1-1-2012

Identification of Influential Social Networkers

Magdalini Eirinaki
San Jose State University, magdalini.eirinaki@sjsu.edu

S. P. Singh Monga
San Jose State University

S. Sundaram
San Jose State University

Recommended Citation

Magdalini Eirinaki, S. P. Singh Monga, and S. Sundaram. "Identification of Influential Social Networkers" *International Journal of Web Based Communities* (2012): 136-158. https://doi.org/10.1504/IJWBC.2012.046256

Identification of Influential Social Networkers

Magdalini Eirinaki, Sumit Pal Singh Monga, Shreedhar Sundaram

April 11, 2012

Abstract

Online social networking is deeply interleaved in todays lifestyle. People come together and build communities to share thoughts, offer suggestions, exchange information, ideas, and opinions. Moreover, social networks often serve as platforms for information dissemination and product placement or promotion through viral marketing. The success rate in this type of marketing could be increased by targeting specific individuals, called "influential users", having the largest possible reach within an online community. In this paper we present a method aiming at identifying the influential users within an online social networking application. We introduce ProfileRank, a metric that uses popularity and activity characteristics of each user to rank them in terms of their influence. We then assess this algorithm's added value in identifying influential users compared to other commonly used social network analysis metrics, such as the Betweenness Centrality and the well-known PageRank, by performing an experimental evaluation on a synthetic and a real-life data set. We also integrate all three metrics in a unified metric and measure its performance.

Keywords: social networks; influential users; profile ranking; link analysis.

1 Introduction

During the past few years, the Internet world has witnessed a staggering growth of social networking websites. A social network can be defined as "a social structure made up of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, [...] knowledge or prestige."[1]. Such media present features unique to the Web, in terms of inherent connectivity between users, rich user profile information, shared authorship, and high update and interaction rate. All these characteristics provide a platform that can be exploited in order to mine interesting information about the dynamics of users' interactions.

One common type of analysis is the identification of communities of users with similar interests (Perugini et al., 2004; Yang et al., 2007), and within such communities the identification of the most "influential" users (Agarwal et al., 2008; Akritidis et al., 2009; Cai and Chen, 2009). Influential users act as hubs within their community

[1] http://en.wikipedia.org/wiki/Social_network

and thus play a key role in spreading information. This has obvious implications on "word of mouth" and viral marketing, as indicated in recent studies (Domingos and Richardson, 2001; Berry and Keller, 2003; Kempe et al., 2003; Gillin, 2007), which in turn makes influential users important for the promotion and endorsement of new products or ideas. The simplest notion of influence is the number of connections a user has within the community (also known as degree centrality), but there are numerous ways to define influence in such rich social graphs.

In this work, we introduce a new metric, named *ProfileRank*, that generates a ranking of all the users within a social networking application, based on their influence. ProfileRank takes into consideration not only the structural characteristics of the graph (as performed in traditional social network analysis), but also the rich information that is available in virtual social networking applications in terms of the users' popularity and activity, depending on the focus of the specific social network (Tables 1 and 2). We perform an experimental evaluation using both synthetic and real-life data collected from a social networking site and compare the ranking with those generated by two other algorithms traditionally used to identify influential nodes within a network, namely Betweenness Centrality and PageRank.

The results of our preliminary analysis indicate that ProfileRank succeeds in finding important users that are identified by the state-of-the-art link analysis metrics, but also manages to identify other users who might not have very strong connections or a central position in their social network but are very active and/or popular and thus can be considered important in terms of information dissemination. We also show that a unified metric, that integrates both link analysis- and activity-based parameters, seems to be the most well-performing indicator of importance within a social networking application.

The remaining of this paper is organized as follows; in Section 2 we review the related literature; in Section 3 we provide some preliminaries on the algorithms that we employ in our study; in Section 4 we introduce the ProfileRank algorithm, followed by an extensive experimental evaluation in Section 5; finally, we conclude in Section 6 with our plans for future work.

2 Related Work

Studying and analyzing Web 2.0 media, such as social networks, blogs, forums, wikis etc. has gained a big momentum, resulting in an increase of research in the related fields. Among the several facets of these social media, influence (Agarwal et al., 2008; Akritidis et al., 2009; Kempe et al., 2003; Kim and Han, 2009; Kimura et al., 2007, 2008; Song et al., 2007), trust (Golbeck, 2009; Liu et al., 2008; Matsuo and Yamamoto, 2009; Walter et al., 2009), and ranking (Adar et al., 2004; Kritikopoulos et al., 2006; Nakajima et al., 2005; Varlamis and Louta, 2009) are receiving a lot of attention. Although, all three aspects can be successfully combined in the context of social networks (Louta and Varlamis, 2010; Varlamis et al., 2010), we regard trust and ranking as orthogonal to our approach, and thus overview only research works focusing on influence.

Influence in social networks, a topic extensively studied in the pre-WWW era

(Wasserman and Faust, 1994), has again emerged as a research topic. The most straightforward approach is to use (real) social network analysis metrics, such as centrality and prestige. An extensive comparative study of such link analysis measures is given by Musiałet al. (2009). Another approach is to model the identification of influencers as a combinatorial optimization problem: given a fixed number of nodes that can be initially activated or infected, find the set of nodes with maximum influence over the entire network - the one that generates the largest cascade of adoptions (Domingos and Richardson, 2001). Several works build on this Information Cascade (IC) notion proposing various machine learning algorithms (Kempe et al., 2003; Kimura et al., 2007, 2008; Estévez et al., 2007; Richardson and Domingos, 2002). Even though such approaches have been shown to improve over traditional social network analysis metrics, they are solely based on the link structure of social networks, and do not take into consideration other important parameters, such as activity and popularity.

In that vein, researchers have investigated the identification of likely influential bloggers or social networkers through a combination of link analysis techniques and other characteristics. More specifically Song et al. (2007) proposed a modified version of PageRank that takes into consideration the novelty of the content posted by a user (computed using cosine similarity) as well as the user's position in the graph. This work is still heavily based on the structure of the graph and does not directly apply to the social networking domain, since the content posted by users is one out of the many parameters that characterize a user. A similar idea appears in Weng et al. (2010) where the authors propose a variation of PageRank that incorporates the content of micro-blogging user's status.

