Software* 9(5) 70-78.

Dawande, M., S. Kumar, V.S. Mookerjee, C. Sriskandarajah. 2008. Maximum commonality problems: Applications and analysis. *Management Science* 54(1) 194-207.

Dingsøyr, T., S. Nerur, V. Balijepally, N.B. Moe. 2012. A decade of agile methodologies: Towards explaining agile software development. *Journal of Systems and Software* 85(6) 1213-1221.

Field, T. 1997. When bad things happen to good projects. *CIO Magazine* 11(2) 54-59.

Kafura, D., G. Reddy. 1987. The use of software complexity metrics in software maintenance. *IEEE Transactions on Software Engineering* SE-13(3) 335-343.

Maruping, L.M., V. Venkatesh, R. Agarwal. 2009. A control theory perspective on agile methodology use and changing user requirements. *Information Systems Research* 20(3) 377-399.

Sethi, S.P., G. Thompson. 2000. *Optimal control theory: Applications to management science and economics*, Kluwer Academic Publishers.

Standish Group. 1994. *The chaos report*. http://www.standishgroup.com/sample_research/chaos_1994_1.php

Wells, D. 2009. *Extreme programming: A gentle introduction*. http://extremeprogramming.org

Optimal Coordination in Distributed Software Development

Hao Xia, Milind Dawande, Vijay Mookerjee
Naveen Jindal School of Management, The University of Texas at Dallas, Richardson, Texas 75080
xiah@utdallas.edu, milind@utdallas.edu, vijaym@utdallas.edu

The construction of a software system requires not only individual coding effort from team members to realize the various functionalities, but also adequate team coordination to integrate the development effort into a consistent, efficient, and bug-free system. On the one hand, continuous coding without adequate coordination can cause serious system inconsistencies and faults that may subsequently require significant corrective effort. On the other hand, frequent integrations can be disruptive to the team and delay development progress. This tradeoff motivates the need for a good coordination policy. Both the complexity and the importance of an optimal coordination policy are further highlighted in the context of distributed software development (DSD), where a software project is developed by multiple, geographically-distributed sub-teams. Coordination in DSD exists both within one sub-team and across different sub-teams. The latter type of coordination involves communication across spatial boundaries (different locations) and possibly temporal boundaries (different time zones), and is a major challenge DSD faces. In this paper, we develop an analytical framework to model these two types of coordination activities in DSD and derive an optimal coordination policy.

Key words: distributed software development; project management; coordination in teams

---

1. Introduction

Starting in 2001 with a budget of $223 billion, the F-35 Joint Strike Fighter (JSF) program, which is the most expensive defense project in US history, has again been reported to be lagging in progress. The major reason for the delay – according to a 2011 GAO report on JSF to congressional committees – is that the development of the JSF software, which is essential for about 80% of its functionality, suffers from unexpectedly low productivity. As stated in the report, "the officials underestimated the time and effort needed to develop and integrate the software substantially, contributing to the program's overall cost, schedule problems, and testing delays" (GAO 2011). This problem of low productivity in the construction of a complex software system is not a rare occurrence. A survey by Keil, Mann, and Rai (2000) suggests that 30-40% of information systems projects had outcomes that were significantly below expectation, in terms of implementation performance, budget, or schedule. Through a review of surveys on software effort estimation, Molkken and Jrgensen (2003) show that about 60-80% projects encounter effort and/or schedule overruns. The 2006 CHAOS report by the Standish Group claims that 19% of the software projects were outright failures (Rubinstein 2007). Together, these observations strongly motivate the need to study approaches aimed at improving productivity in software development.

An incremental development strategy is used in most modern software projects, where the system is broken down into smaller parts and these parts are scheduled to be developed over time and integrated when completed (Cockburn 2008). Accordingly, the construction of a software system using incremental development can be assumed to consist of a series of scheduled *construction cycles*, each of which in turn consists of a development phase followed by an integration phase (Chiang and Mookerjee 2004, Mookerjee 2002). In the development phase, the major activities are coding and unit testing of system function units to realize the functionalities required in the design specification. The workload in this phase is primarily the individual effort of developers; the requirement of coordination among developers is minimal. After development is complete, the outcomes of the individual development efforts need to be integrated into a consistent system, which then serves as a baseline for future construction cycles (Humphrey 1989). The developers gain common understanding of their prior work through team communications and perform system-level testing and inconsistency/fault removal. Naturally, a significant amount of coordination is required in this integration phase.

The following tradeoff in scheduling the construction cycles is critical to the productivity of software construction: On the one hand, developers are typically productive when concentrating on individual development work; therefore, switching to an integration phase too frequently interrupts the fluent progress of their work (Brooks 1995). On the other hand, continuous individual development without adequate coordination may cause serious system inconsistencies and faults that may later require significant corrective effort (Keil, Mann, and Rai 2000). The coordination work in the second phase of construction cycles is often the crux of a software project, especially a large-scale one. Relative to the individual development effort, the coordination effort required often becomes highly unmanageable and escalates over the budget (Brooks 1995). Consequently, designing a good policy – one that balances this tradeoff – to schedule construction becomes essential for the development team.

Coordination becomes even more significant and complex in distributed software development (DSD), where the software system to be constructed is broken down into subsystems that are assigned to geographically-distributed sub-teams. DSD can exploit a variety of economic factors and has become a common practice in the software industry (Aspray, Mayadas, and Vardi 2006, Damian and Moitra 2006). The subsystems, however, cannot be fully isolated from each other and need to be integrated into a consistent system. Accordingly, coordination activities in DSD consist of two types: co-located coordination among developers in the same sub-team and remote coordination among developers dispersed in different sub-teams. Remote coordination in DSD requires communication and cooperation across spatial boundaries and possibly across temporal and cultural boundaries, which further underline the major challenges DSD encounters (Agerfalk, Fitzgerald, and Slaughter 2009, Jimnez et al. 2009).

In DSD, an incremental development strategy is often applied via two-level construction cycles (Sangwan et al. 2006). At a sub-team level, each sub-team schedules its own construction cycles, which we refer to as *local cycles*. The integration phase in a local cycle of a sub-team primarily involves co-located coordination among the corresponding local developers. At the level of the entire team, construction cycles that require participation from all the sub-teams are referred to as *global cycles*. During the integration phase of a global cycle, all sub-teams coordinate together to integrate the work increments of their respective subsystems. We provide a more detailed discussion on the coordination activities in local and global cycles, and the corresponding mathematical expressions in Section 2. A coordination policy for a DSD project schedules the local and global cycles, i.e., prescribes when each sub-team should switch to the integration phase in a local cycle and when the entire team should switch to the integration phase in a global cycle.

2. An Analytical Model for Coordination in DSD

In this section, we present an analytical specification of coordination activities in a DSD project. For simplicity, we discuss the special case of two sub-teams. However, our analysis can be easily extended to $n \geq 3$ sub-teams.

Due to lack of space, we limit ourselves to discussing a symmetric situation: the two sub-teams are of equal size, have identical work capability, and their respective subsystems have identical complexities. Each sub-team has S developers, each with identical work capability. The system is constructed through repeated, two-level construction cycles, i.e., local cycles and global cycles, as discussed in Section 1. A local cycle for a sub-team consists of a development phase followed by a local integration phase. In the development phase, developers work individually on coding and unit testing, to realize or enhance function units (e.g., system modules) as scheduled. Then, they switch to the local integration phase in which they reconcile their previous development work via coordination to obtain a consistent fraction of their sub-system. A global cycle consists of several local cycles of the two sub-teams followed by a global integration phase, in which developers from both sub-teams coordinate to release a consistent fraction of the entire system.

We denote the construction effort that occurs during the development phases, the local integration phases, and the global integration phases, by *development effort*, *local coordination effort*, and *global coordination effort*, respectively. The total development effort is largely predetermined by the software requirements specification. Accordingly, we assume that each sub-team is responsible for a total development effort of L. The total coordination effort, however, is affected by the coordination policy adopted, and is the focus of this study.

We denote a general coordination policy by $(\boldsymbol{\tau}, \boldsymbol{\theta})$, where $\boldsymbol{\tau} = [\tau^1, \tau^2, \ldots, \tau^m]$, $\boldsymbol{\theta} = [\theta^1, \theta^2, \ldots, \theta^n]$. This notation is defined as follows: m is the total number of local cycles for each sub-team and n is the total number of global cycles. The i^{th} local integration of a sub-team is scheduled after τ^i amount of development effort is expended in local cycle i, $i = 1, 2, \ldots, m$. The j^{th} global integration for the entire team is scheduled after θ^j amount of development effort per sub-team is expended in global cycle j, $j = 1, 2, \ldots, n$. It follows immediately that $\sum \tau^i = \sum \theta^j = L$. We now discuss the coordination effort under policy $(\boldsymbol{\tau}, \boldsymbol{\theta})$.

2.1. Coordination Effort under Policy $(\boldsymbol{\tau}, \boldsymbol{\theta})$

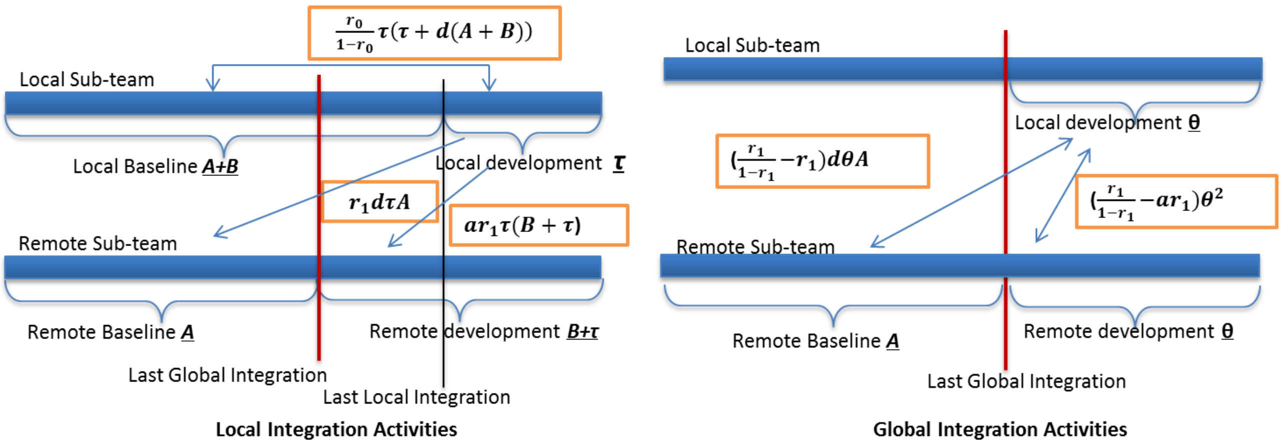
We consider the coordination activities in the local and the global integration phases.

2.1.1. Local Integration Phase. Since the two sub-teams are assumed to be symmetric, it suffices to calculate the local coordination effort for one sub-team. The local coordination effort has one fixed component and two variable components. The fixed component captures the *preparation effort* required for a developer to switch from development to integration. Assuming that, on average, a developer spends k_0 amount of time on this switch, the total preparation effort incurred by a sub-team per local cycle is $k_0 S$.

We refer to the first variable component as the *program comprehension effort*. In order to achieve consensus before actual integration, developers review and comprehend each other's development effort and discuss the integration plan. Naturally, the effort required is proportional to the amount of development effort in the current local cycle. Another factor is the communication overhead, which is typically assumed to be a quadratic function of the group size involved (see, e.g., Brooks (1995)). Thus, the total comprehension effort incurred by a sub-team in local cycle i can be expressed as $k_1 S^2 \tau^i$, where k_1 is a local-comprehension coefficient that is influenced by a variety of factors, including organizational structure of the sub-team, relationships between the developers, the collaboration tools used, etc.

The second variable component is the actual *integration effort*. To integrate several pieces of development efforts that are related but have not been earlier integrated into one consistent system capable of serving as a stable baseline repository, developers need to "adjust" these pieces until they are fully compatible with each other. We now discuss the required effort in detail. As illustrated in Figure 1, assume that in a specific local cycle, a sub-team's development effort is τ, the total amount of development effort before the last global integration is A, and the total amount of development effort expended after the last global integration and before the last local integration is B. For a sub-team, all development effort before the previous local integration (which amounts to $A + B$) was already integrated and can serve as a local baseline repository. The new development effort of τ needs to be made compatible internally (since it may consist of several individually-developed pieces) and also with respect to the local baseline repository. This compatibility may require several rounds of reciprocal adjustment: integration effort done in the first-round of adjustment triggers the necessity to launch a second-round adjustment, and so on. This so-called "ripple effect" yields a potentially infinite sequence of diminishing efforts (Lejter, Meyers, and Reiss 1992). We posit that the first-round adjustment incurs an effort of $r_0 \tau(\tau + d(A + B))$, which can be explained as follows: the first-round adjustment of two pieces (say) of development effort is proportional to the amount of development effort of each of these pieces. A stability discount factor d is applied to the "stable" development effort that already belongs to the baseline repository. The first-round adjustment effort is also affected by the collaboration efficiency of the sub-team and the structural complexity of the subsystem, i.e., strength and density of internal relationships between the functional units. We use

Figure 1 Illustrating Local and Global Integration Activities.



$r_0 < 1$ as a local-integration coefficient to capture these two factors. The second-round adjustment requires an effort of $r_0^2\tau(\tau + d(A+B))$, and so on. Summing these up, the total effort to achieve compatibility is $(r_0 + r_0^2 + r_0^3 + \cdots)\tau(\tau + d(A+B)) = \frac{r_0}{1-r_0}\tau(\tau + d(A+B))$. Furthermore, with the use of distributed revision-control systems and telecommunication tools, part of the integration effort between the local development effort τ and the remote subsystem can also be achieved during local integration. The development effort A in the remote subsystem is globally integrated and serves as a remote baseline repository. Therefore, the local sub-team can adjust its local development effort to adapt to the remote baseline repository, but cannot adjust to the remote subsystem. Thus, the local sub-team can only accomplish the first-round adjustment, incurring an effort of $r_1 d\tau A$. Here, similar to r_0, r_1 is a remote-integration coefficient that captures the collaboration efficiency across different sub-teams and the structural complexity of the entire software system. Note, however, that the remote development effort $B + \tau$ has not been globally integrated earlier. We assume that only a portion $(a < 1)$ of the first-round adjustment can be effectively done, incurring an effort of $ar_1\tau(B+\tau)$.

2.1.2. Global Integration Phase. The global integration phase follows a local integration phase. All the developers in the entire team coordinate, either by coming together at a common location or via telecommunication, both of which incur a much higher cost than coordination within a sub-team. As with local coordination, the global coordination effort for the entire team also consists of a fixed component and two variable components. The preparation effort is $2k_2 S$, given that a developer needs, on average, time k_2 to prepare. Then, each sub-team needs to comprehend the other sub-team's development effort since the last global integration. The total comprehension effort in the j^{th} global cycle is $2k_3 S^2 \theta^j$, where k_3 is a remote-comprehension coefficient.

As illustrated in Figure 1, consider the development effort θ since the last global integration by the local sub-team. This development effort is fully integrated within the local subsystem, but is only partly adjusted with respect to the remote subsystem (the first-round adjustment with the remote baseline repository and part of the first-round adjustment with the remote development effort in the current global cycle were done in the local integrations, as discussed above). The integration effort that remains for integrating the local development effort θ with the remote baseline repository A and the remote development effort θ are $(\frac{r_1}{1-r_1} - r_1)d\theta A = \frac{r_1^2}{1-r_1}d\theta A$ and $(\frac{r_1}{1-r_1} - ar_1)\theta^2$, respectively.

2.1.3. An Optimal $(\boldsymbol{\tau}, \boldsymbol{\theta})$ policy We show that there exists an optimal (i.e., one that minimizes total coordination effort) $(\boldsymbol{\tau}, \boldsymbol{\theta})$ policy that is uniform. That is, a (q,n)-policy that uniformly schedules n global cycles and uniformly schedules q local cycles in each global cycle, i.e., the total number local cycles is $m = qn$ and $\boldsymbol{\tau} = [\tau^1, \tau^2, \ldots, \tau^m] = [\frac{L}{qn}, \frac{L}{qn}, \ldots, \frac{L}{qn}]$, $\boldsymbol{\theta} = [\theta^1, \theta^2, \ldots, \theta^n] = [\frac{L}{n}, \frac{L}{n}, \ldots, \frac{L}{n}]$. For brevity, we avoid providing the proof here and derive the total coordination effort as a function of q and n in the next subsection.

Table 1 Effect of Project Parameters on the Optimal Coordination Policy.

	L	S	k_0	k_1	r_0	r_1	a	d	k_2	k_3
n^*	+	-	\	\	\	+	-	-	-	\
q^*	\	\	-	\	+	-	+	+	+	\
n^*q^*	+	-	-	\	+	+	+	-	\	\

2.2. Coordination Effort under a (q,n) Policy

Since a (q,n) policy is a special type of a $(\boldsymbol{\tau},\boldsymbol{\theta})$ policy, the analytical specification of coordination effort under a $(\boldsymbol{\tau},\boldsymbol{\theta})$ policy still applies. It is easy to obtain the following:

In each local integration phase, the preparation effort and the program comprehension effort for one sub-team are k_0S and $k_1S^2\frac{L}{qn}$, respectively. In the i^{th} local cycle of the j^{th} global cycle, the integration effort is $LI(j^i)=\frac{L^2}{n^2}(\frac{r_0}{(1-r_0)}\frac{1+qd(j-1)+d(i-1)}{q^2}+r_1(\frac{d(j-1)}{q}+\frac{ai}{q^2}))$. Then, the total local coordination effort over all local cycles per sub-team can be expressed as $LI=(k_0S+k_1S^2\frac{L}{qn})qn+\sum_{j=1}^{n}\sum_{i=1}^{q}LI(j^i)$. Similarly, the global coordination effort in the l^{th} global cycle is $GI(l)=2(\frac{L}{n})^2(\frac{r_1}{1-r_1}(dr_1(l-1)+1)-ar_1)$. Thus, the total global coordination effort can be expressed as $GI=(2k_2S+2k_3S^2\frac{L}{n})n+\sum_{l=1}^{n}GI(l)$. The total coordination effort under the (q,n) policy is then $CE=GI+2LI$.

3. An Optimal Coordination Policy

Based on the model of coordination effort above, we can derive an optimal coordination policy for DSD and also compare it with the case of co-located development.

3.1. Optimal Coordination Policy under a Schedule Constraint

Consider the DSD project discussed in Section 2. The project manager wants to choose a team size S and an optimal coordination policy (q,n) to minimize the total coordination effort CE, given that the project should be completed within the schedule constraint T.

To solve this problem, we first fix the team size S and minimize CE with respect to n and q. We relax the constraint that n and q be integers, by allowing them to take any positive value. Then, we obtain $n^*=\frac{L\sqrt{r_3(2-dr_1)-(a+d)r_1}}{\sqrt{2k_2S}}$, $q^*=\frac{\sqrt{k_2(r_2(2-d)+ar_1)}}{\sqrt{k_0(r_3(2-dr_1)-(a+d)r_1)}}$ where $r_2=\frac{r_0}{(1-r_0)}$ and $r_3=\frac{r_1}{(1-r_1)}$. To obtain a feasible and optimal solution, we then compare CE at four points: $(\lceil n^*\rceil,\lceil q^*\rceil),(\lceil n^*\rceil,\lfloor q^*\rfloor),(\lfloor n^*\rfloor,\lceil q^*\rceil),(\lfloor n^*\rfloor,\lfloor q^*\rfloor)$, and select the point that achieves the minimum. For a team size of S, let the minimum value of CE be $CE^*(S)$. Thus, the minimum total construction effort is $TE^*(S)=CE^*(S)+2L$, since the development effort is predetermined. Consequently, the minimum possible time to finish the project, given the team size S, is $T^*(S)=\frac{TE^*(S)}{2S}$.

Increasing the number of developers leads to a reduction in the time required for development, but increases the effort required for coordination due to the communication overhead. It is easy to show that the minimum total construction effort increases monotonically with the team size S, while the minimum possible time first decreases and then increases with S. Therefore, it is possible that the schedule constraint cannot be satisfied, irrespective of the number of developers in the team. It is desirable to choose the smallest team size S that (possibly) achieves $T^*(S)\le T$.

Table 1 illustrates how q^*, n^*, and q^*n^* (total number of local cycles) change with the model parameters. The symbols "+", "-" and "\" indicate, respectively, that a policy parameter increases, decreases, and does not change, with an increase in a model parameter. Observe that when facing a tight schedule that requires more developers, less integrations are scheduled and more effort is spent on coordination activities.

3.2. Comparison with Co-located Development

A natural question arises: what is the difference between the optimal coordination policies under DSD and co-located development? Consider the case when each of the two sub-teams develops its subsystem independently. Then, there is only local coordination involved. Assuming that the total number of local cycles scheduled is m, the total coordination effort (following the same argument as before) is $CE = k_0Sm + k_1S^2L + \frac{r_2L^2d}{2} + r_2L^2(1-\frac{d}{2})\frac{1}{m}$. Minimizing CE with respect to m, we obtain $m^* = \frac{L\sqrt{r_2(2-d)}}{\sqrt{2k_0S}} \leq \frac{L\sqrt{r_2(2-d)+ar_1}}{\sqrt{2k_0S}} = q^*n^*$ where q^*n^* is the total number of local cycles for each sub-team in the DSD case. Thus, the optimal frequency of integration in DSD is higher than that in co-located development, since the remote sub-team also benefits from frequent integrations. We limit ourselves to only this observation here.

4. Summary and Discussion

Our attempt has been to devise an analytical framework for coordination activities in DSD and then derive an optimal coordination policy. We find that the optimal frequencies of local and global integration are affected by a variety of project parameters, and a smaller team size is preferable to the extent permitted by the schedule. The optimal integration frequency in DSD is typically higher than that in co-located development.

There are two natural limitations of our discussion thus far: our model is static and symmetric. The sub-teams working on a DSD project are often non-identical and have differences in size, composition, and the tasks assigned to them. Furthermore, since the progress of development, the level of system consistency, and the required specifications, cannot be fully predicted over the duration of construction of a software project, the consideration of uncertainty gains practical importance. We are actively working on analyzing both asymmetry and uncertainty.

References

Agerfalk, P. J., B. Fitzgerald, S. A. Slaughter. 2009. Introduction to the Special Issue–Flexible and Distributed Information Systems Development: State of the Art and Research Challenges. *Information Systems Research* **20**(3) 317–328.

Aspray, W., F. Mayadas, M. Y. Vardi. 2006 Globalization and Offshoring of Software, a Report of the ACM Job Migration Task Force. *Association for Computing Machinery*, New York.

Brooks, F. P. 1995. *The Mythical Man-Month: Essays on Software Engineering.* Addison-Wesley, MA.

Chiang, I. R., V. S. Mookerjee. 2004. A Fault Threshold Policy to Manage Software Development Projects. *Information Systems Research* **15**(1) 3–21.

Cockburn, A. 2008. Using Both Incremental and Iterative Development. *STSC CrossTalk (USAF Software Technology Support Center)* **21**(5) 27–30.

Damian, D., D. Moitra. 2006. Global Software Development: How Far Have We Come? *IEEE Software* **23**(5) 17–19.

U.S. Government Accountability Office. 2011. Joint Strike Fighter: Restructuring Places Program on Firmer Footing, but Progress Still Lags. *Report to Congressional Committees* GAO-11-325.

Humphrey, W. S. 1989. *Managing the Software Process.* Addison-Wesley, Boston.

Jimnez, M., M. Piattini, A. Vizcano. 2009. Challenges and Improvements in Distributed Software Development: A Systematic Review. *Advances in Software Engineering* **2009** 1-14.

Keil, M., J. Mann, A. Rai. 2000. Why Software Projects Escalate: An Empirical Analysis and Test of Four Theoretical Models. *MIS Quarterly* **24**(4) 601–630.

Lejter, M., S. Meyers, S. P. Reiss. 1992. Support for Maintaining Object-Oriented Programs *IEEE Transactions on Software Engineering - Special Issue on Software Maintenance* **18**(12) 1045–1052.

Molkken, K., M. Jrgensen. 2003. A Review of Surveys on Software Effort Estimation *Proceedings of the 2003 International Symposium on Empirical Software Engineering* 223–230.

Mookerjee, V. S., I. R. Chiang. 2002. A Dynamic Coordination Policy for Software System Construction. *IEEE Transactions on Software Engineering* **28**(7) 684–694.

Rubinstein, D. 2007. Standish Group Report: There's Less Development Chaos Today. Available at http://www.sdtimes.com/link/30247.

Sangwan, R., M. Bass, N. Mullick, D. J. Paulish, J. Kazmeier. 2006. *Global Software Development Handbook.* Auerbach Publications, New York.

Uncertainty, Switching Cost, and Competition in the Software-As-a-Service Market

Robert J. Kauffman and Dan Ma
Singapore Management University
{rkauffman, madan}@smu.edu.sg

Key words: Software-as-a-service, two-stage competition, switching cost, analytical model

Abstract

We construct a two-stage model to study the competition between Software-as-a-service (SaaS) providers who offer both horizontally and vertically differentiated software and IT services. The model captures three key features of SaaS: multi-tenancy, pay-as-you-go pricing, and partnership between users and providers. Users bear additional mismatch costs due to the multi-tenant structure of SaaS, and they also face potential risk of being locked in by one SaaS provider. We identify five different types of market equilibrium, and derive conditions for each to appear. Interestingly, the fact that users might be locked in by one provider could benefit SaaS providers or users themselves, depending on the magnitude of switching cost and the two provider's relative service quality. This finding suggests that a SaaS provider should set its service quality level strategically so that it can leverage on switching cost to get extra competitive advantage.

1. Introduction

Software as a Service (SaaS), a new business model for software industry, has attracted a lot interests. Under the SaaS model, software providers bundle applications with IT services and deliver them through a network. Users subscribe to use the application which is installed in a centralized location, run and maintained by the software provider.

