Exploiting social circle broadness for influential spreaders identification in social networks

Article *in* World Wide Web · May 2014

DOI: 10.1007/s11280-014-0277-1

CITATIONS
9

READS
77

6 authors, including:

Senzhang Wang
Nanjing University of Aeronautics & Astronautics
107 PUBLICATIONS **773** CITATIONS
SEE PROFILE

Fang Wang
Tianjin University
31 PUBLICATIONS **197** CITATIONS
SEE PROFILE

Xiaoming Zhang
Beihang University (BUAA)
63 PUBLICATIONS **479** CITATIONS
SEE PROFILE

World Wide Web (2015) 18:681–705
DOI 10.1007/s11280-014-0277-1

Exploiting social circle broadness for influential spreaders identification in social networks

Senzhang Wang · Fang Wang · Yan Chen · Chunyang Liu · Zhoujun Li · Xiaoming Zhang

Received: 15 May 2013 / Revised: 27 August 2013 /
Accepted: 13 January 2014 / Published online: 11 March 2014

Abstract Influential spreaders identification in social networks contributes to optimize the use of available resources and ensure the more efficient spread of information. In contrast to common belief that highly connected or core located users are most crucial spreaders, this paper shows that both user's local and global structural properties matter in information diffusion. We propose a new metric, social circle broadness, to measure a user's information spreading influence by qualitatively combining the two above properties. Firstly, a definition of social circle diversity is introduced to measure the dispersion extent of a user's friends distribution in the network. Based on it, a method to calculate each user's local social circle broadness is presented. Preliminary experiments on a coauthor dataset demonstrate the effectiveness of social circle broadness in information diffusion. Furthermore, a social circle weighted PageRank (SCWPR) algorithm is proposed to iteratively rank each user's global social circle broadness. We conduct extensive comparison experiments against six state-of-the-art baseline methods on four real social network datasets. The results show that SCWPR outperforms all of them for influential spreaders identification in information propagation.

S. Wang · F. Wang · Y. Chen · Z. Li (✉) · X. Zhang
Beihang University, Beijing, China
e-mail: lizj@buaa.edu.cn

S. Wang
e-mail: szwang@cse.buaa.edu.cn

F. Wang
e-mail: wangfang0325@cse.buaa.edu.cn

Y. Chen
e-mail: chenyan@cse.buaa.edu.cn

X. Zhang
e-mail: yolixs@cse.buaa.edu.cn

C. Liu (✉)
National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
e-mail: lcy@isc.org.cn

Keywords Social circle · Information propagation · Influential identification

1 Introduction

The increasing popularity of online social network has greatly improved the communication and information sharing between users. By connecting to their friends, users in the network can express opinions, post and share pictures or videos. Moreover, they can receive information from their friends and spread information to others. As a new type of social media, social network plays an important role for the spread of various information. For example, innovation, hot topics and malicious rumors can propagate through social network among individuals by word-of-mouth. Meanwhile, as an important marketing and advertising platform, social network allows the marketing message deriving from a small number of users to a large population in a short period of time.

However, a challenge of utilizing social networks as an information diffusion and marketing dissemination platform is how to find influential users.[1] Influential users often greatly influence others in various ways. For example, when users want to buy something or make decisions, they probably consult their friends or experts, and the suggestions or recommendations from influential can significantly affect their final decisions. However, who are the most influential users in the social network and how to find them? This paper focuses on solving this problem. In another word, in order to maximize the number of eventually influenced users, how to select the initial spreaders from the network structure view.

Plenty of prior studies have researched on this problem from the perspective of network structure. However, most of them either only focus on the user's local connectivity property [33] or pay attention to her global structure property [6, 26]. For example, Pastor-Satorras et al. showed that the most connected users are the key spreaders [33]. Similarly, Brandes et al. found that the most important spreaders are usually associated with the betweenness centrality [6]. However, [26] showed both methods may not perform effectively because nodes with high degree may locate at the periphery of a network and nodes with high betweenness may have a small number of neighbors. [26] reported that the most influential spreaders are those located within the core of the network as identified by the k-shell decomposition. However, the main limitation of k-shell method is that it suffers from the ineffectiveness when multiple initial spreaders propagate information simultaneously. In addition, k-shell method can not rank the influence of nodes in the same k-shell layer, though the influence of two nodes in the same k-shell layer but quite different degree may differ considerably.

Different from existing researches, this paper takes both users' global structure property and their popularity (number of connected neighbors) into account. A new metric, socle circle broadness, is proposed to measure a user's information spreading influence by combining the two properties. Generally, a user's social circle is broad if she connects to many neighbors and these neighbors distribute across the network, and vice verse. We investigate whether and to what extent a user's social circle broadness will affect her information dissemination capability. For example, suppose we have two authors A and B in DBLP co-author network, they both co-write 100 papers with others. However, A's papers cover the area of data mining, machine learning, database, pattern recognition and

[1] Here, we define the influential users as the spreaders that can spread information to a large part of the network.

information retrieval, while B's are only about data mining and database. In this case, we say that author A's social circle is broader than author B's as author A's papers spread over different research fields and therefore she has a higher structure diversity. Who is more influential in information dissemination? Compared with user's popularity (number of papers), our research shows that social circle broadness is a much more reasonable measure of spreading influence. It is probably that users with broad social circles are much more likely to spread information to a larger coverage of the network.

In this paper, we address the problem of identifying users with broad social circles and exploiting it for influential ranking in social networks. Hence, we have to investigate and answer the following issues:

– How to define and calculate a user's social circle broadness?
– Whether and to what extent does a user's social circle broadness affect her influence in information diffusion?
– How to identify the most influential spreader in a social network? Is there an effective algorithm to rank all the users' social circle broadness?

The main contributions of this work can be summarized as follows,

- We propose a method to define a user's *social circle diversity* and utilize it to measure her *local social circle broadness*. To the best of our knowledge, this is the first work of measuring a user's social circle broadness.
- To iteratively rank each user's *global social circle broadness*, we further propose a social circle weighted PageRank algorithm. Moreover, the user's social circle broadness is exploited for influential spreaders identification.
- We conduct extensive comparison experiments against six state-of-the-art baseline methods on four real social network datasets, and the results show the effectiveness of our methods.

More specifically, we first utilize the Kullback-Leibler Divergence between the distributions of a user's friends and the network communities to measure her social circle diversity. Based on it, a local social circle broadness calculation method is presented. This method takes both user's social circle diversity and popularity into account. Preliminary experiments demonstrate the effectiveness of user's local social circle broadness in information diffusion. Users with broader social circles are much more likely to spread information to a larger fraction of the network. Meanwhile, those highly connected users with their neighbors clustered in a small part of the network may not be so crucial. Second, as to different social circle broadness, a user's neighbors may have different influence to the user. The neighbors with broader social circles contribute more to the user's social circle broadness than those with smaller social circles. Generally, the social circle of a user is broad because she has many friends whose social circles are also broad. To take it into account, a social circle weighted PageRank algorithm SCWPR is proposed to iteratively rank each user's global social circle broadness. Finally, we compare our methods with six state-of-the-art structure based baseline methods, such as degree centrality, betweeenness centrality, PageRank, k-shell, etc. on four real social network datasets. Experiment results demonstrate SCWPR performs best.

The remainder of the paper is structured as follows. In Section 2, we give our motivation and formally define the problem. In Section 3, we define the social circle diversity. Based on it, a method to measure a user's local social circle broadness is proposed. In Section 4, we further propose a social circle weighted PageRank algorithm to iteratively rank each user's global social circle broadness. We evaluate our method in Section 5. Section 6 discusses related works, and the conclusion is given in Section 7.

2 Motivation and problem formulation

We first briefly describe our motivation. In the social network, if a user has a small number of friends or a user has a large number of friends but they are closely connected in a small part of the network, we think that the user's social circle is small. The information generated by the user is comparatively harder to spread to other parts of the network and thus she may not be a crucial spreader. On the contrary, if a user has many friends and they distribute across different communities of the network, we think that the user's social circle is broad, and the user potentially is a key spreader because she is more likely to spread information to a larger coverage of the network. Furthermore, if a user connects to many users whose social circles are also broad, we believe the user is even more influential. Figure 1 shows an illustration of our idea.

Based on above motivation, we then formally **define the problem** as follows. *Let* $G = (V, E)$ *denote a social network, where* $V = \{v_1, v_2, ..., v_n\}$ *is the set of users, and* $E \subseteq V \times V$ *is a set of undirect relationships between users. Further suppose that the users in the social network can be divided into* K *communities* $\#C_1 = n_1, ..., \#C_K = n_K$ *and* $V = C_1 \cup ... \cup C_K$. *The problem is how to define and measure a user's social circle broadness based on above network structure information, and how to exploit each user's social circle broadness for influential spreaders identification?*

3 Local social circle broadness

In this section, we firstly develop a method to measure a user's *local social circle broadness*. Secondly, an example is given to illustrate our method. Finally, preliminary experiments on a coauthor dataset show the effectiveness of the social circle broadness in information diffusion.

