Analyzing Behavior Of The Influentials Across Social Media

Nitin Agarwal
Information Science Department
University of Arkansas at Little Rock
nxagarwal@ualr.edu

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

July 18, 2011

1 Introduction

Social media, or commonly known as the Social Web, consists of myriad applications including blogs, social networking websites, wikis, social bookmarking or folksonomies, online media sharing, social news, social games, etc. Through reactive interfaces, low barrier to publication, and zero operational costs, which are all made possible by the new paradigm of Web 2.0, social media has observed a phenomenal growth in user participation leading to participatory web or citizen journalism. The blogosphere, for instance, has been growing at a phenomenal rate of 100% every 5 months[1]. BlogPulse, another blog indexing service, says it has tracked over 165 million blogs by July 2011[2]. Facebook recorded more than 750 million active users as of July 2011[3]; Twitter amassed nearly 200 million users in March 2011[4]; and other social media sites like Digg, Delicious, StumbleUpon, Flickr, YouTube, etc. are also growing at terrific pace. This clearly shows the awareness of social media sites among individuals.

Individuals use different social media sites for different reasons. For instance, they use MySpace to keep in touch with their friends, make new ones, share their updates and get updates of their friends; Flickr to upload and share photos with friends and others; Twitter to update their status; Delicious to bookmark, tag and share their favorites with friends or others; Digg to rate and promote the

[1] http://technorati.com/blogging/feature/state-of-the-blogosphere-2008/
[2] http://www.blogpulse.com/
[3] http://www.facebook.com/press/info.php?statistics
[4] http://www.aolnews.com/2011/03/21/twitter-celebrates-5-years-and-200-million-users/

content that they feel is relevant to the society; etc. Some individuals are active on a few social media sites and some are active on many of them. Essentially they try to be sociable or gregarious by making friends on these social media sites. Another type of behavior that can be observed on these social media sites is the influential behavior. Individuals try to lead the community or the conversation in the social media sites. In this chapter, we focus on the later behavior exhibited by individuals.

There has been a lot of existing work on identifying such influential and/or active members [2][15] on one social media site. What is more interesting is to study the behavior of these individuals across different social media sites. Different social media sites could be alike or different in terms of functionality, which is discussed in more detail in Section 2. It would be interesting to study the different types of behaviors users exhibit on sites alike or different. This entails studying the relationship between individuals' behavior and the social media sites in the sense that there could be: same behavior on similar social media site; same behavior on different social media sites; different behavior on similar social media site; and different behavior on different social media site.

Studying these behavior patterns of users across different social media sites have many applications. If an individual exhibits same behavior on various social media sites then it can help predict his behavior on other social media websites by studying one social media site. Various social media sites can be clustered based on the behavior patterns of individuals. These clusters can help discover helpful and valuable trends. The activity of individuals can also help in explaining which social media sites are likely to get more activity for various groups of people. These patterns can be also used to explore marketing opportunities, study the movement of individuals on social media sites to focus on niche sites for unique opportunities.

The rest of the chapter is organized as follows: Section 2 elaborates on the social media taxonomy, Section 3 discusses the influential behavior of individuals along with models to quantify influence, Section 4 highlights the challenges and opportunities of data collection across multiple social media, Section 5 presents experiments and interesting findings on cross media behavior, and Section 6 presents conclusions with future directions for research.

2 Social Media Taxonomy

Individuals participate in different social media applications with different intentions and expectations. Based on a multitude of such functionalities, social media applications can be organized in a taxonomy as illustrated in Table 1. In this section, we briefly describe each category and the functionalities:

Social Signaling refers to a collection of applications that allows individuals to express interactions, opinions, ideas, thoughts, and connect with fellow individuals through blogs, microblogs, and social friendship networks. Blogs, or web logs, are collections of articles written by people arranged in reverse chronological order. These individual articles are known as blog posts. The

Table 1: A taxonomy of social media applications based on their functionalities.