Multi-tenancy is the number one important feature of SaaS (Marc Benioff, seminar talk in Singapore 2010). Providers host "a single instance of the software on a single server and maintain the customer data on a single database" (Hickins 2007), shifting the business model from traditional "one-to-one" (i.e., one provider and one customer) to "one-to-many" (i.e., one provider to many customers) structure. Multi-tenancy is the key technology for the SaaS paradigm to reach economy of scale, both in database management and code maintenance. Meanwhile, it eliminates the possibility to customize applications by changing the source codes. SaaS services are usually designed and developed as standardized functionalities, and users must generally accept such "standardized" applications as provided. The tension between cost efficiency and customizability thus coexist in the SaaS multi-tenancy environment. SaaS is also famous for its "pay as you go" pricing. Unlike traditional perpetual licensing, SaaS users do not need to pay an upfront fee to the provider for software installation and implementation. Rather, they can start the business for free. Users' total payments are increasing in their usage (time), in the form of a monthly subscription fee or fixed fee per transaction. In many cases, using SaaS may prove cheaper than owning and maintaining an in-house IT system – users expect to save money on support and upgrade costs, IT infrastructure, IT personnel, and implementation. Finally, the SaaS model entirely changes the business relationship between the software provider and users, from the conventional seller-and-buyer to close business partners. Using SaaS is not a one-shot business. Instead, a user must work with the provider within a quite long time period. The provider is in charge of all the IT support services, including daily maintenance of the software, data backups, software upgrades, and security. Such close business partnership is good for users; in the sense that now they are relieved of IT maintenance burden. However it also brings potential “lock-in” risk to them. Users lose control over data and application, become dependent on the outside provider to a large extent (IT Business Edge, 2008).

Researchers have studied the SaaS business model from different angles. They study SaaS contracts (Susarla et al. 2003; 2009), compare software publishers’ incentive to invest in software development under SaaS (Choudhary 2007), and analyze competition between the traditional software and SaaS (Balasubramanian et al. 2008; Choudhary 2007; Fan et al. 2009; Ma and Seidmann 2011). This paper, however, targets at a different research question: how SaaS providers compete among each other. Currently there are over 1000 SaaS providers actively operating and competing in different niche markets. The top 25 SaaS, including big vendors such as Oracles, Microsoft, and Salesfore.com, constitute about 50% of the revenue of the whole on-demand software market (Konary and Traudt 2005). The SaaS market is full of competition, far away from a monopoly marketplace. To understand the competition in SaaS market, we construct a game theory model that captures the uniqueness of SaaS setting. It considers both horizontal differentiation in software features/ functionalities and vertical differentiation in IT service qualities. In addition, non negligible switching costs are included in the analysis as well so that the model captures “lock-in” risk faced by SaaS users.

The paper is organized as follows. Section 2 describes the model. Section 3 presents the analysis as well as results. Finally, Section 4 discusses the practical implications of findings and concludes the paper.

2. The Model

Two SaaS providers, denoted by H and L respectively, compete in the market. Each delivers a bundle of software application and IT services to its customers, and charges a fixed subscription price p (per period). The two providers differ in two dimensions. First, their software applications are with different attributes and therefore horizontally differentiated. We adopt the Salop model (1979) which assumes the product space is a unit-length circle. The circle model allows me to ignore providers’ location decisions. Second, the two providers offer IT services at different quality levels: the provider H offers high quality service q_H and L offers low quality q_L. Services offered by the two providers are vertically differentiated. SaaS provider bears initial setup cost S and service cost c. The initial setup cost S is one-time cost incurred for each new customer. Service cost c is on-going cost incurred recurrently for every transaction period. The magnitude of service cost depends on the service quality level: $q = f(c)$ at $q = q_H, q_L$, $c = c_H, c_L$ respectively, and $\Delta c = c_H - c_L > 0$.

Users are assumed to have heterogeneous tastes toward software attributes, which is captured by the circle model (Salop 1979). A user’s location on the unit-length circle represents his ideal product, which is an individually customized application, fits a user’s specific technique requirements and business environment perfectly, and can be integrated with the user’s legacy systems seamlessly. Multi-tenancy of SaaS eliminates this possibility. Following the principle of maximum differentiation, we assume the two SaaS providers’ applications are located at positions opposite to each other. A user incurs a utility loss of tx for not using his ideal product, where x measures the distance between the user’s ideal product and the provider’s offering in the circle, and t is the parameter measuring the user’s unit *unfit costs*. Users all prefer higher service quality but their willingness-to-pay is heterogeneous. The market contains two types of users, h type and l type. An h type user is willing to pay more to obtain high level service quality than an l type user. Therefore, user j of SaaS provider I gains utility $u(\theta^j, q_i, d_i) = \theta^j q_i - p_i - td_i$,(1) where $j = h,l$; $i = H, L$, $\theta^h > \theta^l$; p_i denotes provider i's price (per period); and d_i measures the distance between this user and the provider i in the product circle.

Uncertainty exists on users' side because software is experience goods. Without upfront implementation, test running, and customization process, users are not able to predict precisely to what extent the SaaS application could fit his business needs and how good it can integrate with his legacy systems. Detailed information will be learned by users only after using the SaaS for a period of time. To capture this interesting feature of SaaS, we assume that the two providers' locations on the product circle are initially unknown to all potential users. A user therefore is not able to know his exact distance to one provider, which leads to an *ex ante* inaccurate estimation of unfit cost. In addition, the model also assumes switching cost E for existing SaaS users, and E is exogenously determined.

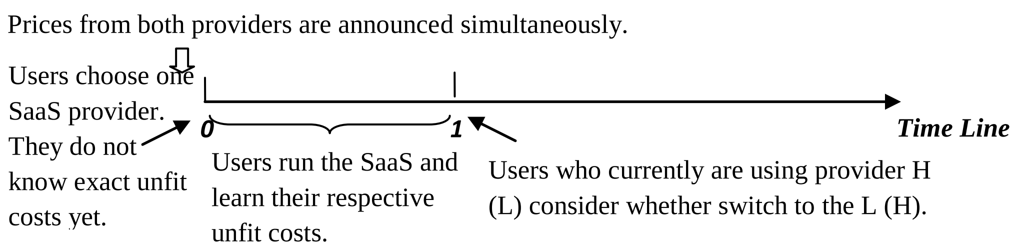
3. The Analysis

Figure1 draws the time line. There are two stages in the competition game, stage I ([0,1]) and stage II ([1,∞]); and two decision-making points, time spot *0* and *1*. The game gives rise in the following orders.

<u>Prior to Time spot 0</u>: SaaS providers post their price p_H and p_L simultaneously. Their service quality is known to potential users but locations on the circle space are unknown to users.
<u>Time spot *0*</u>: Users decide which provider to use with uncertain exact unfit costs.
<u>Stage I ([0,1])</u>: It is the transient stage. Users try the application chosen by them at time spot *0* and learn its full information (the exact unfit cost).
<u>Time spot 1</u>: Users decide to stay with their current SaaS provider or switch to the other SaaS provider by incurring switching cost E on users' side.
<u>Stage II, $[1,\infty]$</u>: Market stabilizes. Users become long-term partners with the SaaS provider.

Stage I is a short "trial" period. For simplicity, we assume both users and SaaS providers are maximizing their long-run utilities and profits from stage II, $[1,\infty]$. There are two discount factors, α for stage I and β for stage II. For example, if a user's total utility from period $[1,\infty]$ is K, it is βK if discounted to time spot 1 and $\alpha\beta\ K$ if discounted time spot 0.

Figure1. Time line for the competition game



Consider the two decision-making time spots 0 and 1. After users made their first choice at time spot 0, the market is segmented. However, this market outcome is temporary because users will switch at time spot 1, which leads to a different market segmentation. In this paper, the first market outcome (before users switch) is called "***intermediate*** market outcome" and the latter one (after users switch) is called "***final*** market outcome (equilibrium)." It is easy to see that there are five types of market outcomes possible to appear.

(1) (*h*H, *l*H): All users, both *h* type and *l* type, choose provider H.
(2) (*h*L, *l*L): All users, both *h* type and *l* type, choose provider L.

(3) (hH, lL): h type users choose provider H and l type users choose provider L.
(4) (hO, lL): l type users choose provider L, while h type users are indifferent between the two providers and therefore will randomly choose one of them.
(5) (hH, lO): h type users choose provider H, while l type users are indifferent between the two providers and therefore will randomly choose one of them.

Proposition 1. The ratio of price difference to quality difference ($\frac{\Delta p}{\Delta q}$) between the two providers determines the intermediate outcome. The market is (hH,lH) iff $\frac{\Delta p}{\Delta q} < \theta^l$; ($h$H,$l$O) iff $\frac{\Delta p}{\Delta q} = \theta^l$; ($h$H,$l$L) iff $\theta^h < \frac{\Delta p}{\Delta q} < \theta^l$; ($h$O,$l$L) iff $\frac{\Delta p}{\Delta q} = \theta^h$; and ($h$L,$l$L) iff $\frac{\Delta p}{\Delta q} > \theta^h$.

Proposition 1 describes the ***intermediate*** market outcome (before users switch). It shows that users' initial choice of SaaS providers critically depends on $\frac{\Delta p}{\Delta q}$. This ratio measures the charge on per unit of quality difference. When it is expensive to get one unit quality increase (a large $\frac{\Delta p}{\Delta q}$), more users tend to choose the lower quality service from provider L. The cutting-off value for h and l type users to choose provider L is just their user type θ: When $\frac{\Delta p}{\Delta q}$ exceeds θ^l , l type users choose provider L; when $\frac{\Delta p}{\Delta q}$ exceeds θ^h , h type users choose provider L.

Users switch if they find their unfit cost with the current SaaS is high and switching cost (E) is not high. Proposition 2 below gives out the threshold value for switching cost, exceeding which a SaaS provider can always exert absolute lock-in power on its existing customers, and describes the unique equilibrium in this circumstance.

Proposition 2. When $E > \frac{\beta t}{4}$, no user switches. The unique equilibrium is (hH,lL).

The threshold value derived here is quite simple; it just equals to a user's ex ante expected unfit costs. This is intuitively correct because users' switching is driven by the fact that unfit cost from the current provider is too high. As a tradeoff, they must balance the switching costs (if switch) and the unfit costs (if not switch).

It is however more interesting to study the case of $E < \frac{\beta t}{4}$. Some users switch at time spot 1; final market equilibrium is different from intermediate market outcome. For notation consistency and convenience, we name the final equilibrium according to its corresponding intermediate market outcome type. For example, an "(hH,lH) equilibrium" is the equilibrium which has an (hH,lH) intermediate market outcome.

Proposition 3 lists a complete set of equilibriums. These equilibriums are mutually exclusive. Due to space limit, detailed derivation steps are omitted here, but available upon request.

Proposition 3 When providers do not have full lock-in power, i.e., $E < \frac{\beta t}{4}$, equilibriums are

(1)if $(2\theta^l - \theta^h)\Delta q - \Delta c > \frac{2E - S}{\beta}$, there is an ($h$H,$l$H) equilibrium.

(2)if $(2\theta^h - \theta^l)\Delta q - \Delta c < \frac{S - 2E}{\beta}$, there is an ($h$L,$l$L) equilibrium.

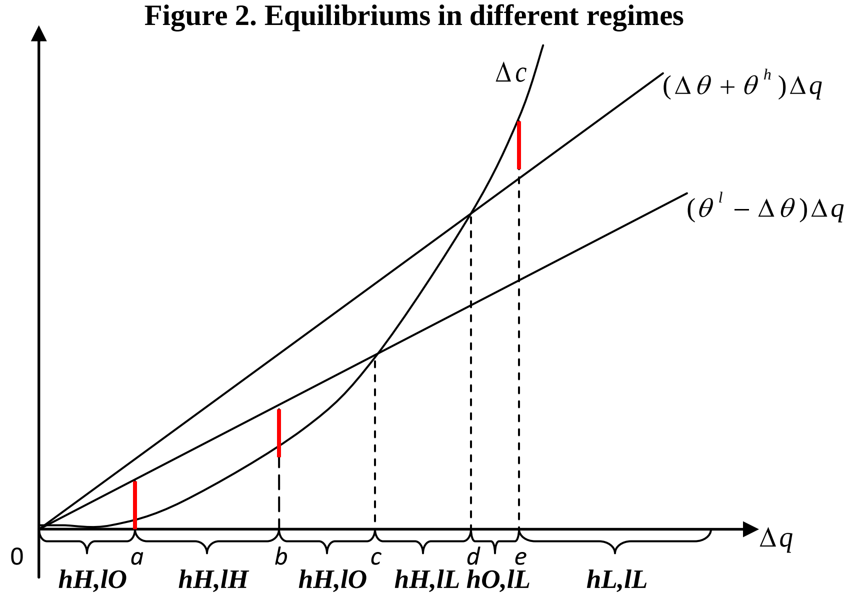
(3)if $(2\theta^l - \theta^h)\Delta q - \Delta c < 0$ and $(2\theta^h - \theta^l)\Delta q - \Delta c > 0$, there is an ($h$H,$l$L) equilibrium.

(4)if $\frac{S - 2E}{\beta} \le (2\theta^h - \theta^l)\Delta q - \Delta c \le 0$, there is an ($h$O,$l$L) equilibrium.

(5)if $0 \le (2\theta^l - \theta^h)\Delta q - \Delta c \le \frac{2E-S}{\beta}$, there is an (*h*H,*l*O) equilibrium.

Findings from Proposition 3 are demonstrated graphically in Figure 2. Three key aspects are noted: switching cost (*E*), providers' vertical differentiation level (Δq), and users' vertical differentiation level ($\Delta\theta$).

Figure 2. Equilibriums in different regimes



(1) Switching cost (*E*). Increase in switching cost shrinks regime of (*h*H, *l*H) and (*h*L, *l*L) equilibriums. Regime of (*h*H, *l*L) equilibrium remains the same, and regime of (*h*H, *l*O) and (*h*O, *l*L) increase. It indicates that when switching cost gets higher, a user becomes more conservative in making its initial choice. Therefore, he is reluctant to choose a provider with un-matched service quality level, namely, *h* type users tend not to choose the provider L while *l* type users tend not to choose the provider H.

(2) Providers' vertical differentiation level (Δq). The SaaS provider H delivers higher service quality. Regime (a,b) with equilibrium (*h*H, *l*H) represents the market that H can deliver such higher quality service at relatively small cost increments. In this case, all users will first try H; H possesses lock-in power. Switching cost benefits provider H **only** and may hurt L. However if H's higher quality service requires large cost increments, the market moves to regime (e,∞) with equilibrium (*h*L, *l*L). In this case, all users first try L; L possesses lock-in advantage. Switching cost benefits provider L **only** and may hurt H. When H can provide higher quality service at a moderate level of cost increments, it is regime (c,d) with equilibrium (*h*H, *l*L). In this case, provider H possesses lock-in power on *h* type users, and L possesses lock-in power on *l* type users. This finding suggests that a SaaS provider should strategically pick its service quality level, depending on its cost efficiency. For example, in a market that a low quality SaaS has existed (q_L has been chosen and fixed), a new entrant, i.e., a high quality SaaS, should consider choosing its service quality to be located in regime (a,b), which gives this new entrant lock-in power, instead of in regime (e,∞). In this example, it is better for the high quality provider to offer customers a small quality improvement (Δq) rather than a large one.

(3) Users' vertical differentiation level ($\Delta\theta$). Users become more vertically differentiated if θ^l decreases and θ^h increases. They tend to choose the provider who offers a matched service quality level. Therefore, (*h*H,*l*L) equilibrium is more likely to appear. In contrast, users become less vertically differentiated if θ^h decrease and θ^l increases. The two types of users are likely to make the same initial decisions. In this case, it is little chance to see (*h*H,*l*L) equilibrium but more likely to have (*h*H,*l*H) or (*h*L,*l*L) equilibrium in which all users will choose the same SaaS provider at the initial stage. When θ^h and θ^l change in the same direction, users' differentiation level may increase or decrease. The effects on equilibrium outcomes may be ambiguous.

4. Conclusion

We analyze duopoly SaaS providers' competition in a two-stage model. The competition takes place in a market with both vertical and horizontal differentiations. The SaaS software applications have different attributes while services are provided at different quality levels. In addition, users have incomplete information and face significant switching costs. We identify all equilibriums in different parameter regimes. We get the conventional results that switching costs can bring lock-in power and hence benefit the providers. The beneficiary of such lock-in power however could be either the high quality or low quality SaaS provider, depending on their relative cost efficiency. We highlight three factors that can affect which types of equilibriums to appear: switching costs, two providers' service quality difference, and users' willingness-to-pay difference. These findings suggest that a SaaS provider should set its quality level strategically so that it can leverage on switching cost to get extra competitive advantage.

References

Balasubramanian, S., Bhattacharya, S.; and Krishnan, V. Pricing information goods: a strategic analysis of the selling and pay-per-use mechanisms. *Working paper,* 2008

Choudhary, V. Comparison of software quality under perpetual licensing and software as a service. *Journal of Management Information Systems*, Vol. 24, No., 2, pp. 141-165, Fall 2007.

Fan, M., Kumar, S.; Whinston, A.B. Short-term and long-term competition between providers of shrink-wrap software and software as a service. *European Journal of Operational Research*, 2009.

Hickins, M. Oracle: On-demand is now on the grid. *Enterprise,* April 2007.

Konary, A., and Traudt, E. Building a superior software ownership experience through software as a service, IDC *White paper,*

Ma, D., and Seidmann, A. Analyze the price competition between SaaS and MOTS software business model. Working paper, 2011.

Salop, S. Monopolistic competition with outside goods. *The Bell Journal of Economics*, vol 10, no. 1, 1979.

Susarla, A., Barua, A.; and Whinston, A. Understanding the service component of application service provision: an empirical analysis of satisfaction with ASP services. *MIS Quarterly* 27(1), 2003.

Susarla, A., Barua, A.; and Whinston, A.B. A transaction cost perspective of software as a service business model. *Journal of Management Information Systems,* Vol. 26, No.2, pp. 205 - 240, Fall 2009.

The future of SaaS and IT infrastructure management: a survey of IT decision makers in North America. *IT Business Edge*, 2008.

Online Opinion Formation and Social Interactions

Lu Yan[†], Roch Guerin[‡], Kartik Hosanagar[‡], Yong Tan[*], Santosh Venkatesh[‡]

[†] Indiana University [‡] University of Pennsylvania [*] University of Washington

Abstract

In this paper we examine how opinions are formed when users are influenced by the opinions of others. Using a 5-year panel data set of online discussions on financial topics, we find the empirical evidence of the social influence of prior opinions on the formation and evolution of subsequent opinions. Specifically, WHO (the social role of the user) said WHAT (the quality of the information) and WHEN (the timing of comment being made) are found to have significant impact. One of the main contributions of this work is that our empirical approach can be used to calculate the "weights" of the influence network and predict the outcomes of social interactions.

1. Introduction

Given widespread access to social media applications, consumers widely seek others' opinions before making their own decisions (Sun 2012). They can easily get access to information from different resources, read specialized reviews, and check comments posted by earlier users. Consumers are thus "networked" in one way or another as they create, share, and consume via blogs, tweets, discussion boards, forums, and so forth. These networks affect the information that individuals access and the opinions they form. Understanding how opinions are formed when they are influenced by the opinions of others in a networked system and predicting the outcomes of those complex interactions are therefore a topic of much interest for researchers and practitioners.

The goal of this study is to empirically investigate the mechanisms of opinion formation and evolution in online communities where social interactions take place. Unlike previous studies which use the distribution and dispersion of online WOM to explain its social impact and consequent outcomes, our focus is on finding the strengths of communications among the individuals in a networked system. Specifically, the econometric framework constructed in this study allows estimation of the strength of influence (or edge weight) between a pair of individuals from the characteristics of their interactions which were captured in the dynamics of their opinion formation and evolution.

Contributing to the theoretical underpinnings of social influence, we examine the roles of three factors: opinion leaders, timing, and information quality. We propose a structural model to replicate the process of opinion formation and its subsequent update in a network environment. In the econometric model that we construct, the probability of a positive sentiment embedded in a message is expressed as a function of the intrinsic characteristics of the user who posted the message and the opinions of others that were exposed to this user. Then, we aggregate all the observations of message exchanges into an overall likelihood function which is maximized to yield the parameter estimates. We find strong evidence for the influence of other users' opinions on the evolution of a focal user's opinion. That is, when a consumer joins an online discussion, the prior opinions she observes affect the formation of her opinion or any subsequent updates of her original opinion. In particular, there is a reinforcing impact from observed prior opinions on her belief if these opinions come from trusted or credible users in the community. Second, the timing of a user's opinion disclosure in online discussion has an important impact on other users' decision making. Lastly, we find that information quality has a positive influence on others.

Our work fits into the literature that studies the propagation of information and resulting opinions across a network. Various experimental and field studies have proved that interpersonal influence has a significant impact on individual opinions. This work extends and contributes to the current research work by measuring the influence of one user on another in a network setting. Our approach is able to estimate the strength of influence between any two users in the network based on observed prior interactions. The calculation of the social influence in social interactions and the prediction of the opinion outcomes can help understand the dynamics of opinion formation – and consequently, the formation of aggregate public opinion – which has obvious economic implications.

2. Theory and Hypotheses

It is well known that changes of opinions are determined by social impact (Sznajd-Weron and Sznajd 2000). We consider a community in which a change in an individual's opinion is caused by information or opinion sharing with other consumers in the community. Thus, the strength of social interactions and the impact of their influence are of particular interest for this study. Specifically, the social pressure generated in the focused networked system is continuous and thus influence is captured from three aspects: opinion leadership, observational learning, and information quality.

Opinion leaders are individuals who are likely to influence others in the communication environment. These relatively small number of people, the 'influentials', are believed to have a substantial influence on the opinions and decisions of the majority. However, prior findings are mixed. While influentials were found to play a critical role in content diffusion and adoption in some studies (for example, Goldenberg et al. 2009), others have also suggested that large cascades of influences may not be driven by the influentials but rather by a critical mass of easily influenced individuals (Watts and Dodds 2007). Thus, to understand the role of influentials in the context of our study, we hypothesize:

H1: *Opinion leaders have a significant impact on others' opinion formation and subsequent opinion updates*.

Second, users exchange their opinions sequentially in online discussion forums. The social effect of an opinion could, therefore, likely depend on the time of its entry to the discussion thread. Thus, it is reasonable to assume that an opinion revealed later has a higher likelihood of being inconsistent with the initial opinion because of the accumulative learning effect in the discussion procedure. In other words, more recent opinions should have a higher weight:

H2: *The recency of an opinion has a positive impact on influencing others' opinions*.

Third, the convenience of applications for users to share their opinions gives rise to variations in the content and quality of opinions in discussion. This creates a substantial obstacle to assessing the usefulness of such shared knowledge (Cao et al. 2010, Christensen et al. 2004). If an opinion is shared by a marplot that always generates spam messages, it is safe to assume that other users will not be influenced and will ignore this message. Similarly, if a user continuously oscillates between the extremes of opinion, her message becomes less convincing in the discussion. Thus we hypothesize:

H3: *The quality of an opinion is positively related to its influence on other users*.

3. Model

Although it is interesting to study how a user's opinion is formed, we instead focus our attention on users' opinion exchanges and how those opinions are influenced by the judgments of others in this study. Specifically, we consider a group of users who participate in various discussions and thus are connected through an online community. They reveal their opinions in sequential messages. In a specific discussion s, i and j indicate the i^{th} and j^{th} message respectively. Since each message carries a time stamp and prior messages are publicly available, at the time a user submits her post j, we know all the opinions before j. The probability of a positive opinion embedded in the j^{th} message is thus expressed as a function of the intrinsic characteristics of the user who posts the message and the opinions of others that she has learned in discussions. The empirical specification of the model is explicitly defined as:

$$\Pr_j = \frac{\exp\left(\alpha_0 + \gamma Y_j + \sum_{i<j} \beta Z_{ij} \cdot x_i\right)}{1 + \exp\left(\alpha_0 + \gamma Y_j + \sum_{i<j} \beta Z_{ij} \cdot x_i\right)}; \qquad L = \prod_j \left(\Pr_j\right)^{\frac{1+x_j}{2}} \left(1 - \Pr_j\right)^{\frac{1-x_j}{2}}$$

where Z_{ij} refers to the social influence factors, and Y_j controls for the user's individual features, x_i is the sentiment (1 or −1) embedded in message *i*. The model includes the potential influence of all messages posted prior to the j^{th} message. It can be inferred from the model that positive impacts of social factors will enhance the content provider's opinion, whereas any negative coefficients will reduce her credence about her belief. As there are multiple messages in a discussion, we aggregate all the observations of message exchanges into an overall likelihood function which is then maximized to yield the parameter estimates.

4. Data and Variables

We collected data from a website that provides a platform for consumers to talk about the performance of stocks and investments. Discussions and messages were collected for a large financial firm from Jan. 2008 to Nov. 2012. Messages within a discussion are listed by time and each message contains information such as user name, content, rating, and time stamp. Users in the community can learn what others are thinking about the firm through collaborative communications. As a result, 24,780 discussions had been generated by 6,584 users and 153,129 messages in total had been posted during the data collection period. Table 1 provides the data description and statistics for this study.

The opinion that consumers reveal in their online text messages is the key item that interests us, and we therefore use LingPipe[1] to conduct sentiment analysis and classify these messages as positive "+" or negative "–" (in probabilities). During message exchanges, users can be "persuaded" and may change their decisions, which is expressed by a user flipping her opinion.

Table 1: Data Description and Statistics

Variables	Description	Mean	Median	Min	Max
x_i	the opinion in the post: -1 – negative, 1 – positive	-0.33	-1	-1	1
$NumPost_i$	the number of posts the individual who uploads post *i* in the discussion has posted for this stock, i.e., including all posts the user has posted for this firm so far	22.21	19	1	627
$AvgOp_i$	the absolute value of the averaged opinion of the individual who posts the post *i*	0.55	0.4435	0	1
$Celeb_i$	binary variable – is the proxy for influential (e.g. if the username has been called by other users in other	0.26	0	0	1

[1] Alias-i. 2008. LingPipe 4.1.0. http://alias-i.com/lingpipe (accessed October 1, 2008)

	discussion titles; 1- yes 0 – no)				
$\Pr ePart_i$	dummy variable – if the user who posts post i has participated in this discussion before: 1- yes 0 – no	0.50	0	0	1
$\Pr eOp_i$	the most recent opinion of the user who posts post i in the discussion	-0.33	-1	-1	1
Pos_i	the position of the ith message in the discussion	4.17	6	1	62
$SDay_{ij}$	binary variable, 1 – if message i and j are posted at the same day, 0 – otherwise	0.51	1	0	1
$LnNumDays_{ij}$	the time difference between message i and j (in days)	2.51	2	1	14
Op_i	is 1 if the opinion embedded in post i is positive, 0 otherwise	0.375	0	0	1

5. Estimation Results

We use a maximum likelihood method to estimate the parameters of the proposed model. Necessary procedures were taken to remove correlation problems. Our data exhibit a long tail pattern in which 50% of discussions have fewer than 4 posts, while only 0.5% of the discussions contain more than 35 exchanged messages. However, the latter small portion of discussions covers 2% of the messages in our dataset. To address these issues, the dataset is partitioned into 3 cases: Case 1 is created by including discussions that have 36 contributed messages or more; Case 2 is formed by discussions that have fewer than 4 message exchanges; Case 3 contains all discussions and messages. We perform our analysis on each case and Table 2 reports the estimation results.

