A Cluster-based Method for Isolating Influence on Twitter

Shawndra Hill[1], Adrian Benton[1], Lyle Ungar[1],Sofus Macskassy[2], Annie Chung[1], John H. Holmes[1]

[1]University of Pennsylvania, Philadelphia, PA 19104

[2]Information Sciences Institute, Marina Del Rey, CA, 90292

Abstract

This paper demonstrates a cluster-based method to isolate influence in social network-based observational data, where "influence" is defined to mean that one person posts about a topic online and a second person posts about the same topic *because* he or she read the first post. Isolating influence in observational data is difficult, because we may observe that connected people discuss the same topic in proximate periods for reasons other than influence, including homophily–connected people are similar–and exogenous shock; they may have learned of the topic from some external source. We employ a matched sample estimation technique that has been used in the past to measure influence by controlling for demographic and usage based homophily, and add to the matching scheme a cluster ID. Our contribution is two-fold: First, we provide preliminary evidence that social network-based clusters capture homophily, indicating that a network-based attribute approach may not only capture homophily but also may be used in lieu of using demographic attributes for matching similar users in scenarios when privacy preservation is a concern. Second, we show that by adding a network position attribute, a cluster ID, when matching similar users, we can isolate influence better. We believe that our approach to isolate influence can have a broad impact on problems where social networks and associated behaviors can be observed over time.

1. Introduction

The study of influence is important because influencers have the ability to impact business outcomes and decisions in both positive and negative ways. Therefore being able to measure influence and identify influentials is a topic that is often studied by academics and practitioners alike. Influence and information contagion on social networks has been studied in the context of technology, healthcare [5, 6], marketing [8, 2], finance, news, and other domains from perspectives originating from epidemiology, business, computer science, and communication science. Most historical studies rely only on aggregate-level networks and outcome data for analysis. Only recently have researchers begun to try to understand the dynamics and behavior of users on social networks, partly due to the availability of massive amounts of individual-level social network data.

On the social networking site Twitter, researchers can observe short messages posted from one user to a set of other users at the individual level. Researchers can also observe the follower network and assess how connected individuals are, based on reciprocal links and how many followers users have in common, as well as whether messages, once broadcasted, get passed on or mentioned at a later point in time. For example, one recent paper explores what makes people "retweet" or repost messages and found that when tracking profiles based on connections between people and connections to topics using psychographic profiles[14], people-people affinity was a much stronger indicator of likelihood of retweeting [11]. The broadcasting mechanism on the Twitter network is a wonderful testbed to study the dissemination of information on social networks [4, 7, 16, 17, 19].

Links between social network-based clusters and topical information have been studied before, usually to identify the "most influential" people on the topic [3], for example, in the blogosphere where topic-specific clusters (clusters formed by considering only the links between bloggers formed in the context of a particular topic) can be derived[12]. The topic-specific clusters have been shown to have different identified influencers than if you looked at people blogging on a topic in the context of the blog-graph as a whole. Many free software tools try to identify *influencers* based on social media indicators. For example, the software Klout.com, Traakr.com, and Technorati.com are all free tools to rank the influence of individuals or blogs. Firms often use these tools to identify people to target for their products and services. In addition, firms use them to measure their own influence.

While publicly available tools aim to identify so-called influencers and measure influence, isolating influence is challenging because the apparent spread of information on a network may not only be the result of influence but also the result of either homophily [13] (connected people are likely to be similar and hence to post similar topics) or exogenous shock (people often learn about a topic from the same external source, for example, the news). Although challenging, isolating influence on networks has been the topic of many recent papers. For example, researchers have explored using spatial statistics models to model influence on social networks [10]. In addition, one social network-based method to identify influence is called the edge reversal test where directed links are reversed to see whether there is influence after the network structure has been perturbed [1]. The reversal test is limited in that it does not control for homophily by using node-level attributes. Another method used with observational data to control for homophily is matched sample estimation; users are matched, usually based on demographic and usage covariates, in order to make the data appear as if users were randomly assigned to experimental conditions [8, 2].

In this paper, we combine a link-based and demographics-based matching approach to improve on techniques for isolating influence. We focus on isolating influence from homophily and show that when we control for the network by clustering users based on their social network, we control for more similarity between users-possibly controlling for both implicit homophily and possibly even exogenous shock. While it is impossible to control completely for how people learn about topics on Twitter from external sources without running a controlled experiment, the problem can be ameliorated using hashtags, which are generally used only on Twitter. Hashtags are text-based tags that people create and reference in posts so the messages can be classified and then indexed by topic. In this paper, we study how these hashtags are passed on over time over the social network. In addition, we show that the social network clusters capture homophily to some degree and can be used as a partial substitute for user attributes when it is important for privacy to be preserved for applications like targeted marketing and advertising [15]. The paper is organized as follows. In section 2, we present our testbed. In section 3, we present our method. In section 4, we present our results, followed by conclusions and limitations in section 5.