3.1 Local social circle broadness measurement

The method of measuring a user's local social circle broadness is as follows. Firstly, the social network is divided into a number of communities by the community detection

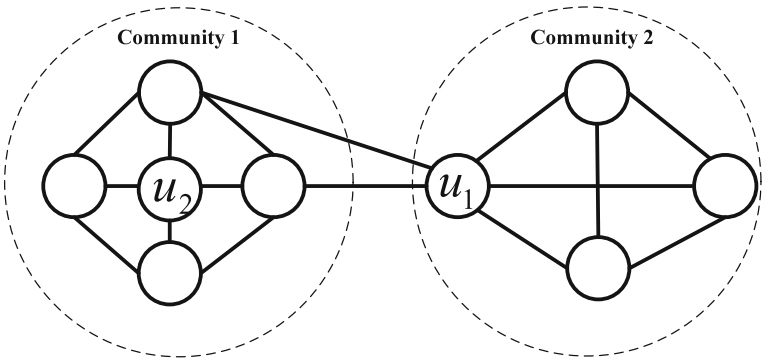

Figure 1 Illustration of social circle broadness. The two dashed circles represent two communities. All the neighbors of u_2 are in community 1 while u_1's neighbors spread over the two communities. Therefore, u_1's social circle is broader

method. Based on the results, the distributions of communities and the user's friends are then defined. Next, the two distributions are used to measure a user's *social circle diversity* by calculating their Kullback-Leibler divergence. By exploring a user's social circle diversity and popularity, a method to measure the user's *local social circle broadness* is finally proposed. Formally, we have the following definitions.

Definition 1 (Distribution of Communities in the Social Network) Given a social network represented as a direced/undirected graph $G = (V, E)$, we define V and $E \subseteq V \times V$ as the sets of users and edges in G, respectively. Assume there are N users $\{user_i | 1 \leq i \leq N\}$ in G, and they are divided into K communities $\{C_j | 1 \leq j \leq K\}$. We define the distribution of communities in the social network as:

$$P(c \in C_j) = p_j = \frac{number\ of\ users\ in\ community\ c}{total\ number\ of\ users} \tag{1}$$

In community detection, each node can be assigned to one community (disjoint community detection) [5, 11] or otherwise be shared between communities (overlapping community detection) [4, 15, 32, 38]. Therefore, we divide the distribution definition of communities into the two categories as follows:

For the **disjoint communities**, we have $\#C_1 = n_1, ..., \#C_K = n_K$, $N = n_1 + ...n_K$ and $V = C_1 \cup ... \cup C_K$. The distribution of communities in the network can be defined as:

$$P(c \in C_j) = p_j = \frac{n_j}{N} \tag{2}$$

For the **overlapping communities and crisp (non-fuzzy) assignment**, we have $\#C_1 = n_1, ..., \#C_K = n_K$, $N_c = n_1 + ...n_K (N_c > N)$ and $V = C_1 \cup ... \cup C_K$. Then the distribution of communities in the network can be defined as:

$$P(c \in C_j) = p_j = \frac{n_j}{\Sigma_{i=1}^{K} n_i} \tag{3}$$

For the **overlapping communities and fuzzy assignment**, $user_i$ is associated with each community with a certain probability $[a_{i1}, ...a_{ij}, ...a_{iK}]$, where $0 \leq a_{ij} \leq 1, 1 \leq i \leq N, 1 \leq j \leq K$, and $\Sigma_{j=1}^{K} a_{ij} = 1$. We use $\Sigma_{i=1}^{N} a_{ij}$ to denote the number of users within communities C_j. Therefore, the distribution of communities can be defined as:

$$P(c \in C_j) = p_j = \frac{\Sigma_{i=1}^{N} a_{ij}}{N} \tag{4}$$

Definition 2 (Distribution of User's Friends) For **disjoint communities**, assume $user_i$ has N^i friends and the numbers of her friends in each community are denoted by ${n_1}^i, {n_2}^i, ... {n_k}^i$. Then the distribution of $user_i$'s friends among communities can be defined as:

$$P^i(Fri^i \in C_j) = p_j^i = \frac{n_j^i}{N^i} \tag{5}$$

where Fri^i is $user_i$'s friends, and $\sum_{j=1}^{K} p_j^i = 1$. For the **crisp assignment of overlapping communities**, the formula is the same to disjoint communities as the relationship between a node and a community is binary. For the **fuzzy assignment of overlapping communities**, the distribution is defined as:

$$P^i(Fri^i \in C_j) = p_j^i = \frac{\Sigma_{user_t \in Neighbor(user_i)} a_{tj}}{\Sigma_{user_t \in Neighbor(user_i)} \Sigma_{j=1}^{K} a_{tj}} \tag{6}$$

where $Neighbor(user_i)$ denotes the set of $user_i$'s friends.

Now we have the distributions of communities and users' friends in the network. Next we use the Kullback-Leibler divergence (KL divergence) between the two distributions to measure a user's *social circle diversity*. KL-divergence is a non-symmetric measure of the difference between two probability distributions[8].

Definition 3 (User's Social Circle Diversity (SCD)) We define $user_i$'s Social Circle Diversity (SCD_i) as follow:

$$SCD_i = \frac{1}{1 + e^{-1/D_{KL}(P^i||P)}} = \frac{1}{1 + e^{-1/\sum_{j=1}^{K} p_j \cdot log \frac{p_j}{p_j^i}}} \tag{7}$$

P is the distribution of communities and P^i is the distribution of $user_i$'s friends.

SCD_i is negatively related to the Kullback-Leibler Divergence between the two distributions. The smaller KL Divergence, the larger probability the distribution of the user's friends obeys uniform distribution. Hence she has a higher social circle diversity. However, the range of KL Divergence is $[0, +\infty)$. To normalize KL value, we use the sigmoid function $S(t) = \frac{1}{1+e^{-t}}$ as a mapping function. Based on the definition of SCD, next we give the definition of a user's *local social circle broadness*.

Definition 4 (User's Local Social Circle Broadness (LSCB)) Based on definition 3, we define $user_i$'s local social circle broadness ($LSCB_i$) as:

$$LSCB_i = N^i \cdot SCD_i = \frac{N^i}{1 + e^{-1/\sum_{j=1}^{K} p_j \cdot log \frac{p_j}{p_j^i}}} \tag{8}$$

$user_i$'s local social circle broadness depends on both her popularity and social circle diversity. We call it local social circle broadness because we just consider the distribution of each user's friends but the distribution of her friends' friends is ignored. It is worth noticing that the friends with broader social circle contribute more to $user_i$'s social circle broadness than those with smaller social circle. In Section 4 we will investigate this problem and propose a method to iteratively ranking each user's global social circle broadness.

Here we give an example in Figure 2 to illustrate the LSCB calculation process. Assume a network is divided into three communities, and two users, u_1 and u_2 are in community 2 and community 3, respectively. To calculate $LSCB_1$ and $LSCB_2$, we first calculate the distributions of communities and u_1, u_2's friends as [5/15 4/15 6/15], [2/5 2/5 1/5] and [0 0 1]. Then the KL divergence of the two distributions can be represented as $D_{KL}(P^1||P) = \sum_{j=1}^{K} p_j \cdot log \frac{p_j}{p_j^i} = \frac{2}{5} \cdot ln(\frac{2/5}{5/15}) + \frac{2}{5} \cdot ln(\frac{2/5}{4/15}) + \frac{1}{5} \cdot ln(\frac{1/5}{6/15}) = 0.0965$. Similarly, we have $D_{KL}(P^2||P) = 0.916$. According to definition 3, u_1 and u_2's social circle diversity can be calculated as $SCD_1 = \frac{1}{1+e^{-1/0.0965}} = 0.9999$ and $SCD_1 = \frac{1}{1+e^{-1/0.916}} = 0.749$, respectively. Finally, according to definition 4, we have $LSCB_1 = N^1 * SCD_1 = 5 * 0.9999 = 4.9995$, $LSCB_2 = N^2 * SCD_2 = 5 * 0.749 = 3.745$.

In order to compare the ranking results between LSCB and degree methods, we also show an example of top ranked users by the two methods. Table 1 lists the top-10 authors identified by two methods on the DBLP DM/DB co-authorship dataset. This dataset is extracted from DBLP, and only the co-author relationships in data mining and database areas are selected. The detail is given in Section 4.1. The result in Table 1 shows that the two lists are

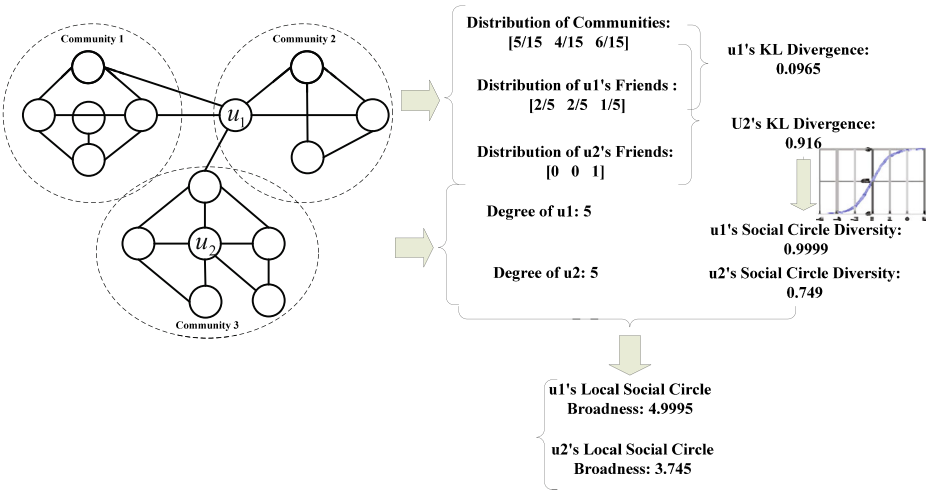

Figure 2 Illustration of *local social circle broadness* calculation process

significantly different and only three authors, Jiawei Han, Philip S. Yu and Michael J.Carey, keep the same ranking positions in two lists.

