Innovating knowledge communities

An analysis of group collaboration and competition in science and technology

S. Phineas Upham · Lori Rosenkopf · Lyle H. Ungar

Abstract   A useful level of analysis for the study of innovation may be what we call "knowledge communities"—intellectually cohesive, organic inter-organizational forms. Formal organizations like firms are excellent at promoting cooperation, but knowledge communities are superior at fostering collaboration—the most important process in innovation. Rather than focusing on what encourages performance in formal organizations, we study what characteristics encourage aggregate superior performance in informal knowledge communities in computer science. Specifically, we explore the way knowledge communities both draw on past knowledge, as seen in citations, and use rhetoric, as found in writing, to seek a basis for differential success. We find that when using knowledge successful knowledge communities draw from a broad range of sources and are extremely flexible in changing and adapting. In marked contrast, when using rhetoric successful knowledge communities tend to use very similar vocabularies and language that does not move or adapt over time and is not unique or esoteric compared to the vocabulary of other communities. A better understanding of how inter-organizational collaborative network structures encourage innovation is important to understanding what drives innovation and how to promote it.

Keywords   Knowledge communities · Innovation · Dynamic clustering

S. P. Upham (✉) · L. Rosenkopf
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6370, USA
e-mail: uphams@wharton.upenn.edu; phin.upham.wg05@wharton.upenn.edu

L. Rosenkopf
e-mail: rosenkopf@wharton.upenn.edu

L. H. Ungar
CIS Department, University of Pennsylvania, Philadelphia, PA 19104-6309, USA
e-mail: ungar@cis.upenn.edu

Introduction

Technical and scientific innovation is accomplished through a joint effort of thousands of researchers working for different kinds of organizations, including firms, universities, hospitals, research think tanks, and government labs (Fleming and Sorenson 2001; McCain 1987; Osareh 1996; Small 2003). In some cases the work of a group of researchers across firms assumes a joint coherency, and the group begins to function, in many ways, as if it were a new emergent organizational form (Crane 1972; Kuhn 1962). These scientific "knowledge communities" comprise an inter-organizational large-scale network in which researchers work together by building on each other's advances. Researchers in knowledge communities tend to produce at higher rates of innovation than less cohesive researchers (Boyack and Borner 2003; Merton 1972; Murray and Stern 2005; Narin et al. 1997). In the context we study in this paper, for example—technical computer science paper publications—a disproportionate 56.61% of citations are received by papers in cohesive knowledge communities, even though only 43.67% of papers are in such communities.

Using citation and rhetorical data to look at the network typology of these innovating communities, we begin to isolate key structural factors that drive their success. We attempt to quantify some of the substantive differences between knowledge communities in both use of previous knowledge and use of rhetoric. First we examine community cohesiveness, or the extent to which communities build on each other's knowledge and language. Next we examine community uniqueness, or the extent to which a knowledge community is different from others in its use of past knowledge and language. Finally we examine community flexibility, or the rate of change a knowledge community has shown over time in use of knowledge and language.

Modern analysis in innovation has largely focused on the study of formal organizations, usually firms. But how researchers communicate and learn *between* these organizations is where much of the value is generated (Crane 1972, 1989; Fleming and Sorenson 2001; Kuhn 1962; Small 2003). Knowledge communities can exist in any specialized areas of research where there is free exchange of information; examples include looking for a cure for Alzheimer's disease, trying to improve Internet search, and looking to improve the gas efficiency of diesel engines (Culnan 1986; Guimera et al. 2005; Hargens 2000; Small 1994). Knowledge communities are by their very nature homogenous, unifying people of similar research interests and specialties to learn from each other and build on each other's ideas each for their own—though similar—purposes (Merton 1972). Members of knowledge communities can work together as closely as the members of most firms, yet may never meet (Crane 1989).

While firms are excellent at facilitating cooperation by unifying incentives, knowledge communities are superb at encouraging collaboration; this embodies the distinction between working together and working jointly. An interactive knowledge community offers a built-in and knowledgeable audience for research, a stimulating intellectual dialogue, and an accelerated technical environment (Crane 1972, 1989; Kuhn 1962). Being a member of a knowledge community is less a conscious choice than a reflection of being a part of a stream of the knowledge that involves the collaboration and cooperation of other researchers in a tight, cohesive pattern of research.

Theory

We draw on two emerging areas of network research that have examined innovation in large-scale networks, though with different emphases: research on small worlds, and research on geographic technology clusters.

In his work on "Small Worlds" (Watts 1999), Duncan Watts observed that the aggregate structure of connections between people and other networks was not random. These ideas have begun to allow network theorists to effectively grapple with large-scale networks of all sorts, including, for example, the small world of the casts of Broadway musicals (Guimera et al. 2005; Uzzi 2005). This research shows the benefits behind cohesive communities for research engaged in collaborative innovation. Michael Porter and others have argued for a different logic behind such communities of expertise. They study geographic clusters of competence, showing how a bubble of geographically close, dense connections, high expert knowledge, and unified interests has led to sustained advantages in innovation and core competencies (Porter 1998; Porter et al. 2004). This strain of thought assumes that communities are geographical, like Silicon Valley in the 1990s—and that it is the social reinforcement, and to some sense external dislocation, of people into the "bubble" that allows for such effective innovation in business.

Both the small world and Porter's lines of research integrate well with the arguments developed within the sociology of knowledge since the 1960s. Merton (1972), Kuhn (1962), and Price (1963) looked at the realm of science and research and argued that paradigms—lenses for viewing the world, frameworks of meaning—created tight clusters of researchers who built on each other's ideas. The stronger the paradigm, the more intellectually coordinated the cluster could be, since it had more clearly defined questions, methodologies, and language (Pfeffer 1993; Yoels 1974). On the other hand, maladjusted and restrictively strong paradigms can lead to myopia, since higher commitment to a framework more strongly excludes other frameworks (Meyer and Zucker 1989; Pfeffer 1993).

Research within the scientific academy and within firms, traditionally studied separately, has begun to be more integrated as university researchers seek more patents, academic and corporate researchers fruitfully collaborate, and firms such as Google emerge from the ivory tower (Henderson et al. 1998; Murray 2005; Murray and Stern 2005; Narin et al. 1997; Shane 2002). Nevertheless, fundamental and proprietary research settings rest on substantively distinct incentive structures (Gittelman and Kogut 2003; Murray and Stern 2005). Fundamental science encourages the broad distribution and sharing of knowledge, since the success of university researchers is bound up with the attention of and usefulness to their fellow researchers. The importance of research within firms, while it builds on fundamental science (and sometimes contributes to it), rests on the ability to extract value from proprietary, and thus private, knowledge.

Search

The search strategies used in research have been found to be very important to the resulting nature and degree of innovation. James March (1991) described two strategies firms could use when searching for new ideas or technologies. An explorative search strategy focuses on searching through unfamiliar material in the hopes of a breakthrough that would expand an area of specialty. An exploitive search strategy looks at areas of existing competence to find ways to better use or extract value from current resources. Studies that have built on the importance of search scope in innovations include Nerkar (2003), Gittelman (2003), Rosenkopf and Nerkar (2001), and Katila and Ahuja (2002).

In measuring innovative performance, this work traditionally focuses on patents, the firms' "claim" on a technology (Cohen et al. 2000; Pakes and Shankerman 1984) and its

incentive to invest in R&D (Trajtenberg 1990). Although they are publicly available, patents are not filed to disseminate knowledge. In contrast, the knowledge communities we study in this paper work primarily by publishing papers on technical subjects in journals or presenting papers in conferences (Garfield 1983, 1988). Citations in papers, like citations in patents, are made when another work had been influential to a discovery, or when a discovery is based on another work (Merton 1965).

Marketing

The well-known theory of "integrated marketing" holds that organizations which present a unified and clear message to customers are more likely to perform well (Phelps and Johnson 1996; Schultz et al. 1994). Consistency and dependability in marketing messages is the essence of brand identity, one of the most robust findings in the marketing literature (Haynes et al. 1999).

A strongly focused, sharing collaborative group culture can help gather and triangulate information about a market to maintain a better window into its needs and preferences (Hurley and Hult 1998; Pfeffer 1993; Powell et al. 1996; Slater and Narver 1995). Such a group has several interlocking abilities that allow it to communicate a message to the market in a unified, coherent way (Slater and Narver 1995). The abilities to market products and ideas are not so different, as argued by literature on "science wars" and discussions on how ideas propagate as "memes" (Kogut and Macpherson 2004). Strategies for selling ideas can be in many ways analogous to strategies for selling tangible products (Downs 1957; Hotelling 1929).

Knowledge communities are interesting in this respect because the "market" for a knowledge community is perhaps primarily internal to those within the community and then secondarily—but necessarily—to those beyond the community. This increases the value of a shared internal form of presentation and uniform norms for use of language and methods (McCloskey 1998).

Hypotheses

We examine two aspects of knowledge communities: how they use and build on previous knowledge, and how they use language (Abrahamson 1996; Braam 1991a, b).

We believe that the way a knowledge community uses previous knowledge will be closely related to the way it has searched its relevant knowledge-space for new ideas (Katila and Ahuja 2002; Nerkar 2003; Rosenkopf and Nerkar 2001). On the other hand, we believe the way a knowledge community uses language will be more closely related to how it communicates internally and externally—how it markets its ideas to attract attention and members (Abrahamson 1996; Powell et al. 1996; Slater and Narver 1995). In explaining differential success, we believe that the existence and stability of community attributes, and their low appropriability, is best explained under a broad resource-based view framework that takes into account interconnected capabilities, norms, and routines.

