Fine-grained Feature-based Social Influence Evaluation in Online Social Networks

Guojun Wang, *Member, IEEE,* Wenjun Jiang, Jie Wu, *Fellow, IEEE,* and Zhengli Xiong

Abstract—The evaluation of a user's social influence is essential for various applications in online social networks (OSNs). We propose a fine-grained feature-based social influence (FBI) evaluation model. First, we construct a user's initial social influence by exploring two essential factors, that is, the possibility of *impacting* others, and the *importance* of the user himself. Second, we design the social influence adjustment model based on the PageRank algorithm by identifying the influence contributions of friends. For the aim of fine-grained evaluation, based on a feature set which includes the related topics and user profiles, we differentiate the feature strength of users and the tie strength of user relations. We also emphasize the effects of common neighbors in conducting influence between two users. Through experimental analysis, our FBI model shows remarkable performance, which can identify all users' social influences with much less duplication (it is less than 7% with our model, while more than 80% with other degree-based models), while having a larger influence spread with top-k influential users. A case study validates that our model can identify influential users with higher quality.

Index Terms—social influence, feature strength, tie strength, common neighbors, online social networks.

1 Introduction

Online social networks (OSNs) [1] have attracted a lot of attention since they allow users to conveniently share ideas, activities, events, and interests within their individual networks. Participating users join a network, publish their profile and any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users [2].

In OSNs, various applications, such as personalized recommendation [3], viral marketing [4], and expertise discovery [5], have motivated the tremendous attention of social influence. A wide range of potential applications also need the evaluation of social influence, e.g., selecting or evaluating some excellent scientists [6] (or employees, specialists, experts, etc.), either for forming a team of specific aim [7], or for recruiting new members. Therefore, to evaluate the social influence of users is becoming an essential technique.

Motivation. Social influence is becoming a prevalent, complex, and subtle force that governs the dynamics of all social networks [8]. A few state-of-the-art literatures have been proposed. Many useful findings have been made from them, such as the following: different relationships play different roles [9]; the effect of the social influence from different angles may be different [10]; social influence actually exists only when a friendship has been built up [11]; people decide to adopt activities based on the activities of the people they are currently interacting with [12].

However, three challenges remain open: (1) It is still not very clear what factors should be considered to construct social influence; (2) It also lacks the ways to properly integrate those factors in order to evaluate each user's influence efficiently and effectively; (3) It is hard to measure a model due to the lack of ground truth and commonly accepted standard metrics, putting aside the diversity of social network applications, as well as the complexity of the concept of social influence.

Our work in this paper tries to address the above challenges. We focus on exploring the essential factors that should be considered to construct a user's social influence, and present a general framework to integrate these factors; we then provide some rational metrics to measure the efficiency and effectiveness.

Main Ideas. To evaluate a user's social influence, the most important thing is to know (1) what is social influence; (2) which factors may impact a user's influence; and (3) how the impact takes place. Social influence is defined as a change in an individual's thoughts, feelings, attitudes, or behaviors that result from interactions with another individual or group [13]. Generally speaking, an online social network consists of users, social ties or relationships between users, and topics they are involved in. All three parts may impact the social influence of users, with different approaches.

Gaining further insight into what happens in our daily lives, we can see that the social ties/user relationships, common interests, and similar experiences between two users can be taken as explicit indications of influence, which we model as the possibility of impact, or impact

- *G. Wang, W. Jiang, and Z. Xiong are with the School of Information Science and Engineering, Central South University, Changsha, Hunan Province, 410083, P. R. China. E-mail: csgjwang@csu.edu.cn, wenjj8a@gmail.com. G. Wang is the corresponding author.*
- *J. Wu is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA. E-mail: jiewu@temple.edu.*

Digital Object Indentifier 10.1109/TPDS.2013.135

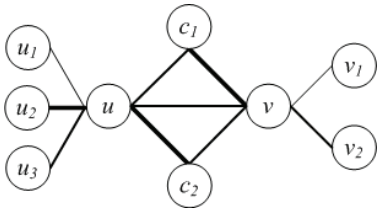


Fig. 1. A simple social network, in which some nodes are common neighbors of two other nodes (e.g., c_1 and c_2 to u and v).

for short. Moreover, there is another essential but implicit indication of influence, that is, the personal importance of the user himself. It is usually overlooked by existing models, but does exist in real life. For instance, in real life, the research experts, the leaders, and the presidents are usually more powerful to influence other people.

We take each related item in user profiles or the topics they are involved in as a *feature*. We further present the concepts of feature strength and tie strength to identify the fine-grained user influence.

We also analyze the interpersonal structures between users. In user-degree based models, all neighbors are taken equally, which is not consistent with real life. We emphasize the effects of the common neighbors. Taking Fig. 1 for instance, when considering the possibility of impact between u and v, in spite of their direct affinity, we believe that the two common neighbors, c_1 and c_2, have more chances and strength to conduct influence between u and v (than other friends of u_1, u_2, u_3, v_1, and v_2), which we model as indirect affinity.

Moreover, as we may find in our physical world, a person's influence can impact his friends' influences, and vice versa. Similar patterns exist in OSNs. Thus, the overall influence of a user should reflect the total influence in corresponding aspects to corresponding friends.

Contributions. We propose a novel model to evaluate a user's feature-based social influence, called FBI for short. Our goal is to develop a fine-grained model, which shows what the social influence of each individual is on a given feature set. Our contributions are as follows:

i) To the best of our knowledge, we are the first to extract the two essential parts of social influence: the possibility of impact between two users, and the importance of each user, himself. We analyze and construct each part. We also conduct experiments to discover how each part affects the social influence.
ii) We differentiate the feature strengths of users and the tie strengths of edges in a social network, based on the features being considered. In addition, we emphasize the effects of common neighbors between two users. Moreover, we design the influence adjustment model based on the PageRank algorithm by identifying friends' contributions.
iii) We identify three metrics to measure a model, which are extracted from related research. We take an evaluation model to be effective if it has less duplication of the value of influence, a larger influence spread with top-k influential users to influence more users, and higher identification accuracy to identify influential users.
iv) We evaluate FBI using three data sets: *HEPTH* [14], *DBLP* [15] and ArnetMiner [16]. Experimental results show that our model can evaluate all users' social influences with less duplication (it is less than 7% with our model, while more than 80% with the user-degree based models), while having a larger influence spread. A case study with the ArnetMiner coauthor data set demonstrates that the FBI model can identify influential users with high accuracy.

The remainder of this paper is organized as follows: Section 2 surveys related work. Section 3 states the problem we address, and presents the overview of our approach. Section 4 describes the details of FBI. Section 5 describes the experimental evaluation. Finally, Section 6 concludes this paper and suggests future work.