An extensive overview of how the various activity parameters are correlated to the influence of bloggers has been performed in Agarwal et al. (2008). Based on the work of Berry and Keller (2003), the authors propose that recognition, activity generation, novelty and eloquence are the main properties that need to be measured when calculating the influence of blog posts, and as a result, of their authors. They propose that the influence score of a blog post should be computed as a linear combination of the aforementioned properties and perform an extensive experimental evaluation of the effects and correlations of various parameters related to these properties.

The same problem, that of identifying influential bloggers, is addressed in Akritidis et al. (2009). Whereas this approach is also based on behavioral characteristics of users as expressed by the number of incoming links to a blog post (defined in both approaches as a strong evidence of influence), the authors challenge the approach of Agarwal et al. (2008) who assume that outlinks is an indication of novelty, and also isolate single blog posts to identify influential users. Instead, they present a more unified model, that also incorporates the element of time. According to their approach, an influential is recognized as such if he/she "has written influential posts recently or if its posts have an impact recently". They also stress that high activity level is a good indication when seeking influential bloggers.

As shown from the analysis above, most of the work in identification of influencers within a social network (real or online) is based on extensions of well-known link analysis algorithms and as such, exploit the structural characteristics of the network. The use of behavioral characteristics as an indication of influencers, has been adopted by the real world to address the problem of identifying influential bloggers. Only

recently, a smaller-scale analysis has been performed in Kim and Han (2009) focusing on social networking applications. In this work the authors propose five activity criteria to rank the users. The algorithm they use is a weighted sum. They then experiment how a game dissemination happens through the users of a specific social network, comparing their approach to the rankings computed by using degree centrality (i.e. the number of connections/friends in a graph).

Our approach focuses mostly on social networking applications, such as Facebook[2] and MySpace[3] and incorporates many parameters related to the user's activity or popularity. Compared to the work of Kim and Han (2009), we propose a much broader metric in terms of parameters it incorporates. Blogs are not included in our analysis since they present a few different characteristics, and they have already been extensively studied in the related literature, as described before (Agarwal et al., 2008; Akritidis et al., 2009). Yet, we build upon the same ideas of Agarwal et al. (2008), Akritidis et al. (2009) and Kim and Han (2009) following the approach of a weighted function. Moreover, we perform an experimental evaluation comparing our metric to two very well known and used social network analysis metrics, and go one step further by incorporating all three in one overall ranking scheme, again comparing it to the three distinct rankings.

3 Preliminaries

Social network analysis is the study of social entities (*actors*) and their interactions and relationships. The interaction and relationships are represented as a graph, where each node represents an actor (user), and the edge between two nodes represents their relationship. Several link analysis algorithms have been proposed, that are applied on such graphs in order to identify and analyze the role, position, and influence of each user.

In our work, we compare and combine our approach, which is based on popularity and activity characteristics of each user, with two well-known social network analysis metrics, namely Betweenness Centrality and Rank Prestige. *Centrality* identifies as important actors (i.e. users) those that are linked (i.e involved) extensively with other actors. *Prestige* is a more refined measure since it differentiates between in-links and out-links, focusing on in-links. In other words, the importance of an actor depends on the opinion of other actors, expressed by their ties to him/her. More specifically, we are interested in *rank prestige*, that also takes into account the prominence of individual actors that participate in this "voting" process, and more particularly the PageRank algorithm (Page et al., 1998).

3.1 Betweenness centrality.

Betweenness Centrality, further denoted as $BCentrality(i)$, signifies the importance of user i with regards to the flow of information in the social network. If the user is between two non-adjucent users j and k then i has control over their interactions.

[2]http://www.facebook.com
[3]http://www.myspace.com

If i is on the paths of many such interactions (i.e. *between* many users), then this is an important user. In essence, a user with high Betweenness Centrality is an important factor determining what information is spread to which other users, and how fast (since it is up to the user to decide whether to forward some piece of information or not). Let sp_{jk} be the number of shortest paths between j and k, and $sp_{jk}(i)$, ($j \neq i$ and $k \neq i$) be the number of shortest paths that pass i. Betweenness centrality of a user i is defined as follows (Liu, 2007):

$$BCentrality(i) = \sum_{j<k} \frac{sp_{jk}(i)}{sp_{jk}}. \quad (1)$$

3.2 PageRank.

PageRank (Page et al., 1998) also identifies "authorities" in a graph. The intuition is that the more actors "endorse" or vote for an actor i (i.e. add a link pointing to i), the more important i is. What is more, prominence of the endorsers is crucial, since the vote of important actors is more valuable. Transferring this notion to the social network paradigm, a user i is considered to be important, or influential (i.e. has a high PageRank score), if a) many other users endorse i (for example by "trusting" i, adding i's blog in their blogroll, or becoming i's followers), and b) these users are in turn influential. The PageRank score $PageRank(i)$ of user i is iteratively computed as follows:

$$PageRank(i) = (1-d) + d \sum_{(j,i)\in E} \frac{PageRank(j)}{O_j}, \quad (2)$$

where O_j denotes the number of out-links of node j and d is the so-called damping factor. In our approach we employ a variation of the PageRank algorithm, adjusted for undirected graphs, since in the social networking applications we are analyzing, the "friendship" link is undirected. We should note at this point, that we assume "friendship" to be the explicit connection between two users, and not any implicit interactions, such as posting comments, liking content, etc.

4 ProfileRank

According to Berry and Keller (2003), one is influential if she is recognized by fellow members of the community, is an activity generator, has novel ideas or perspectives and is eloquent. Those characteristics have been previously embedded in the works of Agarwal et al. (2008) and Akritidis et al. (2009) in the context of identifying influential bloggers/blogs.

Our perspective is broader in the sense that we are interested in covering social networking applications in general. Thus, we translate the aforementioned properties as follows:

- *Recognition*. This property translates to the notion of *popularity* in social networking applications. Several parameters contribute to defining the popularity

of a person within his community in a social network. Such parameters are for example the number of friends, the number of profile views, or the number of application requests one gets.

- *Activity Generation*. *Activity* parameters indicate how actively a person is participating in a social network. There are several ways for a user of a social networking application to generate activities, and these differ depending on the specific context. Such parameters are for example the number of status updates, the time of last login, or the number of installed applications.

- *Novelty*. This property is the most hard to measure in the context of social networking applications. Agarwal et al. (2008) measure the novelty of blog posts as reversely analogous to the number of outlinks of the post, whereas Akritidis et al. (2009) argue that outlinks are not relevant to the post's novelty and all links should have a single semantic, that of implying endorsement (influence). Due to the diversity of the applications we aim at covering, we also opt for not embedding the element of novelty in our model.