Previous research has shown how important to innovation it is how a knowledge community searches for new ideas (Katila 2002; Katila and Ahuja 2002; Rosenkopf and Nerkar 2001). Literature on search strategies implies that broad search will be more likely to result in innovation (Cohen and Levinthal 1990; Fleming 2001; Fleming et al. 2005; March 1991). Therefore, we believe that a firm which looks broadly in its intellectual landscape will be more likely to identify valuable sources of ideas.

H1 Innovating knowledge communities that draw from diverse sources of knowledge will perform better.

Broad search, however, will only be useful to the extent that a knowledge community can exploit the good ideas it finds (March 1991). In quickly moving areas of science, therefore, this ability to draw from sources with valuable ideas will signify itself in a knowledge community's ability to redirect its focus as an intellectual community as its research evolves and new areas of knowledge become more useful (Fleming 2001; Levinthal and March 1981).

The broad range of knowledge collected within a tight and interconnected knowledge community, as hypothesized in H1, allows such a community to scan the intellectual and technical landscape quickly and efficiently (Uzzi 2005; Watts 1999). The ability to search broadly within a tight community and beyond it is crucial to finding valuable resources, and the ability to then reposition to take advantage of these resources is also crucial (Fleming and Sorenson 2001; Rosenkopf and Nerkar 2001).

H2 Innovating knowledge communities that are flexible in their use of knowledge will perform better.

It is important for a firm to search and take advantage not only of standard resources but of non-standard ones (Katila 2003 draft; Katila and Ahuja 2002). Particularly where academia and industry mix, the emphasis is on ideas that contribute something innovative (Gittelman and Kogut 2003; Henderson et al. 1998; Murray and Stern 2005; Pakes and Shankerman 1984).

Therefore, a knowledge community must search in areas that are somewhat unique (Cohen and Levinthal 1990; Drucker 1985; Katila 2003 draft; Schumpeter 1934). Ceteris paribus, the more competitors searching in an area, the less likely it is that undiscovered value can be found there (Porter 1985). Further, it follows that a firm which searches broadly (H1) and is flexible (H2) will more likely be among the first to find valuable untapped resources (Lieberman and Montgomery 1988; Makadok 1998).

H3 Knowledge communities that use unique knowledge will perform better.

Now we turn from the way knowledge communities use knowledge to the way they use, in aggregate, rhetoric. Knowledge communities are largely text-based—that is, the primary form of communication between members is through articles. Knowledge communities have a particularly difficult signaling problem: Members need to identify each other in order to learn from each other and benefit from the advantages of community membership, but such identification is not always easy (Braam 1991a; White 2003). Since articles do not contain explicit school identity labels (though journals can give hints toward potential identities), authors who wish to position their articles within a school may use keywords to identify themselves as a part of that school or may use language specific to that school (Abrahamson 1996; McCloskey 1998).

From a marketing perspective, we expect rhetoric to be used in a knowledge community very differently from the way a community uses previous knowledge. Members of strong knowledge communities will, we believe, for internal and external reasons, tend to converge on shared language both as a way to reduce the ambiguity of communication and because such communities use similar methodologies and similar language in presenting problems and issues (Abrahamson 1996; Bartel and Saavedra 2000; DiMaggio and Powell 1983; McCloskey 1998).

From an external perspective, literature on integrated marketing and branding suggests that firms which use a single consistent message will be more effective (Schultz et al.

1994). This should hold even more strongly for knowledge communities, where the lack of concrete labels causes authors to flag themselves as part of a school as a signaling message (Mizruchi and Fein 1999; Pfeffer 1993).

Lastly, the very diversity of knowledge being presented by knowledge communities gives rise to a need for a single rhetorical lens with which to express these ideas. Successful knowledge communities, we believe, will draw on many sources for their knowledge and then present these diverse ideas under one shared rhetoric or framework of analysis.

H4 Innovating knowledge communities that use consistent rhetoric will perform better.

But the lack of formal leadership that makes knowledge communities so different from firms also limits their ability to coordinate change. The rhetoric of a strong paradigm tends to be stable over time, as it is based on a shared construction of rhetorical meaning (Price 1963). Reconceptualizations of shared mental constructions and changes in meaning of shared language, particularly in a context without any explicit leadership or coordinating mechanism, can result in inefficiency and ambiguity in communication.

The advantages of shifting rhetoric over time are not as clear. Buzzwords and rhetorical fashion are not a good basis for a sustained competitive advantage, particularly in a field that relies on technical work (Abrahamson 1996). Since knowledge communities are amorphous groups without formal boundaries, they benefit from a unified use of language in order to create a recognizable sort of identity or brand (Haynes et al. 1999). Particularly in a technical context, language is not the source of innovation; rather, the ideas underlying language are. Therefore, once a knowledge community constructs a consensus in rhetoric, it remains relatively stable.

H5 Innovating knowledge communities that use stable rhetoric will perform better.

Finally, knowledge communities, if they are to be successful, must appeal to large numbers of people. This implies that schools claim and use common and recognized language as their core set of words, allowing them to grab the metaphorical "middle ground" (Downs 1957; Hotelling 1929).

Knowledge communities whose words and meanings are difficult for others to understand tend to isolate themselves. While this has proven to be a successful strategy for professional fields that desire barriers to entry (e.g., doctors, lawyers), it hampers the cooperation and collaboration that is desirable between knowledge communities (Kripke 1982). Addressing the important issues of the field and communicating good ideas in broadly understood language will lead to the largest possible audience (Downs 1957).

H6 Innovating knowledge communities that use mainstream rhetoric will perform better.

Although knowledge communities do not have formal leaders or formal organizational legitimacy, they do have the preconditions for idiosyncratic capabilities and routines—intense socialization, repeated interactions, mechanisms for punishment and reward, an interconnected incentive structure (Barney et al. 2001; Cohen et al. 2000; Wernerfelt 1984). These differential routines and capabilities could constitute a competitive advantage for communities and lead some communities to have sustained levels of higher innovativeness. Understanding how and why knowledge communities create a culture of innovation that effectively generates and distributes good ideas is our central goal.

The effect of firm involvement in knowledge communities is an interesting one. While it is not a main variable, we include it in order to explore its impact on all our models. Firms place a higher value on proprietary information, while knowledge communities often thrive

on very open sharing of information, at least in the sciences and social sciences. The difference this makes is hard to fully appreciate until our analysis extends to firms or commercial patents.

Methods

Data

The data in this paper were drawn largely from CiteSeer, a digital library of papers from conferences and journals in computer science. We cross-reference all of CiteSeer's more than 700,000 indexed papers with the DBLP Computer Science Bibliography, a European database with over 600,000 papers that indexes a similar group of computer science papers, in order to verify existing information and gather supplemental information on journals and conferences. The match between these two databases is close; indeed, the DBLP links most of its papers to the corresponding papers in the CiteSeer database.

The majority of papers in these databases are from between 1992 and 2003. They give us a rich picture of the field of computer science as it has evolved in the past decade and a half. Some limitations, such demarcation of paper type, remain.

Clustering methodology

To find our knowledge communities we use a clustering algorithm to identify clusters of like papers. Our methodology utilizes the structure of co-citations in paper bibliographies sometimes called bibliometric coupling to group papers that are "similar" in the papers they cite, representing the similar knowledge they are building on (Small et al. 1985). Essentially, therefore, we are comparing the citations of all papers to all other papers to find papers that use similar citations. See Appendix 2 (available online) for technical details and pseudocode on clustering methodology.

When previous research in management and innovation has used algorithms for clustering, two methods have been overwhelmingly prevalent: CONCOR and hierarchical clustering. These methods work best in smaller datasets with low dimensionality and clear cluster divisions. Their chief weakness, as with other ad hoc single pass methods in management (Aharonson et al. 2004), is that their initial clustering choices result in path dependency.

We built on prior clustering methodologies to develop a new clustering approach, StrEMer, that produces high-quality clusters that are dynamic over time and allows papers that are not tightly connected to a cohesive set of research to remain out of a cluster (Kandylas et al. 2005; Zhong and Ghosh 2003). All standard clustering methods create clusters based on citations made both by a paper and by other similar papers published subsequently, conflating the importance of a cluster in the future with its position now. We developed an iterative clustering scheme ("rolling clustering") that successfully resolves this temporal confounding, allowing all cluster assignments in each year to be backward-looking only based on the previous 5-year frame—an appropriate "context" for knowledge development. At the same time we find very high continuity between clusters, since the knowledge landscape we created changes gradually. Another benefit of this method is that our measures of "centrality" at the paper and cluster levels refer to the appropriate frame rather than an aggregate over the entire time range, as with all other standard methods.