2 Related work

A common approach to identifying influential users is to analyze the social network structures [17]. Opsahl et al. [18] presented an evaluation model of a user's reputation, based on degree centrality. Newman [19] discovered the very notion of influential users that is closely related with closeness centrality. Katona et al. [20] presented an evaluation model of a user's reputation using betweenness centrality. How the tie strength relates to influence and information diffusion was studied in [21]. In this paper, we also explore the local topology information to construct user influence.

Fei et al. [11] identified a new factor of social influence, i.e., the "gravitation" between users, which they called "user attractor." Tang et al. [8] proposed a quantitative measure of topic level influence. In this paper, we consider a much broader concept, which we call the feature. It can include all the topics being considered, as well as items in user profiles such as gender, age, special interests, and so on.

Crandall et al. [12] studied the feedback effects between similarity and social influence in online communities. The authors recommended a future direction of combining the two factors, which motivates our work. In our previous work [22], the indirect similarity via common neighbors was used to construct the initial social influence.

In this paper, we aim to identify the factors of the features (which can be chosen from topics and user profiles), the local network topology, and the similarity, as well as the personal importance, to construct a comprehensive social influence evaluation model.

3 Overview

The goal of social influence evaluation is to derive the feature-based social influence, based on the input net-

TABLE 1
Notations.

SYMBOL	DESCRIPTION
$G = (V, E)$	online social network
$u/v/e(u, v)$	node u, v, and the edge between the two nodes
F	$\{f_i \mid i \in [1, n]\}$, the feature set
$\overrightarrow{F_u}$	the feature strength vector of u
t_{uv}	the tie strength of $e(u, v)$
N_u	the neighbor set of u
C_{uv}	the set of common neighbors of u and v
$\overrightarrow{S_{uv}}$	the similarity
$\overrightarrow{A^d_{uv}}$	the direct affinity
$\overrightarrow{A^{id}_{uv}}$	the indirect affinity
$\overrightarrow{I^{in}_{uv}}$	the impact between u and v
$\overrightarrow{I^{out}_u}$	the importance of u
w_u/I^F_u	the initial/general feature-based influence of u

work and the features being considered. We first introduce some terminologies, and then define the social influence evaluation problem. The notations are described in Table 1; and all the variables are normalized into the range of [0,1].

Terminologies. A social network is modeled as an undirected graph $G = (V, E)$, with V indicating the users in the network and E indicating the social ties/relationships between users. Two users are taken as neighbors if there is an edge between them.

According to the features being considered, a *feature set* can be determined, which contains the related topics or user profiles. Based on this, we present the concepts of *feature strength* and *tie strength*, representing the strength of a user on a specific feature, and the strength of a social tie between two users, respectively. Suppose $F = \{f_1, f_2, ..., f_n\}$ is the feature set. For each user u, we define a feature strength vector $\overrightarrow{F_u} = (f_{u_1}, f_{u_2}, ..., f_{u_n})$, with f_{u_i} representing the strength of u on the specific feature f_i. For each edge $e(u, v)$, we use t_{uv} to represent the tie strength between u and v.

We aim to discover the key factors that can be used to construct a user's social influence, and to design a proper framework to integrate the factors in order to evaluate a user's social influence in OSNs efficiently and effectively. To this end, we identify three metrics to measure social influence evaluation models:

- *The percentage of duplication*. It should be able to distinguish the social influence of different users as much as possible, i.e., the values of all users' social influences have little duplication.
- *The influence spread*. The users with high influence (top-k users) should be able to spread the influence information to a large range.
- *The identification accuracy*. It should identify influential users with high accuracy.

Problem Definition. Given a social network $G = (V, E)$, and a feature set $F = \{f_1, f_2, ..., f_n\}$, the problem is: How is it possible to evaluate the social influence of each user with less duplication, a larger influence spread, and a higher prediction accuracy?

Before describing our solution framework, we first introduce the PageRank algorithm, which will be used in our work.

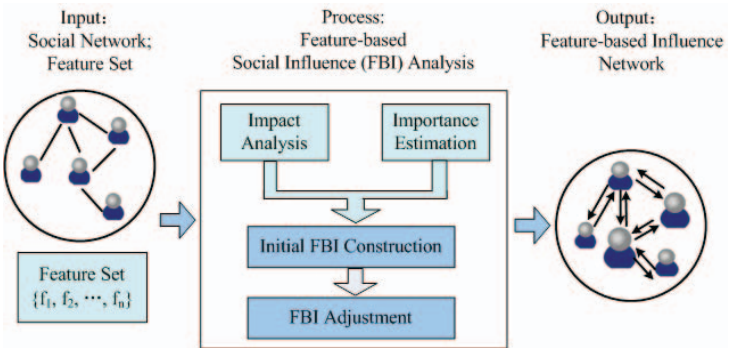


Fig. 2. Solution framework.

PageRank is a link analysis algorithm used by the Google Internet search engine [23]. The rank value indicates the importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively, and depends on the number and the PageRank metric of all pages that link to it. A page that is linked to by many pages with high PageRanks receives a high rank, itself. If there are no links to a web page, then there is no support for that page. Formally, the rank value $PR(s)$ of a given page s is given by:

$$PR(s) = C \cdot \sum_{l \in N_s^{in}} \frac{PR(l)}{N_l^{out}},$$

where C is a normalized constant. N_s^{in} represents all pages pointing to s, and N_l^{out} represents the number of links that l ($l \in N_s^{in}$) points to.

Quite similarly, in real life, a user's social influence also depends on the quantity and quality of his friends. Just as in the "Matthew Effect," the phenomena that "the rich get richer and the poor get poorer" exists universally. This finding will be used in our work.

Our Approach. As shown in Fig. 2, through the proposed feature-based influence (FBI) model, each user is being assigned a numerical value as his social influence, and each edge is labeled with the proportion of influence contributions from one user to another:

i) *Impact analysis and importance estimation*. We extract two essential factors of a user's social influence: (1) The possibility of impacting others. Given two connected users, we first integrate their feature-based similarity and tie strength to construct their direct affinity, then calculate their indirect affinity via common neighbors. Finally, we combine the direct and indirect affinities to get the final impact; and (2) the importance of the user, himself. Many metrics can be flexibly applied, such as the feature-strength, or the role in a network.

ii) *Initial FBI construction*. We integrate the two parts of the impact and the importance using an adjustable weighted sum method to construct a user's initial

social influence. To be specific, we combine (1) the summation of a user's impact with all his neighbors, and (2) the importance of the user, himself, with a weighted sum factor, to measure a user's initial social influence.

iii) *FBI adjustment*. We observe the fact that everyone makes different contributions to their friends, and vise versa, which is very similar to the idea of the PageRank algorithm. Inspired by that, we model the adjustment of all users' feature-based influences with a similar approach.