- *Eloquence*. Again, this property is very difficult to translate in the Web context (Agarwal et al., 2008). Status messages in virtual social networks and microblogging sites, as well as comments are most often short, as implicitly directed by the nature of such applications. We may, however, implicitly infer as eloquent a user who is regularly updating his/her status, writing comments to other users, etc. The proposed model may thus incorporate the notion of eloquence by giving more importance to the related parameters.

Based on the above discussion, we take into consideration two categories of profile characteristics in order to generate a ranking of users based on their influence in the network, namely the *popularity* and the *activity* of the user. In what follows we discuss our intuition and give a brief description of all parameters that should be measured and are supported by our system. We explain in more detail the calculation of the non self-descriptive parameters.

4.1 Popularity parameters

Number of friends (p_f). The first sign of trust to a user's opinion is evident in acceptance of the friend request. This number is also a good measure of the outreach a person has in a community and thus is a good parameter to measure influence.

Communities/groups outreach (p_g). The more communities one is subscribed to, the higher the possibility to reach out to a greater number of people. The effect is amplified analogously to the number of users subscribed to these communities. Hence this is a good parameter to determine the influence of a user. This parameter is affected by the total number of users across all communities the user has subscribed to (normalized by the total number of users subscribed to all communities in the database).

Comments on posts (p_c). The comments are direct indication that people were influenced by the user's actions and so have commented on the user's posts. The score is

computed as the total number of comments c_i user i receives by the total number of i's friends f_i:

$$p_c(i) = \frac{|c_i|}{|f_i|}. \tag{3}$$

In this way, a friend that posts many comments on the same item will not affect drastically the score.

Views of posts (p_v). The number of people spending time to view one's blog/videos/photos/links is direct indication of one's influence. We calculate the influence score of this parameter as the average number of views v_i user i receives over all the media m_i she has posted:

$$p_v(i) = \frac{|v_i|}{|m_i|}. \tag{4}$$

For instance, a user who has posted 100 videos and received a total of 120 views has an average of 1.2 view, whereas a person who has posted 1 video and received 10 views has an average of 10 and is more influential.

Testimonials (p_t). A testimonial is considered to be a stronger parameter of trust and relationship. A high number of testimonials shows that the positive influence one has on the community is very high. Most social networking applications offer nowadays the ability to endorse a user's post/comment/etc. For example, Facebook has the "Like" feature, Google+ has the "+1" feature, and Twitter has the "favorite" feature. We calculate this parameter as the ratio of friends who have given a testimonial (t_i) over all the user's friends:

$$p_t(i) = \frac{|t_i|}{|f_i|}. \tag{5}$$

Number of messages per friend (p_m). The number of messages back and forth between two friends is an indication of how active and responsive their friendship is. The more active a friendship is, the higher the possibility for a message to be passed and adopted.

Number of profile views over a period of time (p_{pv}). One's influence is directly proportional to the number of profile views one is getting, since it shows how many people are interested in the user's opinion/views.

Number of active contacts/friends (p_{af}). The number of active contacts indicate the real reach one has in the community. Contrary to degree centrality that regards all contacts as similarly important, this parameter only considers the active ties within a network.

Number of application requests received (p_a). People usually try to find new ways of social interactions like new applications. For instance, social networking sites have become platforms for collaborative online gaming. Getting more requests usually shows that many people want to interact more with the user.

User ratings (p_{ur}). Social networking sites such as Orkut allow users to rate other users ("karma ratings"). Others, such as LinkedIn, allow users to write lengthy recommendations for other users in their circles. We consider such ratings to be direct

measures of popularity. For instance, a user endorses another user (becomes a "fan") only when he is impressed by that person's work or interested in that person. The score is calculated by dividing the total number of user's i fans $|fan_i|$ by the total number of his/her friends:

$$p_{ur}(i) = \frac{|fan_i|}{|f_i|}. \tag{6}$$

Similar metrics can be calculated for other types of ratings.

Quality of friendship (p_q). Quality of friendship is a direct measure of popularity. For example, Orkut allows a person to categorize a friend as "Friend", "Good friend", "Best friend", etc. If a lot of friends have added the user as a best friend, then there is a greater probability the user can influence those friends.

Responses received on posted questions (p_r). The more the responses per question, the more impacted the users are in that community and hence this is a good measure of influence. However, we don't want to allow users who respond more than once on a question to affect the overall score. Thus this score reflects the total number of distinct users $|u_{q_i}|$ who answered on each question q_i:

$$p_r(i) = \sum_q |u_{q_i}|. \tag{7}$$

Public vs. private profile (p_{pr}). A public profile has more chances of being viewed and so such users can reach a lot of people. Hence this parameter has been included in the influence calculation.

4.2 Activity parameters

Number of posts (a_p). The more a user posts, the more she has a chance to influence others. Activity like updating photo albums, videos etc. or posting a link to an article immediately catches attention and is a good way to make the user's presence felt in the network.

Number of questions posted (a_q). The more questions one posts and more answers they get, the more one is in contact with the remaining network. This also indicates the active status of a user who is in touch with friends who can answer his questions.

Number of status updates (a_{su}). The more frequent the status updates, the more probable that the community will get to know the user and get influenced. This has a great effect in microblogging, where "followers" can be updated about every activity several times throughout the day. This clearly indicates that a user is highly active on the network, thus increasing the probability of his influence again. Moreover, a high number of status updates is also an indication of an eloquent user as previously discussed.

Number of applications installed (a_a). The more applications installed, the higher the probability the user sends an application request. A user who has a large number of applications installed can be considered active and also affects others in doing so.

Number of application requests sent by the user (a_{ar}). A user who sends out a large number of application requests is actively participating in the network and might also convince others to use these applications if influential enough.

Number of profile updates (a_{pu}). Some of the information in a profile can be updated from time to time to making a profile look more attractive and interesting. It is also an indication of an active user.

Last login time (a_t). This score indicates when the user was last online thus giving us the frequency of usage and activity.

4.3 ProfileRank metric

As shown in Tables 1 and 2, the various profile parameters we regard as indicators of a user's influence cover the features of most of the social networking applications currently available. It is evident that since the various social networking sites have overlapping, but not exactly the same features, in order to apply the metric to a specific social network, only the applicable parameters should be used. Moreover, each parameter carries a weight that signifies the importance the analyst wants to assign to it with regards to the rest in finding influentials.