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
1	9.23%	11.85%	10.76%	9.14%	7.06%	7.23%	7.42%	7.34%	6.58%	5.15%	4.57%	4.24%
2	15.85%	10.55%	11.93%	10.44%	7.52%	7.61%	7.70%	8.16%	8.44%	8.36%	7.58%	8.95%
3	7.84%	7.05%	8.38%	6.05%	6.06%	6.73%	5.91%	7.21%	6.22%	5.86%	4.48%	4.91%
4	7.49%	9.09%	5.46%	6.22%	5.84%	4.67%	5.25%	4.23%	1.85%			
5	4.01%	3.42%	7.65%	7.28%	8.92%	9.65%	9.19%	6.50%				
6	1.92%	2.55%	3.63%	3.59%	6.13%	6.33%	5.56%	5.58%	5.78%	6.02%	4.71%	3.70%
7	9.06%	10.76%	5.80%	4.66%	4.08%	3.30%	3.16%	3.29%	2.67%	3.18%	3.04%	2.89%
8	3.31%	3.71%	6.76%	7.21%	8.04%	8.25%	6.97%	6.58%	7.45%	6.63%	4.31%	5.32%
9	17.94%	17.02%	10.01%	8.58%	10.02%	8.02%	7.22%	8.11%	5.67%	7.33%	9.23%	5.85%
10	4.18%	6.91%	8.45%	10.32%	10.23%	9.37%	9.42%	10.29%	8.35%	8.03%	7.51%	5.11%
11	0.17%	0.36%	0.46%	2.15%	4.98%	5.77%	6.15%	6.68%	4.77%	5.85%	6.48%	6.53%
12			3.59%	5.57%	5.86%	7.82%	7.71%	5.30%	4.40%	4.02%	2.44%	
13			2.67%	5.35%	5.02%	5.86%	7.52%	9.59%	13.57%	12.08%	10.76%	10.90%
14	14.29%	12.36%	9.02%	8.76%	6.36%	3.91%	3.92%	1.82%	1.41%			
15	4.36%	2.18%	3.40%	2.45%	1.28%	1.18%	1.03%	0.35%				
16					0.04%	0.29%	0.50%	0.95%	3.91%	5.42%	7.08%	8.55%
17					0.29%	0.79%	1.23%	2.21%	3.97%	5.61%	6.87%	7.60%
18					0.23%	0.45%	0.90%	1.29%	4.32%	5.53%	5.96%	7.20%
19	0.35%	2.18%	2.01%	2.23%	0.51%	0.31%	0.28%	0.12%				
20					1.56%	2.46%	2.25%	2.77%	4.54%	4.56%	6.16%	7.74%
21							0.71%	1.63%	6.11%	6.35%	8.83%	10.50%

Additional Information:
1993-1994: Split of cluster 14 into cluster 14(40%) and 12(23%)
1995-1996: Split of cluster 9 into cluster 9(38%) and 17(16%)
19999-2000: Merge of cluster 6 and 15 into cluster 6
2000-2001: Merge of cluster 12 and 14 into cluster 12 (see previous split)

Fig. 1 Cluster size as percent of total papers in knowledge communities over time

Innovating communities

The evolution of clusters over time can be seen as a result of the choices agents make as they, in aggregate, position themselves on the intellectual landscape. Figure 1 represents the dramatic evolution of the computer science landscape from 1992 to 2003. Appendix 1 details our proposed names for each knowledge community and also provides some details (top three most cited papers, top keywords, etc.).[1]

In 1992 we observe 14 knowledge communities. Between 1992 and 1999 seven new knowledge communities formed and none disappeared. This finding is in keeping with the dramatic growth of computer science in the Internet boom in California's Silicon Valley and around the world. In these years the Nasdaq index, disproportionately heavy with technology and Internet stocks, rose from approximately 600 to 4,000.

From 1999 to 2001 five knowledge communities disappeared and none were created. These broad cluster trends are in keeping with the collapse of the Internet bubble and the fall of the Nasdaq from 1999 to 2001 to almost 2,000. The movement and rates of change of clusters also reflect these changes, with more activity during times of shakeup in 2000–2001. The general trends in our data examining computer science knowledge communities closely track changes in the financial sector.

We see the emergence of a number of clusters that were not present at the start of our study. In 1996 a new cluster emerged on "Internet search." One of its top three most cited papers is by Larry Page and Sergey Brin, the founders of Google, who built their first search engine in 1996, founded a company in 1998, and went on to have a

---

[1] Appendices with other details on the 21 knowledge communities identified are available from the author.

Table 1 Summary statistics for clusters

Cluster #	Total papers	Total cites	Cite/ Paper	InClustBib (%)	CAGR
1	7,892	50,924	6.45	44.29	−6.83
2	9,368	70,414	7.52	41.04	−5.06
3	7,022	46,199	6.58	42.92	−4.16
4	4,144	39,962	9.64	28.39	n/a
5	6,053	50,990	8.42	22.74	n/a
6	6,070	39,854	6.57	37.53	6.17
7	3,968	32,076	8.08	38.23	−9.85
8	7,743	47,861	6.18	38.94	0.40
9	9,002	89,961	9.99	27.49	−9.68
10	10,180	86,158	8.46	40.72	1.85
11	5,890	52,377	8.89	28.85	39.01
12	5,990	32,899	5.49	43.11	−13.59
13	9,566	52,569	5.50	41.58	13.82
14	3,818	39,589	10.37	26.85	n/a
15	950	7,050	7.42	14.18	n/a
16	2,297	12,239	5.33	24.03	76.57
17	2,743	15,890	5.79	23.37	44.07
18	2,428	12,307	5.07	29.32	51.74
19	472	3,886	8.23	10.24	n/a
20	3,289	26,104	7.94	32.88	28.04
21	3,043	11,185	3.68	18.51	11.44

Variables represent totals for clusters from 1992 to 2003 or from emergence to disappearance from data

Cite/Paper is the average citations received by a paper in that cluster

InClustBib is calculated as the % of citations made by a cluster to papers also in that cluster

CAGR of clusters that emerge during data range are calculated beginning their second year in data

multi-billion-dollar IPO in 2004. In 1996 the knowledge community representing "Internet search" comprised only 0.23% of our computer science papers; in 2003 it represents 7.20%. This leads to a CAGR of 51.74% from 1998, when Google was founded, to 2003—the second highest of all clusters after only cluster 16 representing the related topic of "congestion control," which also emerged in 1996 (Table 1).

Variables

Performance variables

Our dependent variable for cluster is a measure of the "vigor" or performance of a cluster at a given time. To model this we use the number of papers presented at conferences or published in computer science journals in a cluster from 1992 to 2003, controlling for the number of papers published the year before. Thus, effectively, we are measuring the performance of an intellectual community controlling for the prior year's number of papers it published. To generate this variable we aggregate number of papers published by each cluster into year-long time periods, consistent with a broad set of network papers in

technology and strategy (Rosenkopf and Nerkar 2001). This provides us with a summary measure of how many "members" an intellectual community was able to both attract and promote. To simplify the relative scale of our data we divide the number of papers by 1,000 in our regressions.

Cohesiveness

We are interested in seeing if the intellectual "cohesiveness" or "overlap" of both the shared rhetoric (words) and shared knowledge (papers cited) of the knowledge community are significant for predicting its performance. Our goal is to find a measure of how paradigmatic knowledge and language are within a school of thought.

For the Knowledge Cohesiveness variable we represent how widely the cluster as a whole searched for knowledge during that year in the intellectual landscape vs. how focused (coordinated) that search was. This is done by computing the average similarity between the citations structure of each paper and the overall citations structure of the cluster, where similarity is as defined in the previous clustering section. We constructed a similar variable for rhetoric in the cluster by taking the title and keywords of each paper and, as is common, removing "stop" words such as "and," "if," and "by" and then "stemming" them so that, for example, *Learning and Artificial Intelligence* becomes *Learn Artifici Intellig*. Keywords are selected by the author and included in the journal to identify the distinctive research focus of the paper (Abrahamson 1996; McCloskey 1998). We then construct the Rhetorical Cohesiveness variable as we constructed Knowledge Cohesiveness above. This is a proxy for how similarly people in a cluster use language.

The average of these for each cluster in each year is a measure of the extent to which a cluster's use of rhetoric is narrow or disperse and the extent to which a cluster's use of knowledge is focused or expansive. This is measured as:

$$\frac{\sum_{i=1}^{n_C} \mathrm{sim}(e_i, \mathrm{cen}_C)}{n_C}$$

where C represents a cluster for a given year; i indexes papers in cluster C, and n_C is the number of papers in cluster C; and $\mathrm{sim}(\cdot)$ is the measure of similarity as previously defined in the clustering methodology. These processes yield measures, relative to other clusters, of how disperse or focused the use of knowledge and the use of rhetoric are for this cluster in this year.

Uniqueness

We are also interested in how different an intellectual community is, either in the knowledge it generates or in the rhetoric it uses, from other intellectual communities. Uniqueness of Rhetoric represents how unique the rhetoric of a school of thought is at a given point in time compared to other clusters. Uniqueness of Knowledge is a measure that represents how unique the sources of knowledge of a school of thought are at a given point in time. The variable is computed in the same way as Uniqueness of Rhetoric, using citation structure rather than words.

In this calculation we focus on the average citation or rhetoric for a cluster, and compare it to all other clusters' average citations or rhetoric. For example, if a cluster generally uses the same keywords or cites the same papers, the average for this cluster will be small.

Rate of change

Adaptability or rate of change for a cluster is an important measure of how flexible a cluster is over time. We assume that over time in a changing environment, flexible clusters move more than less flexible clusters. Since knowledge changes as a function of other knowledge, we use a relative measure of change in constructing this variable. Given our averages or centroids for citation structure and language per cluster, we construct a cosine similarity between each cluster and itself in the previous year. The difference between the cluster average from year t to $t + 1$ is a measure of the "rate of change" of a cluster over time. To smooth out this number over time we take the 3-year running average of this change as our variable of interest. However, a cluster's average rates of change over 1, 2, 5, and all years were comparable. The change in rhetoric represents how much the words a cluster uses change (operationalized as described previously) from one year to the next; the change in knowledge represents how much a cluster's average use of citations (or the knowledge sources it draws from) changes from one year to the next. It is defined as:

$$\frac{\sum_{t=0}^{T} \text{sim}\left(\text{cen}_{T-t}, \text{cen}_{T-(t+1)}\right)}{T}$$

where t indexes the years considered in the formula; T is the span of years we consider; and sim( ) is the measure of similarity as previously defined.