Table 2: Estimation Results

Variables	Case 1		Case 2		Case 3	
	est.	std	est.	std	est.	std
Constant						
α_0	-1.0750***	(0.0038)	-1.1088***	(0.0077)	-0.4220***	(0.0040)
Social Interactions						
β_0	0.4564***	(0.0058)	0.8711***	(0.0049)	0.1339***	(0.0017)
$\beta_1[NumPost_i]$	0.3115***	(0.0039)	0.3798***	(0.0031)	0.5230***	(0.0061)
$\beta_2[AvgOp_i]$	0.4680***	(0.0039)	0.8986***	(0.0033)	0.9853***	(0.0072)
$\beta_3[Celeb_i]$	1.0007 ***	(0.0023)	0.4560 ***	(0.0033)	0.9356***	(0.0050)
$\beta_4[Pos_i]$	0.0543***	(0.0046)	0.5442***	(0.0021)	1.1537 ***	(0.0032)
$\beta_5[SDay_{ij}]$	0.2454***	(0.0024)	0.2160***	(0.0060)	0.6250***	(0.0031)
$\beta_6[LnNumDays_{ij}]$	-0.3229***	(0.0033)	-0.2027***	(0.0052)	-0.7223***	(0.0088)
$\beta_7[Op_i]$	0.6439***	(0.0029)	0.3340***	(0.0037)	0.5548***	(0.0063)
Individual Characteristics						
$\gamma_1[NumPost_j]$	0.5704***	(0.0040)	1.0511***	(0.0051)	0.4548***	(0.0027)
$\gamma_2[AvgOp_j]$	0.3614***	(0.0064)	1.0395***	(0.0042)	0.7960***	(0.0032)
$\gamma_3[Celeb_j]$	0.5432***	(0.0081)	0.5434***	(0.0039)	0.1394***	(0.0011)
$\gamma_4[\Pr ePart_j]$	1.0113***	(0.0024)	-0.0439***	(0.0050)	-1.0550***	(0.0009)
$\gamma_5[\Pr eOp_j]$	1.1755***	(0.0065)	0.4057***	(0.0083)	0.9952***	(0.0055)
$\gamma_6[\Pr eOp_j * \Pr ePart_j]$	0.5648***	(0.0092)	0.9204***	(0.0006)	0.3712***	(0.0046)

Opinion Leader

There are two variables in the model that are used as proxies for opinion leaders, namely $NumPost_i$ and $Celeb_i$, and we measure the coefficients of these two factors to interpret the social impact of opinion leaders. In general, these factors have positive and significant impact across all 3 cases. The variable $NumPost_i$ indicates the number of prior messages that have been contributed by the user. The higher the number, the greater the knowledge that has been shared by the user. In other words, when a user actively participates in various discussions, she is more visible than other users who do not. The number of messages she has contributed to the community not only indicates her expertise but also gives her a greater chance of being recognized as a leader in discussions. Under such circumstances, the opinion she expresses is more persuasive. Similarly, if a user has been called to participate in discussions ($Celeb_i$), this invitation suggests the importance of her opinions to other members in the community. Consistent with other online WOM studies (for example, Goldedberg 2009), these influential users' messages are more attractive and thus impact the diffusion of content which in our study represents the influence on other users' opinions. Hence, Hypothesis 1 is supported by the positive and significant estimates β_1 and β_3.

Timing in Observational Learning

Our results indicate the important impact of the timing of opinion disclosure in online discussions on other users' decision making. We find evidence to support Hypothesis 2 in all 3 cases (positive and significant β_4 and β_5, negative and significant β_6). First, when an opinion has been added to the discussion, it has a significant impact on influencing followers' beliefs. The variable Pos_i controls for the position of the message in the discussion. The higher the number, the later the position of the message in the discussion. A later position reflects the most recent opinions in the discussion. Therefore, the opinion extracted from such a message has a higher chance of being concordant with the newly generated post because it is formed by similar knowledge accumulated from previous discussions and is thus more convincing. $SDay_{ij}$ and $LnNumDays_{ij}$ are variables that control for the time difference between previous opinions and the focal message. The closer the previous post is to the focal message, the more influential the post will be. These two covariates control when the opinion is exposed in the discussion. The more recent the opinion extracted from a posted message, *i.e.*, the closer to the focal opinion, the higher will be the supportive impact on the focal message.

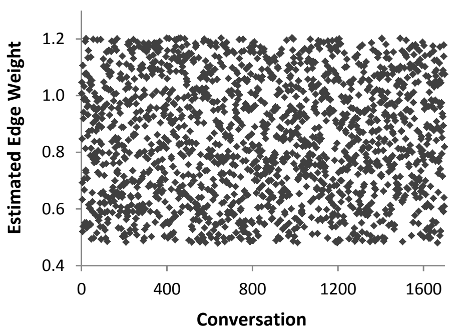
Information Quality

In the model, we construct variable $AvgOp_i$ as a proxy for information quality. $AvgOp_i$ is calculated as the absolute value of a user's previously revealed average opinion in the community. If the number is close to 1, it suggests that the user is consistent in her opinion, whether positive or negative, and she is confident about her decision. However, if a user is easily influenced and often flips her opinions, this variable will be close to 0. As is expected, a confident user is more credible and thus has a positive influence on others (positive and significant β_2). As a result, Hypothesis 3 is supported.

Based on the estimated parameters, we also use the estimated β to calculate βZ_{ij}, i.e., the connection "weights", and these weights can then be used to forecast how opinions evolve in a network. The edge weight between any pair of user who had social interactions appears to be uniformly distributed in the range from 0.45 to 1.2 (as shown in the figure). For example, if a pair of users has an edge weight of 0.8, a positive post by the first user has a 69% chance to induce a positive message from the second user, ignoring all other factors.

6. Conclusion

It is widely believed that social media can have a powerful impact (Kietzmann et al. 2011). Users share feedback about products and services, exchange knowledge and opinions and work together to create the "knowledge of the many" (Dellarocas 2006). This study is intended to measure the social influence of connections in an online community. By applying data collected from a real-world discussion board to the proposed model, we are able to estimate coefficients to evaluate the impact of social influence via social interactions. Added to existing evidence that consumers are influenced by Internet-based opinion forums before making a variety of purchase decisions (Thompson 2003, Senecal and Nantel 2004, Chevalier and Mayzlin 2006), our findings can be used to infer the "weights" of others' opinions on an individual's decision.



One extension of the work is to deploy the model on a dataset with a longer history and study necessary conditions to achieve public opinion agreement. Another extension could be further subdividing the communication network based on users' characteristics and study what key characteristics that increase the probability of realizing opinion agreement.

Acknowledgments. This work was supported in part by the National Science Foundation under Grant No. CCF-1137519 and CCF-1137597.

References

Cao, Q., Duan, W., Gan, Q. 2011. "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," *Decision Support Systems* (50), pp. 511-521.

Chevalier J.A., Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: online Book Reviews," *Journal of Marketing Research* (XLIII), pp. 345-354.

Christensen, H., Griffiths, K. M., & Jorm, A. F. (2004). "Delivering interventions for depression by using the Internet: Randomised controlled trial," *British Medical Journal* (328), pp. 265-268.

Dellarocas, C. 2006. "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science* (52:10), October, pp. 1577-1593.

Godes, D., Mayzlin, D. 2004. "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science* (23:4), pp. 545-560.

Goldendberg, J., Han, S., Lehman, D.R., Hong, J.W. 2009. "The Role of Hubs in the AdoptionProcess," *Journal of Marking* (73), March, pp.1-13.

Senecal S., Nantel J. 2004. "The influence of online product recommendations on consumers' online choices," *J. Retailing* (80), pp.159–169.

Sun, M. 2012. "How Does Variance of Product Ratings Matter?" *Management Science* (58:4) April, pp. 696-707.

Sznajd-Weron, K., Sznajd,J. 2000. "Opinion Evolution in Closed Community," *International Journal of Modern Physics C* (11:6), pp.1157-1165.

Thompson, N. 2003. "More companies pay heed to their "word of mouse" reputation," *New York Times* June 23.

Uzzi, B. 1997. "Social Structure and Competition in Interfirm Networks: The Problem of Embeddedness,"*Administrative Science Quarterly* (42), pp 35–67.

Watts, J. D., Dodds, P.S. 2007. "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research* 34(4), 441-58.

News Article Propagation on Twitter based on Network Measures - An Exploratory Analysis

Devipsita Bhattacharya, Sudha Ram
Department of Management Information Systems, Eller College of Management,
The University of Arizona, Tucson, AZ 85721, USA

Abstract: News agencies regularly use Twitter to publicize and increase readership of their articles. In this research, we use network analysis techniques to compare the diffusion process for articles published by different news agencies. We propose five different measures based on Twitter activity networks and use them to compare the propagation process. Using a dataset of tweets collected over a period of six months, we provide crucial insights into factors that influence the spread of news articles and provide a comprehensive account of news article propagation on Twitter.
Keywords - Social Media; Twitter; Micro-blogging; News Diffusion; Information Diffusion; Social Networks

1. Introduction

Internet users are increasingly turning to social media such as Twitter to access news articles from a variety of sources[1]. Many news agencies, such as NYTimes and Washington Post, use Twitter to post their news articles. These news articles propagate via the tweet-retweet-reply cascades caused by the news sharing activity of users of the micro-blogging platform. This proliferation of user responses provides an excellent insight into the mechanics of online news diffusion and also gives a new perspective for the study of information diffusion in general. Traditionally news diffusion has been studied at the granularity of each individual participant; the Epidemiological models (Kermack and McKendrick 1991; Newman 2002); Product Adoption models (Rogers 1995; Bass 2004); and Word-of-Mouth (WOM) model (Katz and Lazarsfeld 2005) to name a few. Instead of following this approach, we analyze the diffusion of these news articles based on the behavior of the crowd; that is the collective response of users. Both of these concepts (individual and group) have been applied for studying different aspects of information diffusion on Twitter. Most relevant of these include (Bakshy et al. 2011)'s use of a WOM model to explore the factors influencing tweets cascades of article URLs; (Xiong et al. 2012)'s revision of the SIR model to include a new state called "contacted" to predict propagation probability on Twitter; (Lerman and Ghosh 2010)'s exploration of the effect of network structure on cascade patterns by comparing the user activity and friendship network between Digg and Twitter; (Bakshy et al. 2009)'s study of friendship networks formed by members to observe their patterns of gift/asset sharing on a MMORPG called Second Life[2]; and (Bandari et al. 2012)'s prediction of news article popularity based on the characteristics of the content being shared. Our work is different because this research proposes a new technique for studying online diffusion - via topological analysis of news dispersion patterns. We accomplish this by constructing tweet cascade networks derived from user propagation data and timestamps. We also apply the theory of network analysis to develop metrics for diffusion and for gauging audience participation. Specifically we focus on the following aspects:

a. News article cascades caused by Twitter users' participation via retweets and replies

---

[1] Understanding the Participatory News Consumer - http://www.pewinternet.org/Reports/2010/Online-News/Summary-of-Findings.aspx

[2] MMORPG -Massively multiplayer online role-playing game

b. Twitter news source Followers' engagement in the diffusion process.
c. Profile characteristics of the prominent news propagators.

2. Data Collection Process

We collected a dataset from Twitter containing URLs of news articles from 12 specific news sources (see Table 1) during the period from September 2011 to March 2012. From this set we created a subset of 6 million tweets for four random weeks allowing us to analyze weekly snapshots of activity and to smooth out effects of factors such as seasonality of news. Only those tweets containing URLs of articles published by these news sources, including original tweets as well as retweets and replies were included in the dataset.

Table 1: Selected News Sources and Follower Count for Twitter News Diffusion

Arstechnica	BBC News	Bloomberg	Financial Times	Forbes	Guardian
373,569	*808,769*	*603,740*	*680,545*	*872,376*	*412,625*
Mashable	New York Times	NPR	Reuters	Washington Post	Wired
2,933,693	*5,571,425*	*1,069,326*	*1,867,784*	*1,077,115*	*1,522,175*

3. Data Analysis

We used social network and graph theory concepts to construct user cascade networks and to define important network measures for news article propagation. Our User-User network consists of a directed graph G of nodes and weighted edges: G = (N, E, W) (Biggs et al. 1999), where

- N - Set of all nodes. Each node is a Twitter user, including news agency's Twitter account.
- E - Set of all edges. An edge represents a directed communication between two Twitter users, i.e. from source user to the responding user (user who retweets or replies).
- W - Set of all the edge weights. Edge weight represents the number of directed interactions between two Twitter users

In this network, some nodes represent the news agency's Twitter account (we refer to these as the media seeding nodes), while others are general Twitter users who are propagators of news. We constructed one such network for each day of activity for each news media source. Note that we aggregated data for each day of the week using a sample of tweets over four different weeks spread over a 6-month period. We refer to these networks as the *Twitter activity networks*.

a. Analyzing Twitter Activity Network Cascades

As stated earlier, we created networks to analyze the cascade streams for each day of a week. Since this network accumulates all the Twitter posting activity related to a news source for a specific day, there is a significant section of the network containing unconnected nodes. These nodes represent Twitter users whose tweets do not receive any retweets/replies.

Figure 1: Twitter Activity Network for BBCNews for One Week in December 2011

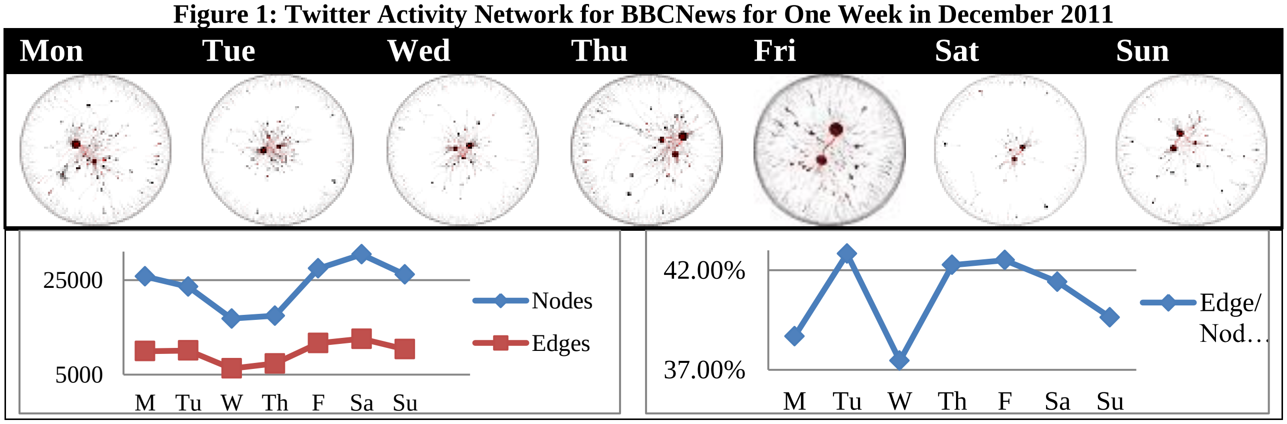


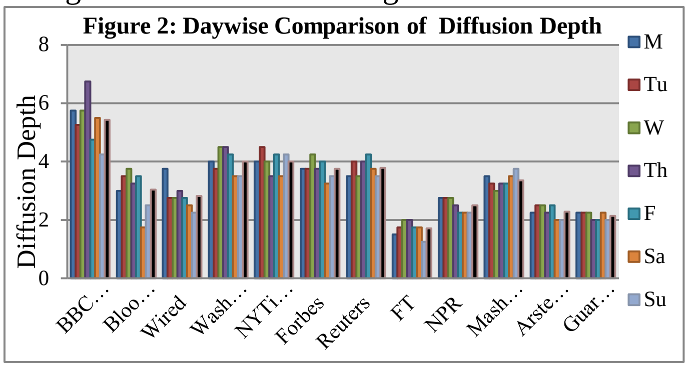
Figure 1 shows an example of the BBC News Twitter activity network during the second week of December 2011. It shows the daily networks as well as the number of nodes and edges and edge/node ratio.

a.1. Network Diameter to Measure Diffusion Depth

Network diameter or maximum eccentricity in a diffusion network represents the longest cascade chain in the network (Hage and Harary 1995). Using this structural property we define the first diffusion metric as:

$$\textbf{\textit{Diffusion Depth}} = \textit{Number of Cascade Levels in the Diffusion Network} = \textit{Maximum Eccentricity}$$

Using the news media seeding node as the initiator of the cascade process, the average diffusion depth of the network for each day of the week (over a 4-week period) is shown in Figure 2. The average weekly diffusion depth for BBC News, Washington Post and NYTimes are among the highest with values of 5.43, 4.00 and 4.00 respectively. The shortest diffusion depth values are for FT, Guardian and Arstechnica at 1.71, 2.14 and 2.29 respectively.

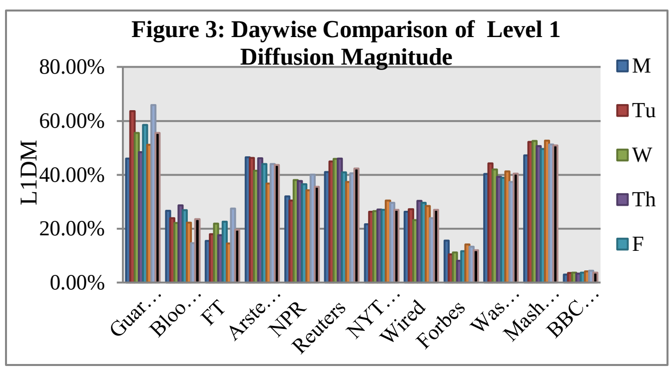


a.2. Ego Network Nodes Concentration to Measure Diffusion Magnitude

Next, to measure diffusion magnitude, we created an ego network for each media seeding node (Freeman 1982). Each such ego network has several levels; level 1 represents all the users who directly retweet or reply to the media seeding node, and level 2 are the users who retweet or reply to level 1 nodes and so on. Thus, node concentration at level 1 (expressed as a percentage of total number of nodes in the Twitter Activity Network) provides a normalized measure of direct retweets and represents a diffusion metric called *Level 1 Diffusion Magnitude*. This metric is generalized for all levels of the news Ego network to yield the metric *Level N Diffusion Magnitude*. This is defined as follows:

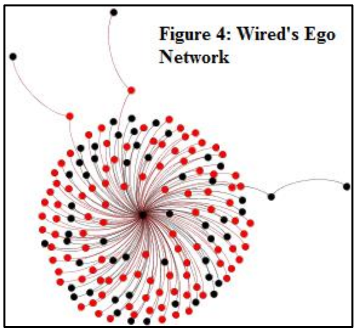
$$\textbf{\textit{Level N Diffusion Magnitude (LNDM)}} = \frac{\textit{Number of Nodes in Level N of Ego Network}}{\textit{Total Number of Nodes in Twitter Activity Network}}$$

Figure 3 shows the average diffusion magnitude at level 1. BBC News has a significantly lower diffusion magnitude at level 1 as compared to all other news sources, averaging at 3.73% while Guardian has the highest average Level 1 diffusion magnitude at 55.62%.



b. Followers Engagement, Density and Participation in the Diffusion Network

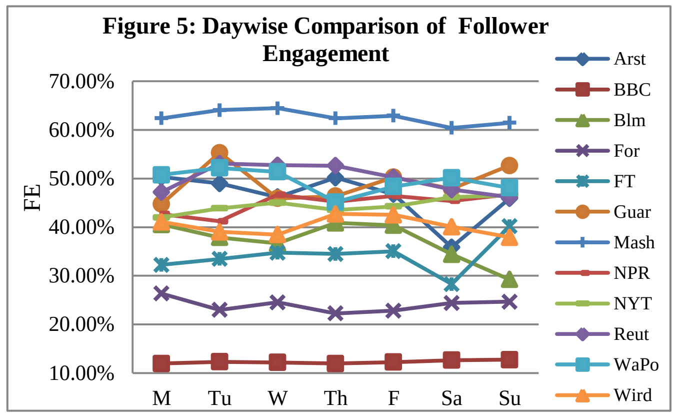
Each news source has a number of followers on Twitter. We analyzed the composition of followers in the Twitter activity networks as well as in the ego networks at various levels using sub-graph based analysis (Biggs et al. 1999). Figure 4 shows the example of an ego network for Wired at level 1 for a single day, with red nodes representing its followers. In this network the



nodes that are also followers of Wired are marked in red. It is obvious that a high percentage of level 1 nodes of Wired are its followers. Followers of each media source were found to exist at different levels of their ego networks.

Some followers also exist in the disconnected part of the Twitter activity network for each news source. We define a measure called *Follower Engagement* as:

$$\textbf{\textit{Follower Engagement (FE)}} = \frac{\textit{Total Number of Follower Nodes in the Twitter Activity Network}}{\textit{Total Number of Nodes in the Twitter Activity Network}}$$

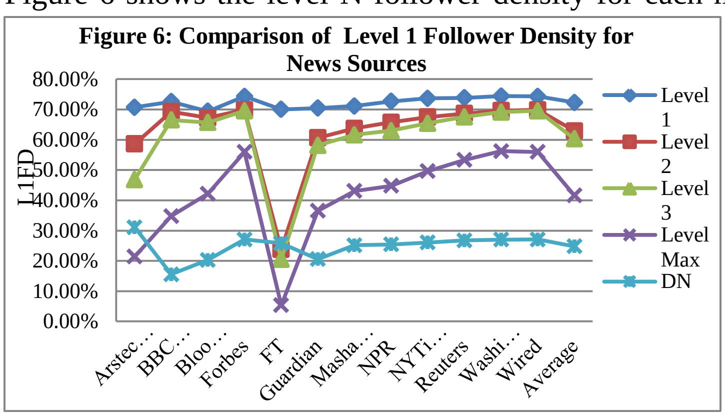


A time wise comparison of follower engagement during each day of the week is shown in Figure 5. For a given news source, the concentration of followers remains fairly constant throughout the week. This is confirmed by the fact that the average variance of follower engagement among 12 news sources is 0.00083. Only in the case of Bloomberg (10% drop from Friday to Sunday) and Guardian (11% increase from Monday to Tuesday and 5% from Friday to Saturday) are any significant changes observed. BBC News (12.3%), Forbes (24%) and FT (34.07%) have the lowest follower engagement in their Twitter activity networks.

To further investigate the role of followers, we analyzed the percentage composition of followers in each level of the ego network. We define a metric called *Level N Follower Density* as follows:

$$\textbf{\textit{Level N Follower Density (LNFD)}} = \frac{\textit{Number of Follower Nodes in Level N of the Ego Network}}{\textit{Total Number of Nodes in Level N of the Ego Network}}$$

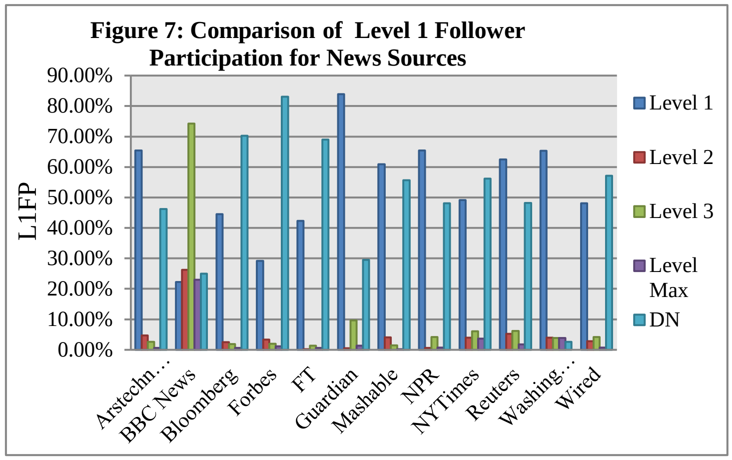
Figure 6 shows the level N follower density for each media source. Note that "DN" represents the follower density in the disconnected part of the Twitter activity network. Level 1 of each news source's ego network has the highest follower density with an average of 72% across all media sources. This implies that around 3/4 of the retweets/replies of the news sources come from their followers. The remaining levels of the ego network follow similar patterns across different news sources. The density at levels 2 and 3 are similar to the level 1 ego network, having average values of approximately 63% and 60% respectively.



Finally, we examined the distribution of followers at different levels and in the disconnected section of the network. The diffusion measure capturing this aspect of network followers is named *Level N Follower Participation*. It is defined as follows:

$$\textbf{\textit{Level N Follower Participation (LNFP)}} = \frac{\textit{Number of Follower Nodes in Level N of Ego Network}}{\textit{Total Number of Follower Nodes in Twitter Activity Network}}$$

As depicted in Figure 7, followers are engaged mostly at the level 1 of the news source ego network and in the disconnected section of the activity network. Guardian (83.95%), NPR (65.41%) and Arstechnica (65.47%) have the highest average level 1 follower participation. Forbes (83.14%), Bloomberg (70.24%) and FT (69.02%) have the highest average disconnected section follower participation. This shows that followers not only re-tweet the news source nodes but also directly tweet about the news articles. Lastly, BBC News (74.29% at Level 3) has significantly higher follower participation in levels 2 and up of its ego network.



c. Twitter Users' Classification Based on Profile Characteristics

Finally, for an all-round understanding of the audience type, major users in the Twitter activity network were identified and categorized into groups based on their profile description. Major nodes are defined as the ones in the 99th percentile of the network based on weighted out-degree ranking. These nodes either act as hubs in news source ego networks or create independent streams in the Twitter activity network for the news source. A significant fraction of the participants in the news propagation network is comprised of personnel directly affiliated with the news sources (such as authors) or members of other professional organizations.