Let $p_1, \ldots, p_n$ and $a_1, \ldots, a_m$ be the popularity and activity parameters used for a specific instantiation of ProfileRank, and $w_1, \ldots, w_{n+m}$ be the respective weights assigned for the particular social network. Then, the profile rank of a user i is given by the following generic formula:

$$ProfileRank(i) = w(e_i) \times f(w(p)P(i), w(a)A(i)). \tag{8}$$

where

$$P(i) = w_1 * p_1(i) + \cdots + w_n * p_n(i), \tag{9}$$

$$A(i) = w_{n+1} * a_1(i) + \cdots + w_{n+m} * a_m(i), \tag{10}$$

and $\sum_{i=1}^{n+m} w_i = 1$. $w(e_i)$ is a weight which rewards or penalizes the ProfileRank of a user depending on the number of his/her status updates a_{su}, thus incorporating eloquence in the overall score. For example, $w(e_i)$ can be defined to be equal to the normalized number of status updates of user i, or take pre-defined discrete values depending on the status update frequency. In case the eloquence is not deemed as an important characteristic for the identification of influencers, then $w(e) = 1$. In turn, $w(p)$ and $w(a)$ are weights that determine the overall importance of popularity and activity parameters respectively. These two weights are not "personalized", i.e. are the same for all users and can be used to adjust the metric depending on the overall objective one has when analyzing the social network's data. For instance, the metric can be adjusted to take into account the popularity of the user when, for instance, activity level is not an important characteristic for the community ProfileRank is applied to, by setting $w(a) = 0$ (or a very small number). On the other hand, there might be cases when activity is especially important with respect to popularity. Think, for example, the case of identifying spam users. Such users have common characteristics in that they're very active (e.g. posting comments) but have usually low number of friends or

Table 1: ProfileRank's popularity parameters in social networking sites (on Oct'11)

	$w(p_f)$	$w(p_g)$	$w(p_c)$	$w(p_v)$	$w(p_t)$	$w(p_m)$	$w(p_{pv})$	$w(p_{af})$	$w(p_a)$	$w(p_{ur})$	$w(p_q)$	$w(p_r)$	$w(p_{pr})$
Facebook	x	x	x	x		x		x	x		x	x	x
Twitter	x	x	x	x		x		x				x	x
MySpace	x	x	x	x		x		x				x	x
Google+	x		x	x		x		x			x	x	x
LinkedIn	x	x	x	x	x	x	x	x				x	x
Orkut	x	x	x	x	x	x	x	x	x	x	x	x	x
Flickr	x		x	x	x			x		x			x
YouTube	x	x	x	x			x	x		x			x

Table 2: ProfileRank's activity parameters in social networking sites (on Oct'11)

	$w(a_p)$	$w(a_q)$	$w(a_{su})$	$w(a_a)$	$w(a_{ar})$	$w(a_{pu})$	$w(a_t)$
Facebook	x	x	x	x	x	x	x
Twitter	x	x	x			x	x
MySpace	x	x	x			x	x
Google+	x	x	x			x	x
LinkedIn	x	x	x			x	x
Orkut	x	x	x	x	x	x	x
Flickr	x					x	x
YouTube	x					x	x

low number of incoming comments and good testimonials. Appropriately adjusting the related parameters can help the analyst identify or eliminate (depending on the overall objective) such users from the top-ranked list. Finally f can be the sum or the product of the two parameters.

It is important to stress here that ProfileRank is a metric that shows who are important users of the social network with respect to their levels of activity and interaction with other users. Thus, the profile-based ranking generated by ProfileRank can be used alone, or in combination with the graph-based social network analysis metrics discussed before. The latter option would result in a score that incorporates both behavioral and structural characteristics of the social network nodes (i.e. the users).

A possible integration of the three can be achieved by computing their linear sum:

$$\begin{aligned} TotalRank(i) = w_P * ProfileRank(i) \\ +w_C * BCentrality(i) + w_{PR} * PageRank(i) \end{aligned} \tag{11}$$

where the respective weights signify the importance we want to attribute to each ranking.

5 Experimental Evaluation

We performed an experimental evaluation using both synthetic and real data. We have evaluated the effect of various parameters and the respective importance assigned to them in correctly identifying the influencers of a social network. We have also performed a series of experiments in order to compare the performance of the three ranking metrics, namely Betweenness Centrality, PageRank, and ProfileRank, as well as the cumulative TotalRank. Finally, we evaluated how changing the parameters' weights affects the final score of each user's ProfileRank. In this paper we present the most interesting results of this preliminary analysis.

5.1 Data sets.

The nature of the problem we are addressing makes the acquisition of data a challenging task; most users have private profiles, whereas the APIs provided by some sites divulge a small amount of popularity-related data, and no activity-related data. Such data is only available to the owners of the social network. In that case, the boundaries of the social structures are clear and the results are expected to be of highest accuracy. However, for our experimental evaluation with a real-life data set, we were limited to collecting information from publicly available profiles. This data set, although appealing, is not complete since most of the users analyzed also connect to users with private profiles for which it was impossible to gather any information. Thus, in order to better understand how the algorithm behaves in different network structures and whether increased popularity and/or activity will affect the ranking that would be generated using structure-based metrics, we created a synthetic social network interconnected in various formations. The details of both data sets are discussed in the sections that follow.

Table 3: Default weight vaules for ProfileRank

Weight	Value	Weight	Value
$w(p)$	1	$w(p_f)$	0.05
$w(a)$	1	$w(p_g)$	0.05
$w(e)$	1	$w(p_c)$	0.07
$w(a_p)$	0.03	$w(p_v)$	0.06
$w(a_q)$	0.015	$w(p_t)$	0.085
$w(a_{su})$	0.04	$w(p_m)$	0.06
$w(a_a)$	0.015	$w(p_{pv})$	0.05
$w(a_{ar})$	0.015	$w(p_{af})$	0.07
$w(a_{pu})$	0.03	$w(p_a)$	0.06
$w(a_t)$	0.04	$w(p_{ur})$	0.07
		$w(p_q)$	0.07
		$w(p_r)$	0.05
		$w(p_{pr})$	0.07

5.2 Small-scale Synthetic Social Network Analysis

In this first set of experiments we created a small social network consisting of 50 users interconnected in various formations (i.e. social structures), as shown in Figures 1 - 6. As we could notice from the graphs, there are various small disjointed groups within the test online social network. By manipulating in advance the structure of the social network, we are able to evaluate the effectiveness of our approach as compared to purely structure-based metrics, such as Betweenness Centrality and PageRank in cases where users might not be well-connected or central but are very popular/active. Thus, we intentionally created popular and active users (e.g. E1, I1, K group's users), whereas we downplayed central (in terms of connectivity) users such as V0 and Z4. Our motivation was to demonstrate that ProfileRank manages to bring to the top of the ranked list users who would otherwise be much lower in the influence ranks based on their position in the social graph. A detailed overview of the synthetic dataset is included in a technical report that cannot be cited due to double-blind constraints.