Leadership/coordination controls

A common way to explain differential performance in firms is to look at the level of leadership or coordination. We test for such effects on three levels—from members of the knowledge community, for concentration of the institutions the members identify with, and for concentration in distribution in the community (Porter 1998).

While an intellectual community does not have CEOs or boards, influential members can act as intellectual leaders (Pfeffer 1993). We identify influence ties between authors of papers and the authors of those papers that they cite and thereby construct an influence network for each cluster. We then run centrality measures on these networks to measure the clusters' eigenvector, degree, and in-degree centrality. We found eigenvector to be the most useful measure of centrality, since it measures both direct and indirect influence, though all measures led to similar results (see robustness, Table 4).

We also wish to control for the potential coordinating influence of institutions. For example, a school such as MIT or a company such as Google might be home to a significant number of members of a school of thought, and thus the formal control, social network, institutional norms, and institutional organization these institutions exhibit may contribute to the de facto coordination of the school of thought. To construct a variable to measure this we first identified the institutions that the authors in the database identified within their papers. We then found the percent of papers for each cluster that came from the most common 10 schools, research institutions, or companies to see if a cluster had concentrated influences by a few institutions. We also measured the concentration of the top 1, 2, 3, and 5 institutions and received similar results. We assume that a higher concentration of control by a few players in a knowledge cluster increases the potential for cluster coordination.

Lastly, we feel a coordinating or leadership role might be played if an intellectual community is dominated by a powerful institution that controls distribution for that

cluster—such as a journal or conference that acts as a gatekeeper for the community. In this case we look at the percent of articles published in the 10 most common venues of the authors (either journals or conferences). We also measured the concentration of the top 1, 2, 3, and 5 institutions and received similar results.

Prestige Controls

Prestige is a powerful factor in explaining differential performance in organizations; we wish to test whether this also holds true of knowledge communities. As with leadership, we therefore control for prestige on the member, journal/conference, and employer/university levels of analysis. For members we wish to control for the prestige that would result from the "top" members of a field preferring to publish in some intellectual communities, leading to superior performance. We constructed this variable by finding the authors who had been nominated to the prestigious post of fellow by the three top societies in computer science—the Institute for Electrical and Electronics Engineers, the Association of Computing Machinery, and the National Academy of Engineering—from 1975 to 2005 and counting the number of these fellows who published in any of our intellectual communities by cluster and year. Next we constructed a variable that counted the number of papers coming from the most prestigious 20 universities in computer science as ranked by the *US News and World Report* graduate school rankings of academic programs. By doing this we help control for the tendency for some intellectual communities to be associated with prestigious institutions. Lastly, we constructed a variable that counted the number of papers published in the top 10 most prestigious journals as ranked by impact factor in Thomson ISI's Impact Factors, which ranks the influence of journals, and the top 10 most prestigious conferences, as ranked by citation impact by DBLP. While the Impact Factor methodology has flaws when used in marginal cases, we feel it is adequate to identify the rough group of very top journals in narrowly defined fields—the top 10 list used was examined by a number of tenured professors in computer science who verified it as reasonable. We took these counts broken up by cluster and by year, and ranked them within year by cluster. This rank-ordered list of clusters by year indicated the relative prestige of knowledge communities on multiple levels.

Industry/academia controls

Each author of each paper is coded as being affiliated with a firm or academic/research institution. We then code each paper as "academic" if all of its authors are affiliated with academic/research institutions, "industry" if all of its authors have firm affiliations, and "mixed" if some of its authors are affiliated with firms and some with academic/research institutions. We entered this information into the regression by including the two categorical variables "mixed" and "industry."

Analysis

Regressions

Since our data encompasses dynamic communities and measures their characteristics over time, we use a cross-sectional time series model to gain insight into the effects of

community attributes. We estimated our models using Generalized Least Squares (GLS), including robust standard errors for determining statistical significance. This approach allows us to investigate the time trends within our data while also adjusting our standard errors for intragroup correlations, since we believe the performance measures of any cluster will be correlated over time. In addition, we evaluated a variety of plausible model estimation methods, explored further in our robustness section.

A Wooldridge test for first-order auto-correlation in panel data (Drukker 2003; Wooldridge 2002) found that, as expected, our data exhibited autocorrelation, which implies that a time lag will be required to ensure independence of residuals ($F(1,21) = 228.358$, $p < 0.0001$). After testing different lag periods for appropriateness using Akaike's Information Criterion (AIC), we include the cluster's prior year performance in the model (1-year lag) as a predictor. A 2-year time lag yielded comparable results.

Our dependent variable is a measure of the performance of the intellectual community over time. Since our dataset consists of papers published in the field of computer science from 1992 to 2003, we choose to use the total number of papers published by each community in each given calendar year as a measure of cluster vigor or "success." We include a 1-year lag in the regression as well, controlling for the prior size of the cluster on year earlier.

The empirical goal of our model is to explore the extent to which we can measure community attributes and use them to predict performance. Specifically, we consider how the community draws on past knowledge and generates persuasive rhetoric by measuring the cohesiveness and uniqueness of both. We estimate our model as follows:

$$y_{it} = \beta \cdot x_{it} + b_i \cdot z_{it} + e_{it}$$

where i indexes our clusters from 1 to n, t indexes years (time), and (1) y_{it} denotes our response variable, (2) β represents the portion of effects that is constant across clusters (the fixed effects), (3) b_i represents the portion of our effects that varies between individuals (the random effects), (4) x_{it} is the vector of our predictors, (5) z_{it} is a subset of our predictors x_{it}, and (6) e_{it} represents the error term for our model.

Our base assumption is that our response is multivariate-normally distributed: $Y \sim \text{MVN}[\beta X, V]$, where V is a block diagonal, symmetric, matrix as $V = \text{diag}[V_1, V_2,\ldots,V_n]$ with each component matrix V_i composed of two components: $V_i = \Sigma_{z_j} + T_{z_j}$.

In our chosen specification, Σ_{z_j} signifies the usual error terms arising from the random effects model, while T_{z_j} is an optional additional term that will reflect the alternative error possibilities we explore later in our robustness section. For our chosen models we have simply $V_i = \Sigma_{z_j}$.

Models

We arrive at our full model by analyzing variables systematically to examine their marginal effects as well as the end joint effects. Model 1 contains only the control variables—leadership, prestige, and the industry/academic dummy variables. In Models 2–4 we consider first our variables for community cohesiveness, then our variables for community uniqueness, and lastly our variables for community flexibility. Finally, in Model 5 we consider our full model with all community attributes simultaneously included (Tables 2, 3).

Table 2 Descriptive statistics ($n = 231$, $i = 22$ clusters, $t = 12$ years: 1992–2003)

	Correlation														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of papers	–														
Knowledge Cohesiveness	(0.16)	–													
Rhetorical Cohesiveness	0.09	0.92*	–												
Knowledge Uniqueness	0.39*	0.26*	0.41*	–											
Rhetorical Uniqueness	0.66*	0.23*	0.45*	0.63*	–										
Know. Comm. Adaptability Rate	0.71*	0.05	0.27*	0.40*	0.75*	–									
Rhet. Comm. Adaptability Rate	0.72*	0.01	0.28*	0.46*	0.80*	0.95*	–								
Journal Leadership	0.01	0.17	0.21	0.07	0.24*	0.25*	0.28*	–							
School Leadership	0.14	0.81*	0.88*	0.53*	0.57*	0.33*	0.35*	0.19	–						
Member Leadership	0.05	0.68*	0.71*	0.47*	0.53*	0.25*	0.28*	0.28*	0.80*	–					
Journal Prestige	0.01	(0.17)	(0.18)	(0.35)*	(0.03)	0.04	0.02	0.14	(0.27)*	(0.13)	–				
School Prestige	(0.55)*	(0.05)	(0.22)*	(0.57)*	(0.54)*	(0.52)*	(0.55)*	0.05	(0.41)*	(0.25)*	0.46*	–			
Member Prestige	(0.46)*	(0.06)	(0.23)*	(0.59)*	(0.54)*	(0.48)*	(0.50)*	0.05	(0.37)*	(0.19)	0.48*	0.82*	–		
Pure Industry Affiliation	0.42*	0.32*	0.46*	0.47*	0.74*	0.54*	0.60*	0.42*	0.49*	0.65*	0.05	(0.27)*	(0.26)*	–	
Mixed Affiliation	0.42*	0.19	0.36*	0.39*	0.67*	0.59*	0.64*	0.37*	0.42*	0.44*	(0.01)	(0.37)*	(0.38)*	0.64*	–
Mean	0.60	0.14	0.19	0.02	0.37	0.62	0.81	0.02	0.36	33.91	6.07	8.57	8.65	0.19	0.08
SD	0.44	0.06	0.04	0.01	0.07	0.18	0.17	0.02	0.08	9.18	4.09	5.12	5.17	0.05	0.03
Min.	0.02	0.06	0.14	0.00	0.18	0.10	0.20	0.00	0.18	20.63	1.00	1.00	1.00	–	–
Max.	2.10	0.38	0.33	0.05	0.54	0.93	0.97	0.19	0.63	62	16	21	21	0.35	0.21