4 The FBI Model In Detail

In this section, we present the details of the FBI model. We first introduce the two factors of impact and importance. We then use them to construct and adjust the feature-based influence.

4.1 Impact Analysis

In this subsection, we analyze the feature-based direct and indirect affinities between users to construct the impact. All three concepts of direct affinity, indirect affinity, and the final impact can be represented as a vector, according to the feature set. They can also be simplified as a simple value if all the features can be treated equally. Moreover, we normalize each item into the range of $[0, 1]$.

Direct affinity. Observing from real life, we know that the larger the similarity is between two users, and the more frequently they contact each other, the larger the affinity is between them. Therefore, we first define the similarity, then combine it with tie strength to gain the direct affinity.

Given a feature set $F = \{f_1, f_2, ...f_n\}$, we consider the feature-based similarity between two connected nodes u and v, denoted as $\overrightarrow{S_{uv}}$. The $\diamondsuit$ operator is used to combine the similarities from all the features:

$$\overrightarrow{S_{uv}} = \omega_u \diamondsuit \omega_v \tag{1}$$

As a simple example, we can calculate the final similarity as follows:

$$s_{uv} = \frac{1}{n} \cdot \sum_{i \in [1,n]} s_i,$$

where

$$s_i = \begin{cases} \lambda & f_{u_i} = f_{v_i} = 0 \\ f_{u_i} \cdot f_{v_i} & \text{other.} \end{cases}$$

Here, the similarity vector $\overrightarrow{S_{uv}}$ is simplified as s_{uv}. In addition, in the case of $f_{u_i} = f_{v_i} = 0$, we regard it as partial similarity. The intuition is that, at least they are not absolutely different, which occurs when they are not involved in common areas. We can treat the case in two ways: (1) taking it as half similarity ($\lambda = 0.5$), or (2) taking it as random similarity, and generating a random number in [0,1) to represent it.

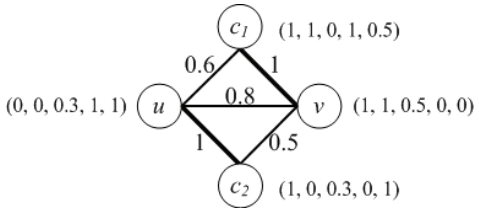


Fig. 3. A social network with feature strengths (of nodes) and tie strengths (of edges).

Taking Fig. 3 for instance, the similarity of each connected pair of nodes will be calculated as follows: $s_{uv} = (0+0+0.15+0+0)/5 = 0.03$, $s_{uc_1} = (0+0+0+1+0.5)/5 = 0.3$, $s_{uc_2} = (0+0.5+0.09+0+1)/5 = 0.318$, $s_{vc_1} = (1+1+0+0+0)/5 = 0.4$, $s_{vc_2} = (1+0+0.15+0.5+0)/5 = 0.33$. Note that, since $f_{u_2} = f_{c_{2_2}} = 0$, and $f_{v_4} = f_{c_{2_4}} = 0$, here, we let $\lambda = 0.5$.

The $\square$ operator is used to combine the similarity and the tie strength into direct affinity for two connected users:

$$\overrightarrow{A^d_{uv}} = t_{uv} \square \overrightarrow{S_{uv}} \tag{2}$$

Again, as a simple example,

$$A^d_{uv} = t_{uv} \cdot s_{uv}.$$

Here, we use product $(\cdot)$ to combine the tie strength and the similarity of two users. Other operators may be flexibly used, according to the specific contexts.

Taking Fig. 3 for instance, the direct affinity of each pair of nodes will be calculated as follows: $A^d_{uv} = 0.03 \cdot 0.8 = 0.024$, $A^d_{uc_1} = 0.3 \cdot 0.6 = 0.18$, $A^d_{uc_2} = 0.318 \cdot 1 = 0.318$, $A^d_{vc_1} = 0.4 \cdot 1 = 0.4$, $A^d_{vc_2} = 0.33 \cdot 0.5 = 0.165$.

Indirect Affinity. A friend of a friend's idea may influence our thoughts. In this case, common friends work as a bridge of propagating information (ideas, news, and influences, etc.). Based on the observation, we emphasize the effect of common neighbors, and consider the indirect affinity. Here, we assume that all common friends have positive effects on the affinity. In fact, it may be much more complex in reality. For instance, the effects of bad-mouthing may lead to negative effects.

Intuitively, there are several conditions that the indirect affinity, denoted as $\overrightarrow{A^{id}_{uv}}$, should satisfy:

- It should be a monotonically increasing function of the number of the common friends. The intuition is that, having one more common neighbor, the chance to influence others, or be influenced, will be increased.
- The same rule of (1) also applies to the common features and tie strengths. The intuition is that, having one more common feature or stronger ties, the strength to influence others, or be influenced, will be increased.
- It should be normalized into the same scale, with direct affinity $\overrightarrow{A^d_{uv}}$. Since $\overrightarrow{A^d_{uv}}$ is the 1-order function of features associated with u and v, then A^{id}_{uv} should also be some 1-order function of that.

- It should have an upper bound, i.e., it will stop increasing when it reaches some threshold. Since we normalize all the variables into [0,1], we use 1 as the upper bound.

Several normalization approaches meet the above requirements. Here, we use the square root normalization as an example in the following:

$$A_{uv}^{id} = min\left\{1, \sqrt{\sum_{c \in N_u \cap N_v} A_{uc}^{d} \cdot A_{cv}^{d}}\right\} \quad (3)$$

Taking Fig. 3 for instance, the indirect affinity of each connected pair of nodes will be calculated as follows: $A_{uv}^{id} = \sqrt{0.18 \cdot 0.4 + 0.318 \cdot 0.165} \approx 0.3528$, $A_{uc_1}^{id} = \sqrt{0.024 \cdot 0.4} = 0.098$, $A_{uc_2}^{id} = \sqrt{0.024 \cdot 0.165} \approx 0.0629$, $A_{vc_1}^{id} = \sqrt{0.024 \cdot 0.18} \approx 0.0657$, $A_{vc_2}^{id} = \sqrt{0.024 \cdot 0.318} \approx 0.0874$. Note that, c_1 and c_2 are the common neighbors of u and v; v is the common neighbor of u and c_1, as well as u and c_2; u is the common neighbor of v and c_1, as well as v and c_2.