Category	**Social Media Sites**
Social Signaling	Blogs (Wordpress, Blogger, Blogcatalog, MyBlogLog), Friendship Networks (MySpace, Facebook, Friendfeed, Bebo, Orkut, LinkedIn), Microblogging (Twitter, SixApart)
Media Sharing	Flickr, Photobucket, YouTube, Multiply, Justin.tv, Ustream
Social Health	PatientsLikeMe, DailyStrength, CureTogether
Social Bookmarking	Del.icio.us, StumbleUpon
Social News	Digg, Reddit
Social Collaboration	Wikipedia, Wikiversity, Scholarpedia, Ganfyd, AskDrWiki
Social Games	Farmville, MafiaWars, SecondLife, EverQuest

blogs are collectively referred to as the blogosphere. A blog can be maintained by an individual, known as an individual blog or by a group of people, known as a community blog. The authors of blogs are known as bloggers. Some blog cataloging services such as BlogCatalog[5] also allow users to create friendship networks. Microblogging sites, as the name suggests, are similar to blogs except for the fact that the articles can only be of limited length. In the case of Twitter[6], the posts (or tweets in this case) can be 140 characters or less. These sites are typically used to share what you are doing and diffuse information simulating a word-of-mouth scenario. Social Friendship Networks allow people to stay in touch with their friends and also create new friends. Individuals create their profile on these sites based on their interests, location, education, work, etc. Usually the ties are non-directional, which means that there is a need to reciprocate the friendship relation between two nodes.

Media Sharing sites allow people to upload and share their multimedia content on the web, including images, videos, audio, etc., with other people. People can watch the content shared by others, enrich them with tags, and share their thoughts through comments. Some media sharing sites allow users to create friendship networks.

Social Health applications strategically use various social media tools in revolutionizing healthcare process and cut costs for both patients and providers by fostering patient communities for psychological support through social networking opportunities, building knowledge portals with vertical search capabilities, and promoting telehealth and telemedicine opportunities. Internet and software giants such as Google and Microsoft have launched health services

[5] www.blogcatalog.com

[6] www.twitter.com

Google Health and Microsoft Health Vault providing interfaces from mobile devices to the cloud.

Social Bookmarking sites allow people to tag their favorite webpages or websites and share it with other users. This generates a good amount of metadata for the webpages. People can search through this metadata to find relevant or most favorite webpages/websites. People can also see the most popular tags or the most recently used tags and emerging website/webpage in terms of user popularity. Some social bookmarking sites like StumbleUpon[7] allow people to create friendship networks.

Social News sites help people share and promote news stories with others. News articles that receive positive votes emerge as the popular news stories. People can tag these news stories as well. They can search for most popular stories, fastest upcoming stories for different time periods, and share their thoughts by commenting.

Social Collaboration applications are publicly edited encyclopedias. Anyone can contribute articles to wikis or edit existing ones. However, most of the wikis are moderated to protect them from vandalism. Wikis are a wonderful medium for content management, where people with very basic knowledge of formatting can contribute and produce rich information sources. Wikis also maintain history of all the changes to a page and aree capable of rollbacks. Popular wiki sites like Wikipedia[8] also allow people to classify articles under one of the following categories: Featured, Good, Cleanup, and Stub.

Social Games offer a medium for individuals to express their interactions and provide an opportunity for researchers to gain detailed insights into their behavior in a simulated environment. These could be casual games where individuals play to achieve an objective governed by incentives or could be as complicated as massively multiplayer online role-playing games (MMORPGs) where users can self-portray as avatars, create objects, and interact with other individuals and objects. Avatars can be displayed in various forms including text, two-dimensional, and three-dimensional images with rich graphics and intricate detail. The almost real like nature of these virtual interactions offers ways which were previously impossible to simulate and explore the unexplained landscape of human behavioral psychology. Social Games have been used in many diverse areas such as military training, movie theaters, and scientific visualization.

Next, we discuss the influential behavior of individuals along with models to quantify influence.