* Significance at the $\alpha = 0.05$ level using the Bonferroni correction for multiple pairwise tests

Note: Numpapers divided by 1,000 to adjust scale and n for adaptability variables is 220

Table 3 GLS estimates (time series GLS estimation)—dependent variable: number of papers published by a community in a given year

	(1)	(2)	(3)	(4)	(5)
Cohesiveness					
Knowledge		(1.090)**			(1.032)**
Rhetoric		1.155*			1.169***
Uniqueness			(4.354)*		(4.040)*
Knowledge			1.536***		1.494***
Rhetoric					
Adaptability					
Knowledge				0.173	0.293*
Rhetoric				(0.004)	(0.272)*
Control variables					
Lagged response					
One year	0.669***	0.632***	0.591***	0.623***	0.557***
Leadership controls					
Journal Leadership	(4.715) *	(4.649)*	(3.323)	(4.809)*	(3.240)
School Leadership	0.316⁺	0.152	(0.249)	0.213	(0.394)
Member Leadership	(0.006)**	(0.004)⁺	(0.005)*	(0.005)*	(0.004)*
Prestige Controls					
Journal Prestige	0.006	0.006	(0.001)	0.005	(0.002)
School Prestige	(0.018)**	(0.018)**	(0.018)***	(0.017)**	(0.019)***
Member Prestige	0.004	0.004	0.010*	0.005	0.011**
Industry/academy affiliation controls					
Pure Industry Affiliation	1.528***	1.454***	0.519⁺	1.428***	0.599 *
Mixed industry/ academy affiliation	(0.340)	(0.518)	(0.823)	(0.508)	(0.858)⁺
Constant	0.176**	0.164**	0.103	0.149*	0.108
N	231.000	231.000	231.000	231.000	231.000
R^2	0.794	0.800	0.828	0.798	0.835
χ^2	1,536.359	1,689.914	2,110.959	1,682.826	2,213.746

*** $p < 0.001$;** $p < 0.01$;* $p < 0.05$; ⁺ $p < 0.10$. Standard errors in parentheses

Note: The number of papers in our regressions divided by 1,000 to simplify the relative scale

Results

Model 1 indicates that among our Leadership Controls both Journal Leadership and Member Leadership are significant and have negative effects on community performance. Only one of our Prestige Controls, School Prestige, is significant and negative.

Model 2 introduces the first of our main variables of interest, community cohesiveness. We find that cohesive rhetoric is associated with improved performance, while a broad use of knowledge maximizes performance.

Model 3 examines the second of our main variables of interest, community uniqueness. We find that a knowledge community maximizes performance when it uses rhetoric that is similar to that of other clusters, while knowledge, as represented by citations, that is gathered from diverse sources predicts superior community performance.

Model 4 uses the third pair of our chief variables of interest, community flexibility. We find that, when taken without our prior variables of interest, the flexibility of knowledge communities, in both rhetoric and citations, is not a statistically significant predictor of community performance.

Model 5 incorporates all the previously discussed variables for a simultaneous examination of their effects on knowledge community performance. We find that all main variables of interest retain the significance and direction found in Models 2–3. Additionally, our variables of interest from Model 4 are now significant predictors and collectively explain 83.5% of the variation in knowledge community performance. The variable for mixed industry/firm affiliation becomes statically significant at the $p < 0.10$ level in this model. Table 2 displays a summary of our model coefficients and their statistical significance, along with model goodness of fit summaries for all five models.

In all our models we examined the coefficients of industry affiliation. We found that throughout Models 1–5 community performance is enhanced by a high percentage of purely industry-affiliated papers, though in Model 3 this effect is only significant at the $p < 0.10$ level. On the other hand, a higher percentage of mixed industry-affiliated papers indicated a slightly negative, though statistically insignificant, impact on community performance. This indicates that the effect of higher proportions of purely academic-affiliated papers is indistinguishable from that of mixed-affiliation papers. Clusters with higher proportions of purely industry-affiliated papers were associated with higher performance than clusters with elevated proportions of either purely academic or mixed-affiliation clusters.

Since our hypotheses examine use of citations and rhetoric for the same three measures, we now examine the correlation between rhetoric and citation structures for each pair of similar variables. There is a significant, positive relationship between citation and rhetoric measures for all three knowledge community measures. For knowledge community cohesiveness, regressing the similar measures for rhetoric on citations yields an r^2 of 0.846, indicating that approximately 85% of the variation in Rhetorical Cohesiveness is attributable to changes in citation cohesiveness. Similarly, for knowledge community uniqueness the r^2 of 0.4008, approximately 40% of the variation in rhetorical uniqueness, is explained by changes in citation uniqueness. Lastly, an r^2 of 0.9004 for knowledge community flexibility indicates that about 90% of the changes in rhetorical flexibility are explained by corresponding changes in citation flexibility. It is important to note that the trends identified by these measures are consistent throughout the dataset; the extremely high correlations are chiefly due to extreme values. The large correlations for community cohesiveness and community flexibility indicate that the inclusion of both knowledge and rhetoric variables might cause problems in our regressions as a result of the "wrong sign" problem (Gugarati 1995). To address these concerns we have examined our models without the paired variables that create these issues.[2]

---

[2] To check Model 2 for this effect we kept the same controls and ran the model with both knowledge and rhetorical cohesiveness alone. Each variable retained its direction but became slightly less significant.

In Model 3 the individual inclusion of each variable sees knowledge uniqueness flip to the positive when included individually; however, it is not statistically significant. This could indicate that our joint significance reveals a secondary trend in knowledge uniqueness that is only evident after controlling for the rhetorical uniqueness of a cluster. Rhetorical uniqueness retains its significance and direction when it is included alone in Model 3.

In our investigation of Model 4 we found that Knowledge Flexibility retained its significance and direction when included individually while Rhetorical Flexibility did flip to the positive direction but without statistical significance. This individual flip explains why the original Model 4 including both variables finds neither to be significant.

The large correlations between rhetoric and knowledge suggest that the use of interaction effects could untangle the nature of the relationships between variables. Specifically, we wonder what the relationship is between the pairs of variables for cohesiveness, uniqueness, and flexibility for knowledge and rhetoric, and also about the relationships between each of these variables and cluster size. We revisited our models including the interaction effects between pairs of knowledge community variables. For uniqueness we found that interaction effects are not significant and thus have no impact on our results. For cohesiveness, on the other hand, the interaction effect was more significant than either rhetorical or Knowledge Cohesiveness variables, and the Rhetorical Cohesiveness effect no longer supports our hypothesis. Since both of our variables for cohesiveness are expressed as decimals, it is not surprising that the product of these two variables demands a larger coefficient to compensate. When the interaction effect for flexibility is added, rhetorical flexibility continues to support our original hypotheses, but citation flexibility now falls slightly short of statistical significance. These findings confirm that knowledge and rhetoric have some differential relationship, while 4 of our 6 hypotheses remain fully supported.

We also examined the interaction of prior-year cluster size with our knowledge and rhetoric variables. These interactions were largely insignificant, with the only notable exception being the interaction between rhetorical uniqueness and prior-year cluster size. This interaction suggests that the effect of rhetorical uniqueness may be related to the size of the cluster. This could be related to an intrinsic property of clusters or an artifact of variable construction.

These correlations do not negate our findings, but they do require us to keep in mind that when interpreting our regression coefficients we assume that all other variables in the model are held constant. That the use of language and the use of citations are related to each other is not surprising, since authors are citing papers they learned from. Language and source of language are inseparable but not identical—a paper, while it relies on citations, is not a function of them. In order to isolate the effects of language and knowledge sources we chose to leave out interaction effects in the main models.

Robustness

Many additional factors in our analysis have not yet been considered, and our preceding models have a number of limitations. Table 4 summarizes the results for the first three robustness checks, and Table 5 summarizes the fourth.

First, our original dependent variable is a count of papers published by a cluster in a given year. To focus on the impact of a cluster, we replace this with the aggregate number of citations received by publications from each cluster in a given year, another powerful measure of the community's success—though we believe this latter quantity emphasizes long-term impact over current performance. As seen in Model 2 in Table 4, the use of Total Citations does not affect the direction of our coefficients; however, only Knowledge Cohesiveness and rhetorical uniqueness retain their significance.

Second, the use of Total Citations above as our dependent variable affords us the unusual ability to subdivide the dependent variable into internal and external measures of impact. We divide the citations received by a cluster into two distinct variables,

---

Footnote 2 continued

  Finally, to verify our results in the final model are not unduly influenced by these we included just one of each pair and found our directions remained fairly consistent with the expected changes in significance already detailed in the earlier models.