We applied all three methods (ProfileRank, PageRank, Betweenness Centrality) on the synthetic social network. For the ProfileRank we used the default weights, set empirically (as shown in Table 3), and summed the activity and popularity parameters. Notice that we did not consider eloquence as an important characteristic (setting $w(e) = 1$) and we also gave overall equal importance to popularity and activity parameters (setting $w(p) = w(a) = 1$). However, since the popularity of a person is a good measure of influence, most of the parameters that come under this category carry slightly more weight than the parameters that come under activity. We should stress that the weights shown here are just an indication of our intuition on each parameter's importance and can be altered to reflect different approaches to what is considered important, as discussed in Section 4.3.

We first wanted to verify that the ProfileRank metric behaves as expected. Indeed, as shown in Figure 7 that depicts the top-10 influencers, the metric identifies users E1

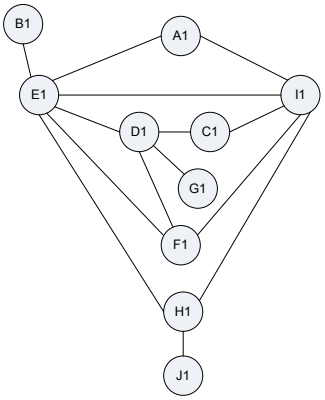

Figure 1: Group A

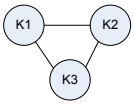

Figure 2: Group B

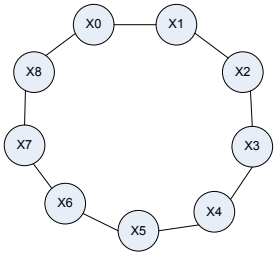

Figure 3: Group C

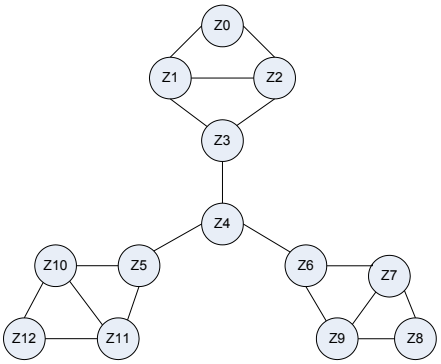

Figure 4: Group D

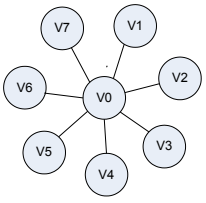

Figure 5: Group E

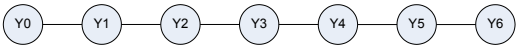

Figure 6: Group F

and I1 as the most influential ones. This is verified by the popularity parameters these users demonstrate, for instance E1 has received testimonials from 4 different users and I1 from 5 different users (note that due to space constraints the complete dataset is not shown here). E1 and I1 have the higher number of friends in their Group (A), high user rating and have received many comments. It is noteworthy that although E1 has received less comments for the media he has posted than I1, he has got comments from different users for different media. Hence E1's influential score for media and the overall ProfileRank is higher than that of I1's.

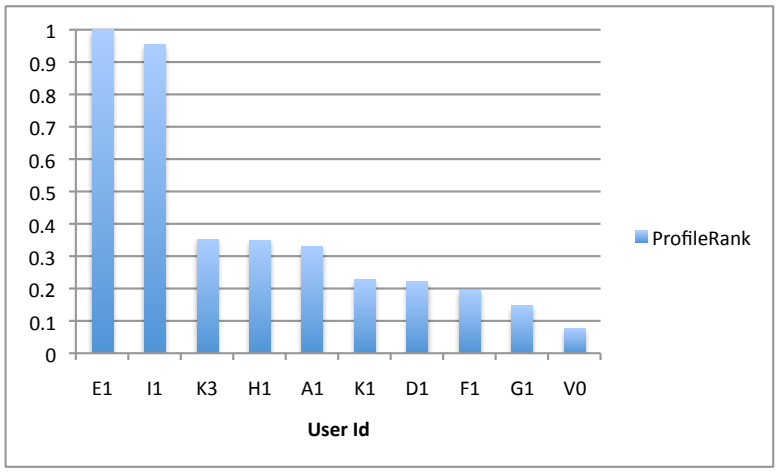

Figure 7: Top-10 influential users according to ProfileRank - Synthetic dataset

We then compared the three approaches. Figure 8 shows the top-10 influential users as computed by each metric. When compared with the rankings generated with PageRank, we observe that there is an overlap of 40% similar users in the top-5 list. On the contrary, the top-5 users according to Betweenness Centrality have no overlap with the other two rankings. The position of the most influential user according to the user's PageRank and Betweenness Centrality score (users V0 and Z4 respectively) as compared to everyone else in the social graph, verifies their high ranking in the two structure-based metrics respectively. None of the two users, however, are in the top-5 list of ProfileRank, since these users presented (intentionally) minimal activity and had low popularity. This demonstrates the need of incorporating activity and popularity characteristics in the ranking function.

Finally, in order to evaluate the effectiveness of all rankings as compared to each one individually, we computed the TotalRank (Equation 11) setting $w_1 = w_2 = w_3 = 1$. The top-5 users are shown in Figure 9. E1 is the most influential user, followed by V0, I1, Z4 and Z5. Although V0 didn't have much activity or popularity, she turned out to be the second most influential user in the network because of the positional advantage (higher PageRank and Betweenness rank). More sophisticated ways to integrate the three rankings can be employed, and we are currently working on this direction.