3 Influence in Social Media

Influence is reflected by the degree to which an individual is able to affect other individuals in the form of shaping or changing their attitudes or overt behavior in a desired fashion with relative frequency [11, 18, 21]. Accordingly, influence is

[7] www.stumbleupon.com

[8] www.wikipedia.org

earned and maintained by the individual's technical competence, social accessibility, and conformity to the social system's norms. Finding influential blog sites in the blogosphere studies how few blog sites influence other blogs [7] and the external world. The blogosphere, however, follows a power law distribution [6] with very few influential blog sites forming the short head of the distribution and a large number of non-influential sites forming the long tail, where abundant new business, marketing, and development opportunities can be explored [4]. Regardless of the blog being influential, there could exist influential bloggers.

Different social media websites provide various types of information including link information and content information. In this chapter, we use a generic model of computing influence scores of individuals based on both link and content information with tunable weights [2]. The choice of model presented in [2] over others like [1] [1] [9] [10] [13] [20] is exercised due to its flexibility to adapt to various social media sites depending on the availability of content or network information. Next we give a brief explanation of the model which uses both content-driven statistics and graph information to identify influential individuals[9]. Some of the desirable properties of an influential individual are summarized as follows:

Recognition: An influential individual is recognized by many. His writings, p are referenced by other individuals. The more influential the referring individuals are, the more influential the referred individual becomes. Recognition is measured through the inlinks (ι). Here ι denotes the set of inlinks to an individual's writings p.

Activity Generation: An individual's capability of generating activity can be indirectly measured by how many comments he receives, the amount of discussion (s)he initiates. A large number of comments (γ) indicates that the individual *affects* many such that they care to write comments, and therefore, the individual can be influential. Some of these comments could be spam which could be eliminated using the existing work in [14] [17].

Novelty: Novel ideas exert more influence as suggested in [12]. Hence, the outlinks (θ) is an indicator of novelty. If an individual's writing refers to many other articles it indicates that it is less likely to be novel. An individual's writing p is less novel if it refers to more influential articles than if it referred to less influential articles.

Eloquence: An influential individual is often eloquent [12]. Given the informal nature of the social media, there is no incentive for an individual to write a lengthy piece. Hence, a long writing often suggests some necessity of doing so. Therefore, we use the length (λ) as a heuristic measure for computing eloquence[10].

[9] Interested readers can find more details in [2].

[10] This property is most difficult to approximate using some statistics. Eloquence of an article could be gauged using more sophisticated linguistic based measures.

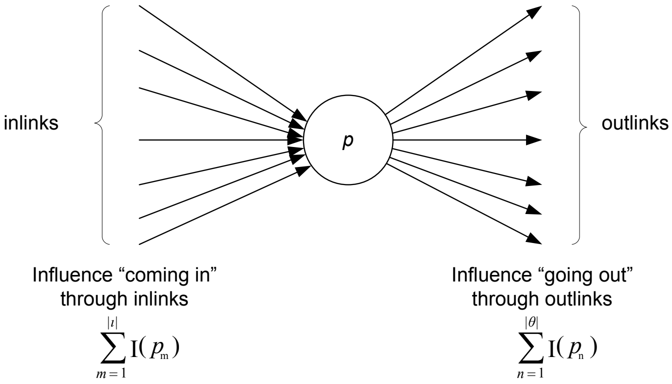
inlinks
p
outlinks
Influence “coming in” through inlinks
$\sum_{m=1}^{|\iota|} \mathrm{I}(p_m)$
Influence “going out” through outlinks
$\sum_{n=1}^{|\theta|} \mathrm{I}(p_n)$

Figure 1: *i-graph* showing the $InfluenceFlow$ across blog post p.

Influence of an individual can be visualized in terms of an influence graph or *i-graph*. Each node of an i-graph represents an individual's writing characterized by the four properties (or parameters): ι, θ, γ and λ. i-graph is a directed graph with ι and θ representing the incoming and outgoing influence flows of a node, respectively. Hence, if I denotes the influence of a node p, then $InfluenceFlow$ across that node is given by,