2. Testbed: Collecting Twitter hashtag outcomes and social networks

Twitter is a social networking microblogging solution where users of the service follow other users who answer the question, "What are you doing?" in 140 characters or less. These short messages are called Tweets. In the Tweets, people can reference other users, as well as links to webpages and webstories. Users can also "retweet" messages other users have posted; these are indicated by the tag "RT." Reportedly, Twitter has more than 200 million registered users sharing information about a wide range of topics daily. We use Twitter as our testbed, because both links between users and tweets about particular topics can be observed in public over time.

The Twitter Search API was queried to collect data relating to the network of users, such as their friends and followers. The Twitter Streaming API was used to collect real-time Twitter posts for a collection of users. We followed Twitter users who self-identified as Pakistani and Israeli from September 2010 to May 2011. As new users became followers of users in our initial set, we followed them as well and collected their social networks. We took a snowball sample of users in order to build a large linked community of users. This network consisted of over 94 thousand users and over 14 million links between users. The data was anonymized by mapping all users in our data to our own set of anonymous IDs. Metadata referring to usernames or Twitter IDs and all "@" and reply-to mentions of other usernames within the body of the statuses were replaced with their corresponding anonymous IDs. The users we followed made millions of posts during the study period, which was a time of great activity in world politics. We extracted news-relevant hashtags indicated by a # sign from the text of the Tweet messages.

For a given hashtag, we tried to isolate the effect of influence by controlling for demographics, location, usage, and network position, using the propensity score matching approach described below.

3. Method

In order to isolate influence from homophily, we employed a method called propensity score matching, which is an approach to making observational data appear as if it came from an experiment, where users would have been randomly assigned to treatment groups. In our case the hypothetical treatment group was "being a follower of a poster of a topic" and the hypothetical control group was "not a follower of a poster of a topic." We matched users in the treatment group to users in the control group based on a set of user-centric covariates related to demographics, geographic location, usage, and the network position in order to reduce any selection bias to the treatment group due to those covariates. We used logistic regression to estimate the propensity of belonging to the treatment, "having friends that have posted a particular hashtag," given a set of

Attribute Set	Descriptions
Non-network	-Location, for example, coded as Egypt/Libya/Tunisia/Israel/Pakistan/Other. Searched for mention of country, large city in the country, or large region in country in the user->location field of Tweets. -Number of Tweets the user made in our corpus. -Number of total Tweets the user made (drawn from total Tweets field in Twitter status) -Proportion of Tweets that contained a hashtag -Proportion of Tweets that were retweets -Proportion of Tweets that were direct replies to other users
Network	-Number of friends that this user has within our corpus -Number of total friends that this user has (drawn from friends count field in Twitter status) -Number of total followers this user has (drawn from followers count field in Twitter status)
Cluster ID	-When partitioning the full graph into 250 clusters using Metis [9], treating each edge as undirected, the ID of the cluster this user was placed in.

Table 1: Description of covariate sets used for matching

user-based covariates. After we resampled the observations by matching, we measured influence by calculating the difference in the repost probabilities of the two treatment groups after matching. Our main contribution is a way to combine clustering on the social network as a proxy for homophily with propensity score matching in order to better isolate influence. Each step of our method is further explained below.

- **Assignment to treatment groups:** For each hashtag, all users in our corpus were placed into two groups, a *Control* group of users who followed no one who tweeted the tag, and a *Treatment* group of users who followed at least one person who tweeted the hashtag.
- **Attribute construction:** We constructed three sets of covariates: non-network, network, and cluster. The non-network attributes included location and usage attributes; the network attributes included Twitter network attributes such as how many followers the user had; and the cluster attribute was the cluster ID. We assigned network positions to each user by segmenting the network into 250 clusters using social network links through use of the Metis algorithm[9]. We chose this algorithm because it was network-based and could easily accommodate the size of our network. The attribute descriptions are presented in Table 1.
- **Match sample estimation:** We generated propensity scores for all users using different logistic regression models based on three different sets of covariates. Each user from the control group was matched with a user in the treatment group who had the closest propensity score. The treatment users with which they were matched were required to have a propensity score within 0.25* (std dev of propensity scores) of the control group user's propensity score (the caliper). All unmatched users were discarded from both groups.
- **Calculation of the difference in the repost probability:** After matching, we calculated the repost probabilities for each resampled group and evaluated the difference between them. In addition, we built a model to estimate the "treatment effect" of the connection to a prior poster after controlling for the different covariate subsets.