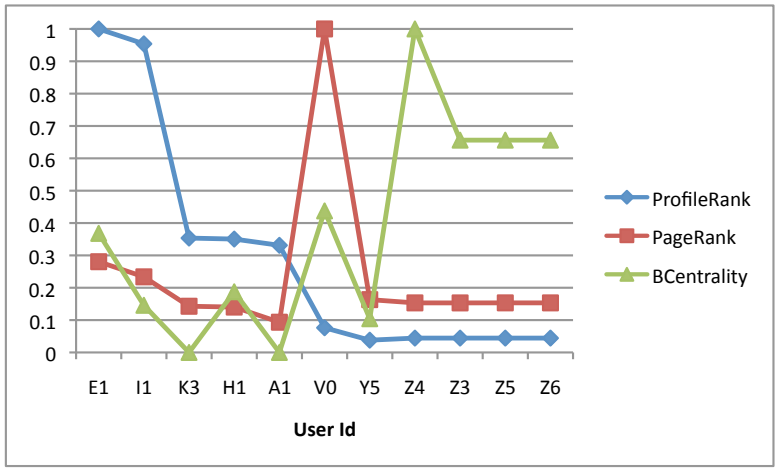

Figure 8: Top-5 influential users of all rankings - Synthetic dataset

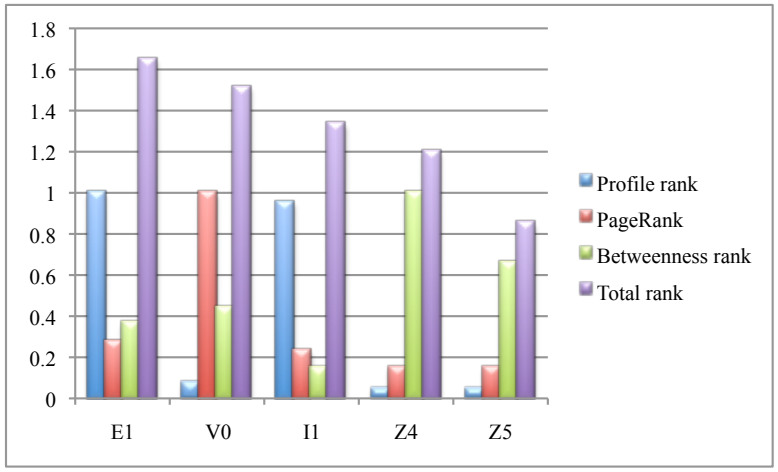

Figure 9: Top-5 Influential Users Rank Comparison - Synthetic Dataset.

5.3 MySpace Public Profile Data

In order to gather real social network data, and due to the absence of any benchmark dataset online, we crawled the MySpace social network, having as initial seed the profiles of about 300 users who had public profiles. We then expanded the network to include their friends reaching a total of 2000 users. We should note here that collecting data from a real web site such as MySpace was a challenging task since the majority of users of such social networking sites have set their profiles to private.

Since some of the profile characteristics for evaluating the popularity and activity parameters are either not available or not visible in MySpace (Tables 1 and 2), we calculated the ProfileRanks based on the available fields having visible values. Unfortunately, there was no straightforward way to compare our findings with a "ground truth" on finding influential users based on their profile for the same reasons described before. Whereas Agarwal et al. (2008) where able to define a baseline using the most commented/reposted blog entries in Digg in order to identify the most influential bloggers, a similar approach is not feasible in the social networks' context since such a "universal" ranking of social networkers cannot be similarly derived. This is due to the nature of social networking (short status updates and comments vs. long blog posts, limited visibility/private profiles vs. public blogs, some reposts/endorsements of existing content vs. original content produced by bloggers etc.).

Thus, we decided to follow a different line of evaluation. We applied the other two metrics, namely PageRank and Betweenness Centrality in order to identify potential correlations and evaluate the added value of the profile-based ProfileRank to the rankings provided by these state-of-the-art structure-based metrics.

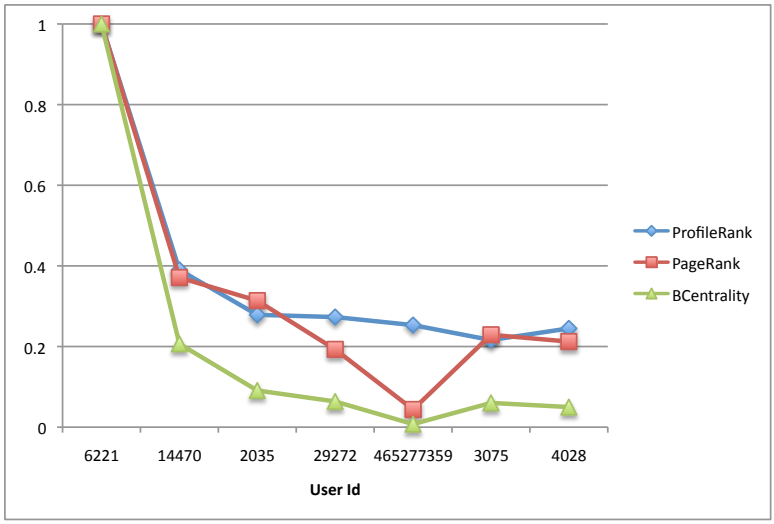

Figure 10: Top-5 Influential Users Rank Comparison - MySpace dataset.

When comparing the ProfileRank ranking with that generated with PageRank, we observe that there is a significant overlap; 100% similarity in the top-3 list, 60% sim-

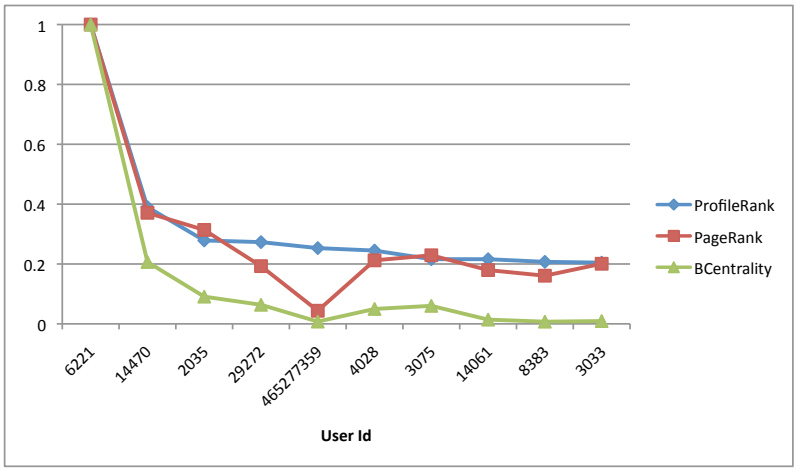

Figure 11: Top-10 Influential Users Rank Comparison - MySpace dataset.