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n) \tag{1}$$

where w_{in} and w_{out} are the weights that can be used to adjust the contribution of incoming and outgoing influence, respectively. p_m denotes all the nodes that link to p, where $1 \leq m \leq |\iota|$; and p_n denotes all the nodes that are referred by p, where $1 \leq n \leq |\theta|$. $|\iota|$ and $|\theta|$ are the total numbers of inlinks and outlinks of p. $InfluenceFlow$ measures the difference between the total incoming influence of all inlinks and the total outgoing influence by all outlinks of p. From Eq. 1, it is clear that the more inlinks a node acquires the more recognized it is, hence the more influential it gets; and an excessive number of outlinks jeopardizes the novelty of the node which affects its influence. We illustrate the concept of $InfluenceFlow$ in the i-graph displayed in Figure 1. This shows an instance of the i-graph with a single node p. Here we are measuring the $InfluenceFlow$ across p. Towards the right of p are the outlinks and inlinks are towards the left of p.

The influence (I) of an individual is also proportional to the number of comments (γ_p) posted on his(her) writing. The influence of p can be defined as,

$$I(p) \propto w_{com}\gamma_p + InfluenceFlow(p) \tag{2}$$

where w_{com} denotes the weight that can be used to regulate the contribution of the number of comments (γ_p) towards the influence of p.

Although there are many measures that quantify the goodness of a writing such as fluency, rhetoric skills, vocabulary usage, and content analysis, for the sake of simplicity, we here use the length of the writing p as a heuristic measure of its goodness. We define a weight function, w, which rewards or penalizes the influence score of a p depending on its length (λ). The weight function could be replaced with appropriate content and literary analysis tools. Combining Eq. 1 and Eq. 2, the influence of p can thus be defined as,

$$I(p) = w(\lambda) \times (w_{com}\gamma_p + InfluenceFlow(p)) \tag{3}$$

The above equation gives an influence score to each writing of an individual. Now we consider how to use I to determine whether an individual is influential or not. An individual can be considered influential if (s)he has at least one influential piece of writing, p. We use the p with maximum influence score as the representative and assign its influence score as the *influence index* or *iIndex* of the individual. For an individual B, we can calculate the influence score for each of B's N writings and use the maximum influence score as the individuals *iIndex*, or

$$iIndex(B) = \max(I(p_i)) \tag{4}$$

where $1 \leq i \leq N$. With *iIndex*, individuals on a social media site can be ranked according to their infleunce. The top k among all the individuals are the most influential ones. Next, we expand our study of identifying influential behavior patterns across multiple social media websites.

4 Data Collection

Collecting data from one social media site is considered to be a straightforward task [19]. In most cases, the social network graphs are collected. In these datasets identities are usually represented using usernames. The identity of users from social media sites and their network information on these sites has been used for tasks, such as movie recommendation [8] and link prediction [16].

In the following subsections, we will present the advantages of cross media information. Then, we will discuss the challenges of collecting corresponding user identities (a mapping between identities) across social media sites, followed by methods to address these challenges.

4.1 Advantages of Cross-Site Information

Individuals use different social media services for varying purposes and exhibit diverse behaviors on every one of them. We use Flickr to share our pictures with friends, Twitter to update our status, Facebook to keep in touch with friends, and Blogs to express our interests, opinions, and thoughts. It is hence evident that by consolidating this complementary information, a more comprehensive profile of an individual can be built. Few existing studies have considered the prospects of utilizing such information on problems in social media, such as recommending new friends and enhancing user experience. Studying behavioral patterns of users across different social media sites has many applications. As an example, if an individual exhibits the same behavior on various social media sites, then it can help predict his or her behavior on other social media websites by studying only one social media site. The activity patterns of individuals can also help explain why some social media sites are likely to get more activity for various groups of people.

4.2 Challenges of Collecting Cross-Site Information

As identified above, there are several advantages of using user information from multiple sites. However, the task of identifying a user across sites is not so straightforward. Websites do not talk to each other and therefore a user has to create separate user credentials on each online social media site. Although many sites now support using credentials from other sites to logon, for example, users on the media sharing website DailyMotion[11] can use their Facebook credentials to log onto the website. But, this information cannot be collected through APIs or other means. Therefore, the task of identifying a user across sites is a challenge.