The Final Impact. The possible impact between u and v, which represents how much affinity exists between the two users, is denoted as I_{uv}^{in}, and is defined as follows:

Definition 1: The possible impact between two connected users is the integrated effect of their direct and indirect affinities. It can be calculated as follows:

$$\overrightarrow{I_{uv}^{in}} = q \cdot \overrightarrow{A_{uv}^{d}} + (1-q) \cdot \overrightarrow{A_{uv}^{id}}, q \in [0,1] \quad (4)$$

Taking Fig. 3 for instance, if $q = 0.5$, the impact of each pair of nodes will be calculated as follows: $I_{uv}^{in} = I_{vu}^{in} = 0.5 \cdot 0.024 + 0.5 \cdot 0.3528 = 0.1884$, $I_{uc_1}^{in} = I_{c_1u}^{in} = 0.5 \cdot 0.18 + 0.5 \cdot 0.098 = 0.1675$, $I_{uc_2}^{in} = I_{c_2u}^{in} = 0.5 \cdot 0.318 + 0.5 \cdot 0.0629 = 0.1905$, $I_{vc_1}^{in} = I_{c_1v}^{in} = 0.5 \cdot 0.4 + 0.5 \cdot 0.0657 = 0.2329$, $I_{vc_2}^{in} = I_{c_2v}^{in} = 0.5 \cdot 0.165 + 0.5 \cdot 0.0874 = 0.1262$.

4.2 Importance Estimation

The data mining literature is rich in problems, asking to assess the importance of entities in a given data set [6]. In this paper, we do not focus on the methods of estimating entity importance, but the role of importance with respect to social influence. Since the importance of u works as a part of influence beyond the group of two users, i. e., u and his neighbor, we denote it as $\overrightarrow{I_u^{out}}$.

We can define the importance according to different situations. For instance, if the features can be treated in the same way, the importance can be the total of the feature strengths, as the following:

$$I_u^{out} = \frac{\sum_{i \in [1,n]} f_{u_i}}{n} \quad (5)$$

Taking Fig. 3 for instance, the importance can be calculated as: $I_u^{out} = 2.3/5 = 0.46$, $I_v^{out} = 2.5/5 = 0.5$, $I_{c_1}^{out} = 3.5/5 = 0.7$, $I_{c_2}^{out} = 2.3/5 = 0.46$. Here, we normalize the importance to $[0,1]$, by dividing the maximum possible importance.

In some other cases, we may measure the importance as a user's centrality, such as degree (the number of his neighbors) or betweenness (how many paths should come across the node for any pair of two other nodes). In the scientific collaboration/citation network, the number of citations, the number of published papers, or the H-index may indicate an author's importance.

Algorithm 1 initFBI(G, F)

Input: G, a social network; F, a feature set.
Output: G', a social network with each node u being assigned an initial feature-based influence w_u.

```
for each node u ∈ G do
    w_u ← 0.
    for each neighbor v ∈ N_u do
        Calculate similarity S_uv using Eq. 1,
        direct affinity A^d_uv using Eq. 2,
        indirect affinity A^id_uv using Eq. 3,
        impact I^in_uv using Eq. 4,
        and importance I^out_u using Eq. 5.
    Update the initial social influence w_u using Eq. 6.
```

4.3 Initial FBI Construction

Generally speaking, the more friends a user has, the larger social influence he has. It is similar to the ranks of web pages in the PageRank algorithm: the more related (incoming) links a web page has, the more important the web page is. Besides that, as what happens in real life, people who are with high personal importance are more powerful in influencing others. Therefore, we use the summation of a user's impact with his neighbors, as well as the importance of the user, himself, to measure his initial feature-based social influence (FBI). It is calculated as follows:

$$w_u = p \cdot I_u^{out} + (1-p) \cdot \sum_{v \in N_u} I_{uv}^{in} \quad (6)$$

Let us take Fig. 3 for instance. Suppose $p = q = 0.5$, the initial FBIs of all users are calculated as follows: $w_u = 0.5 \cdot 0.46 + 0.5 \cdot (0.1884 + 0.1675 + 0.1905) = 0.5032$, $w_v = 0.5 \cdot 0.5 + 0.5 \cdot (0.1884 + 0.2329 + 0.1262) = 0.5238$, $w_{c_1} = 0.5 \cdot 0.7 + 0.5 \cdot (0.1675 + 0.2329) = 0.5502$, $w_{c_2} = 0.5 \cdot 0.46 + 0.5 \cdot (0.1905 + 0.1262) = 0.3884$.

Algorithm 1 shows the complete process of constructing the initial FBIs for all users. We give the time complexity in the following theorem.

Theorem 1: The time complexity of Algorithm 1 is $O(a^2 m)$, where a is the average degree of nodes, and m is the number of nodes (See the Appendix for the proof).

4.4 FBI Adjustment

We present an FBI adjustment model by first identifying the contribution of friends to a user's social influence, then we conduct the adjustments using the iterative approach.

Based on the intuition that the social influence of a friend can impact a user's social influence, we propose

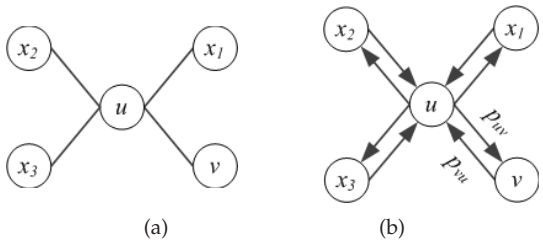


Fig. 4. (a) The initial social network (undirected graph); (b) The influence network (directed graph).

the Feature-based social influence (FBI) adjustment model in the following:

Definition 2: Feature-based influence adjustment: In a social network $G = (V, E)$, the feature-based social influence of a user u, denoted by I_u^F, is the total of influence contributions from his friends, $I_{v\to u}^F$. It is represented as:

$$I_u^F = \sum_{v \in N_u} I_{v\to u}^F$$

Note that, in the initial state, $I_u^F = w_u$.

The Contribution of Friends. However, how can we measure the influence contribution from v to u? Inspired by the idea of PageRank, we argue that the contribution from one user to another should be set with a proper proportion, as follows:

Definition 3: The proportion of influence contribution from one user v to its neighbor u is:

$$p_{vu} = \frac{I_{uv}^{in}}{\sum\limits_{i \in N_v} I_{iv}^{in}} \tag{7}$$

Taking Fig. 4(a) for instance, the contribution from v to u can be measured by $I_v^F \cdot p_{vu}$, and the contribution from u to v can be measured by $I_u^F \cdot p_{uv}$. It is worth noting that, in general, p_{uv} is not the same as p_{vu}. We redefine the adjustment model, as seen below:

$$I_u^F = \sum_{v \in N_u} I_v^F \cdot p_{vu} \tag{8}$$

Now, the undirected social network is evolving into a directed influence network. Each edge is labeled with the proportion of influence contribution from one user to another, as shown in Fig. 4(b).

In summary, the FBI value of u, I_u^F, is mainly dependent on the number of friends, $|N_u|$; and the quality of friends, including the influence I_v^F of each friend v and the proportion of contribution p_{vu}.

The Adjustments. We design the adjustment process of FBI with the iterative method. For the ease of understanding, we first take a global view to describe the process. Then we design a local adjustment algorithm, which can be executed locally by each node.