4. Discussion, Implications and Conclusion

In this study of Twitter-based news diffusion, we derived metrics related to dispersion extent and user influence from structural properties of the news cascade network. As apparent from the analysis presented in the previous section, each of these metrics offer a unique dimension to the group response in news diffusion. But to understand the effectiveness of the diffusion process associated with each news source (Refer Table 1), the collective influence of these measures needs to be considered. Table 2 shows the dashboard containing all these values together for evaluating the efficacy of news diffusion as observed for every news source. Sources with higher tweeting activity have higher diffusion depth for each day. Also, diffusion depth and diffusion magnitude appear to be inversely related, i.e. news sources with low diffusion depths have higher diffusion magnitude and vice versa. On the other hand, follower engagement is directly proportional to the diffusion magnitude. This indicates that involvement of followers in the diffusion process for a news source is more important that the number of followers of that news source. This is counter-intuitive to the idea that higher subscriber count will lead to greater news dispersion. BBC News has the highest diffusion depth and the highest tweeting activity, but its diffusion magnitude and follower engagement are among the lowest. In contrast, Guardian has one of the lowest diffusion depths, but has high diffusion magnitude, follower engagement and participation. Financial Times and Reuters are somewhat different in their diffusion networks. FT has low diffusion depth, magnitude, follower engagement and density; but exhibits high participation of followers in its level 1 ego network. Reuters has a medium to high ranking values for all the different diffusion measures, indicating a successful diffusion process. Finally, Wired has high follower density but performs poorly in follower participation.

Table 2: Comparison of Diffusion Metrics

	Diffusion Depth	Level 1 Diffusion Magnitude	Follower Engagement	Level N Follower Density					Level N Follower Participation				
				Level 1	*Level 2*	*Level 3*	*Level Max*	*DN*	*Level 1*	*Level 2*	*Level 3*	*Level Max*	*DN*
Arstechnica	2.29	43.62%	46.31%	70.71%	58.59%	46.81%	21.43%	31.10%	44.57%	2.52%	1.89%	0.56%	70.24%
BBC News	5.43	3.73%	12.30%	72.54%	69.30%	66.55%	34.80%	15.56%	48.14%	2.84%	4.13%	0.73%	57.16%
Bloomberg	3.04	23.65%	37.14%	69.37%	67.21%	65.67%	42.16%	20.30%	83.95%	0.52%	9.68%	1.30%	29.54%
Forbes	3.75	12.10%	24.00%	74.31%	69.85%	69.55%	55.98%	27.07%	65.41%	0.56%	4.16%	0.70%	48.15%
FT	1.71	19.67%	34.07%	70.00%	23.76%	20.48%	5.36%	25.83%	65.47%	4.71%	2.61%	0.60%	46.16%
Guardian	2.14	55.62%	49.02%	70.45%	60.61%	58.11%	36.54%	20.56%	22.26%	26.27%	74.29%	23.01%	25.00%
Mashable	3.36	50.91%	62.58%	71.14%	63.76%	61.62%	43.10%	25.18%	60.93%	4.04%	1.45%	0.14%	55.65%
NPR	2.5	35.61%	44.90%	72.67%	65.77%	63.02%	44.78%	25.45%	29.16%	3.32%	1.98%	1.14%	83.14%
NYTimes	4	27.00%	44.49%	73.68%	67.50%	65.44%	49.63%	26.07%	62.51%	5.23%	6.16%	1.77%	48.18%
Reuters	3.79	42.42%	49.96%	73.85%	68.63%	67.52%	53.32%	26.73%	49.11%	3.92%	6.11%	3.62%	56.19%
Washington Post	4	40.49%	49.44%	74.42%	69.61%	69.12%	56.25%	27.00%	65.38%	3.97%	3.88%	3.88%	2.59%
Wired	2.82	27.04%	40.26%	74.31%	69.85%	69.55%	55.98%	27.07%	42.32%	0.18%	1.32%	0.63%	69.02%
		Legend		Very Low		Low		Medium		High		Very High	

To summarize, the diffusion measures defined in this paper are useful in analyzing propagation patterns associated with news sources. In our earlier work (Bhattacharya and Ram 2012), we did a half-life and survival analysis of articles for different news sources and found that the rate of spread and propagation extent vary by news source. The implications of this work complement the earlier study and add a set of network-based factors to understand the diffusion process. Our future work will focus on using the results of this study to develop concrete variables for predicting the extent of diffusion.

5. References

Bakshy, E., J. M. Hofman, W. A. Mason, D. J. Watts. 2011. Everyone's an influencer: quantifying influence on twitter. *Proceedings of the fourth ACM international conference on Web search and data mining*. 65–74.

Bakshy, E., B. Karrer, L. A. Adamic. 2009. Social influence and the diffusion of user-created content. *Proceedings of the tenth ACM conference on Electronic commerce*. 325–334.

Bandari, R., S. Asur, B. A. Huberman. 2012. The pulse of news in social media: Forecasting popularity. *Proceedings of the 6th International Conference on Weblogs and Social Media*.

Bass, F. M. 2004. A new product growth for model consumer durables. Management Science. **50**(12) 1825–1832.

Bhattacharya, D., S. Ram. 2012. Sharing News Articles Using 140 Characters - A Diffusion Analysis on Twitter. *IEEE International Workshop on Business Applications of Social Network Analysis*.

Biggs, N., E. K. Lloyd, R. J. Wilson. 1999. *Graph Theory 1736-1936*. Oxford University Press.

Freeman, L. C. 1982. Centered graphs and the structure of ego networks. *Mathematical Social Sciences*. **3**(3) 291–304.

Hage, P., F. Harary. 1995. Eccentricity and centrality in networks. *Social networks*. **17**(1) 57–63.

Katz, E., P. Lazarsfeld. 2005. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers.

Kermack, W., A. McKendrick. 1991. Contributions to the mathematical theory of epidemics—I. *Bulletin of mathematical biology*. **53**(1) 33–55.

Lerman, K., R. Ghosh. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of the 4th International Conference on Weblogs and Social Media*.

Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Physical Review E*. **66**(1) 016128.

Rogers, E. M. 1995. *Diffusion of innovations*. Simon and Schuster.

Xiong, F., Y. Liu, Z. Zhang, J. Zhu, Y. Zhang. 2012. An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A*. **376**(30-31) 2103-2108**.**

Improving correctness of scale-free networks to message distortion

Arie Jacobi (Ono Academic College, jacobi.arie@gmail.com)

Ofir Ben-Assuli (Ono Academic College, ofir@ono.ac.il)

Abstract

Vast numbers of organizations and individuals communicate every day by sending messages over social networks. These messages, however, are subject to change as they propagate through the network. This paper attempts to calculate the distortion of a message as it propagates in a social network with a scale free topology, and to establish a remedial process in which a node will correct the distortion during the diffusion process, in order to improve the robustness of scale-free networks to message distortion. We test a model that we created using a simulation of different types of scale-free networks, and we compared different sets of corrective nodes; hubs and regular (non-hubs) nodes. The findings show that using hubs that correct the distorted message while it's diffused, decrease a global error measurement of the distortion, and as a result improve the robustness of the network.

Introduction

Regardless of era, people communicate by sending messages to each other. The Web 2.0 information revolution has led to an exponential growth in the numbers of organizations and individuals who communicate by sending messages over electronic social networks. A message can be verbal or written, and the mechanisms for sending messages include real-time chatting applications in an electronic social network, Word Of Mouth (WOM) marketing methods, electronic WOMs (Sohn, 2009), or by forwarding a message in an email with an additional text. These messages, however, are subject to change as they are diffused in the network. A practical example is an advertising campaign in an electronic social network that has a clear and concise message statement, and it should be diffused in a social network as accurately as possible with less distortion as possible. An example of distorted advertising campaign in the internet was for Apple's iPad product. Though it is very successful, its name was promptly changed in the internet into 'iTampon' and many jokes and results. Therefore, practical methods for correcting distorted messages while diffused in social networks are needed.

Literature Survey

Propagation and Data Distortion via Social Networks

Social network analysis assesses information opportunities for individuals or groups in terms of exposure to information and its control. By identifying existing information exchange routes, information providers can act on information opportunities and make changes to information routes to improve the delivery of information services. Panzarasa et al. (2009) inferred the processes underpinning dynamics of use of network online communities established over time by online messages. They investigated the role played by hubs, which moderators and managers can appropriate to devise appropriate strategies for improving the security of communication and enhancing the effectiveness of information diffusion. Ma et al. (2010) argued that an information explosion can take place in which the number of distinct pieces of information in the network increases continuously with time, leading to high error probability. They describe a control strategy to maximize the robustness of the network against information explosion. Diffusion of information has been extensively researched and different models have proposed, e.g. Rogers (1995) and Valente (1996).

Crucitti and Marchiori (2001, 2003) introduced a definition of the efficiency of a network to propagate information. They assumed that the efficiency Q_{ij} of the communication between node i and j in a graph G is inversely proportional to the shortest distance: $Q_{ij} = 1/d_{ij}$, $\forall i, j$. The efficiency was defined on global and local scales. Local efficiency was defined as the average efficiency of local sub-graphs.

In this paper, we refer to the concept of robustness with regard to the resilience of a network to distortion of information diffused in it, as opposed to use of robustness in graph theory and network analysis (Crucitti et al. 2003; Singer 2006) that measures the resilience of the network to the removal of edges or vertices. To the best of our knowledge there have been no studies on the robustness of different types of social networks to distortion of information.

Methodology

In this paper we present a study of the robustness of a scale-free network to distortion of information in the form of a verbal or written message, and ways to decrease the total distortion error. A message has a tendency to change when one person transfers it to another person. Our assumption is that this change in the transferred message, what we call its distortion, is usually reflected in different parts of the message, but some of the information remains unaltered. The simulation involved a model of the network and a model of the distortion process propagating in the network. The propagation model was tailored so as to reflect the realities of the dissemination of information in a social network. To ensure accurate message correction, the corrective nodes are paid.

Proposed Model

Given a network *N* with *n* nodes, and a message *m* which represents a sequence of letters, words, or parts of sentences, without loss of generality in this model we define *m* as a sequence of letters $\sigma_i \in \sum, i = 1,\ldots,k$, where $\sum \in \{0,1\}$ is an alphabet. Initially, message *m* is transmitted by *l* different people in *N*. These people are called *initial propagators*. A person P_i in network *N* forwards message *m* to $0 \le q \le k_i$ of his contacts, where k_i is the degree of P_i and q is chosen randomly. The first message that a person P_i receives is denoted as m_1^P and this is the message that P_i remembers. We denote this message as the "*message in memory*" $\hat{m}$.

In order to create a mutated "*message in memory*" $\hat{m}$, we need to consider all the *r* messages received by person *P*, denoted by $\left(m_1^P,\ldots,m_r^P\right)$. For every mutated letter $\sigma_i \in \hat{m}, \quad i = 1,\ldots,k$, σ_i is chosen to be the letter that has the maximum number of occurrences among all letters (the mode) at location *i* in all the *r* messages $\left(m_1^P,\ldots,m_r^P\right)$. In the case where we have an equal number of different letters the original letter in message m_1^P is chosen.

Dependent Variable - Absolute Error

The absolute error EA^i, for a person P^i, represents the number of mutations from the original message *m* that was first propagated in the network. It is calculated as follows:

Let $\mathbf{u} = m - \hat{m}^P$ be the difference vector for person P^i. We calculate $EA^i =$ Number of "1"s in $\mathbf{u}$.

Example: Assuming that the original message $m = [1101111]^T$ and by taking the final message, the difference vector: $\mathbf{u} = m - \hat{m}^P = [1101111]^T - [0101001]^T = [1000110]^T$.

Therefore, $EA^i = 3$ which is the number of "1"s in $\mathbf{u}$. After the propagation of *m* in network *N* that contains *n* people, we can then calculate the average global absolute distortion value N_A^D for *N* as $N_A^D = \sum_{j=1}^{n} EA^j / n$. The absolute error represents the network global average error.

Data Description and Model Simulations

We used network analysis package developed by Barabási's team at Indiana University, and self-written software to produce the different types of networks. The scale-free networks we tested in our simulations have an exponent γ that characterizes human social networks and satisfies $2 \le \gamma \le 3$. The tested networks have 30,000, 50,000, 100,000, 500,000 nodes.

We compared the statistical results using mathematical and statistical tools. The algorithm we used to traverse the undirected graph that represents the network under simulation is the Breadth First Search (BFS) algorithm for graph search and traversal.

Research Hypotheses

We tested the following hypotheses for diminishing message distortion in scale-free networks:

H1: The absolute error will be lower after correction using hubs than after correction using regular nodes for the two researched numbers of correcting nodes (one or two).

H2: There is a positive relationship between the number of the correcting hubs and a decrease of the absolute error.

The above hypotheses were based on the assumption that scale-free networks are less sensitive to data distortion after several hubs, regular nodes, or a mix of them have corrected the message (to its first original values). Moreover, the correction will be more powerful if the correctors are hubs instead of regular nodes.

Results

To test for differences in our dependent (the absolute error) continuous variable between two groups of correctors a paired t-test was performed, since we dealt with the same network and the same nodes. To test for differences in our dependent variable between more than three groups, a General Linear Modeling (GLM) for Repeated Measurements was performed.

As shown in Table 1, there were significant differences between the means of the absolute error for all three types of hub correctors (corrections of Hubs: F=7723.42, P < 0.001; corrections of

Hubs: F=2837.93, P < 0.001). Similar significant findings were obtained in the larger networks of 50,000, 100,000 and 500,000 nodes.

Table 1. GLM for Repeated Measurements for Corrections of (0,1,2) - 30,000 Nodes

Source		**Type III Sum of Squares**	**Mean Square**	**F**	**Sig.**
Tests of Within-Subjects Effects - Measure: Hubs Difference					
factor1	Sphericity Assumed	31088.52	15544.26	7723.42	<0.001
	Greenhouse-Geisser	31088.52	15551.9	7723.42	<0.001
	Huynh-Feldt	31088.52	15550.86	7723.42	<0.001
	Lower-bound	31088.52	31088.52	7723.42	<0.001
Tests of Within-Subjects Effects - Measure: Regular Difference					
factor1	Sphericity Assumed	11311.49	5655.74	2837.93	<0.001
	Greenhouse-Geisser	11311.49	5656.96	2837.93	<0.001
	Huynh-Feldt	11311.49	5656.59	2837.93	<0.001
	Lower-bound	11311.49	11311.49	2837.93	<0.001

Testing Our Hypotheses - Comparison between Hubs vs. Regular Nodes

In the following results we tested for the differences in each pair of correctors to analyze the connections between all types of correctors and to test our hypotheses.

Table 2. Absolute Error Comparison of Regular Nodes (1, 2) and Hubs (1, 2) - 30,000 Nodes

Measured Pair (mean of absolute error)		**Mean difference**	**SD**	**t**	**Sig.**
Pair 1	One Hub Corrector (3.441) vs. One Regular Corrector (4.256)	-.815	2.05	-68.94	<0.001
Pair 2	Two Hub Correctors (3.243) vs. Two Regular Correctors (3.718)	-.475	2.03	-40.52	<0.001

Table 2 shows that the differences between two pairs of means were significant (p < 0.001 in all comparisons). Therefore, the absolute error will be lower after correction using hubs than after correction using regular nodes for one or two correcting nodes (H1 accepted). We expanded the analysis and performed other comparisons between one hub corrector vs. two regular correctors as well, and the results remained significantly in favor of the hub corrections (p < 0.001).

The above results (and others that are not presented here), clearly show that hubs are better correctors than regular nodes or combinations of regular-hub nodes.

Therefore, we will focus on presenting results with respect to correction only using hubs. We describe below the comparisons in the larger networks of 50,000 and 500,000 nodes.

As shown in Table 4, the differences between all three pairs of means were significant subjected to a paired-sampled t-test ($p<0.001$ in all comparisons). Therefore, there is a positive relationship between the number of the correcting hubs and a decrease in the absolute error (H2 accepted).

Table 3. Relationship Between the Correcting Hubs and Absolute Error in 50,000 Nodes

Measured Pair	(mean of absolute error)	Mean difference	SD	t	Sig.
Pair 1	No correctors (4.87) vs. One Hub Corrector (4.83)	.043	1.4	6.92	<0.001
Pair 3	No correctors (4.87) vs. Two Hub Correctors (4.78)	.091	1.4	14.5	<0.001
Pair 3	No correctors (4.87) vs. Three Hub Correctors (4.75)	.121	1.39	19.5	<0.001
Pair 2	One Hub Corrector (4.83) vs. Two Hub Correctors (4.78)	.048	1.43	7.43	<0.001
Pair 5	One Hub Corrector (4.83) vs. Three Hub Correctors (4.75)	.078	1.38	12.6	<0.001
Pair 6	Two Hub Correctors (4.78) vs. Three Hub Correctors (4.75)	.030	1.4	4.84	<0.001

As shown in Table 3, the differences between all three pairs of means were significant subjected to a paired-sampled t-test ($p<0.001$ in all comparisons). Therefore, there is a positive relationship between the number of the correcting hubs and a decrease in the absolute error (H2 accepted).

Table 4. Relationship Between the Correcting Hubs and Absolute Error in 500,000 Nodes

Measured Pair	(mean of absolute error)	Mean difference	SD	t	Sig.
Pair 1	No correctors (4.78) vs. One Hub Corrector (4.75)	.037	1.51	17.15	<0.001
Pair 3	No correctors (4.78) vs. Two Hub Correctors (4.71)	.076	2.03	26.36	<0.001
Pair 3	No correctors (4.78) vs. Three Hub Correctors (4.67)	.110	2.08	37.24	<0.001
Pair 2	One Hub Corrector (4.75) vs. Two Hub Correctors (4.71)	.039	1.99	13.93	<0.001
Pair 5	One Hub Corrector (4.75) vs. Three Hub Correctors (4.67)	.073	2.05	25.22	<0.001
Pair 6	Two Hub Correctors (4.71) vs. Three Hub Correctors (4.67)	.034	2.07	11.51	<0.001

Summary of the Results:

Our results show significant differences between the means of the four types of node corrector sets in terms of absolute error. These differences are summarized in the graph in Figure 3. We added the results of the network with 100,000 nodes. All research hypotheses were confirmed.

Figure 1 shows that when the size of the network increases, all hub corrections are better.

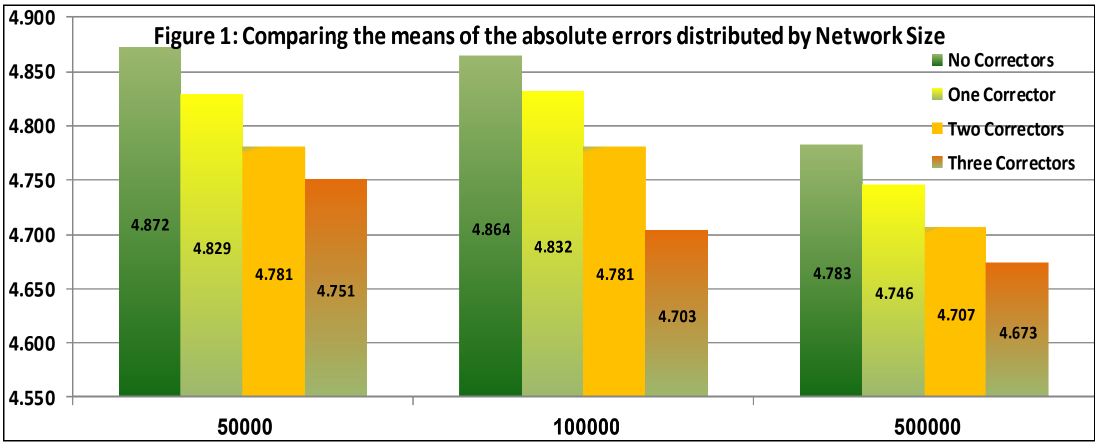


Figure 1: Comparing the means of the absolute errors distributed by Network Size

Discussion and Conclusion

In this study we calculated the distortion of a message as it propagates through a social network using a simulated mathematical model. At the end of the diffusion process of the message in a network, we compared the results of the average global absolute distortion value for all types of corrective nodes, to measure the extent to which the correction process decreased the global absolute error, thus increased the robustness of the network to distortion of information.

The results show that hubs that play the role of message correctors are capable of diminishing the average global absolute distortion error. The hubs and the high degree nodes receive a message during propagation, and after the correction of the distorted message the original one, they affect many contacts in the network.

References

Albert, R., and Barabási, A. L. 2002. "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics* 74, 47-97.

Barabási, A. L., and Albert, R. 1999. "Emergence of Scaling in Random Networks," *Science* 286, 509-512.

Barabási, A. L., and Crandall, R. E. 2003. "Linked: The New Science of Networks," *American Journal of Physics* 71, 409.

Crucitti, P., Latora, V., Marchiori, M., and Rapisarda, A. 2003. "Efficiency of Scale-free Networks: Error and Attack Tolerance," *Physica A* 320, 622-642.

Ma, X. J., Wang, W. X., Lai, Y. C., and Zheng, Z. 2010. "Information Explosion on Complex Networks and Control," *The European Physical Journal B* 76, 179-183.

Panzarasa, P., Opsahl, T., Carley M. K. 2009. "Patterns and Dynamics of Users' Behavior and Interaction: Network Analysis of an Online Community," *Journal of the American Society for Information Science and Technology* 60, 911–932.

Rapoport, A. 1953a. "Spread of Information through a Population with Socio-structural Bias: I. Assumption of Transitivity," *Bulletin of Mathematical Biology* 15, 523-533.

Rapoport, A. 1953b. "Spread of Information through a Population with Socio-structural Bias: Ii. Various Models with Partial Transitivity," *Bulletin of Mathematical Biology* 15, 535-546.

Rogers M. Everett, Diffusion of Innovations, Free Press 1995

Singer, Y. 2006. "Dynamic measure of network robustness," in *Electrical and Electronics Engineers in Israel, 2006 IEEE 24th Convention*, 366-370.

Valente, Computational & Mathematical Organization Theory, Vol. 2, No. 2. 1996

Exploring Crowds' Mean Belief in Fixed Odds Betting for Event Prediction

Weiyun Chen[1], Xin Li[2], Daniel Zeng[1, 3]
[1]Institute of Automation, Chinese Academy of Sciences; [2]Department of Information System, City University of Hong Kong; [3]Department of Management Information Systems, University of Arizona
weiyun.chen@ia.ac.cn, xin.li@cityu.edu.hk, zeng@email.arizona.edu

Abstract *The mean belief of the crowds is a robust and accurate assessment of the state problems in decision-making. Existing auction-based prediction markets general use contract price to capture it as a basis of their predictions. In this study, we propose a model to estimate the mean belief of the crowds in fixed odds betting, which may enable the use of fixed odds betting as a prediction market mechanism. We get an approximate formula to represent the relationship among the crowds' mean belief, fixed odds and the final betting ratio by conducting a simulation of the fixed odds betting. We test the performance of the proposed model on a game betting dataset. The experiments show that our proposed model outperforms linear regression models on various combinations of features when using to predict future events.*

Keywords: prediction markets, mean belief of crowds, fixed odds betting

1. Introduction

The wisdom of crowds is a well-recognized means in decision-making. Particularly, people found that mean belief of the crowds is a robust and accurate prediction of the state and probability for events [1], which led to the recent advances of prediction markets (PMs). To accurately measure crowds' mean belief, however, is not trivial. Existing literature often employ an auction-based market mechanism, in which the price of contracts reflects crowds' mean belief, to estimate the probability for an event to happen. Although being widely used, auction-based market is not immune from its limitations. In general, auction-based markets generate a high cognitive load to participants [2]. It is also vulnerable to market manipulations [3]. Thus, it is necessary to investigate market mechanism that may address these problems.

Fixed odds betting is a very popular game mechanism in sports betting [4]. In fixed odds betting, an odds is set by bookmakers on whether an event will happen. Participants make their individual judgments and choose and how much to bet on the event. If a participants' estimation is correct, s/he will be reward proportionally to the fixed odds. Fixed odds betting is a potential mechanism for prediction markets, which does not have the aforementioned two limitations. However, it does not have an explicit measure on crowds' mean belief. A close analogue, the bet ratio of participants or money on each side, is influenced by both participants' beliefs and the rewarded odds and thus cannot be directly used for this purpose.

In this study, we develop a model to explore the mean belief of the crowds from participants' responses in fixed odds betting. Upon a sports betting dataset, we evaluate whether the proposed model can provide a reasonable prediction of binary event probability as compared with some linear regression models. The experiment shows the potential of using our proposed model to develop a light-weight derivative prediction market upon fixed odds betting.

2. Related Work

PMs are event future markets whose contract payoff depends on related event results. Contract prices in auction-based PMs are found to be a good predictor of event probability [5]. Previous theoretical studies argue that it is because contract price in auction-based PMs reflects the mean belief of crowds. Modeling the PMs' trader behavior as linear utility function, Manski argued suggests that the PMs' equilibrium prices are located in an interval formed by mean belief of traders and are not strictly equal to the mean belief of traders [6]. Wolfers and Zitzewitz further prove that PMs' equilibrium price equal to the mean belief of traders in a group of traders with independent belief and behave to maximize their log utility function [7]. Although the two models have slight differences in detail, they generally indicate mean belief of traders is a good estimation of event probability.

Previous literature on fixed odds sport betting mainly focus on two perspectives, the market efficiency problem and how bettors act under uncertainty [4]. However, they have not provided us a direct measure on mean belief of crowds. As a result, it is seldom used as a mechanism for prediction markets. In this study, we want to fill this research gap and explore mean belief of crowds to estimate the event probability based on observed betting and odds.

3. Methodology

3.1 Problem Setting

For the sake of usability, we here consider only binary game that has two possible results, A or B. The probability for events A and B to happen are p_A and p_B, respectively, where $p_A+p_B=1$. In the beginning, the bookmaker sets the odds (fractional odds) for A to happen as o_A and for B to happen as o_B, where o_A, $o_B>0$ [8]. The bettors can bet on either side with unit cost. If the bet is correct, they will be rewarded with corresponding odds together with their original bets, i.e., $1+o_A$ (or $1+o_B$). It is necessary for $1<d=1/(1+o_A)+1/(1+o_B)<2$, so that bettors can't gain through betting on both sides. Given d, it is easily concluded that $0<o_A$, $o_B<(2-d)/(d-1)$. We denote the ratio $(o_B+1)/((o_A+1)+(o_B+1))$ as p_{odds} which reflects bookmaker' belief on event A's probability[8]. Obviously, $(d-1)/d<p_{\text{odds}}<1/d$. o_A and o_B can also be represented as:

$$o_A = \frac{1}{d\,p_{\text{odds}}} - 1, \quad o_B = \frac{1}{d(1-p_{\text{odds}})} - 1 \tag{1}$$

In this research, we assume there is a large population of bettors and each bettor can switch their bets for dynamic new information incorporation in each game. The shares of votes on A and B are denoted as s_A and s_B, respectively, with $s_A+s_B=1$. The individual's belief on event A's and B's probability are denoted as x_A and x_B respectively.