3.2 Community detection

Our proposed method is based on the result of community detection, and hence the quality of detected communities may have impact on the algorithm performance. To investigate to what extent different community detection methods will affect users' social circle broadness in our approach, we firstly apply four state-of-the-art community detection methods, including disjoint and overlapping ones, to find communities and calculate each user's LSCB. Then we use the Spearman's rank correlation coefficient to measure the similarity of each two ranks of users' LSCB.

Table 1 Top-10 ranked authors by two methods

Method: LSCB			Method: Degree		
Rank	LSCB	Author	Rank	Degree	Author
1	144.4	Jiawei Han	1	225	Jiawei Han
2	138.1	Philip S. Yu	2	217	Philip S. Yu
3	132.1	Gerhard Weikum	3	192	Hector Garcia-Molina
4	128.1	Hector Garcia-Molina	4	186	Michael Stonebraker
5	122.3	Michael Stonebraker	5	178	Gerhard Weikum
6	109.4	Michael J.Carey	6	177	Michael J.Carey
7	109.1	Raghu Ramakrishnan	7	170	Christos Faloutsos
8	108.4	Jeffrey Xu Yu	8	165	Beng Chin Ooi
9	103.5	JianPei	9	159	David J. DeWitt
10	102.8	Aoying Zhou	10	152	Jeffrey Xu Yu

The Spearman's correlation coefficient is defined as follows,

$$\rho = \frac{\sum_i (R_1(P_i) - \overline{R_1})(R_2(P_i) - \overline{R_2})}{\sqrt{\sum_i (R_1(P_i) - \overline{R_1})^2 \sum_i (R_2(P_i) - \overline{R_2})^2}} \tag{9}$$

where $R_1(P_i)$ is the position of user u_i in the first rank list; $R_2(P_i)$ is the position of the specific user in the second rank list; $\overline{R_1}$ and $\overline{R_2}$ are the average rank positions of all users in the first and second rank lists respectively.

For the disjoint community detection methods, we select BGLL [5] and Newman and Girvan's method [11], and for the overlapping ones, we select Clique Percolation Method (CPM) [32] and the method proposed by Baumes et al. [4]. In this paper, we apply the four state-of-the-art methods to test the effect of various community detection methods on users' local social circle broadness.

- **Newman and Girvan's method** It is one of the most popular methods [11]. This method is historically important, because it marked the beginning of a new era in the field of community detection.
- **BGLL method** BGLL method is a a recently proposed hierarchical community detection algorithm [5]. It is applied to extract the community structure. For its effectiveness and efficiency, it is widely used in many community detection applications, especially for large networks.
- **CPM method** CPM method is one of the most popular technique to detect overlapping communities [32]. This method is based on the concept that the internal edges of a community are likely to form cliques due to their high density.
- **Baumes's method** As another popular overlapping method, the Baumes et al. proposed method defines a community as a subgraph which locally optimize a given function W, typically some measure related to the edge density of the cluster [4].

Table 2 shows the experimental results on the DBLP DB/DM co-authorship dataset and Slahdot dataset. From Table 2, we can obtain the following interesting conclusions. First, the results of top ranked users by various methods are very similar. The Spearman's correlation coefficient between each two rank lists of top50 users is around 0.8, which is a very high value meaning a high correlation. It is probably because for the users with broad social circle, their neighbors spread all over the network and therefore, the impact of different community detection methods on top users' social circle broadness is trivial. Second, disjoint community detection methods are more similar to each other and overlapping methods are more similar to each other. Moreover, the result also shows that the similarity of each two rank lists decreases with the increase of their length. For example, the ρ of Top50 users is significantly larger than that of Top1000 users. This is because for the users with a small number of connections, their friends' distribution largely depends on the community detection methods. Hence, such users's social circle broadness varies greatly with different community detection methods.

As we aim to find the most influential spreaders and there is no significant difference among above community detection methods to conduct it, in our paper, we only utilize the BGLL method to divide a graph into communities. Here, we choose the BGLL method due to the following reasons. First, BGLL is a hierarchical community detection method, and it can unfolds a complete hierarchical community structure for the network. Second, It outperforms most existing community detection methods in terms of computation time. It is reported that the size limitation of network this method can deal with is due to limited storage capacity rather than limited computation time [5]. Therefore, this method can handle

Table 2 Spearman's rank correlation coefficient between each two methods on two datasets

	Top50	Top100	Top150	Top200	Top500	Top1000
DBLP DB/DM co-authorship dataset						
Newman vs BGLL	0.814	0.725	0.574	0.527	0.416	0.227
Newman vs CPM	0.782	0.674	0.510	0.495	0.364	0.157
Newman vs Baumes	0.742	0.625	0.508	0.454	0.378	0.237
BGLL vs CPM	0.726	0.623	0.522	0.474	0.365	0.174
BGLL vs Baumes	0.765	0.653	0.517	0.460	0.325	0.247
CPM vs Baumes	0.824	0.736	0.625	0.585	0.415	0.270
Slahdot dataset						
Newman vs BGLL	0.807	0.715	0.620	0.555	0.446	0.235
Newman vs CPM	0.743	0.655	0.515	0.455	0.325	0.195
Newman vs Baumes	0.756	0.645	0.544	0.467	0.324	0.227
BGLL vs CPM	0.752	0.654	0.520	0.455	0.302	0.214
BGLL vs Baumes	0.782	0.617	0.547	0.505	0.325	0.184
CPM vs Baumes	0.815	0.734	0.614	0.523	0.375	0.217

very big datasets which may contains millions of nodes. Moreover, it is also a very effective method with excellent community detection quality. Experimental results show that BGLL outperforms better in terms of modularity than traditional methods. Next, we briefly introduce BGLL method.

This novel method is divided into two phases iteratively. In each iteration, for each node i they consider the neighbors j of i and evaluate the gain of modularity that would take place by removing i from its community and by placing it in the community of j. The gain in modularity ΔQ obtained by moving an isolated node i into a community C can be computed by:

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{10}$$

where $\sum_{in}$ is the sum of the weights of the links inside C, $\sum_{tot}$ is the sum of the weights of the links incident to nodes in C, k_i is the sum of the weights of the links incident to node i, $k_{i,in}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links.

3.3 The information diffusion model

To evaluate the performance of LSCB and SCWPR (described later) in influential spreaders identification, we apply the Susceptible-Infectious-Recovered (SIR) [3, 7, 13] model to simulate the epidemic spreading process. In this model, all nodes are in one of the following three states: the susceptible state (S), the infectious state (I) and the recover state (R). Initially, all nodes are in susceptible state except for one node in the infectious state. At each step, the nodes in the infectious state attempt to infect their susceptible neighbors with a probability β and then enter the recovered state. Nodes in the recovered are immunized and cannot be infected again.

3.4 Preliminary experiments

To evaluate whether a user's local social circle broadness will influence her information dissemination capability, some preliminary experiments are conducted. We first rank the top influential users discovered by LSCB and degree centrality methods, then run the SIR model and report the ultimately infected scope of the network when a epidemic originates from each discovered users.

The parameter of β is a key parameter in SIR model as it determines the probability of an infectious node infecting a susceptible neighbor. Kitsak et al. [26] reported that in the case of large β value, the role of individual nodes is no longer important and spreading would cover almost all the network, independently of where it originated from. Therefore, we set β as a small value 0.1. The dataset is the MathSciNet Co-authorship network dataset (The detail is given in Section 4.1). We design the following two groups of experiments to evaluate our method.

1) We first rank the authors according to their LSCB and degree centrality, respectively. Then every ten successive authors in the two rank lists are grouped. To quantify the influence of each group in a SIR process, we study the size S_i of infected authors in an epidemic originating from each author belonging to the same group. The sizes of infected authors in the same group are averaged using following equation to verify whether top tanked authors are more influential.

$$S(G_i) = \frac{\sum_{u_j \in G_i} S_{u_j}}{Card(G_i)} \tag{11}$$

   $S(G_i)$ represents the average size of infected authors in group G_i, S_{u_j} denotes the size of infected authors originating at author u_j in group G_i and $Card(G_i)$ is the cardinality of group G_i.
2) To study weather LSCB is a better measurement of spreader's influence than degree centrality, we design the second group of experiments. Two groups of authors are selected. The authors in the first group have similar degree but distinctively different LSCB, while the authors in the second group have similar LSCB but different degree.