4. Results

In this section, we compared the repost rate of two different groups of users, those connected to posters of a topic and those who are not, after propensity score matching using five different covariate subsets. The five subsets are unmatched (or none), network, non-network, cluster ID only, and all attributes. The baseline case is the unmatched case. We will present only the results for the hashtag #Egypt because of space constraints[1]. Without matching, the difference between the probabilities of reposting within a certain time window was significantly different between the control

[1] We applied our approach to numerous hashtags, including many compound words only found on Twitter, e.g., #prayforisrael, and found using the network to assess similarity resulted in lower estimates for influence.

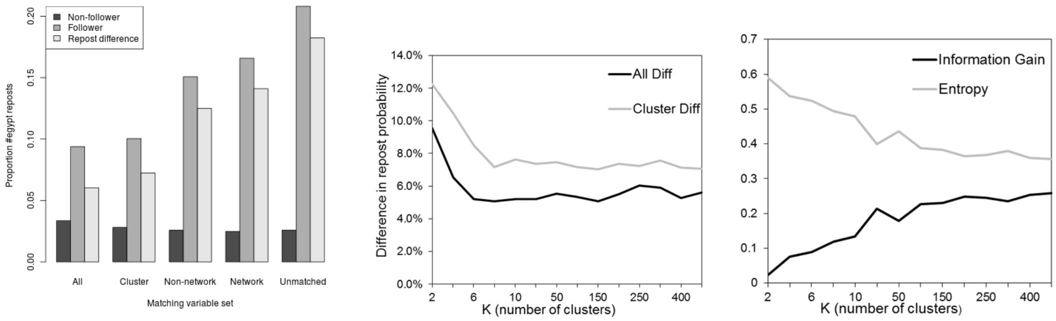

Figure 1: (left) Repost probabilities and differences between them for matched subsets of followers and non-followers. (center) repost probability difference for different clusters (right) entropy and information gain for different number of clusters

and treatment groups for all hashtags. For example, for #Egypt, the probability of reposting within 7 days for the control group was 2.6% and for the treatment group it was 20.8%, a difference of 18.2%, making the treatment group seemingly 8 times more likely to repost the hashtag. Much of this difference may be due to homophily or similar exogenous stimulus, and in fact we did find that when we resampled the data using propensity score matching with the all-attribute model, we seemed to best control for homophily, thereby reducing the difference in the probability of reposting a hashtag between the two groups drastically. For #Egypt, the control group repost probability was 3.3% and the treatment group's was 8.7%, a difference of 5.3%, making the treatment group only 2.6 times more likely to post, compared to 8 times more likely without matching. The different covariate subsets performed differently when used independently with the cluster ID subset, managing to reduce the influence effect the most by itself. See Figure 1 for bar charts showing the different probabilities for treatment (follower repost probabilities) and control (non follower repost probabilities) and the difference between the two for different covariate subsets. In fairness to the non-network attribute subset, the cluster ID may capture homophily that we cannot measure from the available demographic and location data we have; we believe if we were to have richer demographic data, the difference in the reposting probabilities would have been less.

So far, we have reported results when segmenting our network into 250 clusters (Figure 1 left). In Figure 1 (center), we show the results of sensitivity analysis for different numbers of clusters and therefore different-sized clusters. In the center plot, we show the difference in repost probabilities for both the cluster ID model and the all-attribute model. We found that the difference in repost probabilities flattened off at around 8 clusters. In addition, we performed a second sensitivity analysis to measure the homophily in clusters as the number of clusters increased. We calculated the conditional entropy of the users' location given their cluster assignment. If a user's cluster perfectly predicted their actual spatial location, then the conditional entropy of location given cluster ID would be 0. If, however, the user's cluster provided no additional information about their location, then the conditional entropy would be the same as not knowing the cluster ID. We also calculated information gain–the difference in entropy of the starting population and the conditional entropy of knowing the cluster assignment. These plots are shown in Figure 1 (right). The interesting part in the plot is that as the clusters get smaller, there appears to be more homophily for up to about 250 clusters. This is an indication that the clustering algorithm is performing as we intended: it captures homophily.

The main result we found is that by excluding the cluster ID in the matching, we dramatically reduced the ability of the logistic model to reduce the difference in repost probability, in part because it captured homophily. For example, the model yielded a matching where the control group repost probability was 2.6% and the treatment group's was 15%-a difference of 12.4%-which means the treatment group is 5.7 times more likely to repost as opposed to 8 times. All differences are statistically significant.

5. Discussion and Conclusion

In this paper, we have shown a simple result – that when trying to isolate influence in Twitter observational data, if we include network position in a matching scheme to control for homophily, we isolate the influence effect better. In other words, without including the clustering attribute to control for homophily, influence may be overestimated. While these results may have the potential to make a great impact on studies that try to isolate influence in networks, they do not come without many limitations.