We can take the system (an OSN[1]) in a given time as a state. Suppose there are a total of m users in an OSN. Each user is assigned an initial influence. Then, the distribution matrix of all users' FBI can be denoted as $I^F = (I_1^F, ...I_i^F, ...I_m^F)^T$, where I_i^F represents the feature-based influence of a user i.

1. Isolated nodes who have no neighbors are excluded.

Algorithm 2 AdjustFBI(G')

Input: G', a resulting social network from Algorithm *InitFBI*.
Output: G_I, an influence network.

/* Let c count the adjustment iteration. Initially, $I_v^{F^{(0)}} \leftarrow w_v$, which is calculated by Algorithm *InitFBI*; $c \leftarrow 1$. After each iteration, $c \leftarrow c + 1$. */
while true **do**
  for each node $u \in G'$ **do**
    $I_u^F \leftarrow 0$.
    for each neighbor $v \in N_u$ **do**
      Calculate the proportion p_{vu} using Eq. 7.
      Update I_u^F using Eq. 8.
  Calculate $\rho^{(c)}$ using Eq. 9.
  if $\rho^{(c)} - \rho^{(c-1)} \to 0$ **then**
    End the adjustment iteration.

With the adjustment process, each user's current influence is distributed to each of his neighbors with a certain proportion, just like when a page votes for another page in PageRank. Then, the system may transit from one state to another. According to the influence adjustment equation (Eq. 8), we define the state transition matrix, P, as the following:

$$P = \begin{pmatrix} 0 & p_{21} & \cdots & p_{(m-1)1} & p_{m1} \\ p_{12} & 0 & \cdots & p_{(m-1)2} & p_{m2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{1(m-1)} & p_{2(m-1)} & \cdots & 0 & p_{m(m-1)} \\ p_{1m} & p_{2m} & \cdots & p_{(m-1)m} & 0 \end{pmatrix}$$

Here, m is the number of users, and each item is exactly the proportion of contribution from i to j, p_{ij} (Eq. 7), the range of which falls in [0,1]. Moreover, the summation of all items in each column is 1, i.e., $\Sigma_{j\in[1,m]} p_{ij} = 1$, for all $i \in [1, m]$.

Let $I^{F^{(0)}}$ represent the initial matrix, and $I^{F^{(c)}}$ represent the resulting matrix after being adjusted by c steps. It can be calculated as $I^{F^{(c)}} = P \cdot I^{F^{(c-1)}} = ... = P^{(c)} \cdot I^{F^{(0)}}$.

To measure the convergence speed of the adjustment process, we define the average variation [11] of FBI:

$$\rho^{(c)} = \frac{\sum\limits_{u \in V} |I_u^{F^{(0)}} - I_u^{F^{(c)}}|}{m}, \tag{9}$$

where m is the number of total nodes, $I_u^{F^{(0)}}$ is the initial FBI value of u, and $I_u^{F^{(c)}}$ is the updated result. We validate the iteration convergence in the experiments.

We also design a local adjustment algorithm, as shown in Algorithm 2. We give the time complexity of each iteration in the following theorem.

Theorem 2: The time complexity of each iteration in Algorithm 2 is $O(a^2 m)$, where a is the average degree of

TABLE 2
Description of the data sets.

Items	HEPTH	DBLP
Number of nodes	31,163	317,080
Number of edges	120,029	1,049,866
Average degree	8	6.62
Maximum degree	202	343

nodes, and m is the number of nodes (See the Appendix for the proof).

5 Experimental Evaluation

In this section, we evaluate the performance of FBI with experiments in real social network data sets.

5.1 Experimental Design

To validate the effectiveness of FBI, we conduct experiments in three real social network data sets: HEPTH [14], DBLP [15], and ArnetMiner (topic-107) [16].

Data Set and Preprocess. HEPTH is a paper cooperation network between the high-energy physicists, who posted preprints at *arXiv* [14]. In HEPTH, each node represents an author, and each edge represents the cooperative relations; moreover, each edge is weighted with the collaboration strength. We predefined 5 features and crawled the web site to assign the feature strength to each node. The DBLP coauthor data set[2] is published by Leskovec [15], in which nodes are represented by anonymized numerical identifiers without feature information; no edge weight is provided. We generate the missed information by randomly assigning feature strengths and tie strengths. The third data set is a part of the ArnetMiner network, which is about the coauthor relations in the research area of Web services, denoted as topic-107. It has 400 authors and 777 edges. We use it for a case study.

The statistics of HEPTH and DBLP are shown in Table 2. The degrees are distributed exponentially (Fig. 5), which fit with the power-law distribution.

Evaluation Metrics. We consider three metrics:

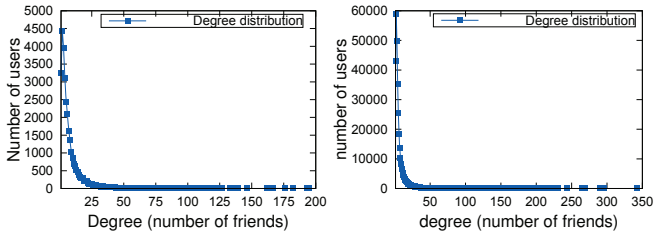
Metric 1: The percentage of duplication. A good model should distinguish all users' influences as much as possible[11]. Thus, less duplication indicates higher efficiency. It is the ratio of influence duplications of all nodes, denoted as η. Suppose ξ is the number of different influence values, it is calculated as

$$\eta = \begin{cases} 1 & if(\xi = 1) \\ 1 - \frac{\xi}{m} & others, \end{cases}$$

where m is the total number of nodes.

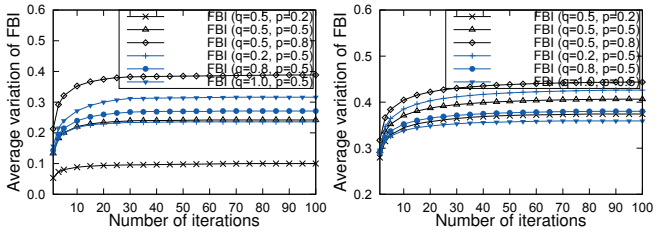
Metric 2: The influence spread. It is a metric to measure how many users can be influenced by k specific users (seeds). To test the influence spread, the methods of diffusing social influence should be determined. We use the Independent Cascade (IC) Model [24]:

2. http://snap.stanford.edu/data/com-DBLP.html



(a) HEPTH (b) DBLP

Fig. 5. Degree distributions of the data sets.



(a) HEPTH (b) DBLP

Fig. 6. Trend of average variation of FBI.