3.2 Model Design

We assume each bettor make betting choice for their best perceived utility. That is, if $U(x_A, o_A)>U(x_B, o_B)$ then choose A, if $U(x_A, o_A)<U(x_B, o_B)$ then choose B. Their perceived utility can be described by Prospect Theory [10]. For event i, $i\in\{A,B\}$, the generalized expected utility for choosing i, $U(x_i, o_i)$, is affected by odds o_i and each person's belief on event probability x_i:

$$U(x_i, o_i)=w^+(x_i)v(o_i)+w^-(1-x_i)v(-1) \tag{2}$$

where $w^+()$ and $w^-()$ are probability weighting functions on belief for event i to happen (x_i) or not ($1-x_i$). $v()$ is the value function for certain gain or loss. In our case, the gain is o_i and the loss is -1 for unit betting. This value function takes form of $v(y)=y^\alpha$ if $y\geqslant 0$ and $v(y)=-\lambda(-y)^\beta$ if $y<0$, where $0<\alpha, \beta<1$. Since the possible loss in this model is always 1, parameter β can be omitted.

In practice, the sum of the two weighed belief is close to 1. Thus, we assume, $w^+(x_i)=w(x_i)$ and $w^-(1-x_i)=1-w(x_i)$. Given the typical form of $w()$ and its corresponding parameters, we find $E(w(x))$ is very close to $E(x)$ for a range of belief distribution by computing these means directly. Thus we take $E(w)$ as a proxy to the mean belief of crowds $E(x)$. We furthermore assume $w(x)$ are in the form of Beta distribution in a bell shape because public opinions often follows a bell shaped distribution [9], specified by its two shape parameters η and υ. The crowds belief expectation $E(w)$ and its variance $D(w)$ are:

$$E(w) = \eta/(\eta+\upsilon), D(w) = \eta\upsilon/(\eta+\upsilon)^2(\eta+\upsilon+1) \tag{3}$$

Given $E(w)$ and $D(w)$, we can also derive η and υ as:

$$\eta = \left(\frac{E(w)(1-E(w))}{D(w)} - 1\right) E(w), \upsilon = \left(\frac{E(w)(1-E(w))}{D(w)} - 1\right)(1 - E(w)) \tag{4}$$

Theoretically, given fixed odds o_A, o_B and collective biased belief distribution $f(w)$~Beta(), we can compute the final bet ratio s_A and s_B on side A and B by comparing $U(x_i, o_i)=wv(o_i)+(1-w)v(-1)$, $i \in \{A,B\}$. Thus we can represent the $E(w)$ as a function of fixed odds, s_A, s_B and $D(w)$. However, we cannot find a close solution for this function. Thus, we turn to a computational experiment approach to estimate this function [11].

3.3 Model Specification

We conduct our computational experiments in the following steps. First, we generate a large amount of heterogeneous bettors by specifying their characteristics on parameters in $v(y)$. Particularly, we fix λ=2.25 and sample α from a reasonable Beta distribution[1] to have different types of bettors. Second we assign these bettors with belief sampled from given Beta distribution. For the same game, we generate various odds o_i. Thus we can compute these bettors' choices based on their behavioral model in section 3.2 and get the share of bettors choosing each side. Based on these simulated data, we estimate the relationship between s_A and odds o_A, o_B, by considering the parameters $E(w)$ and $D(w)$ that specifying distribution of w_i. Because odds o_A, o_B can be represented as function of d and p_{odds} as in (1), we can get odds o_A and o_B by p_{odds} when d is given. In real fixed odds betting d is often set equal to 1.11 [8]. This is also remained in this research. Therefore, particularly we hope to find a compact approximate formula to describe the relationship among s_A, p_{odds}, $E(w)$ and $D(w)$. We include those parameters in a step by step manner. These steps are described in Table 1.

Table 1: Simulation process to find out the influences of model parameters on the experiment result.

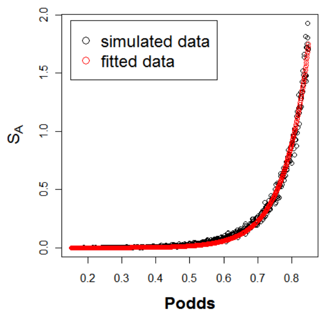
1. Given λ and α~Beta($\cdot$)
2. For w~Beta($\cdot$) in a list of shape parameters generated by (4) given list of $E(w)$ and $D(w)$
3. Sample from these two distributions and produce 5000 bettors
4. For p_{odds} in range of [0.1, 0.9], we get a list of (o_A,o_B) by fixing d=1.11
5. Each produced agent make choice according their decision parameters and beliefs
6. Statistic the final bet share on each side, thus get s_A
7. Fit the curve between p_{odds} and s_A using curve fitting
8.Evaluate how $E(w)$ and $D(w)$ influence the parameter of the formula between p_{odds} and s_A

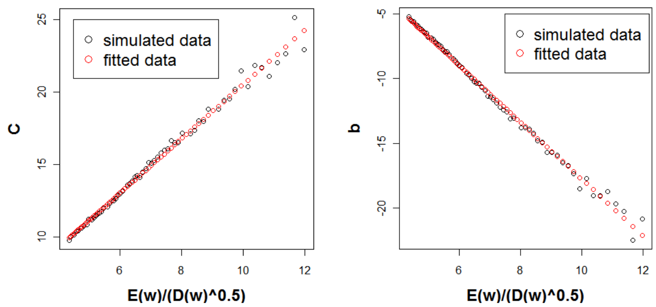
In step 2, we fixed the crowds' belief variance $D(w)$ and compute the two shape parameters of collective belief distribution by pre-assigned $E(w)$. Combined with the generated λ and α in step 1, a large amount of heterogeneous bettors with different belief on this simulated event can produced. From step 4 to 6, these simulated bettors make choices on this event result by giving them different odds for the same event. In step 4, by changing p_{odds} in range [0.1, 0.9], the odds is set by (1). A large amount pairs of s_A and p_{odds} can be produced. By observing the two dimensional curve of s_A to p_{odds} and trying to fit different formula on this curve, we find $log(log(s_A)/log(p_{odds}))=b+c*p_{odds}$. This relationship is shown in figure 1. When we change the collective belief distribution by assigning different $E(w)$, we find this relationship remain the same except that b and c change accordingly. We also find that if we change D(w) in step 2

[1] Such as Beta(32,5) which have mean close to findings in empirical study of Prospect Theory.

abovementioned findings from step 2 to step 7 hold. This means b and c are parameters moderated by $E(w)$ and $D(w)$. By changing the $E(w)$ and $D(w)$ in step 2, we find $b=r+s*E(w)/\sqrt{D(w)}$ and $c=l+m*E(w)/\sqrt{D(w)}$ in step 8. Therefore we finally get:

$$E(w)=\frac{\log\left[\log s_{\mathrm{A}}/\log p_{\mathrm{odds}}\right]-r-l\cdot p_{\mathrm{odds}}}{s+m\cdot p_{\mathrm{odds}}}\sqrt{D(w)} \tag{5}$$





(a) **(b)**

Figure 1: Black points show how Bet ratio on A s_{A} changes with the odds implied probability p_{odds}. Red points are curve fitting result.

Figure 2: Black points in (a) and (b) show how c and b from step 2 to 7 change with $E(w)/\sqrt{D(w)}$. Red points are curve fitting result.

3.4 Model Estimation

Suppose we have N i.i.d. fixed odds sport betting games, we will have observable data set $\{o_{j\mathrm{A}}, o_{j\mathrm{B}}, R_j, m_j, k_j, s_{j\mathrm{A}}, s_{j\mathrm{B}}\}$, $j\in\{1,\ldots, N\}$, where $o_{j\mathrm{A}}$, $o_{j\mathrm{B}}$ are the odds for event A and B in game j respectively; R_j represents jth game's result which take value 1 when A happens and take value 0 when B happens; m_j denotes the number of participants in game j and k_j denotes the number of betting on side A; $s_{j\mathrm{A}}$ and $s_{j\mathrm{B}}$ denote the final bet ratio on each side. $s_{j\mathrm{A}}= k_j/m_j$ and $s_{j\mathrm{A}}+s_{j\mathrm{B}}=1$.

If we employ the mean belief $E^j(w)$ as the jth event probability and assume the belief variances $D^j(w)$ in different games remain in the same, we can estimate the parameters r, s, l, m in (5) by Maximization Likelihood Estimation (MLE) as following in (6).

$$Max\ \log\prod_{j=1}^{N}(E^j(w))^{R_j}(1-E^j(w))^{1-R_j}=Max\sum_{j=1}^{N}R_j(E^j(w))+(1-R_j)(1-E^j(w)) \tag{6}$$

It should be noted that we find the value of r, s, l, m are in different intervals when we change the belief variance $D(w)$ in the analysis of simulated data. Here we do not give these numeric intervals due to page limit. When we estimate the parameters in (5), we should limit the search space in the these intervals determined by $D(w)$. For every given $D(w)$ MLE method could compute a likelihood value for the given dataset. The most possible average belief variance $D(w)$ for the given dataset should be the one which could give the dataset maximal likelihood value.

4. Empirical Study

4.1 Data

In August 2008, www.sina.com.cn held a fixed odd betting game for the 2008 Olympic game on its web page. The game attracts 174,609 participants. We collect the votes on 482 betting games among which 167 are binary betting. We also collect the true result R_j of each game.

4.2 Benchmark Models

To evaluate the performance of our proposed model, we first introduce some benchmark models (BM1-BM7) as in (7)-(13). These models are straightforward if we think the bettors as a intelligent crowd who can tell the prediction accuracy of the odds which is set by bookmakers. If the odds is set wrong the crowd can explore this bias for more profit. Therefore we could expect the final bet ratio s_A could respond to the odds implied probability p_{odds}. This is formulated in (9)-(13). BM1 use bookmakers' opinion as event probability directly. BM2 aims to explore the potential bias in odds by logistic regression. For convenience, here we denote our model in section 3 as MBEP (Mean Belief as Event Probability).

BM1: $$p_A = p_{odds} \tag{7}$$

BM2: $$logit(p_A) = \phi_1 + \phi_2 p_{odds} \tag{8}$$

BM3: $$logit(p_A) = \phi_1 + \phi_2 p_{odds} + \phi_3 s_A \tag{9}$$

BM4: $$logit(p_A) = \phi_1 + \phi_2 p_{odds} s_A \tag{10}$$

BM5: $$logit(p_A) = \phi_1 + \phi_2 p_{odds} + \phi_3 p_{odds} s_A \tag{11}$$

BM6: $$logit(p_A) = \phi_1 + \phi_2 s_A + \phi_3 p_{odds} s_A \tag{12}$$

BM7: $$logit(p_A) = \phi_1 + \phi_2 p_{odds} + \phi_3 s_A + \phi_4 p_{odds} s_A \tag{13}$$

4.3 Model Estimation and Evaluation

For the MBEP model, we compute the likelihood value for different $D(w)$ values. The result shows that when $D(w)$=0.06 the likelihood is greater than those ones where $D(w)$ are assigned to other values. This indicates that the average variance of collective belief on the events in this dataset is about 0.06. In the following model performance comparison, we take $D(w)$=0.06 as a default.

We then divide the dataset into two subsets by randomly sampling from it without putting back. The first one has 110 data points. We take it as training dataset on which the parameters r, s, l, m are estimated. Those parameters in BM2-BM7 are also been estimated on this randomly generated dataset. The second has 57 data points. We take it as testing dataset on which these models' prediction performances are compared. A model's prediction on one event is its probability. This prediction can be evaluated by following proper quadratic scoring rule which is often used to evaluate probability forecasting [12].

$$\begin{cases} \text{If A happens: WinScore} = 100 - 400 * p_B^2 \\ \text{If B happens: WinScore} = 100 - 400 * p_A^2 \end{cases} \tag{14}$$

Higher score means more accurate prediction. For each training dataset and testing dataset, we compute an average score for every model by (14). We repeat this training and testing process for 500 times. Thus we get 500 average score for every above mentioned model. The average scores for these models are given out in Table 2.

Table 2: Model BM1-BM7 & MBEP's prediction performances

Model	Average Score	p-value(paired t-test)
BM1	8.56	< 2.2e-16
BM2	6.14	< 2.2e-16
BM3	11.43	< 2.2e-16
BM4	10.50	< 2.2e-16
BM5	9.54	< 2.2e-16
BM6	9.37	< 2.2e-16
BM7	10.62	< 2.2e-16

MBEP	**13.30**	not involved

The performances of these models are slightly different with what we have expected. BM2<BM1 indicates that the prediction bias in odds cannot be explored simply by logistic regression. BM3 is the most accurate model in BM3-BM7. This indicates BM4, BM5 and BM6 are in the wrong form. This also indicates that BM7 is over fitting to the training dataset. MBEP's average score is the highest in all of these models. We conduct a paired t-test for BM1-BM7 against MBEP on the predictive performance data. The third column of table 2 is the p-value. Our proposed model is significantly better than benchmark models at 95% confidence level. This indicates there is wisdom of crowds in the bet ratio of fixed odds betting. This also indicates that mean belief could be used as event probability if we only have such simple information as betting ratio.

5. Conclusion and Discussion

In this research, we develop a model to estimate crowd's mean belief for the purpose of future event prediction in the context of fixed odds betting. The model is shown effective in experiments on a game betting dataset. This research indicates that fixed odds betting can be used as a mechanism for building prediction markets. In the future, we will continue explore more accurate, consistent, and efficient models on this problem. We plan to combine the dynamics of crowd's behaviors in building this framework.

6. Acknowledgement

This work was partially supported by City University of Hong Kong SRG 7002898 and 7002625. All opinions are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday, 2004.

[2] I. Blohm*, et al.*, "Idea Evaluation Mechanisms for Collective Intelligence in Open Innovation Communities: Do Traders outperform Raters?," in *32th International Conference on Information Systems*, Shanghai, China, pp. 1-24, 2011.

[3] J. Wolfers and E. Zitzewitz, "Prediction Markets," *Journal of Economic Perspectives,* vol. 18, pp. 107-126, 2004.

[4] D. S. Raymond, "The Economics of Wagering Markets," *Journal of Economic Literature,* vol. 36, pp. 2021-2064, 1998.

[5] K. J. Arrow*, et al.*, "ECONOMICS: The Promise of Prediction Markets," *Science,* vol. 320, pp. 877-878, May 16, 2008.

[6] C. F. Manski, "Interpreting the Predictions of Prediction Markets," *Economic Letters,* vol. 91, pp. 425-429, 2006.

[7] J. Wolfers and E. Zitzewitz, "Interpreting Prediction Market Prices as Probabilities," *National Bureau of Economic Research Working Paper Series,* vol. No. 12200, 2006.

[8] T. Kuypers, "Information and Efficiency: An Empirical Study of a Fixed Odds Betting Market," *Applied Economics,* vol. 32, pp. 1353-1363, 2000.

[9] M. Roskin*, et al.*, *Political Science*: An Introduction (10th Edition), Prentice Hall, 2008.

[10] A. Tversky and D. Kahneman, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty,* vol. 5, pp. 297-323, 1992.

[11] F.-Y. Wang, "Toward a Paradigm Shift in Social Computing: The ACP Approach," *Intelligent Systems, IEEE,* vol. 22, pp. 65-67, 2007.

[12] Y. Chen*, et al.*, "Information Markets vs. Opinion Pools: An Empirical Comparison," in *6th ACM conference on Electronic commerce*, Vancouver BC, Canada, 2005.

Learning Sparse Heterogeneous User Preferences

Markus Peters and Wolfgang Ketter
Erasmus University, Rotterdam, Netherlands
peters@rsm.nl, wketter@rsm.nl

Abstract. Models of user preferences will be at the core of the next generation of personalized Information Systems. We propose HPREF, an algorithm for learning a hierarchical, probabilistic preference model that integrates sparse preferences from multiple like-minded users in a principled fashion. Our preliminary experiments indicate that HPREF outperforms previous preference learning approaches and suggest several directions for further improvement.

1 Introduction

Human choices are guided by preferences, and models that explain or predict human choices based on underlying, unobserved preferences have a rich tradition in Marketing and Psychology [4]. More recently, researchers in Information Systems and related disciplines have discovered *preference learning* as one key ingredient to building the next generation of personalized Information Systems [1; 3]. Successful examples of preference learning include personalized search engine rankings, smart home automation systems, and trading agents that require models of the humans they represent or compete against, e.g. [5].

While preference models in Marketing, Psychology and Information Systems share a number of similarities, there are also distinctive differences between the these domains. Information Systems (i) often operate on a *massive scale*, with millions of users and thousands of alternatives to choose from, (ii) are often used anonymously, by individuals with highly *heterogeneous preferences*, (iii) need to learn implicitly from *very few observed choices* to minimize the burden on the user, and (iv) need to learn *online*, integrating new information as it arrives. Current work in preference learning largely ignores settings with heterogeneous preferences, many users, but only few observed choices per user. The key challenge in this setup is to properly *integrate* and generalize from the choices of users with compatible preferences, while learning *separate* preference models for dissimilar users. Note, that preference learning is a harder problem than computing recommendations where items are merely classified as interesting or non-interesting to a user. The goal in preference learning is to predict *all* pairwise preferences between current and prospective items, allowing the learner to actually act on a user's behalf under a wide range of circumstances.

We present a novel algorithm for preference learning based on a nonparametric Bayesian estimation procedure for latent utility functions. Our algorithm explicitly accounts for user heterogeneity, but integrates choices from users with similar preferences to expedite learning. In the process, it automatically infers naturally occurring clusters of users with compatible preferences. Our method is well suited for online learning settings and it inherits many of the notable benefits of the underlying probabilistic model: its immunity to overfitting and high-dimensional input data, its principled approach to estimating predictive uncertainty, and its ability to use powerful kernel techniques for estimating highly nonlinear utility functions.

2 Preference Modeling

Let $X = \{x_1, \ldots, x_n\}$ be a set of *instances*, the objects or actions over which preferences are formed. Each instance is characterized by d features and in the following we will assume

$x_i \in \mathbb{R}^d$. We are interested in learning a partial order over instances s.t. $x_{i_1} \succeq_u x_{i_2} \succeq_u \ldots \succeq_u x_{i_n}$ where $x_i \succeq_u x_j$ means that user $u \in U$ likes instance x_i at least as much as x_j. Instead of operating directly on $\succeq_u$, we introduce a latent *utility function* $f_u : X \to \mathbb{R}$ that assigns a real number to each instance. The learning task then becomes finding f_u s.t. $f_u(x_{i_1}) \geq f_u(x_{i_2}) \geq \ldots \geq f_u(x_{i_n}) \Leftrightarrow x_{i_1} \succeq_u x_{i_2} \succeq_u \ldots \succeq_u x_{i_n}$, at least approximately.

Several challenges result from this formulation. Firstly, users typically do not know how to rank instances directly, especially if the number of instances is large. A learning algorithm will instead need to infer such a ranking from few observed choices of the form $x_i \succeq_u x_j$ without necessarily having further information on how other instances rank relative to this one pair. Secondly, there is mounting evidence from empirical psychologists that real-world preference relations are far from the total orders that one would hope for from a computational standpoint [4]. Real-world preference relations are noisy and inconsistent, and a good preference learning algorithm must provide a principled approach to estimating predictive uncertainty. Finally, and this will be the key contribution of our model, one will observe different, possibly incompatible preference relations for different users or user groups. That is, observations will be of the form $\mathcal{D} = \{(u, x_i, x_j) | u \in U, x_i, x_j \in X, x_i \succeq_u x_j\}$ and we aim to find multiple latent utility functions f_c and an assignment $U \to \{1, \ldots, C\}$ such that the preferences of user u are captured well by the utility function f_c. By modeling users with compatible preferences through a joint utility function, but using separate utility functions for dissimilar users, we aim to make maximal use of the few observed choices available for training. The inference of users with compatible preference is itself a valuable result of the algorithm, as this latent clustering is inferred from observed choices instead of approximating it through, e.g., user demographics.

Consider the example in Figure 1 where the true, latent utility functions of four users are indicated by the dotted lines. For simplicity, we use one-dimensional instances in this example, and we assume that for each user only pairwise preferences among the circled instances are observed. That is, we know that $x_5 \succeq_{u_1} x_2 \succeq_{u_1} x_3$, whereas for user 3 we only know $x_1 \succeq_{u_3} x_5$, and so forth. Based on these observations we aim to learn one or more utility functions and an assignment of users to these utility functions. In the example, we would hope to learn two functions (one increasing and one decreasing in the feature value) and an assignment of the first two users to the further and the last two users to the latter.

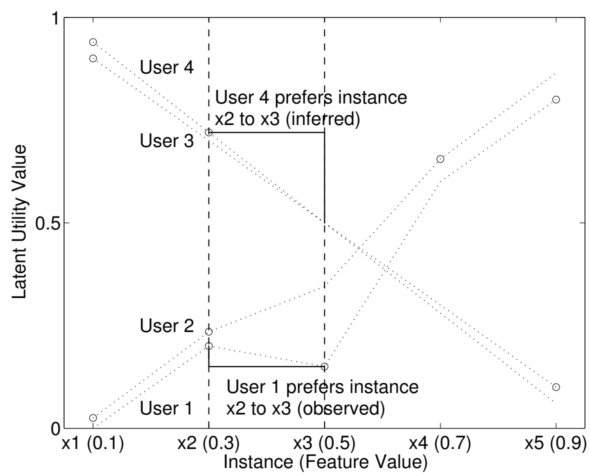


Fig. 1: True latent utility functions (dotted lines). In the example, choices between all pairwise combinations of a user's circled instances are observed.

3 Preference Learning

Probabilistic methods for inferring latent utility functions from observed choices have previously been studied in the Machine Learning community. We make use of GPPREF, a nonparametric Bayesian model [2]. The model starts with a Gaussian Process (GP) prior over the space of all utility functions and infers a posterior distribution over this space after having observed a set of binary choices. The key advantages of the GP approach in our setting are its natural abilities to estimate predictive uncertainty, to capture highly nonlinear utility functions, and to perform incremental updates.

We refer the reader to [2; 6] for a thorough explanation of GPs and of GPPREF, and instead introduce the key idea in the example in Figure 2. In the left panel we estimate a latent utility function based on the 6 observed pairwise choices from users 1 and 2. The maximum a-posteriori (MAP) estimate for the GP posterior is indicated by the dashed line, the shaded area indicates a one standard deviation region around the MAP estimate. GPPREF assigns the highest a-posteriori probability to an almost perfectly linear estimate, which would reproduce 5 of the 6 observed choices. While not precluding an estimate that mimics the outlier from user 1 on instance x_3, the model considers such an estimate at least unlikely given the current evidence. The model is certain in regions where relatively many consistent observations lie and gets increasingly uncertain in regions with sparse or conflicting observations.

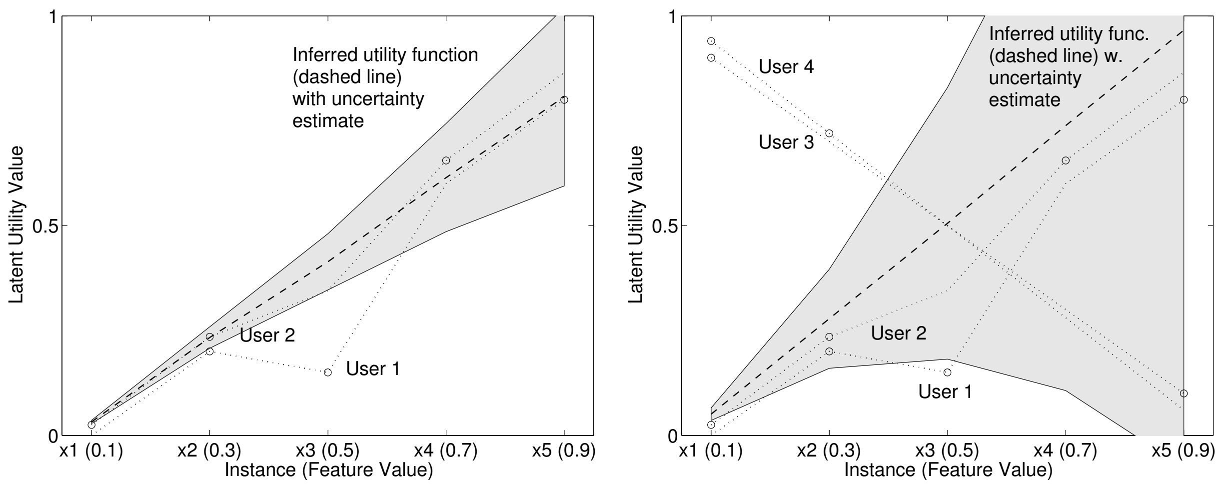


Fig. 2: Gaussian Process posteriors over latent utility functions (dashed line: MAP estimate, shaded area: one σ interval around MAP). Results for largely compatible preferences (left panel) and conflicting preferences (right panel).

The key problem with this approach becomes apparent in the right panel, where we add two conflicting choices from users 3 and 4 but still estimate a single utility function. In this situation, almost all functions on the depicted plane become likely. A naive solution to this problem would be to estimate one utility function per user. However, this is computationally expensive and, more importantly, it does not *generalize* over preferences of like-minded users. For example, we would like to infer that, based on what we know, users 3 and 4 are like-minded: they both have utility that decreases in the feature value. To alleviate this problem, we propose HPREF, a hierarchical preference learning algorithm for heterogeneous user populations. HPREF's core strength is its ability to integrate observed preferences from like-minded users into joint utility functions (thereby making maximum use the relatively few observed choices) while representing the preferences of dissimilar users as separate utility functions.

The idea behind HPREF is the following: Let $C = C_1 \cup \ldots \cup C_{|C|}$ be some partition of U into $|C|$ clusters, and let $\mathbf{f}_c$ and σ_c denote the MAP utility function estimate for cluster

c and its predictive uncertainty, based only on the observed preferences of users in cluster C_c.[1] We reassign each user to the cluster with the highest confidence in regions where the utility function predicts the observed choices correctly and the lowest confidence in regions where predictions are incorrect.[2]

Algorithm 1 The HPREF algorithm.