Table 3 shows the statistics of top authors identified by two methods. AL and AI represent the authors' average LSCB and degree in the same group, respectively. The results of the first group of experiments are shown in Figure 3. It can be seen from the figure that LSCB has a statistical positive relation with spreader's influence. Authors with higher LSCB values always infect more people than those with lower LSCB value. It also shows that degree is not directly related to author's information spreading capability. Authors with higher degree may infect less people than those with lower degree. For example, the average number of infected authors in the top11-20 group is smaller than that in the

Table 3 Statistics of top authors identified by the two methods

		Top10	Top11-20	Top21-30	Top31-40	Top41-50
LSCB	AL	193.9	183.7	118.5	104.6	97.4
	AI	247	174	127	112	116
In-degree	AL	168.5	108.5	114.9	124.5	87.3
	AI	262	168	146	130	119

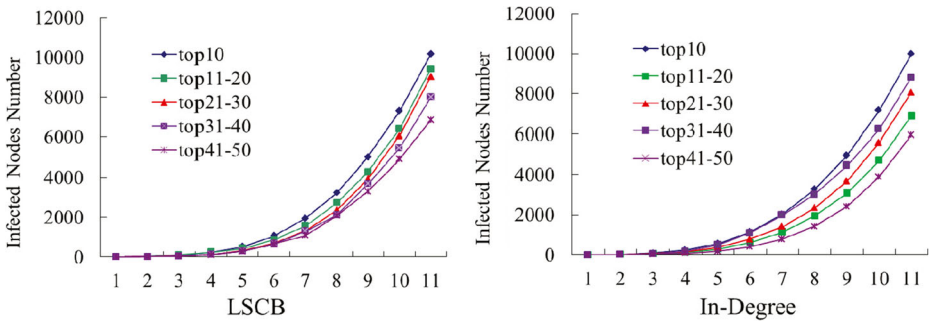

Figure 3 Results of the first group of experiments. *X-axis* represents spreading step/time and *Y-axis* represents the number of infected authors

top21-30 group and top31-40 group. The comparison of infected author size between LSCB and degree centrality methods is shown in Figure 4. This figure shows that the influential spreaders identified by LSCB can almost always infect larger number of authors than those identified by degree method.

Table 4 shows the statistics of selected authors in the second group of experiments. Eight authors are selected to form four pairs of examples. Authors in example 1 and 2 have similar degree but distinctively different LSCB. On the contrary, authors in example 3 and 4 have similar LSCB but different degree. Experimental results shown in Figures 5 and 6 demonstrate that authors with similar degree may have distinctively different influence and authors with different degree may have very similar influence if their LSCB are similar. Therefore, degree centrality may not be a reasonable metric to measure user's information spreading capability. By contrast, LSCB is a more reasonable metric. Authors having similar LSCB always have similar influence independent of their degree.

In this section, we propose a method to calculate users' *local social circle broadness* and evaluate it on a coauthor dataset empirically. Compared with degree centrality, LSCB is a better measurement of users' information dissemination capability. However, due to the limitation of LSCB, we propose a social circle weighted PageRank algorithm to iteratively rank each user's *global social circle broadness* in next section.

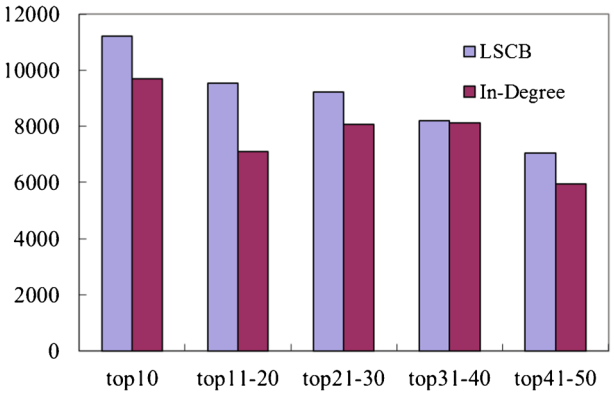

Figure 4 The number of infected authors by two methods. The maximum infected size is reported

Table 4 Statistics of authors selected in experiment

		Author ID	LSCB	In-degree
Similar In-degree, different LSCB	Example 1	3714	122	176
		454	94	177
	Example 2	6548	112	152
		2636	78	151
Similar LSCB, different In-degree	Example 3	1245	168	257
		1233	159	184
	Example 4	2658	180	277
		3166	172	208

4 Global social circle broadness: social circle weighted PageRank

The limitation of LSCB is due to the fact that it only takes the distribution of user's friends among communities into consideration, however, ignores the effect of the user's friends' friends to this user. We give an example shown as Figure 7. Suppose we have two users u_1 and u_2, and they have the same number of friends and the same distribution of friends. Nevertheless, u_2's friends' friends distribute in different communities while u_1's friends' friends are closely connected in a community. Whose social circle is broader and who is more influential in information diffusion? Obviously, the two users should not be considered to have the same social circle broadness.

To address this problem, a revised PageRank algorithm, social circle weighted PageRank (SCWPR) is proposed. PageRank is an effective web page ranking algorithms based on the link structure of the web [31]. The basic idea of PageRank is that if page p_i has a link to page p_j, then page p_i is implicitly conferring some importance to page p_j. Similarly, the influence of a user in social network can be interpreted to the "authority" of a web page: $user_i$ has high influence if $user_i$'s friends are influential. Based on the similarity of the two tasks, some prior work have revised PageRank algorithms for influential users ranking in social networks [39].

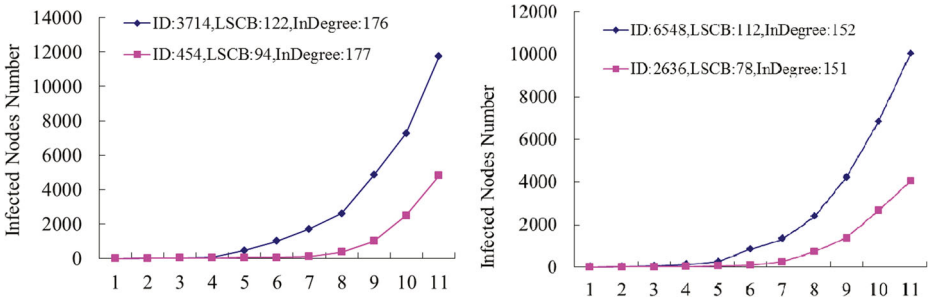

Figure 5 Results of the second group of experiments. Two authors with similar degree but distinctively different LSCB are selected. *ID* represents author ID, *KL* represents LSCB and *InDegree* represents degree

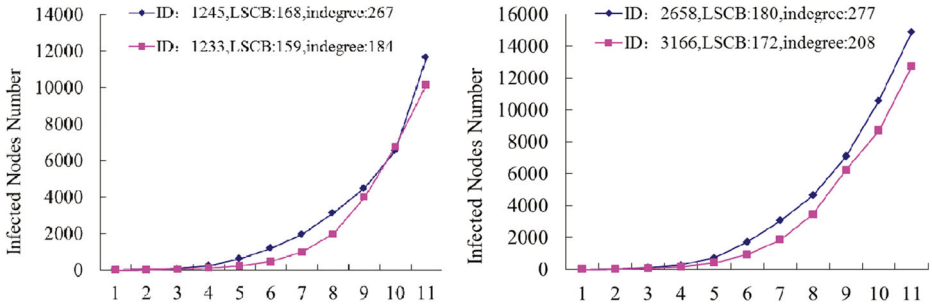

Figure 6 Results of the second group of experiments. Two authors with similar LSCB but distinctively different degree are selected. *ID* represents author ID, *KL* represents LSCB and *InDegree* represents in-degree

The idea of SCWPR is similar to PageRank: $user_i$'s social circle is broad if her friends's social circles are broad. However, SCWPR differs from PageRank in the following two aspects. First, SCWPR aims to rank user's global social circle broadness rather than user's authority. Second, SCWPR is a weighted PageRank algorithm. The edges between two users with considerably different social circles are assigned larger weight, because they are more likely to get more information from each other. Before describing SCWPR in detail, a definition of *social circle divergency between users* are given as follow:

Definition 5 (Social Circle Divergency Between two Users) The social circle divergency between $user_i$ and $user_j$ can be calculated as:

$$dis(u_i, u_j) = \sqrt{2 \cdot D_{JS}(P^i, P^j)} \tag{12}$$

where $D_{JS}(P^i, P^j)$ is the Jensen-Shannon divergence between the two distributions of $user_i$'s friends P^i and $user_j$'s friends P^j. $D_{JS}(P^i, P^j)$ is defined as:

$$D_{JS}(P^i, P^j) = \frac{1}{2}(D_{KL}(P^i||R) + D_{KL}(P^j||R)) \tag{13}$$

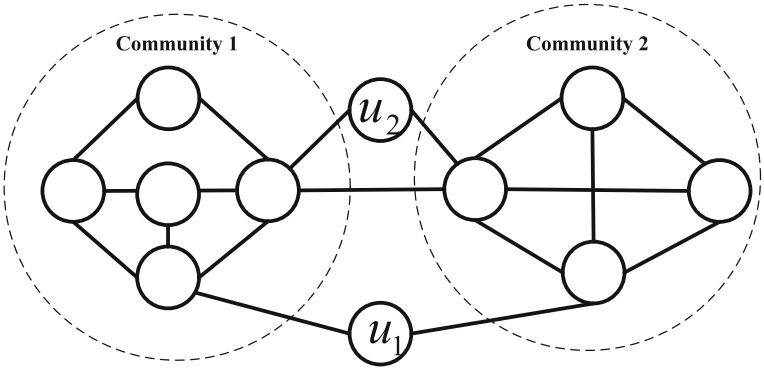

Figure 7 u_1 and u_2 have the same number of friends and the same distribution of friends . Hence their LSCB is the same. However, u_2's global social circle broadness is larger than u_1's, because the neighbors u_2 connected to also connect different parts of the network

where R is the average of the two probability distributions, i.e., $R = \frac{1}{2}(P^i + P^j)$ and D_{KL} is the Kullback-Leibler divergence.

