Kim *et al. Computational Social Networks* (2015) 2:3
DOI 10.1186/s40649-015-0013-8

Computational Social Networks
a SpringerOpen Journal

RESEARCH Open Access

A study on the influential neighbors to maximize information diffusion in online social networks

Hyoungshick Kim[1*], Konstantin Beznosov[2] and Eiko Yoneki[3]

*Correspondence:
hyoung@skku.edu
[1] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Korea
Full list of author information is available at the end of the article

Abstract

The problem of spreading information is a topic of considerable recent interest, but the traditional influence maximization problem is inadequate for a typical viral marketer who cannot access the entire network topology. To fix this flawed assumption that the marketer can control any arbitrary k nodes in a network, we have developed a decentralized version of the influential maximization problem by influencing k neighbors rather than arbitrary users in the entire network. We present several practical strategies and evaluate their performance with a real dataset collected from Twitter during the 2010 UK election campaign. Our experimental results show that information can be efficiently propagated in online social networks using neighbors with a high propagation rate rather than those with a high number of neighbors. To examine the importance of using real propagation rates, we additionally performed an experiment under the same conditions except the use of synthetic propagation rates, which is widely used in studying the influence maximization problem and found that their results were significantly different from real-world experiences.

Keywords: Information diffusion; Information dissemination; Online social networks; Viral marketing

Introduction

In the field of social network analysis, a fundamental problem is to develop an epidemiological model for finding an efficient way to spread information through the model. It seems natural that many people are often influenced by their friends' opinions or recommendations. This is called the 'word of mouth' effect and has for long been recognized as a powerful force affecting product recommendation [1].

Recent advances in the network theory have provided us with the mathematical and computational tools to understand them better. For example, in the *Independent Cascade* (`IC`) model proposed by Goldenberg et al. [2], (1) some non-empty set of nodes are initially *activated* (or influenced); (2) at each successive step, the influence is propagated by activated nodes, independently activating their inactive neighbors based on the *propagation probabilities* of the adjacent edges. Here, activated nodes mean the nodes that have adopted the information or have been infected. This models how a piece of information

will likely be spread through a network over time. It enables us to investigate what sort of information diffusion scheme might be the most effective one under certain conditions.

This model is also highly relevant to security. For example, cyberstalkers might be interested in spreading rumors, gossips, news, or pictures through social networks to damage their victims' (e.g., celebrity, political party, company, or country) reputation. The same model works in social media campaign where spammers and propagandists want to share their advertisements on online social networks; fake accounts with automated bots are often used to amplify advertising campaigns in social media [3-5].

Thus far, however, the models and analytic tools used to analyze epidemics have been somewhat limited. Most previous studies [6,7] aimed to analyze the process of information diffusion by choosing a set of arbitrary k nodes in a network as the initially activated nodes from a bird's eye perspective based on the full control of the entire network, which may indeed be unacceptable in many real-life applications since there is no such central entity (except the online social network service provider itself).

From the point of view of an individual user (e.g., viral marketer) who wants to efficiently spread a piece of information (or a rumor) through a network, a more reasonable epidemiological model would not assume the knowledge of the entire network topology. Kim and Yoneki [8] recently introduced the problem called Influential Neighbor Selection (`INS`) where a spreader s spreads a piece of information through carefully chosen k neighbors of hers instead of a set of any arbitrary k nodes in a network. Under this model, each user can only communicate with the user's immediate neighbors and has no knowledge of the global network topology except for her own connections. However, their work has two limitations: (1) it was simply assumed to use a constant propagation rate, despite variations in user propagation rates in practice. For example, in real-world online social network services such as Twitter or Facebook, each user has a distinct propagation rate for her neighbors on spreading information according to the user's reputation and/or role, such as opinion formers, leaders, or followers [9]; (2) their experimental results were limited to undirected graphs with parameter values chosen in a somewhat *ad hoc* manner.

Recently, Kim [10] extended this model by introducing several parameters (*user propagation weight*, *content interestingness*, and *decay factor*) to provide a more general and practical information diffusion model. This gives much finer granularity than the previous model [8]. However, their experiments still depended on synthetic parameters that might significantly affect the information diffusion process.