1: **function** HPREF($X, U, |C|, \mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_u$) ▷ instances, users, number of clusters, preferences per user
2:   for each $u \in U$: $C^{cand}(u) \leftarrow randi(\{1, \ldots, |C|\})$ ▷ assign each user randomly to one cluster
3:   $objective^{cand} \leftarrow \infty$, $iteration \leftarrow 1$
4:   **repeat**
5:     $objective \leftarrow objective^{cand}, C \leftarrow C^{cand}$
6:     for each $c = 1$ to $|C|$: $[\mathbf{f}_c, \sigma_c] = GPPREF(X, \mathcal{D}|_{C_c})$ ▷ GPPREF on observed pref. from C_c
7:     **for all** $u \in U$, $c = 1$ to $|C|$ **do**
8:       $R(u, c) \leftarrow \{(u, x_i, x_j) \in \mathcal{D}_u | \mathbf{f}_c(x_i) > \mathbf{f}_c(x_j)\}$ ▷ correct predictions for u using $\mathbf{f}_c$
9:       $W(u, c) \leftarrow \{(u, x_i, x_j) \in \mathcal{D}_u | \mathbf{f}_c(x_i) < \mathbf{f}_c(x_j)\}$ ▷ incorrect predictions for u using $\mathbf{f}_c$
10:     **end for**
11:     for each $u \in U$: $C^{cand}(u) \leftarrow \arg\max_c \sum_{R_{u,c}} (\mathbf{f}_c - \sigma_c)(x_i) - (\mathbf{f}_c + \sigma_c)(x_j) - \sum_{W_{u,c}} (\mathbf{f}_c - \sigma_c)(x_j) - (\mathbf{f}_c + \sigma_c)(x_i)$
12:     $objective^{cand} \leftarrow \sum_{u \in U} \sum_{c \in \{1,\ldots,|C|\}} \frac{|R(u,c)|}{|R(u,c)| + |W(u,c)|}$ ▷ predictive accuracy on training sample
13:     $iteration \leftarrow iteration + 1$
14:   **until** $objective^{cand} \geq objective$ or $iteration == MAX$
15: **return** $[\mathbf{f}_c, C_c]$
16: **end function**

Full pseudocode for HPREF is given in Algorithm 1. In line 2, the partition is initialized by randomly assigning each user to one of the $|C|$ clusters. Users are then continually reassigned to *better* clusters in the main loop starting in line 4 until no further improvement is possible. Given a current partition, HPREF first computes $\mathbf{f}_c$ and σ_c for the utility functions of each cluster (line 6). The reassignment in line 11 is then based on an estimate of the *predictive margins* under these estimates. The algorithm terminates when it cannot improve its performance on training data and returns the best partition C along with the utility functions for each cluster.

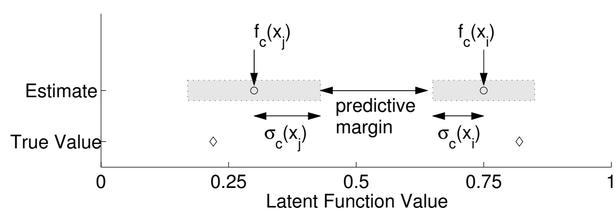


Fig. 3: Illustration of the predictive margin concept

The predictive margin idea is illustrated in Figure 3. Suppose that x_i and x_j are instances with utility values $\mathbf{f}_c(x_i)$ and $\mathbf{f}_c(x_j)$ under the estimate of cluster c, and that $x_i \succeq_u x_j$. If $\mathbf{f}_c$ models this relationship correctly (i.e., $\mathbf{f}_c(x_i) \geq \mathbf{f}_c(x_j)$, as is the case in the figure), then we want the estimate to be as stable as possible. Two factors influence the desired stability: $\mathbf{f}_c(x_i)$ should be much larger than $\mathbf{f}_c(x_j)$, *and* the estimated function values should both be certain, i.e., $\sigma_c(x_i)$ and $\sigma_c(x_j)$ should be low. In effect, Algorithm 1 (line 11) seeks to maximize the predictive margins on already correct predictions, while minimizing the predictive margins on faulty predictions, such that subsequent re-estimations can more easily rectify these faults.

[1] $\mathbf{f}_c$ corresponds to the dashed line and σ_c to the gray area in Figure 2.

[2] The number of clusters $|C|$ is chosen manually in the current version of the algorithm. However, we find that HPREF reacts benign to moderate overestimation of this parameter, leaving the extra clusters empty.

4 Experimental Results

We constructed a series of datasets that allow us to study HPREF's performance while controlling the data's distribution (see Figure 4, right panel). We fix the number $|U|$ of users and partition them evenly into $|C|$ clusters. We then draw a general orientation for the utility functions in each cluster from a d-dimensional uniform distribution on $[-1, 1]^d$. That is, we orient the utility functions in each cluster roughly along a d-dimensional hyperplane. The utility function for each user is finally obtained by adding two levels of noise to the cluster's general orientation: d-dimensional Gaussian zero-mean *inter-relational noise* with standard deviation σ^{inter} is used to tilt the orientation of each user's utility function relative to the cluster's general orientation. The idealized utility values computed from this hyperplane are further distorted using one-dimensional Gaussian zero-mean *intra-relational noise* with standard deviation σ^{intra}.

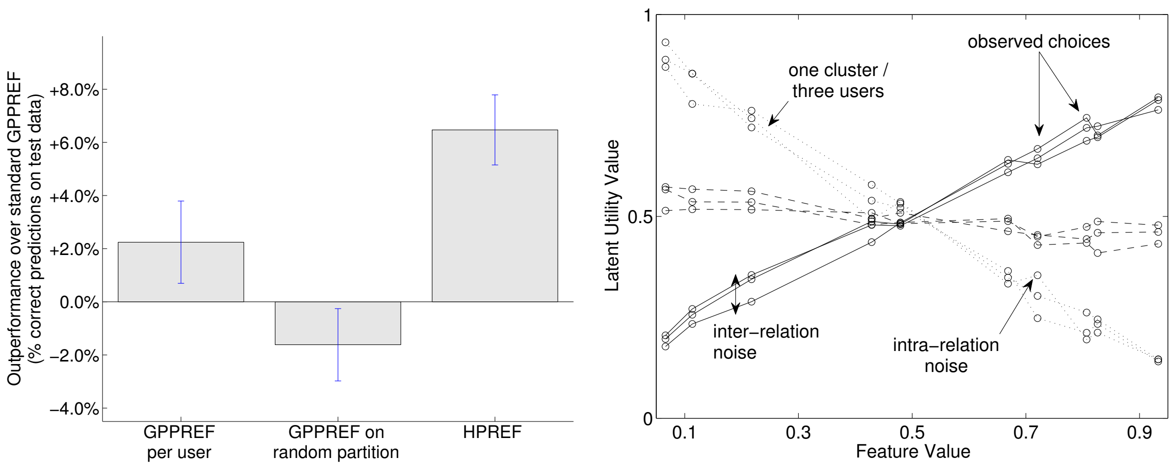


Fig. 4: Left panel: Performance of HPREF and two naive algorithms relative to GPPREF based on 50 synthetic datasets; whiskers indicate 95% confidence intervals. Right panel: Sample synthetic dataset with $|U| = 3$, $|C| = 3$, $d = 1$, $\sigma^{inter} = 0.05$, $\sigma^{intra} = 0.05$.

The left panel in Figure 4 shows the performance of HPREF and two naive algorithms on 50 synthetic datasets relative to GPPREF. The datasets in the experiment were all constructed to mimic a typical small-scale preference learning task with 50 10-dimensional instances, and 30 users clustered into 3 clusters with $\sigma^{inter} = 0.05$, $\sigma^{intra} = 0.10$. Overall, each dataset gave rise to 36,750 binary preferences from which we selected merely 0.5% (or about 6 observed choices per user on average) for training. This reflects our general goal of estimating preferences from very few observed choices per user. On each dataset, we ran HPREF as well as two naive variations on GPPREF[3]: (i) **GPPREF per user**, i.e., where one utility function was estimated for each user, and (ii) **GPPREF on random partition**, i.e., where multiple utility functions were estimated but without using the cluster reassignment procedure of HPREF. HPREF itself, as well as GPPREF on random partition were configured to use the true number of 3 clusters.[4]

[3] We used the reference implementation of GPPREF from `http://www.gatsby.ucl.ac.uk/~chuwei/plgp.htm`.

[4] Note that the actual number of clusters in the dataset can be less if, by chance, the general directions of clusters turn out to be similar.

Our results indicate that learning user preferences from such sparse, heterogeneous data is extremely challenging, but that HPREF is a promising development in this direction. HPREF, on average, predicted 71.1% of the held-out test data correctly (SE=2.1%), whereas the GPPREF benchmark achieved only 64.6% predictive accuracy (SE=2.0%). A naive extension (GPPREF per user) leads to about 2.3% higher predictive accuracy than GPPREF. However, this approach is unable to find naturally occurring clusters of users, and quickly becomes computationally prohibitive when user numbers grow larger. A simple partitioning scheme (GPPREF on random partition) where users are randomly assigned to clusters does not alleviate GPPREF's problems either. In fact, as the *relative* dissimilarities become more pronounced in the resulting smaller training sets, GPPREFs performance deteriorates to about 1.6% below benchmark performance.

5 Discussion

Probabilistic models of user preferences will likely be at the core of personalized, next-generation Information Systems. However, estimating such models from only few observed preference is exceedingly challenging. By combining preferences of like-minded users in a principled way, preference learning models can exploit similarities and achieve better predictive performance more quickly. The results above are based on an early version of HPREF and we are currently exploring various extensions: Firstly, while HPREF's user reassignment procedure *uses* probabilistic information, the procedure itself is non-probabilistic. We think that a fully probabilistic model will lead to an even more useful characterization of users and their preferences. Another important result from our work on a fully probabilistic model will be more theoretical insights into the convergence properties of HPREF, which we have not established at this point. Secondly, we are studying the interaction between HPREF and the underlying utility function estimation. We are interested in ways of improving the underlying estimation in terms of accuracy and runtime performance when used in conjunction with HPREF. And finally, we are exploring other uses of the uncertainty estimates generated by HPREF. These estimates could give rise to elegant and powerful *reject options* (i.e., knowing when not to trust HPREF's predictions) and associated *active learning* schemes.

[1] Bichler, M., Gupta, A., Ketter, W.: Designing smart markets. Information Systems Research 21(4), 688–699 (2010)

[2] Chu, W., Ghahramani, Z.: Preference learning with gaussian processes. In: Proceedings of the 22nd international conference on Machine learning. pp. 137–144. ACM (2005)

[3] Fürnkranz, J., Hüllermeier, E.: Preference learning: An introduction. In: Fürnkranz, J., Hüllermeier, E. (eds.) Preference Learning, pp. 1–17. Springer-Verlag (2010)

[4] Lichtenstein, S., Slovic, P.: The construction of preference. Cambridge University Press (2006)

[5] Peters, M., Ketter, W., Saar-Tsechansky, M., Collins, J.: Autonomous data-driven decision-making in smart electricity markets. In: Proceedings of the European Conference on Machine Learning (2012)

[6] Rasmussen, C.E., Williams, C.: Gaussian Processes for Machine Learning. MIT Press (2006)

Service Systems with Postponable Acceptance and Assignment: A Dynamic and Stochastic Programming Approach

Keumseok Kang[*], J. George Shanthikumar[#], Kemal Altinkemer[#]

[*]Florida International University, [#]Purdue University

ABSTRACT

We study the order acceptance and assignment Problem of a Service System with Postponable Acceptance and Assignment (hereafter we will call it Problem SSPAA). Unlike traditional acceptance and assignment problems, in our settings, the service system can strategically postpone acceptance and assignment decisions for orders in hand while waiting for more profitable orders to come, with the risk of losing orders, low resource utilization, and waiting penalties. We formulate this problem using a dynamic and stochastic programming approach, find the structural properties of the optimal policy, propose approximate policies, and conduct extensive computational experiments. In the experiments, our proposed approximate policies produce near-optimal solutions. Although this paper is motivated by the multi-project selection and assignment problem of IT service companies, same or related problems can be found in other contexts.

1. INTRODUCTION

This study is motivated by the multi-project selection and assignment problem of Information Technology (IT) service companies. Consider an IT service company. The company dynamically receives service orders (e.g., information system implementation projects) from outside customers, accepts or rejects incoming orders based on the availability of the company's resources and the profitability of orders, and provides services for accepted orders. Due to the limited resources and the variety of services requested by customers, the company cannot provide services to all incoming service orders but have to selectively accept a limited number of orders and provide services only to those selected orders in general. Given that accepting an IT service order is a critical business decision for an IT service company, the company typically spends some time evaluating the order to decide whether or not to accept, and this delay is usually accepted by its customer. In other words, the company does not have to decide whether or not to accept a customer's order upon its arrival, but may defer the decision to some extent (e.g., for several days or weeks). However, it is incumbent upon the company to make a decision in a timely manner because customers usually contact more than one IT service company, and a competing company may take the order before the company takes it. However, in many instances the company does not know when the order in hand becomes unavailable to accept (i.e., uncertainty of order abandonment). In addition to uncertain order abandonment, the company suffers from uncertainties about new incoming orders – when, what, and how many new orders will arrive – as well as uncertainties about the availability of their internal resources – when the currently assigned orders will be completed so that the resources can be available for new orders. The objective of this IT service company is to maximize its profits with these uncertainties, by accepting right orders and assigning them to appropriate resources, and sometimes postponing those accepting and assignment decisions. Although this paper is motivated by a particular problem in the IT service domain, similar problems can be found in other domains, especially where postponable acceptance and assignment decisions for incoming orders are allowed. There are many examples such as the patient acceptance problem of a hospital, the employee recruitment issues of a company, the order acceptance and assignment problem of a make-to-order manufacturing company.

One of the unique characteristics of this problem is that the service system can accept and assign orders strategically by postponing order acceptance and assignment decisions, while waiting for more profitable orders to come. In other words, unlike traditional acceptance and assignment problems, such as the secretary (e.g., Ferguson 1989), stochastic or online knapsack (e.g., Kleywegt and Papastavrou 1998), and general queuing problems (e.g., Akçay et al. 2010), in our study the acceptance (or rejection) decision for an order does not have to be made upon the arrival (i.e., request) of that order, and similarly the assignment decision for an accepted order does not have to be made upon the acceptance of that order. Instead, the decision maker may defer the acceptance decision as long as that order remains available to be accepted (i.e., until that order is withdrawn by the customer) and defer the assignment decision until a more appropriate server becomes available.

2. PROBLEM DEFINITION AND FORMULATION

In this section, we formally define Problem SSPAA and formulate it as a dynamic and stochastic programming model. Consider a service system where service orders arrive at discrete times $\{1, 2, \dots, n\}$. Let period k be defined as the time interval $[k, k+1)$. There are I service types ($i = 1, 2, \dots, I$), and each order belongs to one service type. There are J server types ($j = 1, 2, \dots, J$), and each type has the maximal capacity (i.e., the number of servers), $b^j > 0$, $j = 1, 2, \dots J$. All servers are cross-trained so that they can provide any type of service but with different costs, c_k^{ij}, and with different service completion rates, L_k^{ij}. Let Z_k^i be the total amount of i-type orders which arrive at the end of period $k-1$ (i.e., at time k), $k = 1, 2, \dots, n$. Let $X_{k:2}^i$be the total amount of i-type orders in a queue for acceptance, including the ones which just arrived at the end of period $k-1$, for which the acceptance/rejection decision has been postponed by the end of period $k-1$. Let $X_{k:1}^i$ be the amount of orders which are in a queue for assignment but have not been assigned to any server yet by the end of period $k-1$. At the beginning of period k, some part of orders waiting in the queue for acceptance can be accepted or rejected, or any part of them can be postponed. Let A_k^i and R_k^ibe the amount of orders accepted and rejected, respectively, at the beginning of period k. Similarly, some parts of orders in the queue waiting for assignment can be assigned to servers while assignments for some others are postponed. Let U_k^{ij} be the amount of i-type orders which are newly assigned to j-type servers at the beginning of period k. Acceptance and assignment decisions are made almost at the same time; therefore, some orders can be assigned to servers immediately after they are accepted, without staying in the queue for assignment. Thus we have $X_{k+1:1}^i = X_{k:1}^i + A_k^i - \sum_{j=1}^{J} U_k^{ij}, k = 1, 2, \dots, n, ; i = 1, 2, \dots, I$. The capacity of j-type servers which are currently serving i-type orders at the beginning of period k is G_k^{ij}. Therefore, the capacity of available (i.e., idle) j-type servers at the beginning of period k, S_k^j, is defined as $b^j - \sum_{i=1}^{I} G_k^{ij}$. The total amount of orders newly assigned to j-type servers should not be greater than the capacity of available servers (i.e., $\sum_{i=1}^{I} U_k^{ij} \leq S_k^j, j = 1, 2, \dots, J$). Sometimes, some servers may not be assigned and remain idle. A fraction, $1 - L_k^{ij}$, $(0 \leq 1 - L_k^{ij} \leq 1)$ of working (i.e., non-idle) servers complete their assigned orders at the end of period k and become available to be assigned new orders at the beginning of period $k+1$. Thus we have

2

$G_{k+1}^{ij} = \left(G_k^{ij} + U_k^{ij}\right)\left(1 - L_k^{ij}\right)$. Customers who place orders may be impatient; therefore, some parts of orders waiting in the queues for acceptance may be abandoned. At the end of period *k*, only a fraction, W_k^i ($0 \le W_k^i \le 1$), of orders which have not been accepted at the beginning of period *k*, remain waiting for acceptance, while the others are abandoned. Thus we have $X_{k+1:2}^i = \left(X_{k:2}^i - A_k^i - R_k^i\right)W_k^i + Z_{k+1}^i, k = 1, 2, \dots, n, ; i = 1, 2, \dots, I$. Accepting orders are associated with increasing total revenues. Let r_k^i be the revenue earned when an *i*-type order is accepted at time *k*. Waiting and providing services incur associated costs. Let c_k^{ij} be the service cost incurred when a *j*-type server serves an *i*-type order during period *k*. Let $b_{k:1}^i$ and $b_{k:2}^i$ be the waiting penalties for letting an *i*-type order wait during period *k*, in the queue for assignment and in the queue for acceptance, respectively. Without loss of generality, we assume that all costs accrue at the beginning of each period.

Let $\underline{X}_{k:1} = (X_{k:1}^1, X_{k:1}^2, \dots, X_{k:1}^I)$, $\underline{X}_{k:2} = (X_{k:2}^1, X_{k:2}^2, \dots, X_{k:2}^I)$ and $\underline{G}_k = (G_k^{11}, G_k^{12}, \dots, G_k^{IJ})$. Then the state variable at time *k* is denoted by $(\underline{X}_{k:1}, \underline{X}_{k:2}, \underline{G}_k)$. Let $\underline{x}_k = (\underline{x}_{k:1}, \underline{x}_{k:2}, \underline{g}_k)$ represent a state before making decisions – accept, reject, assign, and postpone orders – at time *k* and $\underline{y}_k = (\underline{y}_{k:1}, \underline{y}_{k:2}, \underline{q}_k)$ represent an instantaneous state after making those decisions at the beginning of period *k*. Let us define a set function,
$\Pi\left(\underline{y}_k;\ \underline{Z}_k, \underline{W}_k, \underline{L}_k\right) = \left(\underline{X}_{n+1:1}, \underline{X}_{n+1:2}, \underline{G}_{n+1}\right) = (y_{k:1}^1, y_{k:1}^2, \dots, y_{k:1}^I, y_{k:2}^1 W_k^1 + Z_{k+1}^1, y_{k:2}^2 W_k^2 + Z_{k+1}^2, \dots, y_{k:2}^I W_k^I + Z_{k+1}^I, q_k^{11} L_k^{11}, q_k^{12} L_k^{12}, \dots, q_k^{IJ} L_k^{IJ})$, which represent the state at the end of period *k* (i.e., time *k* + 1), after random events – arrivals of new orders, abandonments of existing orders, service completions of existing orders – occur during period *k*. Then the expected profits function for the remaining periods (including period *k*), by changing the given state $\underline{x}_k$ to $\underline{y}_k$ at the beginning of period *k*, is defined as $\psi_k\left(\underline{x}_k, \underline{y}_k\right) = \varphi_k\left(\underline{x}_k\right) + \phi_k\left(\underline{y}_k\right)$, where $\varphi_k\left(\underline{x}_k\right) = -\sum_{i=1}^I r_k^i x_{k:1}^i - \sum_{i=1}^I \sum_{j=1}^J r_k^i g_k^{ij}$ and $\phi_k(\underline{y}_k) = \sum_{i=1}^I \left(\left(r_k^i - b_{k:1}^i\right) y_{k:1}^i - b_{k:2}^i \mathbb{E}\left[y_{k:2}^i W_k^i\right]\right) + \sum_{i=1}^I \sum_{j=1}^J (r_k^i - c_k^{ij}) q_k^{ij} + \mathbb{E}\left[v_{k+1}\left(\Pi\left(\underline{y}_k;\ \underline{Z}_k, \underline{W}_k, \underline{L}_k\right)\right)\right]$. The maximal expected profits for the remaining period, given state $\underline{x}_k$ at time *k*, $v_k\left(\underline{x}_k\right)$, is defined as: $v_k\left(\underline{x}_k\right) = \max \psi_k\left(\underline{x}_k, \underline{y}_k\right) = \max\left(\varphi_k\left(\underline{x}_k\right) + \phi_k\left(\underline{y}_k\right)\right) = \varphi_k\left(\underline{x}_k\right) + \max \phi_k\left(\underline{y}_k\right)$. Therefore, $v_k\left(\underline{x}_k\right)$ is obtained by solving the following maximization problem [O-1]:

$$\tau_k\left(\underline{x}_k\right) = \max \phi_k\left(\underline{y}_k\right) \text{ s.t.}$$

$$y_{k:1}^i + y_{k:2}^i + \sum_{j=1}^J q_k^{ij} \le x_{k:1}^i + x_{k:2}^i + \sum_{j=1}^J g_k^{ij}, i = 1, 2, \dots, I \tag{1}$$

$$y_{k:2}^i \le x_{k:2}^i, i = 1, 2, \dots, I \tag{2}$$

$$y_{k:1}^i + \sum_{j=1}^J q_k^{ij} \ge x_{k:1}^i + \sum_{j=1}^J g_k^{ij}, i = 1, 2, \dots, I \tag{3}$$

$$\sum_{i=1}^I q_k^{ij} \le b^j, j = 1, 2, \dots, J \tag{4}$$

$$q_k^{ij} \ge g_k^{ij} \tag{5}$$

$$y_{k:1}^i, y_{k:2}^i, q_k^{ij} \ge 0 \tag{6}$$

3

We can find the optimal policy, $\underline{y}_k(\underline{x}_k)^*$, by solving the maximization problem: $\underline{y}_k(\underline{x}_k)^* = \text{argmax}_{\underline{y}_k}\, \phi_k(\underline{y}_k)$ s.t. (1) – (6).

3. PROPERTIES OF THE OPTIMAL POLICY

Theorem 1. *$\phi_k(\underline{x}_k)$ is jointly concave on $\underline{x}_k$, for k = 1, 2, …, n.* [1]

Theorem 2. *$\phi_k(\underline{x}_k)$ is submodular on $\underline{x}_k$, for k = 1, 2, …, n.*

The properties of concavity and submodularity provide useful information about the structural properties of the optimal policy. However, in our model, some decisions such as assignment and acceptance are associated with movement along with the line of - 45 degree (i.e., (1, -1) direction). To find the structural properties of the optimal policy regarding these decisions, we investigate the directional properties of $\phi_k(\underline{y}_k)$.

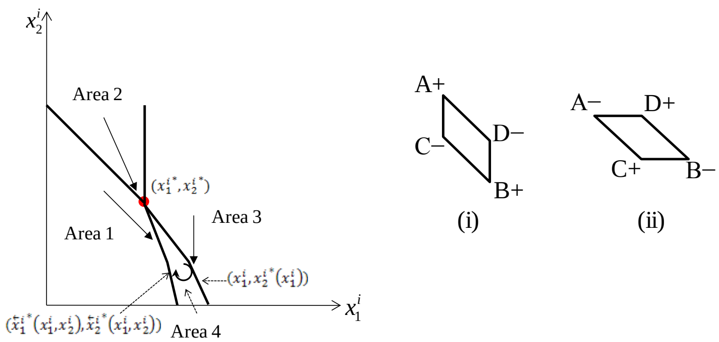


Figure 1. Optimal policy and directional properties in $(\boldsymbol{x^i_{k:1}}, \boldsymbol{x^i_{k:2}})$ – Accept, reject, or postpone

Proposition 3. For k = 1, 2, …, n, if $\phi_k(\underline{x}_k)$ satisfies $\phi_k(A) + \phi_k(B) \geq \phi_k(C) + \phi_k(D)$ in the case of lozenge (i) of Figure 1 (b) and $\phi_k(A) + \phi_k(B) \leq \phi_k(C) + \phi_k(D)$ in the case of lozenge (ii) of Figure 1 (b), then the optimal policy in $(x^i_{k:1}, x^i_{k:2})$ looks like the structure represented in Figure 1 (a) and $v_k(\underline{x})$ satisfies $v_k(A) + v_k(B) \geq v_k(C) + v_k(D)$ in the case of lozenge (i) of Figure 1 (b) and $v_k(A) + v_k(B) \leq v_k(C) + v_k(D)$ in the case of lozenge (ii) of Figure 1 (b).

The structure of the optimal policy in $(x^i_{k:1}, x^i_{k:2})$ consists of four areas (see Figure 3 (a)). In area 1, only acceptance decisions are made; in area 2, some others are accepted and some others rejected so that the optimal point $(x^{i^*}_1, x^{i^*}_2)$ is reached; in area 3, only rejections are made; and, finally in area 4, any order waiting for acceptance is not either accepted or rejected (i.e., stay put).

Proposition 4. For k = 1, 2, …, n, if $\phi_k(\underline{x}_k)$ satisfies $\phi_k(A) + \phi_k(B) \geq \phi_k(C) + \phi_k(D)$ in the case of lozenge (i) of Figure 2 (b) and $\phi_k(A) + \phi_k(B) \leq \phi_k(C) + \phi_k(D)$ in the case of lozenge (ii) of Figure 2 (b), then the optimal policy in $(g^{ij}_k, x^i_{k:1})$ plane looks like the structure represented in Figure 2 (a) and $v_k(\underline{x})$ satisfies $v_k(A) + v_k(B) \geq v_k(C) + v_k(D)$ in the case of lozenge (i) of Figure 2 (b) and $v_k(A) + v_k(B) \leq v_k(C) + v_k(D)$ in the case of lozenge (ii) of Figure 2 (b).