First, it is not clear that propensity score matching using logistic regression that assumes observations are independent is appropriate for networked data where observations are often interdependent. Second, a number of methods for isolating influence from contagion in observational data have come under much scrutiny recently due to the nature of influence and homophily and the fact that the two are often confounded [18]. Without an experiment it is extremely difficult to control for possible bias in the estimate for influence due to both unobserved covariates. Finally, both the context of the discussion and how one constructs and samples from the network makes a difference in the influence effect estimates. Therefore, it is important to understand, in advance, what link types and strengths mean. In our case and many other recent studies [2, 8], it is not clear whether taking a snowball sampling approach to data collection preserves the important properties for analyzing social networks.

For our Twitter network, we compare the results when the network is defined in two ways. In the first case, we consider the network as we have discussed above. The network is directed and influence is directional. In the second case, we consider an undirected network where links have to be reciprocated, this means that we only consider links where both users are followers of one another. One might expect that a reciprocal link is stronger than a non-reciprocal link. In Figure 2, we show the results of repost probabilities for followers and non-followers the directed (left) and undirected (right) networks after matching on all covariates for each pair of followers and non-follower groups. In addition, we show the repost probabilities for different numbers of poster friends in the immediate neighborhood, we consider 1 - 5 friends.

For each of these groups we calculated the probability of mentioning a hashtag within some time window after another user mentioned it, P(User tweets | Someone else tweets). In addition to just using the cluster ID, we also matched users from the control group to ones in the treatment group if there existed an undirected path (edges going both directions between every two nodes along the path) of exactly length 2 between them. We did not match users if they were directly connected, since users in the control group were placed there because none of their friends had ever posted the hashtag, meaning that their matched counterparts in the treatment group never posted this hashtag either. Searching for a length 2 path gets around this and still ensures that the two users are relatively "close" in the network. We matched users that were in the same cluster and were "2 undirected hops" away from each other. After matching, P(User tweets tag | Someone else tweets tag) was generated separately for each hashtag for each group, and the degree to which each of the matching techniques was able to reduce the difference between the two groups was compared.

What we found was that the more connections to people who posted a hashtag, the higher the repost probability. We also found that reciprocal links for the same number of connections to people who posted a hashtag, led to higher repost probabilities. The differences highlight the challenges with dealing with network data and identifying the relevant links for the task and context at hand. In future work, we will further explore the impact of sampling and network definition on measuring influence. In addition, we will explore how well clusters based on different types of social networks (for example, Facebook networks, call networks, instant messenger networks) capture different types of homophily.

ACKNOWLEDGEMENTS

This research was supported by the Office of Naval Research and by Award Number RC1LM010342 from the National Library Of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Office of Naval Research, the National Library of Medicine or the National Institutes of Health.

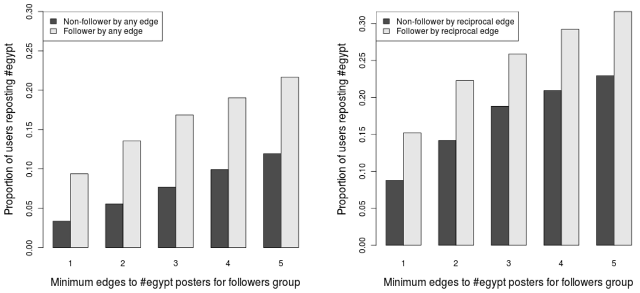

Figure 2: Repost probabilities for directed and undirected networks for different numbers of ties to posters of hashtags. The treatment group was matched based on all covariates to a subset of non-follower users for each number of friend category. The undirected network is comprised only of reciprocal links.

References

[1] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, Las Vegas, Nevada, USA, 2008. ACM.

[2] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[3] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[4] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):e14118, 2010.

[5] N.A. Christakis and J.H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

[6] N.A. Christakis and J.H. Fowler. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358(21):2249–2258, 2008.

[7] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs, 2010.

[8] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.

[9] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359, 1999.

[10] Roger Th. A. J. Leenders. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24(1):21–47, 2002.

[11] S. A. Macskassy and M. Michelson. Why do people retweet? anti-homophily wins the day! In *Fifth International Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011.

[12] Sofus Macskassy. Contextual linking behavior of bloggers: Leveraging text mining to enable topic-based analysis. *Social Network Analysis and Mining*, 1(4):355–375, 2011.

[13] Miller McPherson, Lynn Smith-Loving, and James Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 1983.

[14] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the fourth workshop on Analytics for Noisy Unstructured Text Data*, pages 73–80, Toronto, Canada, 2010. ACM.

[15] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: Privacy-friendly social network targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 707–716, Paris, France, 2009. ACM.

[16] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *20th international conference on World wide web (WWW '11)*, pages 695–704. ACM, 2011.

[17] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188, 2010.

[18] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2).

[19] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM '10*, 2010.