Understanding User Migration Patterns in Social Media

Shamanth Kumar, Reza Zafarani, and Huan Liu
Computer Science & Engineering,
SCIDSE, Arizona State University
{shamanth.kumar, reza, huan.liu}@asu.edu

Abstract

The incredible growth of the social web over the last decade has ushered in a flurry of new social media sites. On one hand, users have an inordinate number of choices; on the other hand, users are constrained by limited time and resources and have to choose sites in order to remain social and active. Hence, dynamic social media entails user migration, a well studied phenomenon in fields such as sociology and psychology. Users are valuable assets for social media sites as they help contribute to the growth of a site and generate revenue by increased traffic. We are intrigued to know if social media user migration can be studied, and what migration patterns are. In particular, we investigate whether people migrate, and if they do, how they migrate. We formalize site and attention migration to help identify the migration between popular social media sites and determine clear patterns of migration between sites. This work suggests a feasible way to study migration patterns in social media. The discovered patterns can help understand social media sites and gauge their popularity to improve business intelligence and revenue generation through the retention of users.

Introduction

Social media has shown a considerable growth over the past years[1]. With numerous social networking sites popping up everyday and the limited amount of time and resources each person has, social media users have to make decisions on which sites to spend their time. It is imperative for social media sites to retain their existing users while continuing to attract new ones. Understanding how users make their choice of social media sites has important implications. Knowing migration patterns can help a social media site to 1) generate revenue from suggested advertising; 2) increase traffic via shared media, which in turn improves marketing outcomes; and 3) grow their base of long-term customers to increase brand loyalty.

Though one may not understand the reasons behind the choices people make, the migration patterns can be invaluable in anticipating user migration and taking actions to prevent it from happening. In this paper, we ask if it is feasible to study user migration in social media and what it takes to study social media user migration. User migration can happen across different social media sites. The study can be complicated by the existence of multiple competing sites of the same social media category. For example, categories (or types) can include social bookmarking, social networking, social media sharing, etc. Seven popular social media sites of different types are used in our study. The contributions of our study include,

- Presenting a feasible way of studying user migration in social media,
- Visualizing clear patterns of user migration across social media,
- Differentiating users based on their importance in the context of migration, and
- Proposing a verification approach based on hypothesis testing.

The rest of the paper is organized as follows: first, we present migration related definitions; second, we introduce how to conduct our study of user migration with 3 key steps (collecting data, acquiring migration patterns, and verifying migration patterns). We present a brief review of related work and conclude the paper with some future work.

Migration Related Definitions

In this section, we give definitions of migration and discuss two types of migration.

Migration Migration can be described as the movement of users away from one location and towards another, either due to necessity, or attraction to the new environment. In the context of social media, we define two kinds of migration, *site migration* and *attention migration*. Let U_{s_1} be the set of all members of site s_1 and U_{s_2} be the set of all members of site s_2. Then, the site migration of user u from social media site s_1 to site s_2 can be defined as follows,

Definition 1 (Site Migration) *Let $u \in U_{s_1}$ and $u \notin U_{s_2}$ at time t_i, if $u \notin U_{s_1}$ and $u \in U_{s_2}$ at time $t_j > t_i$, then user u is said to have migrated from site s_1 to site s_2.*

Site migration of an individual can be determined by checking the presence of a user's profile on sites s_1 and s_2. Reasons for site migration include profile removal, profile

[1] http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/

deletion, and account suspension. In the last case, the account still exists, but the profile information is not accessible. A suspended account may be reinstated later.

The attention migration of user u from social media site s_1 to site s_2 is defined next

Definition 2 (Attention Migration) *Let $u \in U_{s_1}$ and $u \in U_{s_2}$ at time t_i and u is active at s_1 and s_2 at time t_i, if u is inactive at s_1 and active at s_2 at time $t_j > t_i$, then the user's attention is said to have migrated away from site s_1 and towards site s_2.*

Between the two types of migration, the attention migration can be measured by a user's activity. We now define the activity (or inactivity) of a user and related definitions below.