[1] All formal proofs of Theorems and Propositions could not be included in the manuscript due to space limitations. They will be available from the authors upon requests.

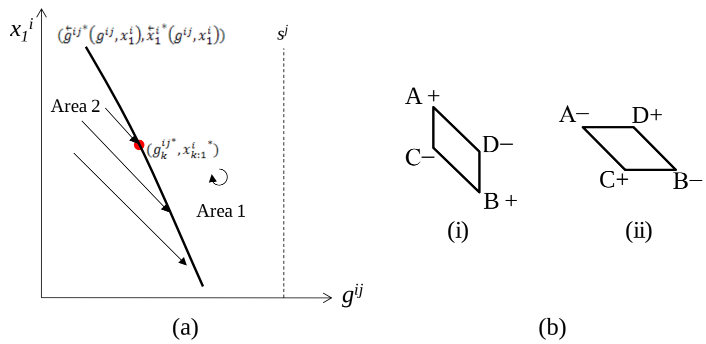


Figure 2. Optimal policy and directional properties in $(\boldsymbol{g_k^{ij}}, \boldsymbol{x_{k:1}^i})$ – Assign or postpone

The structure of the optimal policy in $(g_k^{ij}, x_{k:1}^i)$ consists of two areas (See Figure 4 (a)). In area 1, no *i*-type order is assigned to the *j*-type server (i.e., postpone assignment); and, in area 2, some orders are assigned.

4. APPROXIMATE POLICIES AND COMPUTATIONS

Although Theorems 1 (Concavity) and 2 (Submodularity) and Propositions 3 and 4 (Directional Properties) provide us with the structural properties of the optimal policy, they still cannot tell us an exact optimal policy – i.e., exact functional forms or values of the optimal policy. An exact optimal policy must be obtained by solving the dynamic equations and the optimization problem [O-1], presented in §2, via standard iterative techniques such as a backward induction and value iteration. However, these standard techniques require extensive computation as the size of a problem gets larger. This scalability problem is called *the curses of dimensionality* in the literature (Powell 2007).

In order to provide practical ways to solve relatively large problems, we propose approximate policies. Our goal is not to provide silver bullet approximate policies which perform well in every case, but to provide ones which perform reasonably well on average across various instances. We propose two approximate policies which use the structural properties of the optimal policy that we found in our analytical analysis: They are ASP (Approximate policy using the Structural Properties of the optimal policy) and ASP-Greedy (Approximate policy using the Structural Properties of the optimal policy and a Greedy method). The proposed approximate policies basically can reduce computational time by cutting some search space with the structural properties of the optimal policy. Our structural properties consist of a set of two-dimensional structures. ASP applies these two-dimensional rules in a sequence, and this sequence is also determined by the structural properties as well. Due to the limited visibility of the two dimensional rules, this sequence however may not provide a path to an optimal solution. Therefore, we propose a variant version of ASP, called ASP-Greedy, which determines the sequence of application of structural properties by using a simple greedy method on the fly.

We conduct extensive computational experiments to show the effectiveness of these two proposed approximate policies. The experimental results show that our approximate policies perform relatively close to the optimal policy. The average performance gaps with the optimal policy are 4.0 % and 3.2 % for ASP and ASP-Greedy, respectively, which also indicates that ASP-Greedy performs better than ASP on average. The results also show that all proposed approximate policies perform much better than benchmark policies, the five various versions of

FCFS (First-Come, First-Served) policies. The average performance gap between the best performing FCFS policy among the five versions, and ASP-Greedy is 26.7%, which is up to 89.3% of the gap of the best FCFS. This means ASP-Greedy improves the profits by eliminating 89.3% of the performance gap between the best FCFS and the optimal policy. The experiment results also show that our proposed approximate policies perform relatively well in most problem scenarios. The performance gaps of ASP and ASP-Greedy are less than 10% in 92.4 % and 94.9 % of problem scenarios, respectively. The median gaps of ASP and ASP-Greedy are 2.44% and 1.54%, respectively, which are smaller than the mean values. These median results also suggest that our proposed approximate policies perform relatively well in most problem scenarios.

Table 1. Effectiveness of the Approximate Policies

Size	Performance gap (%) = (Performance of the optimal policy – Performance of the focal approximate/benchmark policy) ÷ Performance of the optimal policy		
QC×SC	ASP	ASP-Greedy	FCFS
2×2	4.5%	3.2%	21.3%
4×1	3.4%	3.3%	39.7%
3×3	4.9%	2.6%	14.6%
10×1	3.4%	3.1%	39.7%
Overall	4.0%	3.2%	29.9%

Notes: QC = Queue Capacity; SC = Server Capacity; ASP = Approximate policy using the Structural Properties of the optimal policy; ASP-Greedy = Approximate policy using the Structural Properties of the optimal policy and a Greedy method; FCFS = First-Come, First-Served. It shows the gap performance of the best FCFS policy among five variants used in the experiment.

5. CONCLUSION

We study the dynamic order acceptance and resource assignment problem of a services system which provides heterogeneous services using heterogeneous resources. This problem is formulated as a dynamic and stochastic programming model, and the structural properties of the optimal policy are illustrated by analyzing the concavity, submodularity, and directional properties of the profit function of the problem. We also propose two approximate policies – ASP and ASP-Greedy – to solve this problem and conduct extensive computational experiments. The experimental results show that our approximate policies produce near-optimal solutions efficiently and perform much better than the FCFS policies, simple but widely used policies in practice. One of the unique contributions of this study is to incorporate the concept of postponable decisions in the order acceptance and assignment problem. This provides significant practical implication. Although this study is motivated by the multi-project selection and assignment problem of IT service companies, postponed decisions (or deferred decisions) can be easily observed in our daily life.

6. REFERENCES

Akçay, Y., Balakrishnan, A. and Xu, S. H. 2010. Dynamic assignment of flexible service resources. *Production and Operations Management* **19**(3) 279-304.

Ferguson, T. S. 1989. Who solved the secretary problem? *Statistical Science* **4**(3) 282-289.

Kleywegt, A. J. and Papastavrou, J. D. 1998. The dynamic and stochastic knapsack problem. *Operations Research* **46**(1) 17-35.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, Wiley-Interscience, Hoboken:NJ.

Using Virtual Worlds as an Innovative Technology for Teaching[1]

Rui Huang
Binghamton University
rhuang@binghamton.edu

Rebecca Jestice
Earlham College
jestire@gmail.com

Virtual worlds are garnering a lot of attention from educators as a new tool of teaching. Educators take advantage of the immersive experience enabled by virtual worlds, such as teaching classes within a virtual world, and using virtual worlds as a supplement to traditional classroom environments to facilitate experiential learning, simulation, role play, and collaboration, etc.

Proponents of the use of virtual worlds in education cite benefits such as cutting costs, flexibility for learners, diversity in classes, increased sense of presence and immersion over other online learning media, decreased anxiety for students, and the ability to construct or reconstruct experiences that is not possible in real world classrooms. However, despite their advantages, virtual worlds also pose some problems that may hinder their effectiveness such as cost to set up and run and poor usability for teachers and students. Students may be exposed to inappropriate individuals or information. Using technology for education may also be harmful to learning because of a lack of richness or naturalness that one experiences in face-to-face classrooms. These challenges associated with teaching in virtual worlds require instructors to have technological knowledge and virtual classroom support and management skills.

We have conducted several studies using the virtual world for teaching. For example we used the virtual world Second Life for teaching students about data center design. Based on the research we have conducted involving the use of virtual worlds for teaching, we highlight the point that because the learning curve for using virtual worlds is very steep, it is important to make sure that the characteristics and features of virtual worlds are being used in a unique way to add value. Otherwise the cognitive load necessary to maneuver through and use the virtual world is additional load that does not benefit learning. For example, students can learn about data center design and operations through a simulation that shows potential configurations and effects as opposed to listening to a lecture within the virtual world about data centers. Learners need to be given some direction for their learning to aid them in using the technology effectively. They can look at models of hardware components or run through a disaster planning scenario within a virtual world environment. With knowledge of the learning rubric, learning experience (cognitive flow) and learning outcome will improve in virtual worlds. Simply assuming that an immersive experience will lead to better outcomes without helping the students with their use of the technology will likely be less successful.

[1] Authors are listed in alphabetical order denoting equal contributions.

It's Not Just a Class, It's an Adventure:
Teaching Web Development Through Game Creation

George M. Wyner
Boston University School of Management
gwyner@bu.edu

Benjamin Lubin
Boston University School of Management
blubin@bu.edu

It is increasingly accepted among business professionals that an understanding of the technologies behind web applications is an essential portion of their skill set (Wortham 2012). Accordingly, information systems (IS) faculty have expanded IS curricula to include courses on modern dynamic web application development. Consistent with this trend, we have created an instructional module on web application development targeted at a group of students concurrently pursuing an MBA and a Master of Science in Information Systems.

Our method consists primarily of a hands-on exercise in which students create their own web application. This exposes them to an important set of web technologies — HTML, CSS, web server, and application server — without plunging them into an overwhelming morass of technical detail.

For the exercise, we provide student teams with a set of tools for building a text-based adventure game (Jerz 2007; Lebling et al. 1979) using the Google AppEngine hosting service and application framework. Students then write their project using the Python programming language. Because the specific nature of the game is left completely unspecified, students develop something that reflects their own imagination even as they are learning web application software engineering and architecture. The code for our framework and other teaching materials, including a demonstration game, have been made available to the IS community at code.google.com/p/msmba-ae-game/.

We chose a text based approach because, while less flashy than a modern interactive game, it builds on previous classwork in which students learned how to work with text in Python and does not require a knowledge of animation techniques or complex user interfaces. Instead, the interface is a fairly straightforward use of basic HTML form elements. We do provide students the ability to include graphics such as images of rooms and the objects in them. This gives students a chance to learn how to use images in a dynamic web page and introduces options for visual creativity.

We provide lectures that cover the key building blocks of web applications including HTML and CSS, as well as the architecture of web applications including databases and web servers. The lectures are self-contained but students are strongly encouraged to rely on *Using Google AppEngine* as a reference (Severance 2009). We also found great utility in the slides, screencasts, and sample code provided by Chuck Severance on his website www.appenginelearn.com.

References

Jerz, D. G. 2007. "Somewhere Nearby Is Colossal Cave: Examining Will Crowther's Original ‘'Adventure' ’In Code and in Kentucky," *Digital Humanities Quarterly* (1:2).

Lebling, P. D., Blank, M. S., and Anderson, T. A. 1979. "Special Feature Zork: A Computerized Fantasy Simulation Game," *Computer* (12:4), pp. 51-59.

Severance, C. R. 2009. *Using Google App Engine*," Sebastopol, CA: O'Reilly Media.

Wortham, J. 2012. "A Surge in Learning the Language of the Internet," in *The New York Times.*

"Interactive Classroom Modeling of Key Information & Technology Management Issues"

(Or, how to better engage the MBA students of the post PC generation)

Professor Abraham Seidmann

Xerox Professor of Computers & Information Systems, Electronic Commerce, and Operations Management
W. E. Simon Graduate School of Business Administration
University of Rochester
Rochester, NY 14627
Tel: (585) 275 – 5694

Avi.seidmann@simon.rochester.edu

In this session we plan to discuss and demonstrate key methodological issues in the innovative use of IT in the classroom as well as showcase interactively new tools for conducting web-based experiments focusing on the Economics of Information Technology. We then plan to summarize some of our recent advances in the creative use of laboratory experiments in the MBA Classroom, at the University of Rochester, and at several other leading programs.

There is a growing body of research focusing on Empirical or Behavioral Aspects of Information Systems and management. Recognizing that laboratory and classroom experiments give researchers a great deal of control, makes them useful for testing analytical models, and for experiential learning. The topics of experiential research and teaching has been gaining a lot of attention in recent years and several interest groups have formed around it already. This will be a hands-on session, communicating our own experience in the topic, and the **Tradewind platform** that I have been developing at the Simon School. The session will also summarize some of our recent advances in the creative use of laboratory experiments in the MBA Classroom. We are going to demonstrate the WWW.Tradewindbusiness.com software that has been already in use (at NO charge), by close to ten leading universities and international organizations. The most popular games include the **Fishing Game**, the **Pricing of Information Goods Game**, **The Stock Trading Game**, and the **Newsvendor Game**. All these web-based games incorporate mathematical models, the latest Web-based tools, and interactively illustrate theories underlying digital market competition, decision optimization, pricing, and the resulting competitive equilibrium. Students and executives who have used these competitive games found them to be extremely useful in developing their technological and information-economics insights, well beyond the typical classroom lectures, case studies, or a general discussion. The games are web based and run on any standard browser; users do not have to download, or to install any software on their computer. A typical classroom session runs for 80 minutes, including initial setup, competitive engagement, and a post-game analytical briefing. A detailed CSV file is available for an optional follow-up assignment by the teaching faculty.

Wearing Your Heart on Your Sleeve: The Effects of Forums and Search on Sales

Tomer Geva[a], Gal Oestreicher-Singer[b], Niv Efron[c], Yair Shimshoni[d]

The availability, level of detail and scale of online data have encouraged researchers as well as practitioners to explore means of using such data to explain offline economic outcomes. Two streams of research have attempted to investigate the informativeness of two different kinds of online data: The first focuses on abstracting data from social media websites such as blogs or forums as a proxy for the volume of word of mouth (WOM) a product enjoys. The second focuses on using search engine logs, which have recently been made public through tools such as Google Trends, as a proxy for consumers' interest in various products. While the broad goal of those two streams is similar, the two sources of data are very different in nature. Opinions posted on social media sites are made publicly visible to all, whereas search is conducted privately. Clearly, publicly broadcasted opinions and conversations may influence other consumers. On the other hand, it has been argued that search data reflect the "true intentions" of consumers. Interestingly, the interplay between these two important sources of data has not been studied. This research seeks to close this gap.

Specifically, as both social media mentions and search volume have separately been shown to have high correlation with offline sales, and are also found in this study to be highly correlated with each other, we ask: Are the two sources complements or substitutes? That is, can current and future sales be better explained using both, or does the information conveyed by one data source render the information conveyed by the other source redundant?

To address this question we use monthly data from the automotive sector between January 2009 and December 2010. We utilize three data sets: mentions of automobile brands in forum data extracted from Google's comprehensive scan of the internet; Google's search engine query logs; and sales data. Using the Arellano-Bond Panel VAR estimation method, we show that the two sources carry complementary information which is useful for better explaining current and future sales.

This research is still in progress. Nevertheless, to the best of our knowledge, this study is the first to use both search engine query logs and forum posting information and to provide evidence that using both WOM data and search logs better explains current and future sales.

Having provided evidence that both data sources contain complementary information, we are currently working to expand this research in several directions. First, we are working on a theoretical model that may explain the dynamics between search volume, forum mentions, and sales. Second, we are working to introduce WOM valence (or sentiment) into our modeling and data. We further plan to work on obtaining additional insights regarding the extent to which search and forums provide complementary information under various conditions. This includes evaluating specific models for different kinds of car brands and different time periods. We will also evaluate the effect of introducing into the model "external" factors such as customer incentives and promotions. Last, we plan to enhance our research in terms of its predictive capabilities and test out-of-sample predictive models using both theory-driven models and data-mining methods.

[a] Google, Tel Aviv, Israel. tomergev@post.tau.ac.il
[b] Recanati Business School, Tel Aviv University, Israel. galos@post.tau.ac.il
[c] Google, Tel Aviv, Israel. niv@google.com
[d] Google, Tel Aviv, Israel. shimsh@google.com

Virtual World for Advanced Cardiac Life Support Training*

Prabal Khanal, ME[1], Akshay Vankipuram, MS[1], Aaron Ashby, MS[1],
Karen Josey, MEd, BSN, RN [2], MSL , Ashish Gupta, PhD[3], Marshall Smith, MD, PHD[2]
[1]Arizona State University, Scottsdale, AZ; [2]Banner Health SimET Center, Phoenix, AZ
[3]Minnesota State University Moorhead

*This project has been funded by Telemedicine and Advanced Technology Research Center (TATRC), DoD (Award # W81XWH-08-2-0186)

Use of virtual worlds in the healthcare contexts that require team coordination and communication offer tremendous opportunities for improvements. Specifically, virtual worlds may alter how medical training is currently being delivered that require dealing with team based scenarios and can lead to significant improvements in various performance outcomes, cognitive and psychomotor learning skills of users. However, the application and utility of virtual world in specific medical traning programs remains nascent and ill-understood at best. In this study, we present a design framework for a system to train care providers on complex, time-critical and collaborative medical procedures. We demonstrate a virtual reality simulator developed on the *Unreal Engine* platform using *Unreal Development Kit*® for providing Advanced Cardiac Life Support (ACLS) training for care providers in simulated environments. ACLS is a series of team based medical interventions performed during the emergency medical resuscitation of an individual in a state of cardiac and/or respiratory failure. Team members, while executing ACLS guidelines on the patient, are assigned specific roles and requires communication and coordination of activities.

The core architecture of the system (Figure 1 and 2) consists of five components: visual, haptics, auditory, database, and performance evaluation. Visual and haptics components allow users to interact with the system to perform required tasks in the simulated environment. The auditory component provides better communication channel between the users during training sessions. The database component is responsible to maintain records of all the activities and communication performed during the training session for post-hoc. Finally, the performance evaluation component provides real time instructions and performance feedback based on the rules that we define during the training session. We use the ACLS guidelines provided by American Heart Association (AHA) to generate a rule-list for performance evaluation. The components described in this architecture are reusable for designing training system for other complex team-based procedures provided that the rules can be generated for evaluation.

The virtual world for ACLS training that we have developed offers numerous advantages over the traditional approach used for conducting ACLS training in terms of feedback immediacy, ability to conduct frequent remote training sessions, convenience and schedule flexibility while freeing up hospital resources. A large-scale comparitive study is underway to compare the efficacy of ACLS training in simulated environment.

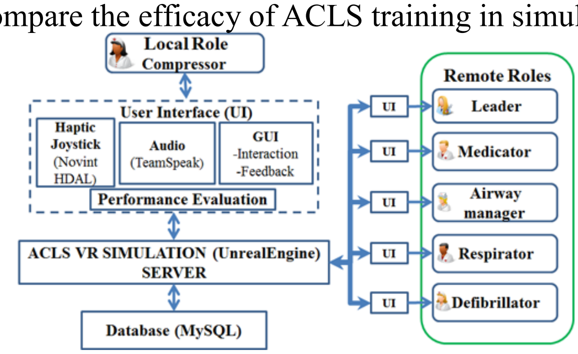


Figure 1. System Architecture

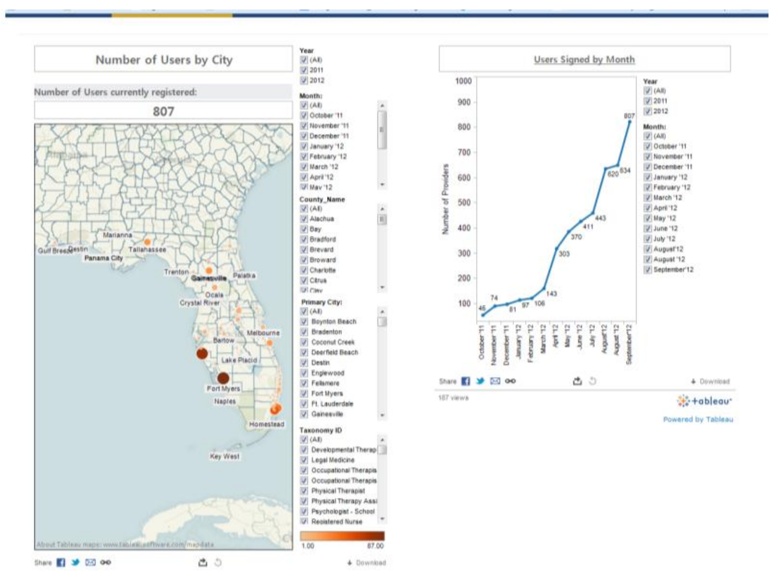
Figure 2. User interface for different roles

Research Prototype: A Knowledge Management System to track the Evaluation of the Implementation of a Statewide Health Information Exchange

Monica Chiarini Tremblay PhD, Arturo Castellanos MSMIS, Gloria Deckard PhD
Florida International University - Miami, FL

Florida International University (FIU) was awarded the contract for the process evaluation of the implementation of the Florida Health Information Exchange (HIE) which was funded by the Office of the National Coordinator for Health Information Technology (ONC) with dollars from the HITECH act (Obama 2008). As part of the contract, a Knowledge Management System (KMS) was developed to display dashboards of key evaluation measures and to provide links to Florida HIE resources. The objective was to provide an assessment tool which would allow the citizens of the State of Florida to monitor the implementation and adoption of the HIE by healthcare providers. The KMS can be found at http://floridahie-eval.fiu.edu.

Development of a Knowledge Management System. The KMS was developed using Microsoft SharePoint, a content management system (CMS) that allows administrators to update the site easily, while keeping a log of all the changes that are made. In order to create an interactive and user-friendly environment, different web technologies and libraries were used to complement the CMS. Tableau Public, a data visualization tool, was utilized to create dashboards that facilitate interaction with the progress of the HIE implementation. These dashboards allow the user to drill-down from a high to a lower level by applying filters (e.g. month, county, city, etc.). Figure 1 provides an example that displays adoption of Direct Secure Messaging, one of the HIE implementations in the state of Florida.



The KMS contains dashboards with key performance indicators and reports that the State of Florida can utilize for benchmarking and displaying metrics and measures requested by the ONC. The metrics included as evaluation measures focus particularly on the ONC priority areas of:
1) Laboratories participating in delivering electronic structured lab results
2) Pharmacies participating in e-Prescribing
3) Providers exchanging patient summary of care records

Conclusion. The KMS system and metrics demonstrate the progress of the State HIE implementation in an easy to use environment. It allows its users to monitor the pace in which health care providers are onboarding into the use of HIE through both PLU and DSM in the State of Florida. The state of Florida has been utilizing the KMS as a marketing tool that encourages adoption.

A Prototype of a *PATIENT SAFETY KNOWLEDGE MANAGEMENT SYSTEM (PSKMS)*

Parameswaran, S., Valecha, R., Sharman, R., Rao, H.R., Singh, R and Singh, G.
State University of New York at Buffalo, NY 14260

Brief description of the prototype system: This prototype (PSKMS) is a software system that captures medical errors and creates knowledge to facilitate mitigation of medical errors. There are existing patient safety systems such as Medical Event Reporting System (MERS), US Patient Safety Reporting System (PSRS), UK National Reporting and Learning System (NRLS), etc. that have been designed to capture medical errors (MERS International, 2007; NASA, 2005; NPSA, 2007). However these systems just capture the tip of the iceberg in terms of medical errors, and majority of them go unreported (Singh et al. 2007; Singh et al. 2012). Moreover these systems do not provide a mechanism to generate knowledge about the errors. To our knowledge the PSKMS is the first software that builds a knowledge base of errors to help us understand the cause of errors, trajectory of errors and its impact. PSKMS is installed in the hospitals for use by patients, physicians, nurses, staffs and other decision-makers. It allows the users to report the errors, decision-makers to analyze these errors, and also suggest corrective actions. Thus, PSKMS facilitates knowledge creation by providing a platform for error reporting, incident analyzing and decision making. Such a system that is helpful in management of errors reduces the burden of patient liability and financial liability.

Features of the Software: The basic feature of the software is that it allows users like patients, caretakers, staff and physicians to report medical errors, and aids decision makers to generate interventions as preventable actions. Special features include (a) front-end interfaces ranging from user-centric to expert-centric for better accessibility, (b) decision makers dashboard to analyze error incident and its impact, and (c) integration in other healthcare settings.

Research Foundation for the Prototype: The system design strictly adheres to the design science research guidelines (Hevner et al. 2004; Peffers et al. 2006), which include step-by-step approach in development of PSKMS. We adapt the Knowledge Management Process (KMP) devised by Allavi and Leidner (2001), and use it as the guiding basis in our design process. We use Reason (2000) Swiss cheese model to categorize errors in order to design our menus and submenus. The work also leans on other research in the area of patient safety.

Future Research: The prototype is intended to serve as a platform for further research in patient safety incident management; for example to study issues relating to error reduction and decision making under time sensitive conditions. As a part of future research, we plan to evaluate our system from patient and physician perspective. The implementation of the system using existing health information exchange standards is also a possible extension.

Conclusion: The prototype allows for testing theories in the IS area that cannot be fully enumerated here given space constraints. Practice in the field will be more informed as a consequence of the developments.

References

Alavi, M and Leidner, D. (2001). “Review: knowledge management and knowledge management systems: conceptual foundations and research issues,” MIS Quarterly 25 (1) (2001) 107–136.

Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004). “Design Science in Information Systems Research,” Mis Quarterly. (28:1), pp. 75-105.

MERS International. Medical event reporting system, 2007. http://www.mers-international.com/index.html (accessed 28 Jul 2008).

NASA. Patient safety reporting system, 2005. http://www.psrs.arc.nasa.gov (accessed 28 Jul 2008).

NPSA. National reporting and learning system, 2007. http://www.npsa.nhs.uk/patientsafety/reporting (accessed 28 Jul 2008)

Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V., and Bragge, J. (2006). “The Design Science Research Process: A Model for Producing and Presenting Information Systems Research,” Proceedings of the 1st DESRIST. Claremont, CA pp. 83-106.

Singh, R., Naughton, B., Singh, A., Anderson, D. R., and Singh, G. (2007). The Safety Journal: Lessons Learned With an Error Reporting Tool to Stimulate Systems Thinking. Journal of Patient Safety, *3*(3), 135-141.

Singh, R., Sharman, R., Anderson, D., Singh, A., and Singh, G. (2012, March). Healthcare simulation for excellence in in-patient safety. In Proceedings of the 2012 Symposium on Emerging Applications of M&S in Industry and Academia Symposium (p. 11). Society for Computer Simulation International.