Table 4 Robustness table

	(1) Base	(2) Totcites	(3) Endo-cites	(4) Exo-cites	(5) Degree	(6) In-Degree
Cohesiveness						
Knowledge	(1.032)**	(11.343)**	(6.745)***	(4.599)*	(1.842)***	(2.186)***
Rhetoric	0.169**	8.188	5.045*	3.143	0.167**	1.586***
Uniqueness						
Knowledge	(4.040)*	44.069$^+$	(24.015)**	(20.054)	(3.308)$^+$	(2.999)
Rhetoric	0.494***	13.931***	5.713***	8.218***	0.515***	1.297***
Flexibility						
Knowledge	0.293*	1.877	1.071$^+$	0.806	0.377**	0.294*
Rhetoric	(0.272)*	(1.360)	(0.838)$^+$	(0.522)	(0.339)**	(0.283)*
Centrality						
Member Leadership (Degree)					0.008	
Member Leadership (In-Degree)						0.015*
Control variables						
Lagged response						
One Year	0.557***	1.566**	0.403	1.163**	0.550***	0.570***
Leadership controls						
Journal Leadership	(3.240)	(35.194)	(15.535)	(19.658)	(3.140)	(3.338)
School Leadership	(0.394)	(1.215)	0.053	(1.269)	(0.590)*	(0.447)$^+$
Member Leadership (Eigenvector)	(0.004)*	(0.040)$^+$	(0.017)$^+$	(0.023)$^+$		
Prestige Controls						
Journal Prestige	(0.002)	(0.058)	(0.032)	(0.026)	(0.002)	(0.002)
School Prestige	(0.019)***	(0.152)**	(0.060)*	(0.092)*	(0.018)***	(0.016)***
Member Prestige	0.011**	0.014	0.043*	(0.029)	0.011*	0.011**
Industry/academy affiliation controls						
Pure Industry Affiliation	0.599*	5.997*	2.335$^+$	3.662	0.391	0.316
Mixed industry/ academy affiliation	(0.858)$^+$	(10.761)$^+$	(3.275)	(7.485)*	(0.917)$^+$	(0.981)*
Constant	0.108	2.651***	0.524	2.127***	0.104	0.070
N	231	231	231	231	231	231
R^2	2 ,213.746	765.696	5 08.747	819.769	2 ,191.543	2,274.363
χ^2	0.835	0.634	0.528	0.641	0.834	0.837

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $^+$ $p < 0.10$. Standard errors in parentheses

Note: Numpapers, Totcites, endo-cites and exo-cites divided by 1,000 to adjust scale

representing citations received from other papers within the community and citations received from outside the community, and examine these two quantities individually. The results of these analyses are reported in Models 3 and 4 in Table 4. While several variables lose significance, in general the directionality of the coefficients confirms our primary

Table 5 Evaluation of model specifications

	(1) RE (main model)	(2) MLE	(3) GLS—panel data	(4) GEE
Cohesiveness				
Knowledge	(1.032)**	(1.032)*	(1.053)***	(1.271)**
Rhetoric	1.169**	1.169*	0.969**	0.588**
Uniqueness				
Knowledge	(4.040)*	(4.040)[+]	(0.847)	(1.410)
Rhetoric	1.494***	1.494***	1.196***	1.279***
Flexibility				
Knowledge	0.293*	0.293*	0.281***	0.337*
Rhetoric	(0.272)*	(0.272)*	(0.249)**	(0.233)*
Control variables				
Lagged response				
One Year	0.557***	0.557***	0.478***	0.409***
Leadership Controls				
Journal Leadership	(3.240)	(3.240)***	(1.591)**	(1.522)
School Leadership	(0.394)	(0.394)[+]	(0.322)[+]	(0.400)
Member Leadership (Eigenvector)	(0.004)*	(0.004)*	(0.003)[+]	(0.005)*
Prestige Controls				
Journal Prestige	(0.002)	(0.002)	(0.000)	0.001
School Prestige	(0.019)**	(0.019)***	(0.014)***	(0.017)***
Member Prestige	0.011**	0.011**	0.005	(0.001)
Industry/academy affiliation controls				
Pure Industry Affiliation	0.599*	0.599*	0.331[+]	0.257
Mixed industry/ academy affiliation	(0.858)[+]	(0.858)	(0.865)**	(0.806)**
Constant	0.108	0.108	0.155**	0.155**
N	2 31	231	2 31	231
χ^2	2,213.746	416.196	950.264	1 ,110.231
LL	n/a	61.448	126.399	n/a
R^2	0.835	n/a	n/a	n/a

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; [+] $p < 0.10$. Standard errors in parentheses

Note: Numpapers divided by 1,000 to adjust scale

findings. Interestingly, the endo-cites analysis seems to parallel our findings more strongly. These two observations, coupled with our prior checks regarding collinearity, may imply that a knowledge community's success is self-driven. If cluster success is indeed endogenous, the high r^2 values of our internal measures, community cohesiveness and flexibility, would be unremarkable because we would expect internal measurements to change together as the community evolves. Perhaps some additional mechanisms influence endocite patterns. For example, we notice an unexpected increase in the prestige variables in this analysis, implying that, after controlling for other factors, cluster prestige is deleterious

for attracting external citations. We are also struck by the loss of significance of the time-lag variable in the endo-cites analysis. We speculate that perhaps same-cluster authors may have pre-publication access to articles from within their field as a result of informal exchanges of working papers, conference presentations, and generally quick diffusion of ideas from their scholarly networks.

Third, we wish to further examine the author leadership measure used in the main analysis. While an eigenvector measure of network centrality, which we used in our main analysis, is a good measure of direct and indirect influence, it uses non-directional ties. We therefore also measure for degree centrality, which includes only direct influence and in-degree centrality measures that represents uni-directional influence. These results are displayed in Models 5 and 6 of Table 4. We find that the coefficients for all but one of our variables of interest are perfectly consistent, both in direction and significance, with our main model. Knowledge Cohesiveness is not statistically significant in these models. The only other noteworthy change is that in-degree centrality, similar to eigenvector centrality, achieves statistical significance, while degree centrality is not itself statistically significant.

Fourth, we realize that numerous techniques could be validly utilized in modeling our particular form of time-series data. Furthermore, parametric assumptions relating to the variance-covariance structure are specific to each model estimation technique. In order to test the robustness of our results to our choice of models we fit three statistically viable alternative models, which differ chiefly in their error and variance-covariance structures. The Random Effects model, estimated in both standard and MLE manners, makes general assumptions common to all least squares methods. This broader assumption is less specific than the panel-data GLS model, which does not force conformity upon correlations and error terms where there could be a more complex structure. The GEE approach avoids considering the variance-covariance structure as a necessity (though it remains an option) in the correct specification of the distribution mean (Diggle et al. 2002; Long 1997) In fitting our GEE model we utilized the option of specifying the within-group correlation structure as AR(1) based on our prior results of the Wooldridge test for first-order auto-correlation in panel data. A further positive attribute of both the GEE and the standard Random Effects approaches is the ability to report significance using the robust modified standard errors.

In terms of our earlier model specification, the panel-data GLS model takes on an identical formulation to our GLS random effects model in all regards except for the second term of the errors, denoted T_{z_j} in the prior formulation above. This second term is no longer ignored and features the auto-regressive terms based on panel-specific auto-correlations, as shown in the following formula:

$$T_{z_i} = \begin{bmatrix} \sigma_1^2 & \rho_1 \sigma_1^2 & \cdots & \rho_1^{n-1} \sigma_1^2 \\ \rho_2 \sigma_2^2 & \sigma_2^2 & \cdots & \rho_2^{n-2} \sigma_2^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_n^{n-1} \sigma_n^2 & \rho_n^{n-2} \sigma_n^2 & \cdots & \sigma_n^2 \end{bmatrix}$$

The GEE model is also identical in terms of the specification for the mean; however, the T_{z_j}s listed above for the panel-data GLS model would remain in the same general form but with two slight modifications. First, the auto-regressive coefficients are constrained to be identical across all panels (rows) such that $\rho_1, \rho_2, ..., \rho_n = \rho$. Second, the values of the σ_i^2 terms are adjusted for group membership, our cluster assignments, according to the

Huber-White robust standard error calculations. For a full comparison of these alternative model specifications via the results they generated, see Table 5.

Our alternative model specifications yield remarkably strong support for the robustness of our findings. Across all alternative specifications the direction and statistical significance of our variables of interest are nearly unchanged. Only Knowledge Uniqueness loses significance when we utilize a different modeling technique. Interestingly, in both our GLS and GEE models the measure of Pure Industry Affiliation loses statistical significance, while the measure of mixed industry/academic affiliation becomes negative and significant.

Our robustness analyses taken together provide additional support for our variable choices and add nuance to our previous analyses. In our model comparisons the consistency of the significance and direction for the coefficients gives us greater confidence that our results are not artifacts of variable construction or model specification. In the next section we discuss the significance and contributions of our statistical findings.

Conclusion

We offer a macro-level framework for organizing large-scale innovation networks—called knowledge communities—and attempt to show how persistent community-level characteristics explain their differential success. Knowledge communities are a valuable level of analysis for studying innovation in science and technology.

Not all scientists are part of such "clusters" of cohesive research. Only about 40% of the work in computer science emerges from clusters of scientists who collaborate in producing knowledge. Knowledge communities produce, however, a disproportionate amount of the knowledge in computer science. In our dataset of computer science publications in technical journals, 56.61% of citations are received by papers in clusters even though only 43.67% of papers are in clusters. This trend is more dramatic within clusters, where 76.16% of citations go from one paper in a cluster to another paper in a cluster. On the other hand, papers not in a cluster cite almost proportionately to the ratio of papers in and out of cluster, with 41.32% of citations going to the 43.67% papers in a cluster and 58.68% of citations going to the 56.33% papers not in a cluster.