Definition 3 (User Activity) *Given a site s, a user $u \in U_s$, time $t_j > t_i$, and time interval $\delta = t_j - t_i$, u is considered to be active on s at time t_j, if the user has performed at least one action on the site since time t_i. Otherwise, the user is considered inactive.*

The interval δ could be measured at different granularity, such as *days*, *weeks*, *months*, and *years*. The user's actions could be one of the many user actions possible on that social media site, such as submitting a news story, posting a status message, and uploading a video. For example, a user *'test'* would be considered active on Delicious in February, 2010 if he has submitted at least one bookmark since January, 2010. Here, $\delta = 1\ month$. Attention migration can be considered as short-term migration of individuals, which might lead to site migration later. In the rest of the paper, we focus on attention migration. We discuss some corresponding measures related to activities.

User activity $\mathcal{A}(u, s)$ is determined from features of a social media site that reflect publishing or communication activities, such as the number of tweets posted by a user on Twitter, the number of bookmarks on Delicious, or the number of photos uploaded on Flickr, among others. Mathematically, the user activity can be represented as

$$\mathcal{A}(u, s) = \frac{f(u, s)}{\max_{u'}(f(u', s))}, \tag{1}$$

where $f(u, s)$ is a linear function of user u's activities on site s.

User Network Activity $\mathcal{N}(u, s)$ is determined from the networking activities of user u on site s, such as the number of friends and the number of followers on Twitter, and the number of subscribers and the number of subscriptions on StumbleUpon. Social media sites adopt different policies towards the formation of network links between users. For example, Twitter allows users to have directed links to other users in the form of followers and friends, while Digg only supports undirected network relationships in the form of friends. To determine the network activity of a user, we consider the size of his immediate network on the site. In the case of Twitter, this would be the sum of the number of friends and the number of followers, while for Digg it would just be the number of friends of the user. The user network activity can be represented as

$$\mathcal{N}(u, s) = \frac{g(u, s)}{\max_{u'}(g(u', s))}, \tag{2}$$

where $g(u, s)$ is a function of user u's network activity on site s.

User Rank $\mathcal{R}(u, s)$, the rank of a user can be defined as the value of user as perceived by other individuals who may or may not be on the same site. Here, this rank is calculated using Normalized Google Rank ($\mathcal{NGR}$). $\mathcal{NGR}$ is computed by identifying the number of Google hits for a user's profile page as returned by Google Search. A user is provided with a unique profile page on every social media site. For example, a user *"test"* on StumbleUpon will have *http://test.stumpleupon.com* as his profile page. Searching for *"link:http://test.stumbleupon.com"* on Google Search gives us the number of links to his profile on StumbleUpon and therefore a measure of how popular the user is in an environment that is external to the site. The number of hits for a user is normalized using the maximum of this value for all the users under consideration. The rank can be represented as

$$\mathcal{R}(u, s) = \mathcal{NGR}(u, s) = \frac{r(u, s)}{\max_{u'} r(u', s))}, \tag{3}$$

where $r(u, s)$ is a function that returns the number of hits for user u's profile page on site s.

Studying Migration Patterns

We address three key challenges in studying migration patterns: (1) how to determine social media sites for data collection, (2) how to acquire migration patterns, and (3) how to verify migration patterns that are different from random changes.

Collecting Data

There are hundreds of social media sites of different types such as social bookmarking and social networking. It is impractical to study all. The selection of social media sites should at least consider the following:

- There should at least be more than 1 type of social media sites to allow for studying attention migration from one type to another.
- At least for one type of social media, there should be more than one site to allow for studying attention migration from one site to another.
- They all exist for a period of time to allow for observation of different time points.

Another problem is to resolve user identities across social media sites as discussed in (Zafarani and Liu 2009). Currently, there are two ways to accurately identify users across social media sites. One is to elicit user identities from the users themselves through surveys, but this method is not scalable. The other method to identify users across sites is to use the services of blog directory sites. One of such sites is BlogCatalog. At a blog directory site, users have the

Table 1: Migration Dataset: Amount of information gathered from the selected social media sites

Site	No of Users	Profile Attributes
Delicious	8,483	10
Digg	9,161	20
Flickr	5,363	11
Reddit	2,392	5
StumbleUpon	8,935	13
Twitter	13,819	15
YouTube	7,801	19

freedom to publish their identities or usernames from other social media sites on their profile pages to connect with their readers and other blog authors. Since users have a motivation to publish their identities from other sites, we trust these identities to be accurate. Using the BlogCatalog API [2], we collected more than 96,000 user profiles. From this dataset, we separated those users who were active on more than one site among the 7 popular social media sites. They are *Delicious, Digg, Flickr, Reddit, StumbleUpon, Twitter, and YouTube*, which gave us 17,798 user profiles. We performed this step to avoid including those users in the study who had limited activity beyond BlogCatalog or whose identities on these seven sites was not sufficiently known.