We use $dis(u_i, u_j)$ instead of Kullback-Leibler divergence, because it was proved in [14] that $dis(u_i, u_j)$ is symmetric and fulfills the triangle inequality. To normalize $dist(i, j)$, the sigmoid function is also used as a mapping function. The normalized social circle divergence between two users can be represented as:

$$dis_{sig}(u_i, u_j) = \frac{1}{1 + e^{-dis(u_i, u_j)}} = \frac{1}{1 + e^{-\sqrt{2 \cdot D_{JS}(P^i, P^j)}}} \tag{14}$$

An example of social circle divergence between u_1 and her neighbors is shown as Figure 8. The social circle divergence between u_1 and u_2 is large, and thus the transition probability between them is high. u_4 and u_5 are in the same community with u_1, hence the transition probability from u_1 to them is low.

It is worth pointing out that the purposes of utilizing K-L divergences for a user's *social circle diversity* definition and two users' *social circle divergency* definition are different. In Definition 3, smaller K-L divergence between the distribution of user's friends and distribution of communities means larger *social circle diversity* for this user. Nevertheless, in Definition 5, larger K-L divergence between two users means less overlap between their social circles. Thus both users can obtain more information from each other, and expand their social circles through each other's social circle. Therefore, the more friends a user has and the larger K-L divergence between the user and her friends, the broader her global social circle is.

Based on above definition and analysis, $user_i$'s *global social circle broadness* depends on both her popularity (number of friends) and her friends' *global social circle broadness*. Hence, we utilize the following formula to compute $user_i$'s *global social circle broadness*:

$$auth(u_i) = d \times \sum_{u_j \in In(u_i)} \frac{(1 + dis_{sig}(u_j, u_i)) \cdot auth(u_j)}{N_j + \sum_{u_k \in Out(u_j)} dis_{sig}(u_j, u_k)} + (1 - d) \tag{15}$$

Here, $d \in (0, 1)$ is a dumping factor similar to PageRank, $dis_{sig}(u_j, u_i)$ is the social circle divergence between $user_j$ and $user_i$. $In(u_i)$ is the set of users pointing to $user_i$, and $Out(u_j)$ is the set of users u_j points to.

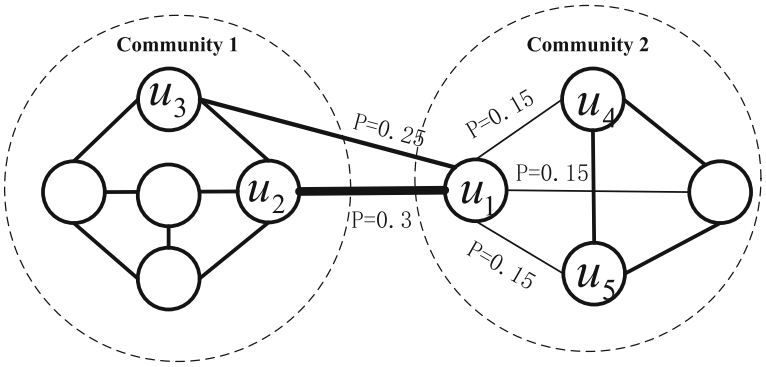

Figure 8 Illustration of transition probabilities between u_1 and its neighbors in social circle weighted PageRank

$user_i$'s *global social circle broadness* is calculated by an iterative algorithm. At the initial instant, each authority value is initialized to:

$$auth^0(u_i) = \frac{1}{N} \tag{16}$$

where N is the total user number in the social network. At each step, the algorithm computes $user_i$'s *global social circle broadness* iteratively as:

$$auth^t(u_i) = d \times \sum_{u_j \in In(u_i)} \frac{(1 + dis_{sig}(u_j, u_i)) \cdot auth^{t-1}(u_j)}{N_j + \sum_{u_k \in Out(u_j)} dis_{sig}(u_j, u_k)} + (1 - d) \tag{17}$$

In matrix notation, formula (14) can be rewritten as follows:

$$\mathbf{Auth} = \mathbf{d} \cdot \mathbf{D} \times \mathbf{Auth} + (\mathbf{1} - \mathbf{d}) \cdot \mathbf{E} \tag{18}$$

Auth is the *global social circle broadness* vector of all the users, **D** is the column normalized transition probability matrix and **E** is the unite vector. Transition probability matrix is defined as:

$$D_{ij} = \frac{1 + dis_{sig}(u_i, u_j)}{N_j + \sum_{u_k \in Out(u_i)} dis_{sig}(u_i, u_k)} \tag{19}$$

5 Experiment and evaluation

In this section, we empirically evaluate LSCB and SCWPR on four real social network datasets. The SIR model described in Section 2.3 is applied as the information diffusion model, and six structure based baseline methods are compared with the two methods. We first introduce the datasets and the baseline methods, then conduct experiments to evaluate our methods about the scope of information propagation on these datasets.

5.1 Datasets

The following four real social network datasets are used to empirically evaluate our methods. More detail information about them are shown in Table 5.

Twitter dataset[2] We use the twitter dataset published by Tang et al. [35, 36]. It was crawled from Twitter.com. The sub-network is comprised of 112,044 users and 468,238 links among them.

Slashdot social network dataset[3] Slashdot is a technology-related news website. The site features user-submitted and evaluated news stories about science and technology related topic. Each story has a comments section attached to it. The network dataset contains friends/foe links between users of Slashdot. It contains 82,168 users and 948,464 links.

MathSciNet co-authorship dataset[4] This dataset is a co-authorship dataset from the Mathematical review collection of the American Mathematical Society. This network contains 873,775 links of collaboration between 391,529 mathematicians.

[2] http://arnetminer.org/structural-hole

[3] http://snap.stanford.edu/data/soc-Slashdot0902.html

[4] http://hal.elte.hu/cfinder/wiki/?n=Main.Data.

Table 5 Statistics of the four real social network datasets

Datasets	N	E	$< k >$	k_{max}	C_N
Twiter dataset	112,044	468,238	4.18	233	1684
Slashdot social network dataset	82,168	948,464	11.54	2532	1622
MathSciNet co-author dataset	391,529	873,775	2.23	496	16,484
DBLP DB/DM co-author dataset	48,500	133,111	2.74	225	3044

N is the number of nodes, E is the number of edges, $< k >$ is the average degree in the network, k_{max} is the max degree in the network and C_N is the number of communities

DBLP DB/DM co-authorship dataset[5] The whole dataset contains 1,572,277 papers and 2,084,019 citation relationships. We only extract papers appeared in database and data mining related conferences. Co-author relationship are then extracted from these papers. The smaller dataset contains 48,500 authors and 133,111 co-author relationships.

To demonstrate the effectiveness of our approaches, we compare them against the following baseline methods:

- **Degree Centrality** As a widely used centrality based measure of user's influence, Kempe et al. have shown that high degree nodes may outperform other centrality-based methods [24]. Hence, we use it as a comparison.
- **Betweenness Centrality** Betweenness is equal to the number of shortest paths from all nodes to all others that pass through that node. Recently, betweenness centrality has become an important centrality measure in social network [6, 16].
- **PageRank** PageRank method is a representative eigenvector centrality method. For its excellent performance in the web pages ranking [31], we also use it as a comparison.
- ***k*-shell+Degree Centrality** [26] reported that k-shell method can mining the nodes in the core of the network. The most efficient spreaders are those located within the core of the network as identified by the k-shell decomposition analysis. However, it can not rank nodes in the same k-shell. To tackle this problem, we rank the nodes in the same k-shell layer according to their degree.
- ***k*-shell+Betweenness Centrality** Similar to k-shell+Degree Centrality method, we rank the nodes in the same k-shell layer according to their betweenness.
- **Hubs Identification** Similar to our idea, [18] proposed a method to classify nodes in a network as non-hubs that have a low within-module degree, and hubs that have high within-module degree. Additionally, hubs can be further divide as provincial hubs, connector hubs, and global hubs according to their participation coefficient. However, this method does not aim to rank nodes' influence, and hence can not be directly used as a baseline to evaluate our approach. Therefore, to conduct a comparison with it, we first identify the hub nodes according to their within-module degree z, then these hub nodes are ranked according to their participation coefficient. By doing so, the top ranked nodes are those global or connector hubs with their most links connecting to most of the communities.

[5]We use the DBLP dataset published by Jie Tang et al in Arnetminer [20, 27]. The whole dataset can be downloaded from http://arnetminer.org/citation.