- For the social network G = (V, E), each node has only two states (active or inactive); each node can only transit from the inactive state to the active state.
- At some time t, if a node u is active, then it will have the ability to activate its connected node v, only if v is in the inactive state. Node u can successfully activate v with some probability, which is called the activation probability.
- If v is activated successfully, it will have the ability to activate its connected nodes.
- Repeat the above steps until there is no new node that can be activated.

Metric 3: The identification accuracy. Due to the lack of the ground truth, there is no fixed answer to who are more influential in reality. As an alternative, we turn to measure the quality of the selected top-k users. Higher quality indicates higher accuracy.

Models of Comparison. We compare our model with user attractor-based social influence model [11] (UAI for short) and the user degree centrality based influence model [18] (UDI for short). Similar to FBI, they can also be taken as local topology-based models.

UAI first calculates "user attractor" using $w_{uv} = G^s \frac{|N_v||N_u|}{(cost(u,v))^2}$, where G^s is a constant and has a suggested value of 6; $|N_v|$ and $|N_u|$ are the degrees of v and u; $cost(u, v)$ is the connection cost, which is defined as the length of the shortest path between u and v. Then, all "user attractors" of neighbors are summarized as the influence, $I_u^A = \sum_{v \in N_u} w_{uv}$.

UDI only considers the local structure around a node. Formally, it is calculated as $I_u^D = |N_u|$.

5.2 Experimental Results and Analysis

Convergence of FBI. Fig. 6 shows the average variations of FBI with different settings of p and q. The result

TABLE 3
Parameter settings for duplication of FBI.

Parameters	Settings							
	1	2	3	4	5	6	7	8
Tie Strength	Y	Y	Y	Y	N	N	N	N
Value of p	0	0	0.5	0.5	0	0	0.5	0.5
State	I	S	I	S	I	S	I	S

Y: Considering tie strength
N: Without considering tie strength
I: Initial state
S: Stable state

validates the convergence of the FBI adjustment model, i.e., after some iterations, the FBI value becomes stable. In fact, after about 15 iterations, the change of the average variations $\rho^{(c)} - \rho^{(c-1)} < 0.0001$. Moreover, p takes a more significant effect on the average variation than q. In both HEPTH and DBLP, if we keep $q = 0.5$, the average variation is increased with p.

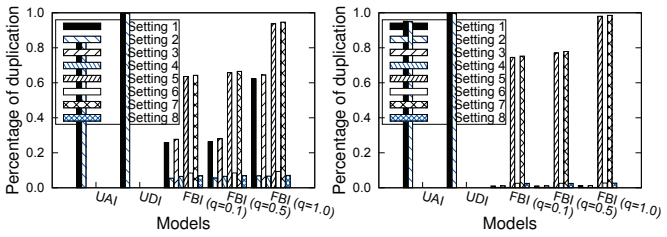
The Percentage of Duplication. On parameters settings for UAI and UDI, we consider the initial state and the stable state.

Fig. 7 shows the results: (1) the percentage of duplication of the UDI model is the highest, being more than 99% duplication. The second highest model is UAI, which is more than 80% duplication. (2) As for FBI, if considering both direct and indirect affinity, or even only considering indirect affinity, the percentage of duplication is lower (less than 25% in the initial state). However, if only considering the direct affinity ($q = 1.0$), the duplication is much higher, which is more than 50% in the initial state. (3) The duplication in the stable state is much lower for FBI with $q \in [0, 1)$. That is, less than 7% in HEPTH, and less than 3% in DBLP. (4) The duplications of FBI that consider the factor of tie strength are much lower, especially in the initial state.

From the comparison, we can see that FBI shows a better and more stable performance in identifying a user's social influence. In addition, simply considering the degree (like UDI), or simply considering the direct affinity (like FBI with $q = 1.0$), cannot distinguish a user's social influence from the others. We show that the reasons are: (1) the degree distribution of online social networks is known to be a power-law distribution, which indicates that most of the users have the same degree. (2) in a large social network (millions of users), the feature set is relatively small (it is 5 in our experiments); then, it is with high probability that multiple pairs of nodes have the same direct affinity. The findings validate the effectiveness of considering common neighbors.

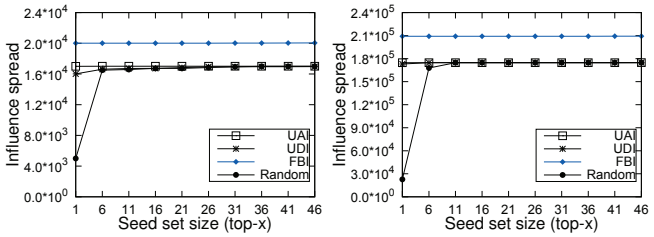
The Influence Spread. As mentioned before, we use the IC model to propagate influence. To conduct the influence propagation, the probabilities of propagation in each edge should first be determined. Much work took non-uniform activation probability, such as [25] and [26]. We define the following strategies to generate asymmetric and nonuniform propagation probabilities:

- The similarity based cascade (SC) model: Each edge is assigned an activation probability with the similarity calculated by our FBI model.
- The TRIVALENCY model: We modified the model in [27]. For every edge $e(u, v)$, we uniformly and randomly select a probability from the set of $\{0.5, ave, 0.001\}$, corresponding to high, medium, and low influences. Here, ave represents the average similarity calculated with FBI, which is 0.37 in HEPTH, and 0.34 in DBLP.



(a) HEPTH (b) DBLP

Fig. 7. The percentage of duplication. The parameters of FBI are set in Table 3. The two bars of UAI and UDI represent the duplication in the initial and stable states, respectively.



(a) HEPTH (b) DBLP

Fig. 8. Influence spreads of top-k nodes.

We use the SC model for FBI and the TRIVALENCY model for UAI and UDI. To obtain the influence spread of each model, we select top $k = (1, 6, 11, ..., 46)$ influential nodes as seeds, simulate the IC propagation 10,000 times, and take the average results. Besides UAI, UDI, and FBI, we also implement a random algorithm as the baseline, which selects seeds randomly.

Fig. 8 shows that the influence spread of FBI is larger than the other three models. UAI and UDI have almost the same spread, while the random method initially has less spread. Then, when the size of seeds gets larger (about 10 or 15 seeds), the performance of random method becomes almost the same as UAI and UDI. Moreover, the change of influence spread with an increasing number of seeds is insignificant for FBI, UAI, and UDI. The reason is because, the most influential users (top-1) can activate most of the other seeds. This finding is not surprising since our focus is not on influence maximization, but the influence of each user.