The activity and user profile information of the users on these 7 social media sites is obtained either using APIs or screen scraping when APIs are not available. Note that not all the users had usernames on all 7 sites. In this migration dataset, more than 7,225 users have user accounts on 5 or more sites. The collection of user profile information on these sites was carried out in March 2010, April 2010, and May 2010. The data for each month corresponds to a snapshot and the value of the time window parameter δ can be used to control the time difference between two snapshots. In this paper, we set $\delta = 1\ month$. We obtain two phases of user data across these social media sites, where each phase is defined as the data from two consecutive snapshots. In this case, Phase 1 spans March and April data while Phase 2 spans April and May data. The user profile information includes real name, age, location, status messages, friends, followers, etc. Information and statistics for each site are presented in Table 1. Next, we discuss how to find migration patterns from the data.

Obtaining Migration Patterns

We study the attention migration of users between the 7 social media sites. We first need to quantify the number of users whose attention migrates away from a site to another. For this purpose, we use data from the three snapshots to identify the trend of attention migration in each of the 7 social media sites. We choose one site as the base and observe how its users' attention moves to the other six sites. Our results are presented in Figure 1 in the form of radar charts. A radar chart is a plot of variables in the form of equi-angular spokes, called radii, with each spoke representing a variable. Radar charts are very useful in determining which variables are dominant for a given observation and hence well suited for representing the study of movement of individuals. In our case, each radar chart corresponds to the migration of individuals from the base site towards other sites. Each spoke in the chart represents a social media site and the radius represents the amount of migration towards the site connected by the spoke. The charts tell us two findings:

- Attention migration does exist between the social media sites. If it does not exist, all the points in the corresponding radar chart would just be a dot in the center, with a 0 radius.
- Migration patterns are shown as which site incurs most significant migration and which sites gain the most users.

It is noted that the summation of the fraction of these radii do not necessarily sum up to 1 as a user can migrate from the base site to multiple sites. From the results in Figure 1, it is clear that the general trend of attention migration is migrating towards Twitter and StumbleUpon; Reddit users have the highest amount of migration to other sites, and the least number of users migrate to Reddit. The most significant fraction of Reddit's population (16% of the users) migrated to Digg. Digg is another social news site where users can "digg" a news story and make it popular and a popular story can be promoted to the front page. This shows that Reddit loses a significant fraction of its population to a competing social media site like Digg, which offers similar functionality but seems more attractive than Reddit. Similarly, we see a significant amount of migration of users between StumbleUpon and Delicious which belong to the same category of social media. So far, we observe attention migration within the same social media category, in other words, as users tend to migrate between competing sites within the same category. Another factor that could be responsible for this migration is "herd effect", when one's friends move, he follows.