A simple method for gathering data across social networks is to conduct surveys and ask users to provide their usernames across social networks. Using these usernames, data can be collected across social networks. However, in addition to being expensive, small data size and population sampling are big challenges with this method. Another method for identifying user identities across sites is finding users on these websites manually. Users more often than not provide personal information such as their real names, E-mail addresses, location, sex, profile photos, and age on these websites. This information can be employed to map users on different sites to the same individual. However, even manually, finding users on these sites can be quite challenging. Many users intentionally hide their identities by limiting the amount of personal information they share or by providing fake information, as reported in our prior research [22]. In [3], the authors address a similar problem in the context of cross-social media sites by looking at the content generation behavior of the same individuals on different social media sites. They define this category of users as “serial sharers” and claim that the compression signature of an author of multiple pages on web is unique across all his authored pages. By computing this signature using the measure Normalized Compression Distance (NCD) as described in [5], the authors show that it is indeed possible to identify pages from the same author on the web.

Fortunately, there exist websites where users have the opportunity of listing their identities (user IDs or screen names) on different social networks. Below, we describe two of these websites, the type of information they provide, and our data collection procedures:

[11] http://www.dailymotion.com

Table 2: Information gathered from the selected social media sites

Social media site	No of Users	Profile Attributes
Delicious	8,483	10
StumbleUpon	8,935	13
Twitter	13,819	15

1. **BlogCatalog[12]:** BlogCatalog is a comprehensive directory of blogs that not only provides useful information about various weblogs, but also comprises of different facilities for users to interact within its community. Users in BlogCatalog are provided with a feature called "My Communities". This feature enables users to list their usernames in other social networks.

2. **MyBlogLog[13]:** MyBlogLog is a social network for the blogger community. It provides a popular web widget that many members have installed on their blogs and is essentially a site based on the interactions that are facilitated by this widget. Users have the "My Sites and Services" feature in their profile for listing their usernames on different social networks.

Users on these websites voluntarily disclose their identities from other websites. This provides blog authors with an opportunity to interact effectively on appropriate channels with their readers. Thus, users on these websites have a valid motivation to publish their identities and these identities can be considered to be reliable. For the experiments described in the next section, we collected user indentities of 96,000 users from BlogCatalog. We identified three popular social media sites, viz., Delicious, StumbleUpon, and Twitter. Using APIs when available and screen scraping in other cases, we collected the activity and profile information of the users on these sites. Note that, not necessarily all the users collected from BlogCatalog had usernames in all of these sites. In Table 2, we present a brief overview of the information collected from these sites. Further, it should be recognized that only publicly shared information was collected from these sites. No private or protected information was collected. The dataset was anonymized after collection for privacy reasons.

In the next section, we present preliminary evaluations of the data and envisage future directions to cross social media studies and influence in social media.

5 Studying Influentials across Sites

Past studies have concentrated on identifying these individuals in a single network, however in this book chapter we present a study of their behavior across a some popular online social networks. Influential behavior here refers to their actions towards their network, and the difference in the amount and the type of

[12] http://www.blogcatalog.com/

[13] http://www.mybloglog.com/

their activity. This section will categorize the behavior of influential individuals across various social media sites and attempt to address issues such as, sustenance of influence, differences in the sphere of influence, and differences in the influence homophily across different sites.

5.1 Sustenance of Influence

In this section, we study the tendency of an influential on one network to remain influential on another social network. This is defined as the sustenance of the influence of an individual on one site across other sites where he is also a member. The motivation behind doing this study is to identify if there exists a pattern in the characteristics of those people who are influential in a network. After reading this section a reader will have a better understanding of starting points to search for people who might be influential within a network given some their identifiable information from other networks. This can be very useful in tasks such as finding influential individuals who can help promote a product in a network where accessing the network information of a large number of individuals is not easy.