5.2 Correlation

We first study the correlation between every two rank lists generated by degree centality (DE), PageRank (PR), local social circle broadness (LSCB) and social circle weighted PageRank (SCWPR) methods. We conduct this experiment because we want to show: 1) how different the results of some methods are, and 2) how similar the results of some methods are. For example, a low correlation between DE and LSCB methods implies that users with very similar degree may have significantly local social circle broadness, and vice versa.

Here, τ proposed by Kendall [25] is used to measure the correlation. The value of τ is in the range of [-1, 1]. If two lists are exactly the same, $\tau = 1$; whereas $\tau = -1$ if one list is the reverse of the other. Larger τ implies higher agreement between the two lists. The Kendall τ coefficient is defined as:

$$\tau = \frac{n_{con} - n_{dis}}{\frac{1}{2}n(n-1)} \tag{20}$$

where n_{con} is the number of concordant pairs, n_{dis} is the number of discordant pairs, and n is the size of the rank list. Table 6 shows the results of τ values by comparing every two rank lists by the four methods.

From the result we can see that 1) first, degree centrality has a lower agreement with LSCB and SCWPR, which implies that a user's social circle broadness is not entirely determined by her degree. As show in the preliminary experimental results in Subsection 3.4, users with similar degree may have distinctively different LSCB or/and SCWPR, and those with similar LSCB or/and SCWPR may have remarkably different number of neighbors; 2) second, LSCB has a higher agreement with SCWPR as both methods take the social circle diversity into account; 3) additionally, PR also has a higher agreement with SCWPR than with LSCB. This is probably due to the fact that SCWPR is an revised algorithm of PR and both methods calculate the user's authority iteratively in the same way.

5.3 Experiment on SIR model

In this section, we apply the SIR model to evaluate our methods from two perspectives: conduct the comparison experiment over different Top-K spreaders and over different spreading probability β.

Similarly to the preliminary experiments in Section 2.4. We first rank the users by various methods, then every ten or fifty successive users are grouped. For different datasets, we set different group sizes because of their different structure properties. For twitter and MathSciNet co-authorship datasets, the group size is set to ten because the difference of users' average spreading influence between groups is significant in this case. However, for slashdot and DBLP DB/DM co-authorship datasets, the group size is set to fifty because the users' average spreading influence between groups is not significantly different if the group size is small. For each group, we run the SIR model originating from each user in the same group separately, and average the results. We conduct the experiments in this way for two reasons. Firstly, the difference of influence between users in the same group is not significant. Therefore, it is hard to evaluate the influence of single spreader in the same group. Secondly, it is fair that we compare the average results for users in the same group with the results of other groups.

Table 6 Correlation between every two rank lists by different algorithms on the four datasets. Four algorithms are used: Degree Centrality (DE), PageRank (PR), local social circle broadness (LSCB) and social circle weighted pagerank (SCWPR)

	Top-10	Top-20	Top-30	Top-40	Top-50
MathSciNet co-author dataset					
DE vs LSCB	0.3146	0.4233	0.4720	0.4473	0.4540
DE vs SCWPR	0.4123	0.3430	0.3890	0.4215	0.4200
PR vs LSCB	0.3145	0.4656	0.3674	0.4851	0.4655
PR vs SCWPR	0.6435	0.6827	0.6370	0.6245	0.7270
LSCB vs SCWPR	0.5150	0.5247	0.6135	0.6325	0.6125
Twitter dataset					
DE vs LSCB	0.3745	0.4257	0.5570	0.5445	0.5648
DE vs SCWPR	0.4255	0.4145	0.4045	0.3865	0.4531
PR vs LSCB	0.4667	0.5025	0.5245	0.5540	0.5510
PR vs SCWPR	0.6487	0.6543	0.6280	0.6149	0.6685
LSCB vs SCWPR	0.6025	0.5420	0.5324	0.5579	0.5820
DBLP DB/DM co-author dataset					
DE vs LSCB	0.4222	0.4656	0.4480	0.4560	0.4435
DE vs SCWPR	0.3560	0.3585	0.4245	0.4415	0.4250
PR vs LSCB	0.4456	0.4875	0.4425	0.4852	0.5245
PR vs SCWPR	0.6758	0.6525	0.5880	0.6224	0.6150
LSCB vs SCWPR	0.5567	0.5640	0.5245	0.5145	0.5612
Slashdot social network dataset					
DE vs LSCB	0.5545	0.5650	0.5420	0.4875	0.4645
DE vs SCWPR	0.3850	0.4550	0.4045	0.4815	0.4650
PR vs LSCB	0.50506	0.5275	0.5420	0.4750	0.5100
PR vs SCWPR	0.7015	0.6150	0.5900	0.6145	0.6250
LSCB vs SCWPR	0.5675	0.5340	0.5050	0.4840	0.4815

5.4 Comparison experiment over different Top-K spreaders

In this subsection, we do analysis of the average infected percentage of the users in the network for different groups. For each group, we run the SIR model and record the number of ultimately infected users when a virus originates from each user in the group, then we average the numbers and use it as the average size of infected users of this group. Figures 9–12 show the experimental results.

First, we clarify how to set the diffusion probability β. When β is too small, only a few nodes will be infected or even no nodes will be infected besides the initial infected node. Conversely, in the case of large β, the role of individual nodes is no longer important and spreading would always cover a large part of the whole network [26]. Therefore, it is

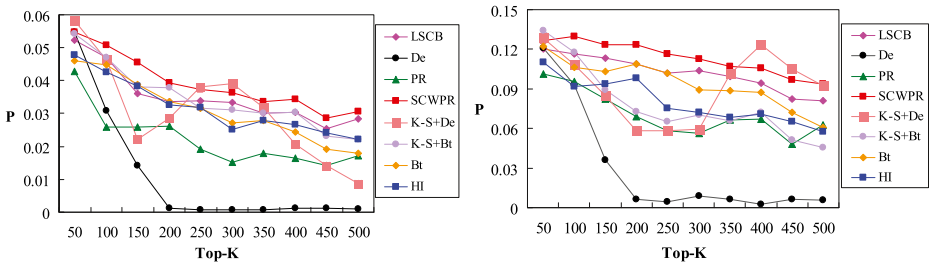

Figure 9 Experimental results on DBLP DB/DM co-authorship dataset. The infection probability β is set to 0.08 and 0.12

unavailable to evaluate our approach in both cases, and usually small β values are selected when SIR model is used to simulate information or virus spreading [23, 26, 34]. Meanwhile, depending on the specific network structure, the information diffusion precess may be vastly different for different datasets. Hence, for different datasets, we evaluate them with different infection probabilities, but on the whole, all the probabilities are smaller than 0.15. Additionally, as there are too many β values, we conduct experiment on various β values but only report two representative experimental results for each dataset.

The x-axis represents Top-K users identified by different methods and the y-axis represents the average infected percentage P of users when a virus originates from Top-K users in the same group. For simplicity, we use De, PR, Bt, K-S+De, K-S+Bt and HI to denote degree centrality, PageRank, betweenness centrality, k-shell+degree centrality, k-shell+betweenness centrality and hubs identification methods, respectively.

Experimental results show that SCWPR almost always outperforms other methods, for its infected percentage is always higher. LSCB is slightly inferior than SCWPR. However, LSCB is competitive with all the baseline methods. Among the baseline methods, betweenness centrality is much more robust than the others. Performance of degree centrality, PageRank, hubs identification and k-shell based methods differs greatly for different datasets. For example, degree centrality method is almost ineffective for the DBLP DB/DM co-authorship dataset when Top-K is higher than 200. Similarly, the performance of k-shell+betweenness centrality is also not inspired for the MathSciNet co-authorship dataset.

Experiment results demonstrate the effectiveness of LSCB and SCWPR. It implies that there exists some relationship between user's social circle broadness and information diffusion. We also observed that SCWPR outperformed LSCB. This is due to the fact that

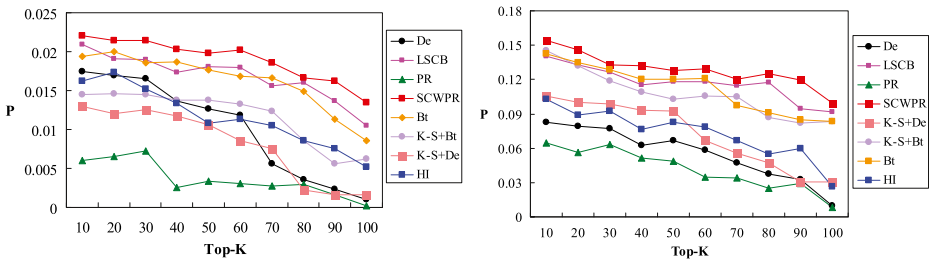

Figure 10 Experimental results on twitter dataset. The infection probability β is set to 0.1 and 0.2

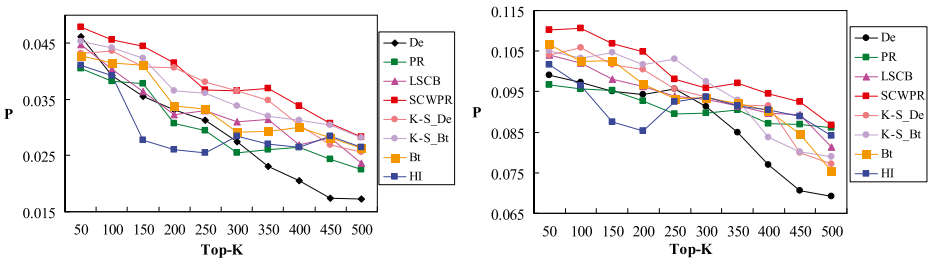

Figure 11 Experimental result on slashdot dataset. The infection probability β is set to 0.02 and 0.03

global social circle broadness is better to measure user's information propagation capability comparing with local social circle broadness.