Table 4: Top-5 influential users

User Id	**ProfileRank**	**PageRank**	**BCentrality rank**
6221	**1**	**1**	**1**
14470	**2**	**2**	**2**
2035	**3**	**3**	**3**
29272	**4**	7	**4**
465277359	**5**	24	30
3075	33	**4**	**5**
4028	6	**5**	7

ilarity in the top-5 list and 90% similarity in the top-10 list. On the other hand, Betweenness Centrality ranking has an overlap of 100% similarity in the top-3 list, 80% similarity in the top-5 list and 60% similarity in the top-10 list. Figures 10 and 11 show this comparison. We should note that in Figure 10 we have included the top ranked users of all rankings, whereas in Figure 11 we only show those included in the top-10 ProfileRank ranking.

From the analysis above, It can be seen that all three metrics coincide in the top-3 users. In fact, the top-ranked user receives significantly higher rank than the remaining ones. Table 4 shows the details of the top-5 influential users according to each ranking algorithm (both overlapping and non-overlapping users are included). By looking at the profile of the most influential user online, we observed that this user id corresponds to the founder of MySpace, Tom Anderson, who automatically becomes friend with everyone who joins the social network. Except for being popular and very central, he is also a very active user, sending/receiving messages to/from everyone in the network on a daily basis. This is another indication that our approach manages to identify the influentials of a social network since activity and popularity are equally important

Table 5: Profile parameter weight values for different iterations of testing

Run #	**Comment**
Run 1 (Number of friends less important)	$w(p_f) = 0.015$ (initial value: 0.05)
Run 2 (Media posts more important)	$w(a_p) = 0.08$ (initial value: 0.03)
Run 3 (Media comments less important)	$w(p_c) = 0.03$ (initial value: 0.07)
Run 4 (Messages received less important)	$w(p_m) = 0.02$ (initial value: 0.06)

characteristics as location in the graph. Another interesting finding is that the 5th most influential user (based on the user's ProfileRank) is much lower in the ranks generated by the structure-based metrics. This shows that these metrics perform well but might miss important users that perhaps are not very central/well-connected, however are very active in their social network.

We finally performed some evaluation of the effect the various weights of the activity and profile parameters have in the final ranking. As discussed before, in our initial experiments, we manually assigned weights based on our intuition on which parameters should be given more priority as compared to others. The algorithm allows for each implementation to define application-specific weights. Other methodologies (such as statistic-based heuristics) are also an option for defining the parameter weights.

In this set of results, we demonstrate the effect of the activity and popularity weights in the final outcome of the algorithm. We performed four different runs, changing drastically the weights of some parameters (with the remaining adjusted uniformly from their default values), as shown in Table 5, and observed how the ProfileRank of randomly selected users changes. In what follows, we discuss the findings for two users with regards to each user's profile characteristics. This discussion provides a closer look on how the rank of each user can be affected/tuned depending on their individual profile. The two examples demonstrate that appropriate tuning is needed when applying ProfileRank.

User Id 8581984. Figure 12 shows the ProfileRanks of this user, depending on the different tunings of the algorithm.

Run 1 (Number of friends less important): This run gives the highest ProfileRank overall. The reason for this is that the rank of this user depends less on the number of friends whereas the large number of media comments this user has received increases the user's ProfileRank.

Run 2 (Media posts more important): Run 2 also gives a higher rank than the Original run for this user since the weight for number of media posts almost doubles. This user has posted a significant number of media and thus gets a higher rank.

Run 3 (Media comments less important): The weight for media comments has been decreased in this case and still this run has the same rank as the original run for this user. Since the user has only very few media comments the rank does not get affected much.

Run 4 (Messages received less important): This user has a huge number of messages and the weight for comments is significantly decreased. However, the user's rank is not decreased as compared to the Original run. We assume that this occurs because all ranks are normalized by dividing them with the highest rank. So, we can infer that

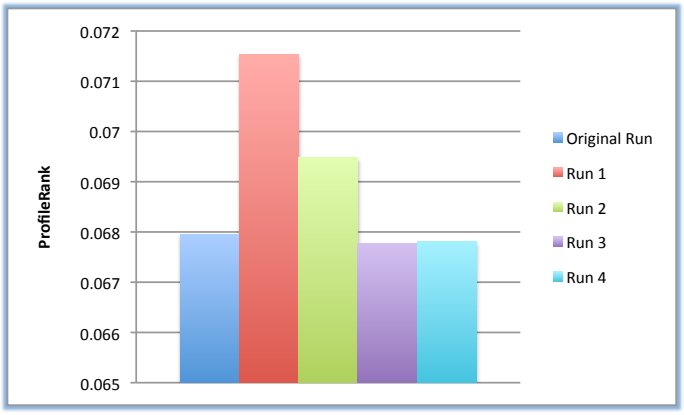

Figure 12: Tuning ProfileRank for User Id 8581984

the highest rank also decreased by the same rate to make the normalized rank score for run 4 similar to that of the original one.

User Id 1048. Figure 13 shows the ProfileRanks of this user, depending on the different runs of the algorithm.

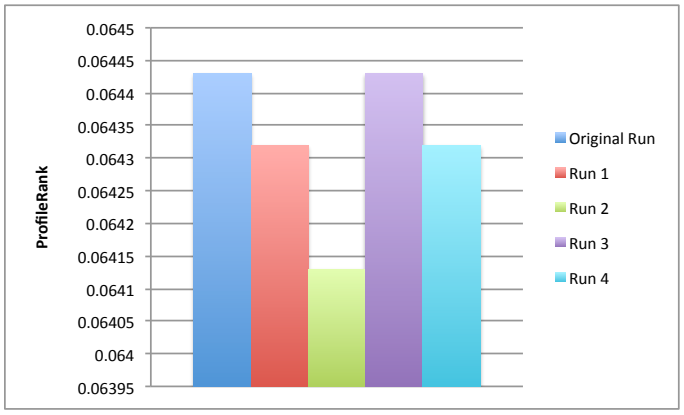

Figure 13: Tuning ProfileRank for User Id 1048

Run 1 (Number of friends less important): This user has a few friends and a few comments that support the user's rank. When the importance of number of friend parameter gets decreased, the rank score also decreases since his other fields already have lower values.

Run 2 (Media posts more important): This user has not posted any media. Increasing the importance of the number of media parameter has no effect on the user's rank. The highest rank overall must have increased due to increase in weight of this parameter, thereby decreasing the normalized rank for this user during this run.