We investigate user's influence sustenance across a pair of different social media sites. Three pairs of social media datasets are used in this experiment: Delicious $\mathcal{VS}$ StumbleUpon, Delicious $\mathcal{VS}$ Twitter, StumbleUpon $\mathcal{VS}$ Twitter. We capture the sustenance of influence through a influence intersection ratio, which is defined as the proportion of users who have the same influence position (i.e. top 10% of the influence list) across a pair of datasets. As an example of capturing influence sustenance on site A and site B, we first calculate the influence score of each user on both sites. Then, we obtain two ranking lists based on these influence scores. After that, we compute the influence intersection ratio among different proportion of the ranking lists, which is the ratio of users whose influence falls into the top x% ($x \in [10, 20, 30, 40, 50, 60, 70, 80, 90]$) of both ranking lists. Here, a user $\mathcal{U}$'s influence score is defined as:

$$I(\mathcal{U}) = S_d(\mathcal{U}) + S_m(\mathcal{U}) \tag{5}$$

$S_d(\mathcal{U})$ is the degree score of user $\mathcal{U}$, which is the ratio of user $\mathcal{U}$'s indegree over the maximum user indegree. $S_m(\mathcal{U})$ is the message score of user $\mathcal{U}$, which is the ratio of message amount published by user $\mathcal{U}$ over the maximum message amount published by other users.

To evaluate the influence behavior, we compare the influence intersection ratio with the null model. Here the null model consists of two shuffled lists of users not ranked by any measure and then compute the intersection ratio for users in these lists. The results are presented in Figure 2, Figure 3, and Figure 4. We observe that the observed influence intersection ratio is always higher than the null model on the three social media sites studied. Our results indicate that a user who is influential on one site has a tendency to be influential on other sites where he is a member, as well.

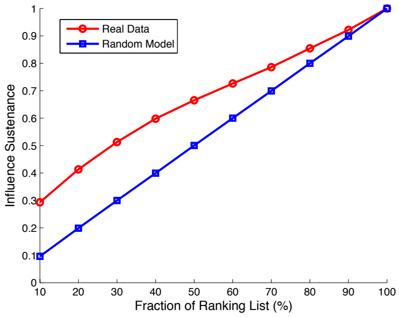

Figure 2: Delicious VS StumbleUpon

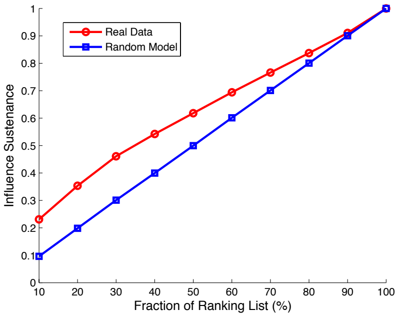

Figure 3: Delicious VS Twitter

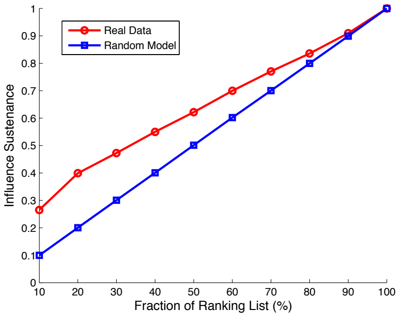

Figure 4: StumbleUpon VS Twitter

5.2 Sphere of Influence

The sphere of influence of an influential individual, would be his closest connections within a network. The influence of a user can be assumed to be strongest on those users who are directly connected to the user. Previous studies [10] have modeled the diffusion of a user's influence beyond his neighbors. However, in this study we will concentrate our efforts on the immediate neighbors of the user. Intuitively, we expect to observe a significant overlap in the spheres of influence of a user across social networks. This can be explained by the tendency of a user to connect with his established friends on other networks.

The sphere of influence of a user can be defined differently depending on the nature of the network. Some networks permit the formation of directed links, such as Twitter, where a person whom you consider your friend may not reciprocate the same feeling towards you. Other networks only consist of undirected links where the feeling of friendship towards another individual is considered to be mutual, as long as both users agree to connect. Given this context, The sphere of influence of an individual would consist of his friends(outlinks) in a directed network and his contacts in the case of an undirected network. The

Table 3: Sphere of influence on Delicious, StumbleUpon and Twitter

Dataset	**Delicious**	**StumbleUpon**	**Twitter**
Delicious	-	0.2031	0.4241
StumbleUpon	0.0166	-	0.3101
Twitter	0.0058	0.0422	-

goal of this part of the study is to analyze the retention of a user's network when we observe him on different sites.