5.5 Comparison experiment over different β

To further investigate the effectiveness of our methods, we conduct experiments to observe the average infected percentage of the users in the network when a virus originates from the Top-K users over different spreading probability β. In many scenarios, only the $Top-k$ most influential users are required and meaningful to some applications. Therefore, we only report and compare the average results of the authors in the first two groups. For MathSciNet co-authorship and Slashdot datasets, we report the average results of the Top50 and Top50-100 users, and for DBLP DM/DB co-authorship and twitter datasets, we report the average result of the Top10 and Top10-20 users.

As detailed in the last subsection, the performance of a method can be significantly different for different datasets. A method may be effective for some datasets, but not so inspired for others. Therefore, for each dataset, we drop the ineffective methods and only select three most effective methods as baselines. Figures 13–16 show the experimental results. The x-axis is the information diffusion probability β and the y-axis is the average infected percentage P of users.

The experimental results show that SCWPR outperforms all the other methods in the infected scope of the network, independently of the spreading probability β. For the MathSciNet co-authorship and DBLP DM/DB datasets, the performance of LSCB and betweenness centrality method is very similar, and both methods are much better than

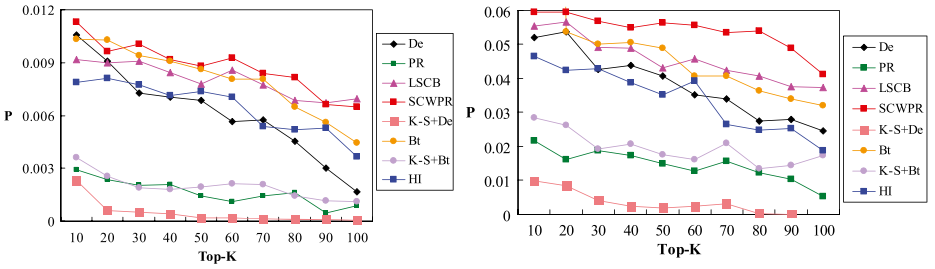

Figure 12 Experimental result on MathSciNet co-authorship dataset. The infection probability β is set to 0.08 and 0.12.

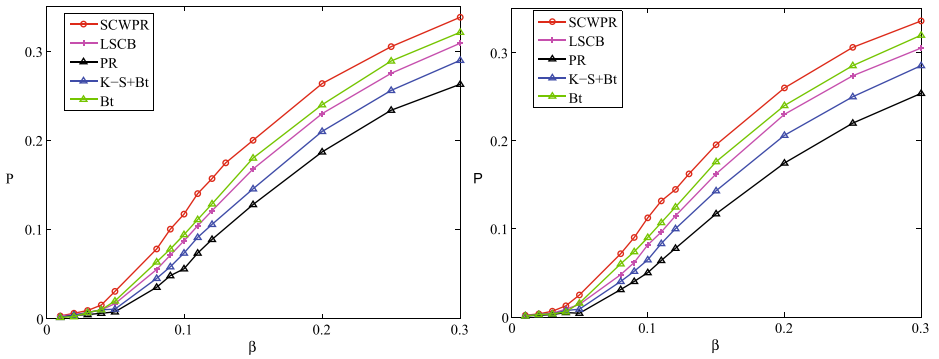

Figure 13 Experimental results on DBLP DM/DB co-authorship dataset over different β. We report the average results of Top50 (*left*) and Top50-100 (*right*) authors

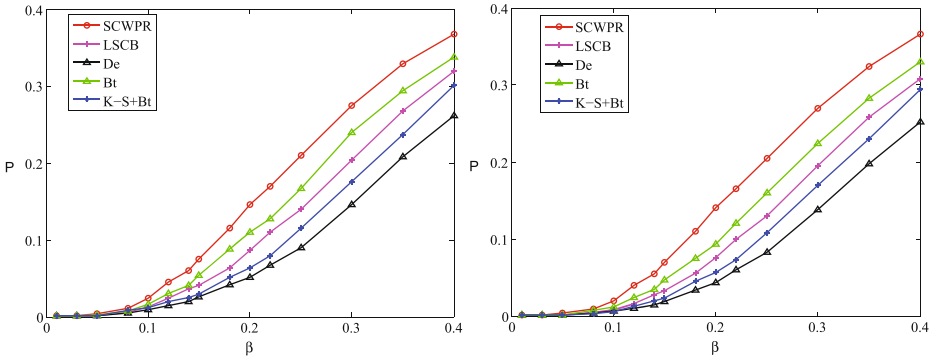

Figure 14 Experimental results on twitter dataset over different β. We report the average result of Top10 (*left*) and Top10-20 (*right*) users

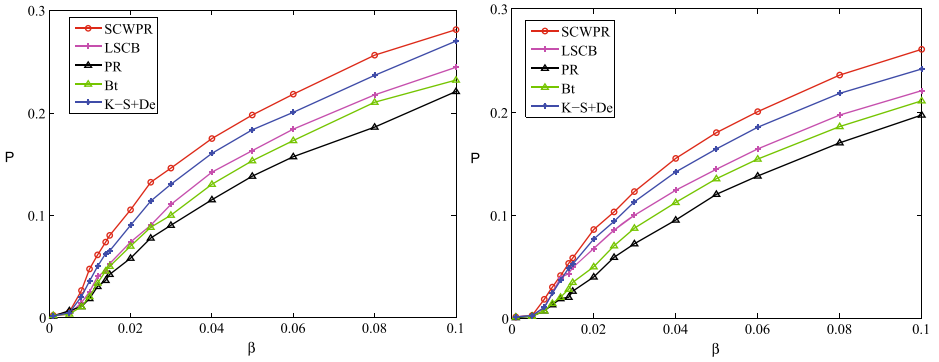

Figure 15 Experimental results on slashdot dataset over different β. We report the average result of Top50 (*left*) and Top50-100 (*right*) users

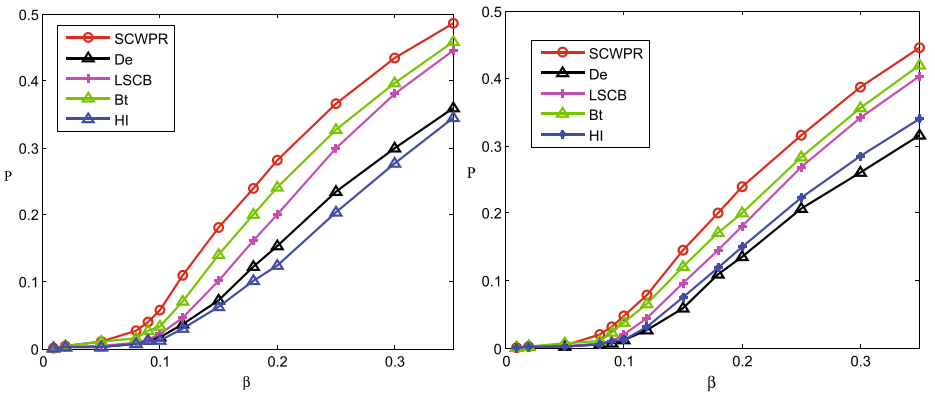

Figure 16 Experimental results on MathSciNet co-authorship dataset over different β. We report the average result of Top10 (*left*) and Top10-20 (*right*) users

other comparison methods. For twitter datasets, LSCB is slightly inferior than betweenness centrality but better than k-shell+betweenness centrality. For Slashdot dataset, LSCB is inferior than k-shell+betweenness centrality but superior than betweenness centrality. This Experiment also shows the effectiveness of LSCB and SCWPR.

6 Related work

Influential mining in social networks has been studied extensively in recent years for the increasing popularity of online social networks [12, 29, 40]. The threads of previous researches can be categorized into two aspects: network structure information and user generated content information.

Plenty of previous studies focused on discovering influential spreaders or important nodes in networks from the perspective of node connectivity property and the network structure. Early research found that the most connected people play a key role in the information dissemination process [34]. Furthermore, in the context of social network theory, the importance of a node in the network is often measured in its betweenness centrality, which means how many shortest paths cross through this node [6]. People with high betweenness centrality in a social network are believed to be more influential in information diffusion. However, these methods ignore the location information of the nodes in the social network. Kitsak et al. reported in [26] that the topology of the network organization plays a more important role in information diffusion. The nodes with high in-degree or high betweenness may have little effect in the range of a given spreading process. They used the k-shell decomposition [19] method to find core nodes in network. Haythornthwaite in [17] presented that users have greater access to information if their network has a greater range, that is, if they are a member of more and larger networks, and if their contacts are themselves members of large networks which do not overlap with their own networks. Goldenberg et al. in [10] discovered that influential people should have the following three important traits: (1) They are convincing, (2) they know a lot (i.e., are experts), and (3) they have a large number of social ties (i.e., they know a lot of people). Chung et al. in [37] developed a theoretical model based on social network theories and the social influence model to understand how knowledge

professionals utilize technology for work and communication. The association between egocentric network properties (structure, position and tie) and information and communication technology (ICT) is investigated. This paper also discussed how network structure, network position and network tie diversity influence information diffusion in ICT. Ugander et al. in [30] discovered the possibility of contagion in Facebook is tightly controlled by individual's structural diversity rather than by the size of the neighborhood. Though they did not tend to mining influential spreaders, their research showed the importance of individual's structural diversity in social contagion.