Recently literature on small worlds (Watts 1999) and geographic clustering (Porter 1998; Porter et al. 2004) has begun to address how large-scale networks contribute to productivity. We focus on the way knowledge communities use knowledge and rhetoric to help explain why some of these knowledge communities flourish and grow (Pfeffer 1993). Building on theory from exploratory search and marketing, we find that the patterns for knowledge use and use of rhetoric are very different. A broad-searching, far-ranging, and flexible use of knowledge maximizes community performance, while a shared, common, and stable rhetoric is most beneficial to community performance. We did not find support for the proposition that the use of unique knowledge benefits knowledge communities. Increased work by authors associated with firms had an overall positive effect on knowledge community performance, but an increase in work done jointly by researchers from firms and academic institutions led to an overall negative effect on knowledge community performance.

How do these characteristics lead to the functioning of knowledge communities? We speculate that in situations of large-scale collaboration and low coordination, a shared technical language helps minimize the cost and complexity of communication. Using a unified and consistent rhetoric, we believe, allows researchers who are intensely collaborating to more efficiently exchange ideas and collaborate. A knowledge community's level of innovation flourishes, on the other hand, when the best knowledge is identified, no matter where it comes from. Therefore, knowledge communities that search broadly and remain intellectually nimble perform best. Particularly in the face of very diverse ideas, expressing these innovations in a unified rhetorical and intellectual framework allows many ideas to be absorbed by a successful school and translated into a unified, explanatory, efficient, and shared rhetorical framework.

The logic of knowledge communities applies broadly to technology and science. Indeed, clusters of cohesive research have been identified in almost every research field (Aharonson et al. 2004; Braam 1991a; McCain 1987). Future research can more clearly delineate the important relationship between firms and knowledge communities, not just firms and scientific research. Many applications of this research are possible. We predict that firms with researchers participating specifically in successful and innovative knowledge communities will be more likely to generate successful innovation as valuable knowledge is transferred back into the firm. Similarly, we believe that the ability to identify and predict community success will be a valuable tool for governments trying to encourage promising nascent technology, or even in guiding the rule-setting bodies of government to help encourage productive collaborative networks. Venture capitalists may be interested in identifying what areas of technology are most likely to be productive in order to more efficiently allocate investments. Lastly, given the intertwined success and failure of knowledge communities, researchers may be able to make more intelligent and productive choices when embedding themselves in a research community.

Patterns of success and failure in science have explanatory consequences for the way science as a whole develops. For example, we were able to note the rise of the knowledge community for search technology in 1996 that preceded the growing importance of this technology in the evolution of computer science—a community that from its founding exhibited extremely focused rhetoric and very wide patterns of knowledge exploration. Further research into the differential success of knowledge communities will give us a better understanding of what guides the development and direction of innovation. Most importantly, continued understanding of the underlying causes of differential innovation in large-scale network structures should make it easier to encourage successful collaboration between researchers and improve the functioning of such communities, and lead as well to an increase in the overall velocity of research and innovation.

Appendix 1

Table 6 Cluster descriptions

Cluster #—Proposed name for cluster topic / 5 Most common words (frequency)	Titles of three most cited papers	CAGR, 1992–2003 (%)	Cumulative number of papers, 1990–2003
Cluster 1—Machine Learning and Neural Networks Learn (1,390)/Network (691)/Robot (649)/ Neural (606)/Model (506)	Finding Structure in Time An Information-Maximization Approach to Blind Separation and Blind Deconvolution Learning to Predict by the Methods of Temporal Differences	−6.83	7,892
Cluster 2—Object Oriented Languages Type (835)/Object (821)/Program (800)/ System (709)/Language (624)	Aspect-Oriented Programming A Hierarchical Internet Object Cache Logical Foundations of Object-Oriented and Frame-Based Languages	−5.06	9,368
Cluster 3—Model Verification System (1,267)/Time (826)/Model (783)/ Verification (435)/Specification (434)	Symbolic Model Checking for Real-time Systems The Algorithmic Analysis of Hybrid Systems STATEMATE: A Working Environment for the Development of Complex Reactive Systems	−4.16	7,022
Cluster 4—Design of Cryptographic Systems System (557)/Distribute (524)/Protocol (279)/ Base (222)/Fault (220)	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing Random Oracles are Practical: A Paradigm for Designing Efficient Protocols	n/a	4,144
Cluster 5—Machine Vision/Graphics Image (717)/Model (506)/Base (483)/ Recognition (350)/Motion (327)	Complements to 'Pattern Recognition and Neural Networks Performance of Optical Flow Techniques Progressive Meshes	n/a	6,053

Table 6 continued

Cluster #—Proposed name for cluster topic / 5 Most common words (frequency)	Titles of three most cited papers	CAGR, 1992–2003 (%)	Cumulative number of papers, 1990–2003
Cluster 6—Constraint Satisfaction Model (471)/Constraint (425)/System (398)/Base (389)/Algorithm (380)	Symbolic Model Checking: 10 20 States and Beyond Symbolic Boolean Manipulation with Ordered Binary Decision Diagrams A New Method for Solving Hard Satisfiability Problems	6.17	6,070
Cluster 7—Real Time Networks Time (1,146)/Real (908)/System (731)/Schedule (638)/Network (302)	Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism The BSD Packet Filter: A New Architecture for User-level Packet Capture Service Disciplines For Guaranteed Performance Service in Packet-Switching Networks	−9.85	3,968
Cluster 8—Programming Languages Logic (756)/Program (700)/System (595)/Proof (485)/Type (445)	Proof-Carrying Code Exokernel: An Operating System Architecture for Application-Level Resource Management A Framework for Defining Logics	4.4	7,743
Cluster 9—Internet Traffic Management System (1253)/Distribute (743)/Network (651)/Mobil (474)/Perform (461)	Congestion Avoidance and Control Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications On the Self-Similar Nature of Ethernet Traffic	−9.68	9,002
Cluster 10—Data Mining Data (887)/Queries (886)/System (777)/Base (767)/Database (762)	Fast Algorithms for Mining Association Rules Mining Association Rules between Sets of Items in Large Databases The Anatomy of a Large-Scale Hypertextual Web Search Engine (by Google founders Sergey Brin and Lawrence Page, 1998)	1.85	10,180

Table 6 continued

Cluster #—Proposed name for cluster topic 5 Most common words (frequency)	Titles of three most cited papers	CAGR, 1992–2003 (%)	Cumulative number of papers, 1990–2003
Cluster 11—Network Routing Network (1060)/Multicast (750)/Service (444)/Base (407)/Protocol (404)	Random Early Detection Gateways for Congestion Avoidance RTP: A Transport Protocol for Real-Time Applications Dynamic Source Routing in Ad Hoc Wireless Networks	39.01	5,890
Cluster 12—Parallel Computing Parallel (1212)/Perform (553)/Distribute (533)/Computing (524)/System (458)	Myrinet A Gigabit-per-Second Local-Area Network Implementation and Performance of Munin Parallel Programming in Split-C	−13.59	5,990
Cluster 13—Machine Learning Learn (1226)/Model (911)/Network (664)/Base (597)/Data (543)	Bagging Predictors Support-Vector Networks Experiments with a New Boosting Algorithm	13.82	9,566
Cluster 14—Shared Memory/Parallel Processing Parallel (479)/Memory (466)/Cache (297)/Perform (281)/Share (255)	Design and Evaluation of a Compiler Algorithm for Prefetching TreadMarks: Shared Memory Computing on Networks of Workstations Lazy Release Consistency for Software Distributed Shared Memory	n/a	3,818
Cluster 15—Optimization Algorithm (134)/Genet (104)/Problem (64)/Optimization (57)/Network (56)	Davenport-Schinzel Sequences and Their Geometric Applications Performance of Dynamic Load Balancing Algorithms for Unstructured Mesh Calculations	n/a	950

Table 6 continued

Cluster #—Proposed name for cluster topic / 5 Most common words (frequency)	Titles of three most cited papers	CAGR, 1992–2003 (%)	Cumulative number of papers, 1990–2003
Cluster 16—Congestion Control	Learning Long-Term Dependencies with Gradient Descent is Difficult	76.57	2,297
Network (514)/Tcp (347)/Control (324)/ Service (273)/Congest (194)	A Crash Course on Markov Chains and Stochastic Stability		
	Equation-Based Congestion Control for Unicast Applications		
	GPSR: Greedy Perimeter Stateless Routing for Wireless Networks		
Cluster 17—Distributed Computing	Wide-area cooperative storage with CFS	44.07	2,743
Network (379)/Web (352)/Traffic (253)/ Cache (213)/Service (196)	Accessing Nearby Copies of Replicated Objects in a Distributed Environment		
	Overcast: Reliable Multicasting with an Overlay Network		
Cluster 18—Internet Search	The PageRank Citation Ranking: Bringing Order to the Web (by Google founders Larry Page, Sergey Brin, et al, 1998)	51.74	2,428
Mine (366)/Data (342)/Web (229)/Base (206)/ Algorithm (187)	Discovery of Multiple-Level Association Rules from Large Databases		
	A Incremental Multi-Centroid, Multi-Run Sampling Scheme for k-medoids-based Algorithms		
Cluster 19—Rewrite Systems	Rewrite Systems	n/a	472
Rewrite (44)/System (43)/Program (42)/ Constraint (36)/Logic (31)	A Survey of Program Slicing Techniques		
	A Needed Narrowing Strategy		

Table 6 continued

Cluster #—Proposed name for cluster topic	CAGR, 1992–2003 (%)	Cumulative number of papers, 1990–2003
5 Most common words (frequency)	Titles of three most cited papers	
Cluster 20—Cryptography	28.04	3,289
Secure (413)/Key (238)/Protocol (200)/ Computing (195)/Scheme (184)	Proof Verification and the Hardness of Approximation Problems	
	A Digital Signature Scheme Secure Against Adaptive Chosen-Message Attacks	
	Simulating Physics with Computers (by physicist Richard Feynman)	
Cluster 21—Image Analysis/Tracking	11.44	3,043
Base (304)/Image (298)/Model (290)/ Recognition (245)/Track (216)	Fast Anisotropic Gauss Filtering	
	The Use of Active Shape Models For Locating Structures in Medical Images	
	Location Systems for Ubiquitous Computing	

References

Abrahamson, E. (1996). Management fashion. *Academy of Management Review, 21*(1), 254–285.