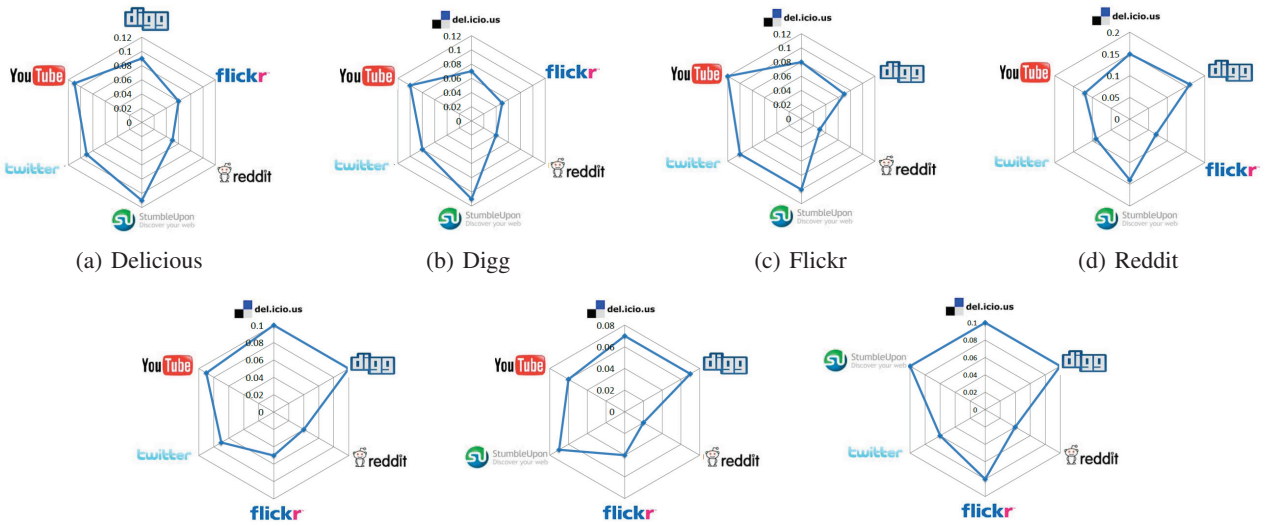
Verifying Migration Patterns

In order to verify these patterns shown in Figure 1, we resort to some statistical test. As the first step, we create a reference point to compare our results with. In our case, this would be the migration of random individuals as we can safely assume that users do not randomly select a new site and then leave for it from the current site. Given any other set of randomly picked migrating individuals we would not expect to observe patterns such as migration to competing sites, like Delicious and StumbleUpon. Hence we have a null hypothesis as

H_0: The migration of individuals is random and no correlation exists between their attributes such as general or network activity on a site and their migration.

To create the reference dataset, we use the methodology of the shuffle test proposed in (Anagnostopoulos, Kumar, and Mahdian 2008). The shuffled dataset for each site is constructed by randomly picking users from the potential migration population, which consists of the overlapping

[2] http://www.blogcatalog.com/api/



(a) Delicious (b) Digg (c) Flickr (d) Reddit

(e) StumbleUpon (f) Twitter (g) YouTube

Figure 1: Pairwise attention migration patterns between different social media sites

users in both Phase 1 and Phase 2. The number of migrated users in a shuffled dataset is kept the same as the number of migrated users we observed in the real datasets. We create 10 such shuffled datasets for each site. To compare the shuffled datasets and the observed data we need to measure the distance between them. In order to measure the relationship of a user's attributes to his migration behavior, we can use techniques such as logistic regression which can be formulated as follows,

$$Y = \frac{e^z}{1 + e^z}, \tag{4}$$

where $z = w^T X + w_0$. Here, the boolean variable Y, is the class attribute for a user, which indicates whether the user has migrated away from a site. Each $x_i \in X$ is a feature whose coefficient is $w_i \in w$, which represents its correlation to the class attribute. In our case, we used user's Activity A (e.g., number of tweets), user's network activity N (e.g., number of friends), and user's Rank R (user's ranking in Google search results) as the attributes. This procedure can be similarly applied to each shuffled dataset for a site. We can then obtain the average of the coefficients for each attribute from the 10 shuffled datasets for each site as the representative. We formulate the distance between the shuffled dataset and the observed dataset as the χ^2-statistic. Using the observed regression coefficients, we evaluate the null hypothesis using the χ^2-statistic as follows,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

where n is the number of regression coefficients, O_i is the coefficient values obtained from the real dataset, and E_i is the coefficient values obtained from the shuffled dataset. Table 2, shows the results of applying chi-square test on the observed and the shuffled dataset. Missing coefficients for the Google rank of users is represented using the symbol $-$, because for some sites all the users had a value of 0 for this attribute. The degree of freedom for the χ^2 test is $n - 1$ (2, in this case).

The results can be interpreted in the following way: p-value for a site tells us the probability of the observed dataset being similar to the shuffled dataset which is obtained by randomly picking individuals from the potential migration population. Here, we consider the result to be statistically significant if $p < 0.05$. From Table 2 we notice that the migration patterns for users from sites Delicious, Digg, and Reddit are very similar to the shuffled dataset. On further exploration, we found that this was due to the small size of the potential migration population which was used to select the individuals who migrated. We also notice that Flickr dataset, although not statistically significant, is still quite different from the shuffled datasets and has a low p-value. StumbleUpon, Twitter, and YouTube on the other hand, very strongly reject our null hypothesis and the patterns from these datasets are clearly very distinct from those of the shuffled dataset. These results also support our earlier observations which show that a majority of the user migration is towards StumbleUpon and Twitter. In addition, during our experiments we observed that user activity has a high correlation with the migration of an individual away from a site.