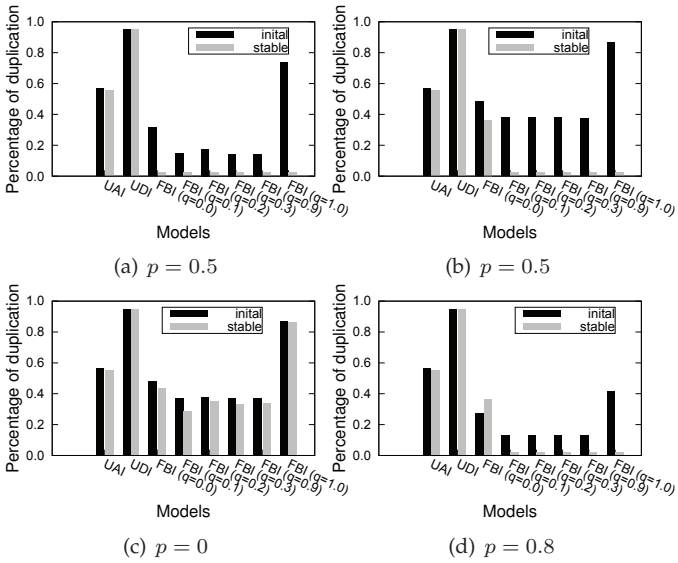
In conclusion, taking both the percentage of duplication and the influence spread into consideration, FBI beats the other three models. It has less duplication, as well as a larger influence spread.

TABLE 4
The top-10 influential users identified by different models.

Top-10	UAI	UDI	FBI				
			$q = 0.2$	$q = 0.5$	$q = 1.0$	$p = 0.4$	$p = 1.0$
1	307	307	328	328	328	328	328
2	213	213	51	51	213	213	51
3	328	328	307	213	307	51	213
4	51	289	213	307	239	307	307
5	239	39	370	239	51	239	239
6	370	51	239	370	113	370	370
7	289	122	399	399	58	399	399
8	39	239	113	113	370	113	113
9	399	370	254	254	316	254	316
10	122	226	79	316	200	316	58

TABLE 5
Quality of top-10 users in topic-107.

Model	Number of Papers	Average Degree
UDI	56.9	19.5
UAI	58.6	19.4
FBI(q=p=0.5)	74.4	17
FBI(q=p=0)	69.7	16.7



(a) $p = 0.5$ (b) $p = 0.5$ (c) $p = 0$ (d) $p = 0.8$

Fig. 9. The percentage of duplication in topic-107. (a) and (d) consider feature strength while (b) and (c) do not.

5.3 Case Study

The Quality of Top-k Influential Users. Table 4 shows the top-10 influential users in topic-107, selected by UAI, UDI, and FBI, respectively. For FBI, we take it as default that $p = 0.5$, $q = 0.5$. For instance, the 4^{th} column, $q = 0.2$, indicates that $p = 0.5$. Just taking the most influential users for instance, UAI and UDI select user 307, while FBI selects user 328. User 307 published 72 papers, and has 27 coauthors, while user 328 published 125 papers, and has 23 coauthors. We manually searched their H-index in Google Scholar and they are 12 and 46, respectively. From the comparison, we can say that user 328 is more influential. Moreover, Table 5 validates that FBI can identify more influential users (whose importance, in terms of the number of published papers, is larger) than UAI and UDI. The results indicate that FBI can identify influential users with higher quality.

The Effects of Feature Strength. Fig. 9 shows the percentage of duplication of the topic-107 data set. Figs. 9(a) and 9(b) indicate that considering the feature strength will reduce the percentage of duplication in initial state significantly. Moreover, feature strength can also be used to construct the importance.

The Effects of p and q. As we have mentioned before, p is the proportion of the importance within the total influence, while q is the proportion of the direct affinity within the total impact. Therefore, with the increase of p, the importance takes more effect on a user's social influence; while, with the increase of q, the direct similarity takes more effect on a user's impact with his neighbors, and then on the final social influence.

Fig. 6 shows the effects of p and q on the convergence of FBI, which shows that a larger value of p leads to a larger variation of FBI. From Fig. 9, we can see that different settings of p and q will lead to different duplications. For instance, Figs. 9(a) and 9(d), a bigger value of p leads to a larger duplication in the initial state. Table 4 shows that FBI with different p and q select different top-10 influential users.

In summary, the selection of p and q will affect all three metrics of influence duplication, influence spread, and identification accuracy. However, it is difficult to provide a general rule to determine proper values of p and q for all scenarios. A suggested solution is selecting proper values according to the specific context.

5.4 Summary of Experiments

The above experiments validate the effectiveness of FBI: it can identify all users' social influences with less duplication than existing user degree-based models, and lead to a larger influence spread. The case study demonstrates that all the factors of the feature strength, the tie strength, the proportion of direct and indirect affinity (q), and the proportion of impact and importance (p) can impact the percentage of duplication, as well as the accuracy of identifying the most influential users.

6 Discussion and Future Work

In this paper, we propose a fine-grained feature-based social influence evaluation model, FBI for short. We explore the two essential factors, the impact and the importance, to construct a user's social influence, present a general framework to integrate those factors, and provide some rational metrics to measure the efficiency and effectiveness. For the aim of fine-grained evaluation, we differentiate the feature strength of users and the tie strength of user relations. We also emphasize the effects of common neighbors in conducting influence between two users. Experimental results show the effectiveness of FBI.

Generality. The FBI model is more general and more powerful than existing models. Our previous work [22]

can be taken as a special case in which we only consider the indirect similarity between two users (disregarding the tie strength, the importance of a user, and the direct affinity), and each feature is equally treated, which leads to higher duplication (about 42%) than FBI (less than 7%); moreover, the top-k users selected by the previous model are not as accurate as in the FBI model, since it neglects some key factors of social influence, such as the tie strength, the importance of a user, and the direct affinity. Some other models can also be taken as a special case of FBI. For instance, FBI becomes a topic-based model if the feature set contains only topics; it becomes a community influence evaluation model if the feature set contains features in a single community; it can be taken as a unified model if the feature set contains features from multiple communities or social networks.

Scalability. First, the complexity of the proposed algorithms are proportional to the number of nodes, which indicates that they have good scalability. Second, the algorithms in FBI are local algorithms, in which the computation is conducted on each user and his neighbors. Therefore, the efficiency can be improved by distributing the calculation into several parts. Last but not least, although FBI can be applied to large social networks, we believe it is more common that in daily life applications, e.g., selecting or evaluating some excellent scientists (or employees, specialists, experts, etc.), the network scales are usually not very large.

Future Work. In this paper, we provide a general framework to integrate the essential factors of social influence. More specific context-based rules can be explored in the future. Another interesting direction is to measure the tie strengths based on different features. Moreover, in a real social network, we have found that influential users are usually taken as more trustful than other users, and vice versa. Therefore, identifying the bidirectional effects between influence and trust [28], and constructing a comprehensive "reputation and trust-based system" [29], is also meaningful work.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful suggestions. This work is supported by NSFC grants 61272151 and 61073037, ISTCP grant 2013DFB10070, the China Hunan Provincial Science & Technology Program under Grant Number 2012GK4106, the Ministry of Education Fund for Doctoral Disciplines in Higher Education under Grant Number 20110162110043, NSF grants ECCS 1231461, ECCS 1128209, CNS 1138963, CNS 1065444, and CCF 1028167.