Aharonson, B., Baum, J., & Feldman, M. (2004). Industrial clustering and the returns to inventive activity: Canadian biotechnology firms, 1991–2000. *Draft*.

Barney, J., Wright, M., & Ketchen, D. J. (2001). The resource-based view of the firm: Ten years after 1991. *Journal of Management, 27*(6), 625–641.

Bartel, C. A., & Saavedra, R. (2000). The collective construction of work group moods. *Administrative Science Quarterly, 45*(2), 197–231.

Boyack, K. W., & Borner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology, 54*(5), 447–461.

Braam, R. R., Moed, H. F., & Vanraan, A. F. J. (1991a). Mapping of science by combined cocitation and word analysis. 1. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Braam, R. R., Moed, H. F., & Vanraan, A. F. J. (1991b). Mapping of science by combined cocitation and word analysis. 2. Dynamic aspects. *Journal of the American Society for Information Science, 42*(4), 252–266.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive-capacity—A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not). *National Bureau of Economic Research Working Paper 7552*.

Crane, D. (1972). *Invisible colleges diffusion of knowledge in scientific communities*. Chicago, IL: University of Chicago Press.

Crane, D. (1989). How scientists communicate—A citation classic commentary on invisible-colleges—Diffusion of knowledge in scientific communities by Crane, D. *Current Contents/Arts & Humanities, 22*, 14-14.

Culnan, M. J. (1986). The intellectual-development of management-information-systems, 1972–1982—A cocitation analysis. *Management Science, 32*(2), 156–172.

Diggle, P., Heagerty, P., Liang, K., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford, UK: Oxford University Press.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited—Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*(2), 147–160.

Downs, A. (1957). *An economic theory of democracy*. New York: Harper.

Drucker, P. F. (1985). *Innovation and entrepreneurship: Practice and principles* (1st ed.). New York: Harper & Row.

Drukker, D. M. (2003). Testing for serial correlation in linear panel-data models. *Stata Journal, 3*(2), 168–177.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132.

Fleming, L., Mingo, S., & Chen, D. (2005). Brokerage versus cohesion and collaborative creativity: An evolutionary resolution. *Draft*.

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy, 30*(7), 1019–1039.

Garfield, E. (1983). How to use citation analysis for faculty evaluations, and when is it relevant 2. *Current Contents, 45*, 5–14.

Garfield, E. (1988). Information technology and the social-sciences. *Current Contents, 46*, 3–9.

Gittelman, M. (2003). Does geography matter for science based firms? Epistemic communities and the geography of research and patenting in biotechnology. *Draft*.

Gittelman, M., & Kogut, B. (2003). Does good science lead to valuable knowledge. *Management Science, 49*(4), 366–382.

Gugarati, D. (1995). *Basic econometrics* (3rd ed.). New York: McGraw Hill.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science, 308*(29), 697–702.

Hargens, L. L. (2000). Graphing micro-regions in the web of knowledge: A comparative reference network analysis. In *Web of knowledge—a Festschrift in Honor of Eugene Garfield* (pp. 497–516). Medford: ASIS.

Haynes, A., Lackman, C., & Guskey, A. (1999). Comprehensive brand presentation: Ensuring consistent brand image. *Journal of Product and Brand Management, 8*(4), 28–30.

Henderson, R., Jaffe, A. B., & Trajtenberg, M. (1998). Universities as a source of commercial technology: A detailed analysis of university patenting, 1965–1988. *The Review of Economics and Statistics, 80*(1), 119–127.

Hotelling, H. (1929). Stability in competition. *The Economic Journal, 39*, 41–57.

Hurley, R. F., & Hult, G. T. M. (1998). Innovation, market orientation, and organizational learning: An integration and empirical examination. *Journal of Marketing, 62*(3), 42–54.

Kandylas, V., Ungar., L., & Forster, D. (2005). Winner-Take-All EM Clustering. Unpublished.

Katila, R. (2002). New product search overtime: Past ideas in their prime? *Academy of Management Journal, 45*(5), 995–1010.

Katila, R. (2003). Innovation product portfolio's by Navigating markets and technology. *Draft*.

Katila, R., & Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. *Academy of Management Journal, 45*(6), 1183–1194.

Kogut, B., & Macpherson, J. M. (2004). The decision to privatize as an economic policy idea: Epistemic communities, palace wars, and diffusion. Unpublished.

Kripke, S. (1982). *Wittgenstein in rules and private language*. Cambridge, MA: Harvard University Press.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago and London: The University of Chicago Press.

Levinthal, D., & March, J. G. (1981). A model of adaptive organizational search. *Journal of Economic Behavior & Organization, 2*(4), 307–333.

Lieberman, M. B., & Montgomery, D. B. (1988). First-mover advantages. *Strategic Management Journal, 9*(Special issue), 41–58.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Makadok, R. (1998). Can first-mover and early-mover advantages be sustained in an industry with low barriers to entry/imitation? *Strategic Management Journal, 19*(7), 683–696.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87.

McCain, K. W. (1987). Citation patterns in the history of technology. *Library & Information Science Research, 9*(1), 41–59.

McCloskey, D. N. (1998). *The rhetoric of economics*. Wisconsin: University of Wisconsin Press.

Merton, R. K. (1965). *On the shoulders of giants: A Shandean postscript*. New York: The Free Press.

Merton, R. K. (1972). *The sociology of science*. Chicago, IL: University of Chicago Press.

Meyer, M., & Zucker, L. G. (1989). *Permanently failing organizations*. Newbury Park, CA: Sage Publications.

Mizruchi, M. S., & Fein, L. C. (1999). The social construction of organizational knowledge: A study of the uses of coercive, mimetic, and normative isomorphism. *Administrative Science Quarterly, 44*(4), 653–683.

Murray, F. (2005). Exchange relationships & cumulative innovation: Standing on the shoulders of the oncomouse. Unpublished.

Murray, F., & Stern, S. (2005). *Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge?* Paper presented at the NBER Academic Science and Entrepreneurship Conference.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy, 26*(3), 317–330.

Nerkar, A. (2003). Old is gold? The value of temporal exploration in the creation of new knowledge. *Management Science, 49*(2), 211–229.

Pakes, A., & Shankerman, M. (1984). The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In Z. Griliches (Ed.), *R & D, patents, and productivity. A National Bureau of Economic Research Conference Report* (pp. 339–374). Chicago: University of Chicago Press.

Pfeffer, J. (1993). Barriers to the advance of organizational science—Paradigm development as a dependent variable. *Academy of Management Review, 18*(4), 599–620.

Phelps, J., & Johnson, E. (1996). Entering the quagmire: Examining the 'meaning' of integrated marketing communications. *Journal of Marketing Communications, 2*(3), 159–172.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York, London: Free Press/Collier Macmillan.

Porter, M. E. (1998). Clusters and the new economics of competition. *Harvard Business Review, 76*(November–December), 77–90.

Porter, M. E., Ketels, C., MIller, K., & Bryden, R. (2004). *Competitiveness in Rural U.S. regions: Learning and research agenda*. Paper presented at the Economic Development Administration, Washington, D.C.

Powell, W. W., Kogut, K. W., & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly, 41*, 116–145.

Price, D. D. (1963). *Big science, little science*. New York: Columbia University Press.

Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: Boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal, 22*(4), 287–306.

Schultz, D. E., Tannenbaum, S. I., & Lauterborn, R. F. (1994). *The new marketing paradigm: Integrated marketing*. Chicago: NTC Business Books.

Schumpeter, J. A. (1934). *The theory of economic development, an inquiry into profits, capital, credit, interest, and the business cycle*. Cambridge, MA: Harvard university press.

Shane, S. (2002). Selling university technology: Patterns from MIT. *Management Science, 48*(1), 122–137.

Slater, S. F., & Narver, J. C. (1995). Market orientation and the learning organization. *Journal of Marketing, 59*(July), 63–74.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations. I. Comparison of methods. *Scientometrics, 7*(3–6), 391–409.

Small, H. G. (1994). A SCI-Map case study: Building a map of AIDS research. *Scientometrics, 30*, 229–241.

Small, H. (2003). Paradigms, citations, and maps of science: A personal history. *Journal of the American Society for Information Science and Technology, 54*(5), 394–399.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *RAND Journal of Economics, 21*(1), 172–187.

Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem. *Administrative Science Quarterly, 111*(2), 447–504.

Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology, 105*(2), 493–527.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal, 5*(2), 171–180.

White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology, 54*(5), 423–434.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Yoels, W. C. (1974). Structure of scientific fields and allocation of editorships on scientific journals—Some observations on politics of knowledge. *Sociological Quarterly, 15*(2), 264–276.

Zhong, S., & Ghosh, J. (2003). A comparative study of generative models for document clustering. In *Proceedings of the workshop on clustering high dimensional data*. Citeseer.