In this section, we observe the overlapping of a user's friendship sphere on both sites. As before, three pairs of datasets are tested. We are interested in a user's common friends on both sites, i.e., how many friends of user $\mathcal{U}$ on site A are also his friends on site B. To get this property, we first extract a user $\mathcal{U}$'s friends on the two sites A and B, namely, $F_A(\mathcal{U})$ and $F_B(\mathcal{U})$. We then calculate the intersection of $F_A(\mathcal{U})$ and $F_B(\mathcal{U})$, represented by $C_{AB}(\mathcal{U})$, which consists of the common friends of user $\mathcal{U}$ on the two sites. The overlapping ratio is defined as:

$$O_{A\to B}(\mathcal{U}) = \frac{C_{AB}(\mathcal{U})}{size(F_B)} \tag{6}$$

Similarly, the overlapping ratio of user $\mathcal{U}$'s friendship from site B to site A is:

$$O_{B\to A}(\mathcal{U}) = \frac{C_{AB}(\mathcal{U})}{size(F_A)} \tag{7}$$

Note that the ratio is not symmetric because of the difference in the size of a user's network on sites A and B, respectively. The total overlapping ratio from site A to site B is:

$$O_{A\to B} = \frac{1}{n}\sum_{i=1}^{n} O_{A\to B}(\mathcal{U}_i) \tag{8}$$

Here, n is the total number of users on site A. The results are shown in Table 3. The overlapping ratio is very low especially on Twitter and Delicious. It indicates that people tend to have different friends on different sites, due to the different site styles. For example, Delicious may contain more information about life and entertainment, while Twitter may serve as a real-time news channel. Therefore, a user's friends on these sites likely match his interests and the style of the site.

5.3 Influence Homophily across Different Sites

In this section, we investigate the influence of an influential user's friends. We want to anaylize whether an influential user tends to connected to influential friends on a social web site. Furthermore, we want to investigate whether an

Table 4: Influence Position Across Two Sites

Dataset	**Delicious**	**StumbleUpon**	**Twitter**
Delicious	0.7289	0.8974	0.7999
StumbleUpon	0.7193	0.9016	0.7928
Twitter	0.7516	0.8843	0.8249

influential user on one web site tends to connect to influential friends on other sites. We define this phenomenon as influence homophily. A individual influence hompliphy is a directed property of a user $\mathcal{U}$ from site A to site B. That is, for every influential user $\mathcal{U}$ on site A, we compute the proportion of his influential friends on site B. Then we get the average influence homophily from site A to site B by averaging all the influential users on site A. Here, user $\mathcal{U}$ is considered to be influential on a site as long as his influence score exceeds the average influence score of users on that site. An individual's influence homophily from site A to site B is:

$$IH_{A \to B}(\mathcal{U}) = \frac{IF_B(\mathcal{U})}{size(F_B(\mathcal{U}))} \tag{9}$$

$IF_B(\mathcal{U})$ is the number of user $\mathcal{U}$'s influential friends on site B, and $F_B(\mathcal{U})$ is user $\mathcal{U}$'s friends on site B.

The results are shown in Table 4, all the influence homophily scores are higher than 70% for Delicious, StumbleUpon and Twitter datasets, which indicates that an influential user on site A tends to have influential friends on site B. In the particular situation that A=B, i.e. A and B are the same site (diagonal entries of Table 4), it indicates an influential user on one site tends to also have influential friends on that site, which is consistent with homophily that explains the tendency of individuals to associate and bond with similar others.

So far, we've observed that an influential user $\mathcal{U}$ on site A is likely to be influential on another site B in Section 5.1. In Section 5.2, we observed that an individual generally doesn't have many common friends across two social media sites. From the diagonal entries of Table 4, we find that an influential user is likely to be connected to other influential individuals on a site. Therefore, the results of the non-diagonal entries of Table 4 indicate that influence homophily does exist, whereby, influential users are more likely to be connected to other influential users of the network and even across network. This observation can be used to design advertising strategies in virtual marketing, in whom only a selective set of influential users need to be targeted to propagate the news of the product across sites through their network.