Probing Further

Social media sites typically have large populations. After acquiring attention migration patterns, we would like to ask if we can make use of the patterns by following a smaller group of users such that (1) their migration patterns are reflective of the general population of the site, and (2) this group can be studied further to see if some kind of intervention can deter or encourage the migration. We borrow the concept of High Net Worth Individuals (HNWI) (Freear, Sohl, and Wetzel Jr

Table 2: χ^2 test results on the observed and shuffled data

Site	Observed Coefficients			Shuffled Coefficients			p-value	Statistical Significance
	N	A	R	N	A	R		
Delicious	0.2858	0.4585	-	0.6029	0.5921	-	0.65	Not significant
Digg	0.4796	0.8066	-	0.52	0.5340	-	0.70	Not significant
Flickr	1	1	0.9797	0.2922	0.2759	0.4982	0.13	Not significant
Reddit	0.5385	0.6065	-	0.4846	0.6410	-	0.92	Not significant
StumbleUpon	1	1	-	0.4191	0.2059	-	0.0492	Significant
Twitter	0.5215	1	0.5335	0.2811	0.0365	0.4009	0.0001	Extremely significant
YouTube	0	1	0.1644	0.7219	0.0040	0.4835	0.0001	Extremely significant

Table 3: Activity Patterns of top 1,000 High Net-Worth Individuals

Social media site	No of users who were inactive (Snapshot 1)	No of users who were inactive (Snapshot 2)	No of users who were inactive (Snapshot 3)
Delicious	426	458	479
Digg	705	717	715
Flickr	424	431	431
Reddit	487	447	605
StumbleUpon	308	358	359
Twitter	29	25	43
YouTube	314	420	464

1992) from the banking industry to show how to search for a group with manageable size of representative users of a site. HNWIs have reasonably high investable assets. In social media sites, high "Net-Worth" individuals can be deemed as those who are important to the site for its development and growth. Their continued support brings more traffic and thus they constitute the social capital of a site. We propose that the HNWIs for a site constitute the representative and manageable set of users to study migration patterns. One straightforward way to define social media HNWI based on our earlier definitions of different kinds of user activity, network activity, and rank.

Definition 4 ("Net-worth" of an individual) *Given a site s and a user u, the "Net-Worth" ($\mathcal{W}(u,s)$) of u on s is*

$$W(u,s) = w_{\mathcal{A}}.\mathcal{A}(u,s) + w_{\mathcal{N}}.\mathcal{N}(u,s) + w_{\mathcal{R}}.\mathcal{R}(u,s), \quad (6)$$

where $w_{\mathcal{A}}$, $w_{\mathcal{N}}$, and $w_{\mathcal{R}}$ are the respective weights.

For each site we rank the users based on our definition of "Net-Worth" and identify the top 1000 HNWIs for this study. The weights $w_{\mathcal{A}}$, $w_{\mathcal{N}}$, and $w_{\mathcal{R}}$ are set to 1.

Basically, if this group shows some patterns which are consistent with the patterns shown in Figure 1, we can monitor them for tasks aiming to improve business intelligence and user retention or recruitment. Table 3 shows activity behavior of high "Net-Worth" users for snapshots 1, 2 and 3, respectively. Each cell in the table represents the number x of inactive high "Net-Worth" users in each snapshot, and $(1000 - x)$ is the number of active users. An inactive user at one snapshot can be active at another snapshot. In general, the more inactive users a site has, the likelier attention migration can occur. For instance, social news sharing site Digg has a particularly low number of active HNWIs and Twitter has a very low number of inactive users (3% across the three snapshots). The inactive patterns correspond well to the migration patterns of the general population of the sites. As suggested in Figure 1, more users migrated to Twitter from other sites.