With a real dataset (Twitter users and messages related to the 2010 UK election campaign), we revisited the `INS` problem and evaluated the performance of four spreading schemes from the simple random neighbor selection to a sophisticated neighbor selection scheme using both the 'number of friends' and 'user propagation rate' each neighbor has. To measure the performance of these schemes, we used the conventional *Independent Cascade* (`IC`) model [2], which is widely used for the analysis of information diffusion [2,7,11].

In particular, we demonstrated the importance of using real propagation rates by comparing the simulation results with those using the randomly assigned synthetic parameters which were often used in studying the influence maximization problem. Our comparison results show that their results were significantly different, which indicate that the use of such synthetic propagation rates might be undesirable to understand the characteristics of information diffusion on a real-world social network (e.g., Twitter).

We also performed a simulation with various parameters. Our experimental results suggest that the scheme to select neighbors who wrote popular posts produced the best overall results, even without consideration of the 'number of friends'. Moreover, we found that the information diffusion speed of some schemes (e.g., random neighbor selection) in the previous study [10] was quite exaggerated and thus contributed to the reduction of the performance gap between information diffusion schemes. For example, we observed that the `random` selection scheme is not practically effective even with a high number k initially activated nodes; this is quite different from previous studies [8,10], which showed that the `random` selection scheme achieved reasonable performance when $k \geq 3$.

The rest of this paper is organized as follows. Related work is discussed in Section 'Related work'. In Section 'Influential neighbor selection problem', we formally define the `INS` problem and notations. Then, we present the four reasonable neighbor selection schemes in Section 'Neighbor selection schemes'. In Section 'Experimental results', we evaluate their performance through simulation with a real dataset collected from Twitter and recommend the best neighbor selection scheme with various conditions. We conclude in Section 'Conclusions'.

Related work

Influential maximization (`IM`) problem has recently received increasing attention, given the growing popularity of online social networks, such as Facebook and Twitter, which have provided great opportunities for the diffusion of information, opinions, and adoption of new products.

The `IM` problem was originally introduced for marketing purposes by Domingos and Richardson [6]: The goal is to find a set of k initially activated nodes with the maximum number of activated nodes after the time step t. Kempe et al. [7] formulated this problem under two basic stochastic influence cascade models: the *Independent Cascade* (`IC`) model [2] and the *Linear Threshold* (`LT`) model [7]. In the `IC` model, each edge has a propagation probability and influence is propagated by activated nodes independently activating their inactive neighbors based on the edge propagation probabilities. In the `LT` model, each edge has a weight, each node has a threshold chosen uniformly at random, and a node becomes activated, if the weighted sum of its activated neighbors exceeds its threshold. Kempe et al. [7] showed that the optimization problem of selecting the most influential nodes in a graph is NP hard for both models and also proposed a greedy algorithm that provides a good approximation ratio of 63% of the optimal solution. However, their greedy algorithm relies on the Monte Carlo simulation on influence cascade to estimate the influence spread, which makes the algorithm slow and not scalable.

A number of papers in recent years have tried to overcome the inefficiency of this greedy algorithm by improving the original algorithm [12,13] or proposing new algorithms [13-15]. Leskovec et al. [12] proposed the *cost-effective lazy forward* (`CELF`) scheme in selecting new seeds to reduce the number of influence spread evaluations, but it is still slow and not scalable to large graphs, as demonstrated in [15]. Kimura and Saito [14] proposed shortest-path-based heuristic algorithms to evaluate the influence spread. Chen et al. [13] proposed two faster greedy algorithms called *MixedGreedy* and *DegreeDiscount* for the `IC` model where the propagation probabilities on all edges are the same; MixedGreedy removes the edges that have no contribution for propagating influence, which can reduce the computation on the unnecessary edges; DegreeDiscount

assumes that the influence spread increases with node degree. Chen et al. [15] proposed the *Maximum Influence Arborescence* (MIA) heuristic based on local tree structures to reduce computation costs. Wang et al. [16] proposed a community-based greedy algorithm for identifying the most influential nodes. The main idea is to divide a social network into communities and estimate the influence spread in each community instead of the whole network topology.