A Prototype of a Local Emergency Response System based on a Multi-agent conceptual Modeling Language

Rohit Valecha, Swati Subhedar, Kshitj Agrawal, Raj Sharman, Raghav Rao, and Shambhu Upadhyaya,
State University of New York at Buffalo, NY 14260

Brief description of the prototype system: This prototype is a software system to deal with local day-to-day emergencies. It facilitates management of emergency resources such as equipment, responders, etc. to better mitigate local emergencies. There are several software systems in the market for dealing with local emergencies such as 9-1-1 call monitoring, resource dispatching, and radio communication amongst others – however these systems are difficult to interoperate due to lack of common design, varying responder requirements, etc. To our knowledge this is the first software system that is developed based on a multi-agent conceptual modeling language. Further this prototype system provides an interoperable solution for integrating systems and responders from multiple agencies such as fire, EMS, law enforcement, hazmat teams, and other local, county, state and federal agencies. The goal is to demonstrate system design and development over a common modeling language for improved communication and coordination among local fire departments, law enforcement agencies, medical services and other agencies. The interactions among responders are useful in co-creating a common operating picture to improve response operations, and to better apply resources to mitigate the incident.

Features of the Software: The basic feature of the software is that it allows emergency system developers to design front-end screens and back-end data tables using a common multi-agent conceptual modeling language. This can provide interoperable modules from entering information to mitigating the fire related incident ranging from resources on location, responding personnel and mutual aid resources based on pre-plans after availability confirmation. Special features include (a) incident command center dashboard and (b) back-end database design diagrams.

Research Foundation for the Prototype: This software was developed to serve as a prototype for the paper by Valecha et al. (2012). It strictly adheres to the design science research guidelines (Hevner et al. 2004; Peffers et al. 2006), which include step-by-step development approach. It builds upon the multi-agent conceptual modeling language and grammar (MLERS) specified in these papers. The multi-agent conceptual modeling language is derived using more than 10,000 real emergency messages and Speech Act Theory (Searle, 1975). The work also leans on other research in the area of emergency response.

Future Research: The prototype is intended to serve as a platform for future research in the local emergencies. It will facilitate in depth research relating to issues such as resource allocation (example Rolland et al., 2010), conflict resolution and decision making under time sensitive conditions, etc. The software will also be migrated to the cloud to allow for investigating other research issues stemming from migration and implementation, and cloud security. A future extension of this prototype will be to make twitter and social networking.

Conclusion: The prototype allows for developing interoperable systems over a common multi-agent conceptual modeling language. It also allows testing theories in the IS area that cannot be fully enumerated here given space constraints. Practice in the field will be more informed as a consequence of the developments.

References

Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004). “Design Science in Information Systems Research,” Mis Quarterly. (28:1), pp. 75-105.

Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V., and Bragge, J. (2006). “The Design Science Research Process: A Model for Producing and Presenting Information Systems Research,” Proceedings of the 1st DESRIST. Claremont, CA pp. 83-106.

Rolland, E., Patterson, R. A., Ward, K., and Dodin, B. (2010). “Decision support for disaster management”, Operations Management Research (2010) Vol. 3, pp 68–79.

Searle, J. (1975). “A taxonomy of illocutionary acts. In Language, Mind and Knowledge,” Minnesota Studies in Philosophy of Science, 11, Gunderson, Ed. Uni. of Minnesota Press

Valecha, R., and Sharman, R., “Modeling Language for Emergency Response Systems (MLERS)”, 12th Workshop on E-Business, Orlando, Florida, Dec. 2012

An Analytics Platform for Mobile Applications

Anindya Datta,Sangaralingam Kajanan
Department of Information Systems, National University of Singapore
{datta,skajanan}@comp.nus.edu.sg

1 Introduction

Mobile smart-device platform ecosystems (Apple iOS, Android, Microsoft Phone 7, etc.) are the emergent digital innovation value chains in today's E-businesses. These mobile platform ecosystems and their consumer software applications (simply known as "apps") represent the fastest growing consumer product segment in the annals of human merchandising . Innovation activity in mobile platform ecosystems and the rate of growth of apps available for download are incredible. For example, each week about 100 new movies and 250 new books are released worldwide. In contrast, over 15,000 new apps, on average, are launched weekly. Currently there are over 1.2 million apps on the Apple, Android, Blackberry and Microsoft native app markets [1]. Each one of these applications generates continuous and ever increasing stream of data such as performance statistics of individual apps, user reviews and social media mentions. The volume of this data is over whelming very rapidly. This data reveals the crucial intelligence for end users, developers and mobile app ad-networks. Consumers need to discover the most suitable app for their need and developers and owners want to derive the right business strategies for their apps. In addition,advertisers and agencies need to pick up the right app to promote their advertisements.

"Mobilewalla Analytics"(MWA) is the app industrys first "big data" and "deep analytics" solution, which tracks all available information about apps as well as data extracted from social media sites for consumers. MWA accumulates and analyze this over whelming data in real time and be the most trusted source of information for the app ecosystem like the Nielsen and ComScore for television and web industries respectively. The platform, MWA, adds about 1 terabyte of data weekly. MWA provides two types of analytics for consumers, developers and mobile advertising networks to ease their decision making.

- **Performance Analytics:** How apps are performing with respect to a variety of observable metrics, such as rank, ratings, reviews, over time
- **Audience Analytics:** Deep profiles of app audience at both micro (individual app) as well as macro (category) levels.

2 Key Features of Mobilewalla Analytics

Mobilewalla Analytics platform can be accessed through `http://analytics.mobilewalla.com`. The web site presents the user with the list of hottest apps organized by country and by category. In addition, MWA provides lot of insights about an app for the developers and end users. Further, the developer of an app can use "Daily App Rank" in MWA to get a summarized report of the rank movement of his/her app across different countries. Moreover, "App Rank History" in MWA helps developer to check his/her app's overall or category specific rank change during a period of time. Furthermore, MWA helps app developer to figure out his/her app's overall and category specific daily gross rank and gross rank history. Further, developers and consumers can check the variation of Mobilewalla Score[1] of an app across days using MWA platform. "Features" and "Summary of features" in MWA provides details of special features of an app and summary of features respectively. Another significant feature of MWA is, app review analytics. Moreover, MWA platform analyzes the overlay of different app features. In addition to the above mentioned key features, Mobilewalla's noteworthy feature, "Social Media mentions" enable users to view an app's social media mentions from Facebook, Twitter and Youtube. In addition to all of the above mentioned key features, MWA's developer forum provides list of APIs and platform specific forums for developers to discuss the issues among them.

References

[1] A. Datta, K. Dutta, S. Kajanan, and N. Pervin, "Mobilewalla: A mobile application search engine," in *Mobile Computing, Applications, and Services*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, 2012, vol. 95, pp. 172–187.

Demonstration of an Analytical Software Feature in Generalized Audit Software: Use of Benford's Law Software Feature for Fraud Detection Audit Tasks

Hyo-Jeong Kim, Michael Mannino
University of Colorado Denver

Advanced analytical features such as Benford's Law are very useful for detecting fraudulent transactions in accounting data. However, those features were less utilized by both auditing students and professional auditors, in part because they perceived that advanced features are useful but not easy to use. To diminish the learning difficulty of advanced auditing features, we provide a demonstration for Benford's Law software feature, a built-in feature in Generalized Audit Software (GAS) and an option in statistical software and spreadsheet modules. The demonstration includes a conceptual background of Benford's Law and the use of Benford's Law software feature[1] to perform fraud detection procedures, leading to increased use of analytical software features in GAS.

Benford's Law is based on a finding that low first digits occur more frequently in most sets of data (Benford 1938). It has been used to detect potential anomalies in many data sets such as disbursements, sales, expenses, accounts receivable, and accounts payable. A deviation from Beneford's Law can mean that the data has been invented, contrived, or manipulated. Although the Benford's law feature is not difficult to use, novice users of Benford's Law still need comprehensive demonstration of Benford's Law covering both conceptual and product details to correctly apply and interpret the analysis. This software demonstration can benefit not only auditing students and professionals but also IT professionals who work on data mining projects especially with accounting data sets.

To facilitate the learning process of this target audience, the demonstration will offer a short lecture on Benford's Law, an exercise with guided instructions, and a case-based assignment to apply the Benford's Law feature of ACL, a popular GAS. The lecture on the conceptual background of Benford's Law provides the foundation for target audience to effectively use the Benford's Law feature of ACL. In the initial exercise, audience will practice the Benford's Law feature of ACL with expense reimbursement data[2]. In the case-based assignment, they can apply their knowledge from the exercise to answer the questions provided in the case assignment with invoice data[3].

This demonstration was used by professional internal auditors in the data collection of a study focusing on the impact of advanced feature training for GAS. Fifty six internal auditors completed the lecture, guided exercise, and assignment. Prior to conducting the study, the instructional materials were developed through pilot tests using a master's level advanced auditing students and several internal auditors.

The lecture and software demonstration are available at the following link:
http://ouray.ucdenver.edu/~h1kim/BenfordsLaw/Benford.htm

---

[1] ACL is one of most commonly used Generalized Audit Software, so we used it for our demonstration.
[2] Data file from "Auditing: A Business Risk Approach; 6th edition; South-Western; 2008; L. E. Rittenberg, B. J. Schwieger, K. Johnston."
[3] http://www.ezrstats.com/CS/CS_Wake_County.htm (April, 2011)

Face Recognition Enabled Avatar

Doug Derrick
dcderrick@unomaha.edu

Aaron Read
aread@unomaha.edu

Description

Figure 1: Avatar Screenshot

We have combined an embodied intelligent agent (or Avatar—see Figure 1) with face recognition software (PittPatt) to enable the Avatar to speak the name of an individual who has been enrolled in the recognition database. Individuals may be enrolled in PittPatt's face recognition database remotely via a mobile app developed for Apple iOS and Android OS mobile devices. See Figure 2 for a screenshot of the enrollment app for IOS devices. Enrollment includes entering the individual's name, taking a front face picture as well as a left and right picture. The face recognition device is an Xbox 360 Kinect Sensor which has a 1.3 megapixel camera. The PittPatt face recognition software utilizes the image taken from the sensor and processes it in real time, allowing the Avatar to respond to the individual upon "seeing" them from the front of the sensor.

Goal of the Prototype

The primary goal of the prototype is to demonstrate the increased sociability of an avatar achieved through the ability of an avatar to speak someone's name to them upon visual recognition of the person. Sociability of agents is important for increasing the trust and enjoyment of humans when interacting with computers to make decisions based off of the computer's recommendations, or perform other computer-guided tasks.

The prototype also demonstrates the convenience of being able to enroll individuals in the face recognition database on site. Such a tool would be useful in both practical and experimental settings to enable ad-hoc additions to a face recognition database.

Intended Impact

A face-recognition enabled avatar is a useful component in many applications where interaction with a human is automated. In the near future we will test this component in the context of decision support. The face recognition component can already be added to existing avatar-based applications, such as a border patrol entry kiosk that has already been developed by Dr. Derrick.

Figure 2: Enrollment Tool Screenshot

Gamifying Collaborative Decision Making

Mohammad Ali Moradian*, Kelly Lyons*, Maaz Nasir*, and Rock Leung^
*University of Toronto ^SAP Canada
This work was funded by an NSERC Collaborative Research and Development Grant with SAP.

Engaging people to participate fully in online collaborative decision-making can be challenging because people are very busy juggling competing demands for their time. Gamification has been used in a variety of environments to incent and increase participation [3]. We implemented and gamified two decision-making tools in SAP's StreamWork collaborative platform [2]. In a comparison of the gamified and non-gamified versions, we found that groups in the gamified version contributed more ideas, engaged in more discussion, selected a greater number of final ideas, and were equally satisfied with the decision-making outcome and process.

We implemented two decision-making tools (Brainstorming & FastFocus) based on the concept of ThinkLets [1] and integrated them into SAP's collaborative platform, StreamWork [2]. SAP StreamWork is a social media platform that supports enterprise-wide and inter-organizational group collaboration through common tools such as pro/con lists, ranking lists, SWOT tables, and polls. The Brainstorming tool enables members of a group to anonymously generate a list of ideas on a pre-selected topic. The FastFocus tool guides the group through several rounds of discussion about selected ideas from the Brainstorming activity. The result is a refined final list of ideas. We added the following individual and group game dynamics to the Brainstorming and FastFocus tools: i) Points for group and individual accomplishments, ii) Achievements for group goals, iii) Goals for the group, iv) Progress bars for group goals, v) Feedback for groups and individuals, and v) Leader boards for individuals.

We used the StreamWork OpenSocial API to integrate new tools into the StreamWork platform. Our custom tools were mostly written in C#, using an MSSQL database, and were published on a commercial server. We also made use of JQuery and AJAX to asynchronously load content and create a seamless user experience.

In two sections of a graduate level project management class, we asked groups of 3-4 people to participate in a project selection decision-making task using Brainstorming followed by FastFocus. Of the 18 groups, half of them used the non-gamified versions of the tools and the other half used gamified versions. We compared the two sets of groups according to their interactions with our system and through a brief post-activity survey. We found that the groups that used the gamified version contributed more Brainstorming ideas per person, nominated a greater number of ideas for the final list, spent more time discussing each idea, and generated more discussion text. All groups were equally satisfied with the decision process and outcome.

References:

[1] Briggs, Robert O., Gert-Jan De Vreede, and Jay F. Nunamaker. "Collaboration engineering with ThinkLets to pursue sustained success with group support systems." Journal of Management Information Systems 19.4 (2003): 31-64.

[2] Kanaracus, C. "SAP's 'Virtual War Room' tool gets a name: StreamWork." http://www.pcworld.com/businesscenter/article/192903/saps_virtual_war_room_tool_gets_a_name_streamwork.html, May 2010.

[3] Thom, Jennifer, David Millen, and Joan DiMicco. "Removing gamification from an enterprise SNS." Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, 2012: 1067-1070.

The Role of Product Variety and Maturity in the Market Valuation of IT Intensive Firms

Wael Jabr and Eric Zheng

1. Motivation

The pervasiveness and intensive use of Information Technology (IT) in every aspect of firm processes have enabled firms to maintain a portfolio of diverse products. By the same token, firms face a big challenge of how best to manage this portfolio under disruptive IT environments (portfolio defined here as a variety of products at different stages of maturity).

2. Research Focus

Our work extends a recent stream of research which established that IT investments contribute to firm performance, in terms of productivity (Mittal and Nault 2009) and profitability (Mithas et al. 2012). IT investments have also been found to increase the level of product variety that firms maintain (Gao and Hitt 2012).

We focus on IT intensive industries and hypothesize that diversity and maturity of products play an important role in the firm's performance and its market value. At this preliminary stage, we define product diversity as the extent to which a firm operates multiple products and the extent to which these products are different relative to products developed by own firm and other firms. We also define product maturity as the extent to which a firm's products change relative to own past products.

The complexity of the research question we attempt to address is multifaceted. First, while the concepts of product diversity and maturity seem quite common-sensical, it is in reality very hard to pin down. Second, similar complexity issues are related to the concept of firm performance. Several studies have shown the limitations of relying on accounting-based performance measures to identify the impact of IT investments. Third, there is a difficulty in identifying the relationship between product portfolio (namely diversity and maturity) and firm performance due to various confounding issues the least of which is the lagging impact of IT.

To address these challenges, we study firms that manage a portfolio of products at different stages of their life cycle. We investigate the role that the portfolio plays on firm performance. We identify multiple confounding factors to this relationship including firm alliances, R&D investments, IT investments, level of outsourcing, patent production, word-of-mouth generated, market competition, and overall market performance (e.g., cyclicality).

3. Model and Findings

We adopt the following model:

$$TobinQ_{it} = \alpha + \sum_j \beta_j Variety_{it,j} + \sum_j \beta_j Maturity_{it,j} + \sum_j \beta_j Sales_{it,j} + \sum_j \beta_j\, R\&D_{it,j} + \sum_j \beta_j\, Size_{it,j} + \beta_j MarketShare_{it} + \beta_j IndusType_{it} + \beta_j IndusConc_{it} + \mu_{it}$$

Using text mining, 10K reports and financial ratios for IT intensive firm, we extend the current and often conflicting literature on the role that IT plays in firm performance, by offering a new perspective from the angle of diversity and maturity of IT products. We quantify the impact of product portfolio on firm performance and market valuation. This offers new insights for firms to make decisions on the selection of portfolio of IT products and provides a new tool for potential shareholders to determine the value of the firm.

Fixed, Spot and/or Flexi Pricing: An Integrated Prototype for Examining Alternate Pricing Mechanisms in Cloud Computing

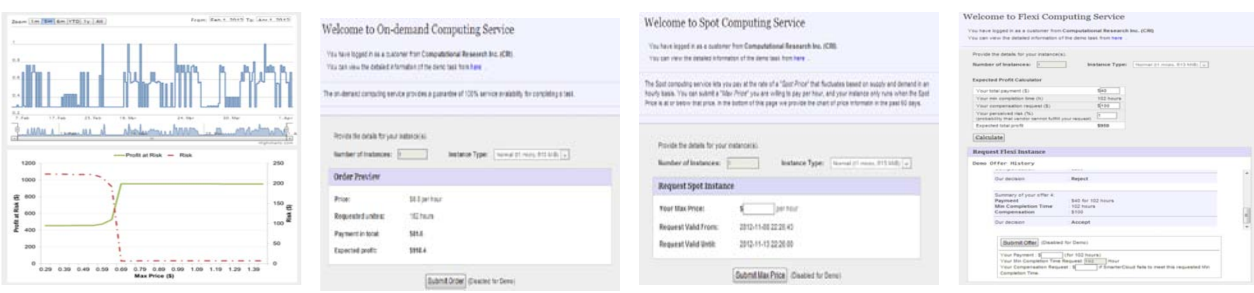
YANG Yinping [†Δ], Richard SHANG[†], HUANG Jianghui [Δ]

[†] Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore
[Δ] School of Information Systems, Singapore Management University (SMU), Singapore

yangyp@ihpc.a-star.edu.sg, richard-shang@ihpc.a-star.edu.sg, jhhuang.2009@phdis.smu.edu.sg

Pricing mechanisms in Infrastructure-as-a-Service (IaaS) cloud computing have evolved over years. Amazon.com's spot IaaS services market (*"Spot"*), launched in December 2009, has become an alternative source to its fixed-price on-demand services market (*"Fixed"*) for users who wish to leverage their budgets and service-level preferences. Nevertheless, the price volatility of the spot market subjects users to a risk of service interruptions. Grounded on our empirical understanding of the Amazon's spot market, we performed an economic analysis to evaluate users' job risk due to the spot market's price volatility (Shang et al. 2012a). We conclude that providing cloud computing services based on a more flexible, customized service level agreement (SLA) (*"Flexi"*) can lead to an increase in user-vendor joint utility and a practically reasonable negotiation interval between users and the service provider.

To allow for empirical examination, we created a research prototype in the form of a hypothetical cloud broker called "SmarterCloud Computing Services" (see URL: www.smarter-cloud.biz) which provides IaaS services with three types of pricing options: Fixed/On-demand, Spot, and Flexi. The site has three functional features to support the comparison of the alternative pricing options to help cloud users maximize their net profits (see demo task scenario for a typical cloud user: http://www.smarter-cloud.biz/biz_demoTask). **1) Risk Assessment**. The user is given an analytics tool to assess the risk associated with using the *Spot* market based on historical spot price traces, and with using the *Fixed* and *Flexi* markets based on SLA terms. **2) Profit Estimation**. The user can calculate and compare her expected net profit based on her own acceptable risk level (for *Spot* and *Flexi*). **3) SLA Negotiation**. If the comparison suggests that *Flexi* is more profitable, the user can negotiate a SLA with the vendor. The vendor is implemented using a software agent based on a simple zero-profit threshold decision algorithm derived from our analytical model. Figure 1 shows the main user interfaces of the prototype.



a. The risk assessment tool for Spot market

b. User interface via one-demand fixed pricing

c. User interface via spot pricing

d. User interface via flexi SLA-based pricing

Figure 1. User Interfaces of the SmarterCloud Prototype Website (www.smarter-cloud.biz)

This prototype is currently employed as an experimental platform that explores cloud computing users' willingness-to-pay for SLA-based service provisioning (Shang et al. 2012b). For WITS 2012, we will demonstrate its functionalities and robustness, and explore interests from the WITS community for collaboration opportunities on the use of this prototype.

Shang, D., Huang, J., Yang, Y. and Kauffman, R. J. (2012a). "Analyzing Customized Service Level Agreements for the Cloud Computing Spot Market", *Working Paper.*

Shang, D., Yang, Y., and Huang, J. and Kauffman, R. J. (2012b). "Exploring Spot Users' Willingness to Pay for SLA-based Service Provision in Cloud Computing", In *Proceedings of 11th Workshop on e-Business (WEB 2012)*, Orlando, Florida, December 15, 2012. *Forthcoming*.

Buy it now or later: the Impact of *Mari* on Multi-unit Sequential Dutch Auctions

Yixin Lu, Paul van Iterson, Alok Gupta*, Wolfgang Ketter, Jan van Dalen, Eric van Heck

Rotterdam School of Management, Netherlands *Carlson School of Management, United States

1 Introduction

While many conventional auction settings have been studied in detail in economic literature, various modifications made in the real-world operating environment often change the incentive structure of different parties participating in the auctions and thus create challenges for the market maker.

In this research, we investigate the performance of two auction formats which are widely used for selling multiple homogeneous perishable goods (e.g., flower, fish, tobacco), the traditional multi-unit sequential Dutch auction (MSDA) mechanism and a variation of MSDA which involves an additional stage called *Mari* where non-winning participants are allowed to purchase goods at the same price as the winner paid in the previous round.

According to Kitahara and Ogawa (2006), *Mari* can substantially speed up the market clearing procedure at the cost of sufficiently small loss of allocative efficiency. However, such findings are purely theoretical and not validated in any empirical or experimental setting. In view of the absence of appropriate empirical data, we conduct a realistic laboratory experiment to examine the impact of *Mari* in terms of throughput and allocative efficiency.

2 Experiment Overview

Our experiment design manipulates two factors: the auction mechanism, which is either the traditional MSDA or the variation which involves *Mari* stage, and bidder's demand (single-unit demand vs. multi-unit demand). In each auction eight bidders compete for a given amount of artificial commodity, and bidders' value of the commodity are randomly drawn from a uniform distribution. Figure 1 provides an overview of the software interface[1].

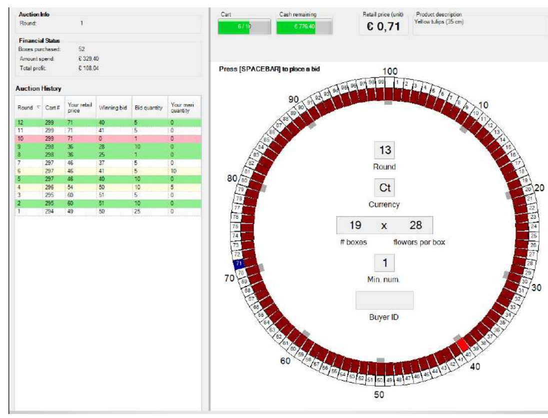


Figure 1: Auction Screen

3 Expected Contribution

The results and findings will contribute to both theory and practice of auction design by shedding lights on the broader impact of *Mari* on auction outcome in terms of throughput and allocative efficiency.

References

Kitahara, M. and R. Ogawa (2006). Efficiency versus economy of time in multi-unit descending auction: The role of “mari” at flower markets in japan. ISER Discussion Paper 0774, Institute of Social and Economic Research, Osaka University.

[1] For detailed information regarding the operation of the auction software, please check the demo at `http://www.youtube.com/watch?v=4DZldR0oS2k` and `http://www.youtube.com/watch?v=IwcMh5U2gXo`.

Increasing Social Welfare and Individual Savings using Economic Incentives for Electric Vehicles

Konstantina Valogianni§ Wolfgang Ketter § Mathijs de Weerdt* John Collins†

† University of Minnesota, Computer Science Department
§ RSM Erasmus University, Department of Decision and Information Sciences, The Netherlands
∗ Delft University of Technology, The Netherlands

Introduction

Electric Vehicles (EVs) comprise a valuable tool towards a sustainable solution for coping with high volatility and uncertainty in the Smart Grid, since they have storage features. According to IEA (International Energy Agency), 95% of the transportation is dependent on oil and by 2050 the number of cars will be tripled. Therefore the use of EVs becomes a necessity for reducing energy prices, making use of their batteries. Energy Informatics [2], plays major role in this effort. We identify the information needed for customer's consumption and consequently energy prices reduction with the use of EVs. The unconstrained use of EVs, will lead to high price peaks during the charging time. On the other hand, if the penetration of EVs is planned appropriately, the prices can be reduced, making use of the storage feature of EV. We model the customer's driving behavior and household consumption, taking into account the heterogeneity in driving attitude among different social groups. Furthermore, we design targeted incentives, which according to customer's risk attitude ensure individual savings, peak and average price reduction.

Simulation Environment

As testbed we use Power TAC, which is an agent based simulation testbed which allows for realistic representation and large scale experimentation [1]. Power TAC's environment leads the brokers to develop strategies that will be able to cope with the challenge combination of maximizing profit and maintaining a balanced portfolio. The competitive character of Power TAC, makes it a valuable research tool, as every year top research groups compete and test their strategies. This allows for building and verifying computationally-tractable customer models of large, diverse populations. We propose a customer model for the EV owners which reflects their actual behavior. For the precise creation of the customers' driving profiles we use mobility data from the Dutch Statistics Office(CBS)[1]. The population is divided according to sex and the social groups that comprise the total population. Those social groups with their special characteristics are: people younger than 15 years, part-time employees, full-time employees, students and pupils, unemployed, disabled and retired persons.For each group there are different activities accompanied with the kilometers needed per day for each activity.

Results

We present an EV customer model that reflects the reality with high precision, as it models every particular daily activity of each social group and the exact kilometers needed for each activity. As a second step we express risk attitude of the customers in terms of percentage of the nominal battery capacity that they tend to charge. Finally, we prove that the presented targeted incentives improve the social welfare, ensuring at the same time bill savings for each individual. In this demo we will show a simulation of Power TAC and how the customers interact with the brokers in the decentralized Smart Grid (http://powertac.org/).

Figure 1: Power Trading Agent Competition.

References

[1] Wolfgang Ketter, John Collins, Prashant Reddy, and Mathijs de Weerdt. The 2012 power trading agent competition. Technical Report ERS-2012-010-LIS, RSM Erasmus University, Rotterdam, The Netherlands, 2012.

[2] Richard T. Watson, Marie-Claude Boudreau, and Adela J. Chen. Information systems and environmentally sustainable development: energy informatics and new directions for the IS community. *Management Information Systems Quarterly*, 34(1):4, 2010.

[1] www.cbs.nl