6 Conclusions

In this chapter, we studied the influential behavioral patterns of individuals spanning across multiple social media websites. Although the problem of identifying influential individuals has been extensively studied in sociology literature, the problem is relatively new in the context of online/virtual spaces, especially social media. These peculiarities specific to online environments is introduced in the chapter along with models to quantify influence. We introduce a formal definition of influence and also propose a model that uses user content and network information to measure the influence of an individual.

This chapter primarily focuses on a new avenue of research, namely, cross site study of behavior in online social media websites. It is known that individuals use different social media sites for different purposes. Some individuals are active on a few social media sites and some are active on many of them. Essentially they try to be sociable or gregarious by making friends on these social media sites. Another type of behavior that can be observed on these social media sites is the influential behavior. Individuals try to lead the community or the conversation in the social media sites. In this chapter, we focus on the influential behavior exhibited by individuals across multiple social media sites.

We provide a novel approach to address the challenges of cross site data collection through the use of blog directory sites, where users voluntarily provide their online identities. We introduce the idea of cross site study to the problem of analyzing behavior of influential individuals through a case study on three popular social media sites, viz., Delicious, StumbleUpon, and Twitter. From our study, we conclude that:

1. Influential individuals have a higher probability to remain influential at most of the sites they are a member of.

2. The principle of homophily, especially the 'influence homophily' exists in the formation of ties on social media sites.

3. Influential individuals are more likely to befriend other influential individuals.

Influential individuals are also commercially important nodes in a network because of their information diffusion capabilities. Analyzing the influence of these influential nodes across social media sites gives us a good starting point in the analysis of an unknown social media site. The study could have far-reaching implications on targeted advertising and social customer relationship management (Social CRM) at large. We envisage the challenges, opportunities, analysis, and findings presented in this chapter will open doors for innovation in this burgeoning area of social media analytics, especially across social media studies with significant contributions to various disciplines such as, computational sociology, cultural anthropology, and behavioral psychology, among others.

7 Acknowledgments

This research was funded in part by the National Science Foundation's Social-Computational Systems (SoCS) Program within the Directorate for Computer and Information Science and Engineering's Division of Information and Intelligent Systems (Award numbers: IIS - 1110868 and IIS - 1110649) and by the U.S. Office of Naval Research (Grant number: N000141010091). We gratefully acknowledge this support.

References

[1] Eytan Adar and Lada A. Adamic. Tracking information epidemics in blogspace. In *WI '05: Proceedings of the The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 207–214, Washington, DC, USA, 2005. IEEE Computer Society.

[2] N. Agarwal, H. Liu, L. Tang, and P.S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pages 207–218. ACM, 2008.

[3] Einat Amitay, Sivan Yogev, and Elad Yom-Tov. Serial sharers: Detecting split identities of web authors. In *Workshop on Plagiarism Analysis, Authorship Identification, And Near-Duplicate Detection*, Amsterdam, Netherlands, July 2007.

[4] C. Anderson. *The long tail: Why the future of business is selling less of more*. Hyperion Books, 2008.

[5] Rudi Cilibrasi and Paul M. B. Vitanyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, April 2005.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262. ACM, 1999.

[7] K.E. Gill. How can we measure the influence of the blogosphere. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, New York*. Citeseer, 2004.

[8] J. Golbeck and J. Hendler. Filmtrust: Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer communications and networking conference*, volume 96. Citeseer, 2006.

[9] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A.s Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

[11] E. Katz. The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, 21(1):61–78, 1957.

[12] Ed Keller and Jon Berry. *One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influentials.* The Free Press, 2003.

[13] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[14] P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[15] R. Kumar et al. On the bursty evolution of blogspace. In *12th International Conference on World Wide Web*, 2003.

[16] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[17] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Transactions of the Web*, 2(1), 2008.

[18] R.K. Merton. *Social theory and social structure.* Free Pr, 1968.

[19] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, page 42. ACM, 2007.

[20] Mathew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. 2002.

[21] E.M. Rogers and F.F. Shoemaker. Communication of innovations; a cross-cultural approach. 1971.

[22] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM09)*, 2009.