Some work utilized both the network structure information and the content information for topic level influential identification and influence mining. Nitin et al. in [1] proposed a method to discover influential bloggers in a blog community. They investigated what constitutes influential bloggers from the perspective of their posted blogs and presented a preliminary model attempting to quantify an influential blogger. A novel finding topic-sensitive influential twitters method TwitterRank is proposed by Weng et al. in [39]. As an extension of PageRank, TwitterRank measures the influence taking both the topical similarity between users and the link structure into account. To find the most interesting and authoritative MicroBlog authors for any given topical metrics, Aditya et al. proposed a set of features for characterizing social media authors, including both nodal and topical metrics in [2]. To estimate the influence probabilities between users in influence propagation process, Amitl et al. proposed a algorithm to build influence models by learning the model parameters in [28]. They also developed techniques for predicting the time by which a user may be influenced. Michael et al. in [22] proposed that the level of expertise of a user with respect to a particular topic is mainly determined by two factors in a collaborative tagging system. A graph-based algorithm, SPEAR, which implements the two factors for ranking users in a folksonomy is proposed. Jie et al. in [9] presented a topic level expertise search framework for heterogeneous networks. A unified topic model to simultaneously model topic aspects of different objects in the academic network is presented. Hui et al. in [21] proposed a novel algorithm CASINO (Conformity-Aware Social INfluence cOmputation) to study the interplay between influence and conformity of each individual. Compared with other influence models, their model utilized both positive interaction (e.g., agreement, trust) and negative relationships (e.g., distrust, disagreement).

7 Conclusion

In this paper, we studied such an interesting problem: whether and to what extend the user's social circle will affect her information diffusion capability, and proposed a method exploiting the user's social circle broadness for influential spreaders identification in social networks. Firstly, the distributions of communities and user's friends are defined based on the results of community detection, then the user's social circle diversity is defined by calculating the K-L divergence between the two distributions. Secondly, based on the user's social circle diversity, a method to measure the user's local social circle broadness is presented. Preliminary experiments have shown that there exists some relationship between the user's local social circle broadness and her information diffusion capability. Furthermore, to iteratively rank each user's global social circle broadness, a social circle weighted PageRank algorithm is developed. By exploiting users' global social circle broadness, we rank their influence in information spreading. Finally, we verified the effectiveness of our method by extensive comparison experiment on four real social network datasets.

This paper studied influential spreaders mining in social networks from the network structure view. As for the future work, we will focus on (**a**) how to combine the user generated content information with the network structure information and design a unified model for influential spreaders mining, (**b**) now to apply our approach to the problem of identifying multiple initial spreaders, and (**c**) how to apply the influential mining results to facilitate other applications, such as recommendation and link prediction.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant Nos. 61170189, 61370126, 61202239), the Research Fund for the Doctoral Program of Higher Education (Grant No. 20111102130003), and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2013ZX-19).

References

1. Aditya, P., Scott, C.: Identifying topical authorities in microblogs. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'2011)
2. Amitl, G., Francesco, B., Laks, V.S.L.: Learning influence probabilities in social networks. In: Proceedings of the 3th ACM International Conference on Web Search and Data Mining (WSDM'2010)
3. Anderson, R.M., May, R.M., Anderson, B.: Infectious Disease of Humans: Dynamic and Control. Oxford Science, Oxford (1992)
4. Baumes, J., Goldberg, M., Krishnamoorthy, M., Magdon-Ismail, M., Preston, N.: Finding communities by clustering a graph into overlapping subgraphs. In: IADIS International Conference on Applied Computing (2005)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebrve, E.: Fast unfolding communities in large networks. In: Journal of Statistical Mechanics: Theory and Experiment, P10008 (2008)
6. Brandes, U.: On variants of shortest-path betweenness centrality and their generic computation. Soc. Networks **30**, 136C145 (2008)
7. Burt, R.S.: Structure Holes: the Social Structure of Competition. Harvard University Press, Massachusetts (1992)
8. Cao, J., Wu, Z.A., Wu, J.J., Xiong, H.: SAIL: summation-bAsed incremental learning for information-theoretic text clustering. IEEE Trans. Syst. Man Cybern. B **43**(2), 570–584 (2013)
9. Carmi, S., Havlin, S., Kirkpatrick, S., Shir, Y., Shir, E.: A model of Internet topology using k-shell decomposition. Proc. Natl. Acad. Sci. U.S.A. **104**, 11150–11154 (2007)
10. Chung, K.S.K., Hossain, L.: Network structure, position, ties and ICT use in distributed knowledge-intensive work. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW'2008)
11. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)
12. Cui, Y., Pei, J., Tang, G.T., Luk, W.S., Jiang, D.X., Hua, M.: Finding email correspondents in online social networks. World Wide Web J. **16**(2), 195–218 (2013)
13. Diekmann, O., Heesterbeek, J.A.P.: Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation. Wiley Series in Mathematical & Computational Biology, New York (2000)
14. Endres, D.M., Schindelin, J.E.: A new metric for probability distributions. IEEE Trans. Inf. Theory **49**(7), 1858–1860 (2003)
15. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
16. Goh, K.C.I., Oh, E., Kahng, B., Kim, D.: Betweenness centrality correlation in social networks. Phys. Rev. **E**, 61 (2003)
17. Goldenberg, J., Han, S., Lehmann, D.R., Hong, J.W.: The role of hubs in the adoption. J. Mark. **73**, 1–43 (2011)
18. Guimer*à*, R., Sales-Pardo, M., Amaral, L.A.N.: Classes of complex networks defined by role-to-role connectivity profiles. Nat. Phys. **3**, 63–69 (2007)
19. Haythornthwaite, C.: Social network analysis: an approach and technique for the study of information exchange. Libr. Inf. Sci. Res. **18**(4), 323–342 (1996)
20. Hopcroft, J., Lou, T., Tang, J.: Who will follow you back? reciprocal relationship prediction. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM'2011), pp. 1137C–1146 (2011)

21. Hui, L., Sourav, S.B., Aixin, S.: CASINO: towards conformityaware scoial influence analysis in online social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'2011)
22. Jie, T., Jing, Z., Ruoming, J.: Topic level expertise search over heterogeneous networks. Mach. Learn. **82**, 211–237 (2011)
23. Keeling, M.J., Rohani, P.: Modeling Infectious Diseases in Humans and Animals. Princeton University Press, Princeton (2008)
24. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD Internatioanl Conference on Knowledge Discovery and Data Mining (KDD'2003), pp. 1175C–1180 (2003)
25. Kendall, M.: A new measure of rank correlation. Biometrika **30**(1–2), 81–93 (1938)
26. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identifying influential spreaders in complex networks. Nat. Phys. **6**, 888C893 (2010)
27. Lou, T., Tang, J., Hopcroft, J., Fang, Z., Ding, X.: Learning to predict reciprocity and triadic closure in social networks. In: TKDD (2013)
28. Michael, G.N., Ching-man, A.Y., Nicholas, G., Christoph, M., Nigel, S.: Telling experts from spammers: expertise ranking in folksonomies. In: Proceedings of the 32th Annual International ACM SIGIR Conference (SIGIR'2011)
29. Musial, K., Kazienko, P.: Social networks on the internet. World Wide Web J. **16**(1), 31–72 (2013)
30. Nitin, A., Huan, L., Lei, T.: Identifying the influential bloggers in a community. In: Proceedings of the 1th ACM International Conference on Web Search and Data Mining (WSDM'2008)
31. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference (WWW'1998).
32. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814 (2005)
33. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**, 3200–3203 (2001)
34. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scalefree networks. Phys. Rev. Lett. **3200–3203**, 86 (2002)
35. Tang, J., Zhang, J., Yao, L.M., Li, J.Z., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)
36. Tang, J., Zhang, D., Yao, L.M.: Social network extraction of academic researchers. In: Proceedings of 2007 IEEE International Conference on Data Mining (ICDM'2007)
37. Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J.: Structural diversity in social contagion. Proc. Natl. Acad. Sci. U. S. A. **109**(16), 5962–5966 (2012)
38. Wei, F., Qian, W.N., Wang, C., Zhou, A.Y.: Detecting overlapping community structures in networks. World Wide Web J **12**(2), 235–261 (2009)
39. Weng, J.S., Lim, E.P., Jiang, J.: TwitterRank: finding topic sensitive influential twitterers. In: Proceedings of the 3th ACM International Conference on Web Search and Data Mining (WSDM'2010)
40. Zhang, R.C., Tran, T., Mao, Y.Y.: Opinion helpfulness prediction in the presence of words of few mouths. World Wide Web J. **15**(2), 117–138 (2012)