REFERENCES

[1] H. Quan, J. Wu, and J. Shi. Online social networks and social network services: A technical survey. *Handbook of Pervasive Communication, M. Ilyas and H. Mouftah (eds), CRC Press, November 14, 2011, ISBN-13: 978-1420051094* .

[2] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. *Proc. IMC*, pages 29–42, 2007.

[3] X. Song, Y. Chi, K. Hino, and B. L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. *Proc. IEEE WWW*, pages 191–200, 2007.

[4] 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.

[5] J. Davitz, J. Yu, S. Basu, D. Gutelius, and A. Harris. ilink: Search and routing in social networks. *Proc. ACM KDD*, pages 931–940, 2007.

[6] A. Gionis, T. Lappas, and E. Terzi. Estimating entity importance via counting set covers. *Proc. ACM KDD*, pages 687–695, 2012.

[7] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. *Proc. ACM KDD*, pages 467–476, 2009.

[8] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. *Proc. ACM KDD*, pages 807–816, 2009.

[9] E. Gilbert and K. Karahalios. Predicting tie strength with social media. *Proc. ACM CHI*, pages 211–220, 2009.

[10] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[11] F. Hao, Z. Pei, C. Zhu, G. Wang, and L. T. Yang. User attractor: An operator for the evaluation of social influence. *Future Generation Computer Systems*, doi: org/10.1016/j.future.2012.04.005, 2012.

[12] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. *Proc. ACM KDD*, pages 160–168, 2008.

[13] L. Rashotte. Social influence. *A.S.R. Manstead, M. Hewstone (Eds.), The Blackwell Encyclopedia of Social Psychology, Malden: Blackwell Publishing*, pages 562–563, 2007.

[14] J. Gehrke, P. Ginsparg, and J. M. Kleinberg. Overview of the 2003 KDD cup. *SIGKDD Explorations*, 5(2):149–151, 2003.

[15] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *CoRR*, abs/1205.6233, 2012.

[16] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. *Proc. ACM KDD*, pages 990–998, 2008.

[17] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks (Elsevier)*, 32:245–251, 2010.

[18] T. Opsahl, V. Colizza, P. Panzarasa, and J. J. Ramasco. Prominence and control: The weighted rich-club effect. *Physical Review Letters*, 101(16):168702, 2008.

[19] M. E. J. Newman. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *PHYSICAL REVIEW E*, 64:016132, 2001.

[20] Z. Katona, P. P. Zubcsek, and M. Sarvary. Network effects and personal influences: Diffusion of an online social network. *Journal of Marketing Research*, 48(3):425–443, 2011.

[21] E. Bakshy, I. Rosenn, C. Marlow, and L. A. Adamic. The role of social networks in information diffusion. *Proc. IEEE WWW*, pages 519–528, 2012.

[22] Z. Xiong, W. Jiang, and G. Wang. Evaluating user community influence in online social networks. *Proc. IEEE TrustCom*, pages 640–647, 2012.

[23] L. Page. Pagerank: Bring order to the web. *Stanford Digtial Libraries Working Paper*, 1997.

[24] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

[25] G. Wang, Q. Hu, and P. S. Yu. Influence and similarity on heterogeneous networks. *Proc. ACM CIKM*, pages 1462–1466, 2012.

[26] S. Tang, J. Yuan, X. Mao, X. Li, W. Chen, and G. Dai. Relationship classification in large scale online social networks and its impact on information propagation. *Proc. IEEE INFOCOM*, pages 2291–2299, 2011.

[27] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. *Proc. ACM KDD*, pages 1029–1038, 2010.

[28] A. Srinivasan, J. Wu, and J. Teitelbaum. Distributed reputation-based secure localization in sensor networks. *Journal of Autonomic and Trusted Computing*, pages 1–13, 2007.

[29] A. Srinivasan, J. Teitelbaum, H. Liang, J. Wu, and M. Cardei. Reputation and trust-based systems for ad hoc and sensor networks. *Algorithms and Protocols for Wireless, Mobile Ad Hoc Networks, A. Boukerche (ed.), Wiley, 2008, ISBN: 978-0-470-38358-2*.

Guojun Wang received his B.Sc. in Geophysics, M.Sc. in Computer Science, and Ph.D. in Computer Science from Central South University, China. He is now Chair and Professor of the Department of Computer Science at Central South University. He is also Director of Trusted Computing Institute of the University. He has been an Adjunct Professor at Temple University, USA; a Visiting Scholar at Florida Atlantic University, USA; a Visiting Researcher at the University of Aizu, Japan; and a Research Fellow at the Hong Kong Polytechnic University. His research interests include network and information security, Internet of things, and cloud computing. He is a senior member of CCF, and a member of IEEE, ACM, and IEICE.

Wenjun Jiang received her Bachelor's degree in Computer Science from Hunan University, P. R. China, in 2004; and her Master's degree in Computer Software and Theory from Huazhong University of Science and Technology, P. R. China, in 2007. She has been a PhD candidate at Central South University since September 2009. Currently, she is a visiting Ph. D student at Temple University. Her research interests include trust and social influence evaluation models and algorithms in online social networks.

Jie Wu is the chair and a Laura H. Carnell Professor in the Department of Computer and Information Sciences at Temple University. Prior to joining Temple University, he was a program director at the National Science Foundation and Distinguished Professor at Florida Atlantic University. His current research interests include mobile computing and wireless networks, routing protocols, cloud and green computing, network trust and security, and social network applications. Dr. Wu regularly published in scholarly journals, conference proceedings, and books. He serves on several editorial boards, including IEEE Transactions on Computers, IEEE Transactions on Service Computing, and Journal of Parallel and Distributed Computing. Dr. Wu was general co-chair/chair for IEEE MASS 2006 and IEEE IPDPS 2008 and was the program co-chair for IEEE INFOCOM 2011 and program chair for CCF CNCC 2013. Currently, he is serving as general chair for IEEE ICDCS 2013. He was an IEEE Computer Society Distinguished Visitor and the chair for the IEEE Technical Committee on Distributed Processing (TCDP). Dr. Wu is an ACM Distinguished Speaker and a Fellow of the IEEE. He is the recipient of the 2011 China Computer Federation (CCF) Overseas Outstanding Achievement Award.

Zhengli Xiong received his Bachelor's degree and Master's degree in Computer Science and Technology from Central South University, P. R. China, in 2009 and 2012, respectively. Currently, he is a software engineer at Tencent. His research interests include community discovery and social influence evaluation models and algorithms in online social networks.