Related Work

The problem of information diffusion is relevant to the problem of migration in many aspects. In this case, it is the information that can be considered to migrate from one site to another instead of people. The problem of diffusion and propagation in social networks has been studied from many perspectives. One of the early works on information diffusion includes (Granovetter 1978) in which the author introduced a model of collective behavior based on the concept of an aggregate threshold that must be overcome for individual behavior to spread to other actors. In (Gruhl et al. 2004), the authors study the diffusion of information in the blogspace. The diffusion of information is possible in the blogosphere due to the support of social networking by most blogging platforms. The study specifically concentrates on investigating the short term topics, or "snapshot models", and presents the study of long term topics of discussion, or the "horizon topics", as an open question to the community. Another interesting study on the diffusion of information in the blogosphere is presented in (Adar and Adamic 2005). In this work, the authors use an influence based model to study the propagation of topics in the blogosphere and also present a visualization tool that can be used to visually analyze the spread of infection in blogs starting from a seed node. Another prominent model in this area, Independent Cascade Model (ICM), is alluded to in (Kempe, Kleinberg, and Tardos 2003). ICM models diffusion on a stochastic process whereby behavior spreads from one actor to another with a given probability. Studies such as (Pasto-Satorras and Vespignani 2001; Moore and Newman 2000; Newman 2002) model the spread of epidemic diseases in the social networks. An epidemic can be described as the spread of a disease at a rate that exceeds the expected rate. These models provide an effective method of gauging how an epidemic would spread in the real world.

Conclusion and Future Work

In this study, we show that (1) studying migration across social media is feasible, (2) patterns can be identified in migration, and (3) it is possible to act on the migration patterns by monitoring a group of high net-worth users. To study migration patterns, we define two types of migration and analyze the migration of user attention between 7 popular social media sites. Using a variety of social media sites, we present some interesting migration patterns which could facilitate further research on solutions to prevent or encourage such migration. For example, social news sites, such as Digg and Reddit have the highest number of users migrating away, i.e., low user retention rates. Identifying these factors could be valuable to social media sites in several ways: e.g., designing features to recapture user attention before the exodus begins and learning to avoid similar pitfalls when launching new social media sites.

After demonstrating the feasibility of studying migration patterns in this work, we can embark on more extensive investigation. One of the principal challenges of this study was to obtain the mapping of users across the different social media sites. Although we use BlogCatalog to get the user mapping in this study, the scope is still limited to these particular users. As an illustration of this limitation, we present a preliminary study of the users in our dataset to observe what is known as the "herding" behavior in migration. Herding can be defined as using the information from other individuals to make a rational choice (Easley and Kleinberg 2010). The user and the herd can exhibit two types of migratory patterns in terms of the herding behavior: either the user migrates when the herd (e.g., friends) moves (*herd-initiated migration*) or the herd moves when the user migrates (*user-initiated migration*). In a preliminary effort to investigate this behavior, we created a dataset of users and their friend network on Blogcatalog. For all these users, we find their identities on all the other social media sites. For this experiment, we could only use those friends who also had valid usernames on all the sites the user himself had a valid username, making possible analyzing the herding behavior and identifying user's network migration across sites. Though this restriction was essential, it significantly reduced the size of the dataset. For example, the average network size of users at Delicious became 10.32. As another example, we found that for users in our dataset that demonstrated migration, only 13.10% of their network migrated when the user himself had migrated. Another challenge is to determine if a user has moved to another site when we cannot uniquely identify him on that site.

Acknowledgements

This work was supported, in part, by the Office of Naval Research grant: N000141010091.

References

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *WI*, 207–214.

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *KDD*, 7–15.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Freear, J.; Sohl, J.; and Wetzel Jr, W. 1992. The investment attitudes, behavior and characteristics of high net worth individuals. *Frontiers of Entrepreneurship Research* 16:374–387.

Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83(6):1420–1443.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW*, 491–501.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *KDD*, 137–146.

Moore, C., and Newman, M. 2000. Epidemics and percolation in small world networks. *Physical Review E* 61:5678–5682.

Newman, M. 2002. The spread of epidemic disease on networks. *Physical Review E* 66(1):16128.

Pasto-Satorras, R., and Vespignani, A. 2001. Epidemic spreading in scale-free networks. *Physical Review E* 86(14):3200–3203.

Zafarani, R., and Liu, H. 2009. Connecting corresponding identities across communities. In *ICWSM*.