Run 3 (Media comments less important): Decreasing the media comments had no effect on the rank of this user since the user has received no media comments.

Run 4 (Messages received less important): This user has some messages and hence the rank decreases. Normalizing it with a decreased value of highest rank reduces the amount by which it decreased.

6 Conclusions

Social networking applications present characteristics that can be exploited in order to mine interesting information about the dynamics of users interactions. Our objective is to identify the users who have maximum influence in the social network's user base. We define the term "influence" in terms of two properties, namely popularity and activity. Users who are popular with other users have the power to affect group behavior and are also predictive of group choices. If these popular users are active then their actions are constantly being noticed by a big majority of users and there is a high probability these actions would influence their network. Thus, these popular and active users are good candidates of becoming "trend-setters". Such users are the entry point of information diffusion through the network. Information flow and word-of-mouth recommendations could be triggered through this key set of users. Likewise, the inverse is also true since we can learn about group trends and behavior from a few key nodes representative of the whole network.

We are proposing a set of profile-based characteristics that can be used as indications of popularity and activity. We have designed the ProfileRank algorithm that calculates the "influence" scores of users. The weights assigned to each metric are parameters of our algorithm and can be empirically or heuristically tuned. We have performed an empirical evaluation of the added value our profile-based metric provides when used complementary to, or instead of, the structure-based state-of-the-art metrics PageRank and Betweenness Centrality using a synthetic and a real-life dataset. Our main conclusion was that ranks from all the three algorithms give comparative and sometimes complementary ranks. Moreover, the difference of ProfileRank's outcome depending on the parameter weights brought forward an interesting assumption that the algorithm not only takes activity and popularity into account but also can rank based on position of user with some fine tuning of the weights of the parameters. Another interesting observation is that, because of the high degree of parametrization of ProfileRank, it can be tuned appropriately to address other problems, for example finding spam users of a social network.

The results presented in this paper consist an initial study on the proposed algorithm. We intend to perform a more in-depth study of the ProfileRank algorithm, on a bigger user base. We are currently working on extending the TotalRank algorithm to integrate all three algorithms in a more elaborate way. Finally we plan to work on methods for heuristically estimating the best parameter weights depending on the data set that needs to be analyzed and the overall objective of the analysis (e.g. identifying spam users).

References

Adar, E., Zhang, L., Adamic, L., and Lukose, R. (2004). Implicit structure and the dynamics of blogspace. In *Workshop on the Blogging Ecosystem: Aggregation, Analysis and Dynamics, WWW 2004*.

Agarwal, N., Liu, H., Tang, L., and Yu, P. S. (2008). Identifying the influential bloggers in a community. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 207–218.

Akritidis, L., Katsaros, D., and Bozanis, P. (2009). Identifying influential bloggers: Time does matter. In *Web Intelligence*, pages 76–83.

Berry, J. and Keller, E. (2003). *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. Free Press.

Cai, Y. and Chen, Y. (2009). Mining influential bloggers: From general to domain specific. In *KES (2)*, pages 447–454.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66.

Estévez, P. A., Vera, P. A., and Saito, K. (2007). Selecting the most influential nodes in social networks. In *the International Joint Conference on Neural Networks, IJCNN*.

Gillin, P. (2007). *The New Influencers: A Marketer's Guide to the New Social Media*. Linden Publishing.

Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *TWEB*, 3(4).

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *in Proc. of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

Kim, E. and Han, S. S. (2009). An analytical way to find influencers on social networks and validate their effects in disseminating social games. In *ASONAM 2009*, pages 41–46.

Kimura, M., Saito, K., and Nakano, R. (2007). Extracting influential nodes for information diffusion on a social network. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 1371–1376.

Kimura, M., Yamakawa, K., Saito, K., and Motoda, H. (2008). Community analysis of influential nodes for information diffusion on a social network. In *the International Joint Conference on Neural Networks, IJCNN*.

Kritikopoulos, A., Sideri, M., and Varlamis, I. (2006). Blogrank: ranking blogs based on connectivity and similarity features. In *Proceedings of the 2nd international Workshop on Advanced Architectures and Algorithms For internet Delivery and Applications AAA-IDEA'06*. ACM.

Liu, B. (2007). *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Springer.

Liu, H., Lim, E.-P., Lauw, H. W., Le, M.-T., Sun, A., Srivastava, J., and Kim., Y. A. (2008). Predicting trusts among users of online communities: an epinions case study. In *ACM Conference on Electronic Commerce (EC2008)*, Chicago.

Louta, M. and Varlamis, I. (2010). Blog rating as an iterative collaborative process. In et al., M. W., editor, *Semantics in Adaptive and Personalized Services, Springer Series on "Studies in Computational Intelligence" SCI 279*, pages 187–203. Springer-Verlag Berlin Heidelberg.

Matsuo, Y. and Yamamoto, H. (2009). Community gravity: measuring bidirectional effects by trust and rating on online social networks. In *WWW 2009*.

Musiał, K., Kazienko, P., and Bródka, P. (2009). User position measures in social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09.

Nakajima, S., Tatemura, J., Hino, Y., Hara, Y., and Tanaka, K. (2005). Discovering important bloggers based on analyzing blog threads. In *2nd Annual Workshop on the Blogging Ecosystem: Aggregation, Analysis and Dynamics*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Technical Report. Stanford InfoLab*.

Perugini, S., Goncalves, M., and Fox, E. (2004). Recommender systems research a connection-centric survey. *Journal of Intelligent Information Systems*, 23(2):107–143.

Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70.

Song, X., Chi, Y., Hino, K., and Tseng, B. (2007). Identifying opinion leaders in the blogosphere. In *CIKM 2007*, pages 971–974.

Varlamis, I., Eirinaki, M., and Louta, M. (2010). A study on social network metrics and their application in trust networks. In *The 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2010)*.

Varlamis, I. and Louta, M. (2009). Towards a personalized blog site recommendation system: A collaborative rating approach. *the International IEEE Workshop on Semantic Media Adaptation and Personalization, SMAP*.

Walter, F. E., Battiston, S., and Schweitzer, F. (2009). Personalised and dynamic trust in social networks. In *RecSys*, pages 197–204.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *ACM WSDM 2010*.

Yang, B., Cheung, W., and Liu, J. (2007). Community mining from signed social networks. *IEEE Trans. on Knowl. and Data Eng.*, 19(10):1333–1348.