As a variant of the conventional IM problem, Kim and Yoneki [8] introduced the problem called *Influential neighbor selection* (INS) to select the most influential neighbors of a node, rather than the most influential arbitrary nodes in a network. Kim [10] extended this epidemiological model by introducing several parameters (*user propagation weight*, *content interestingness*, and *decay factor*) to provide a more general and practical information diffusion model. However, they still used synthetic parameters that might significantly affect the information diffusion process. More recently, Kim et al. [10] found that the information diffusion speed of some schemes (e.g., propagation weight and random) in the previous study [17] was quite overestimated. In this paper, we extend their work by analyzing the effects of real propagation rates compared with the synthetic propagation rates.

Many studies noted that the levels of information-sharing activity varied greatly between users in social networks. Romero et al. [18] argued that a majority of Twitter users might be passive, not engaging in creating and sharing information. Cha et al. [9] found that users with many followers do not necessarily influence in terms of spawning retweets or mentions – the Spearman's rank correlation coefficient between the 'ranking by followers' and 'ranking by retweets' for all users was 0.549. Zhou et al. [19] showed that in Twitter, the content of a tweet might be an important factor in determining the 'retweet rate' – the mean retweet rate was 0.0136 but standard deviation was as high as 0.0501. Also, they observed that cascades tend to be wide and not too deep suggesting that the retweet rate may decay as the cascades spread away from the source – the mean of decay factors was about 0.2.

Influential neighbor selection problem

We begin with the definition of the *Independent Cascade* (IC) model [2] and then introduce the *Influential Neighbor Selection* (INS) problem, which will be used in the rest of the paper.

We model an *influence network* as a directed graph $G = (V, E)$ consisting of a set of nodes V and a set of ordered pairs of nodes E called the edge set, representing the communication channels between node pairs. A directed edge (u, v) from node u to node v of G is associated with a *propagation probability* $\lambda_{u,v}$, which is the probability that v is activated by u through the edge in the next time step if u is activated. Here, v is said to be a *neighbor* (or successor) of node u. For node $u \in V$, we use $N(u)$ to denote the set of u's neighbors. The *outdegree* of node u is denoted as $d(u) = |N(u)|$, which could be used simply in estimating the node u's influence on information propagation.

In the IC model [2], we assume that the time during which a network is observed is finite; without loss of generality, the time period is divided into fixed discrete steps $\{1, \ldots, t\}$. Let $S_i \subseteq V$ be the set of nodes that are activated at the time step i. We consider the dynamic process of information diffusion starting from the set of nodes $S_0 \subseteq V$ that are initially activated until the time step t as follows: At each time step i where $1 \leq i \leq t$,

every node $u \in S_{i-1}$ activates its inactivated neighbors $v \in V \setminus S_{i-1}$ with a propagation probability $\lambda_{u,v}$. The process ends after the time step t with S_t. A conventional *Influential Maximization* (IM) problem is to find a set S_0 consisting of k nodes to maximize $|S_t|$.

The *Influential Neighbor Selection* (INS) problem [8] is a variant of the IM problem: Given a spreader $s \in V$ and a budget constraint k, we aim to maximize the number of activated nodes in a network after the time step t by selecting s's $\min(k, d(s))$ neighbors only (rather than any subset of k nodes), as the set of nodes $S_0 \subseteq V$ that are initially activated. Compared to the conventional IM problem, the INS problem has three additional requirements: (1) each node only communicates with its immediate neighbors; (2) each node has no knowledge about the entire network topology except for its own connections; and (3) each message size is bounded to $O(\log |V|)$ bits (more intuitively, each message can only contain the node identity and some constant values of the node properties).

However, the initial INS problem in [8] – every edge has the same propagation probability – is too simple to correctly reflect the characteristics of the information diffusion process in real-world situations. Clearly, in the most popular online social network services such as Twitter or Facebook, each user has a different propagation rate for her neighbors on spreading information in a network according to the user's reputation or role such as opinion formers, leaders, or followers [9]. Figure 1 shows an example of the INS problem where each user has a different propagation rate. In this figure, a number in each node indicates the node's propagation rate. Here, for example, the node g's propagation rate is 0.4. Given this graph with the spreader node s, when $k = 1$, we might choose g as an initially activated node to maximize the number of activated nodes in the future; however, when every edge has the same propagation probability, the optimal choice might be either f rather than g.

We used the three important parameters (user propagation weight ω, content interestingness ϕ, and decay factor γ) to establish a more general and practical information propagation model. The details are as follows:

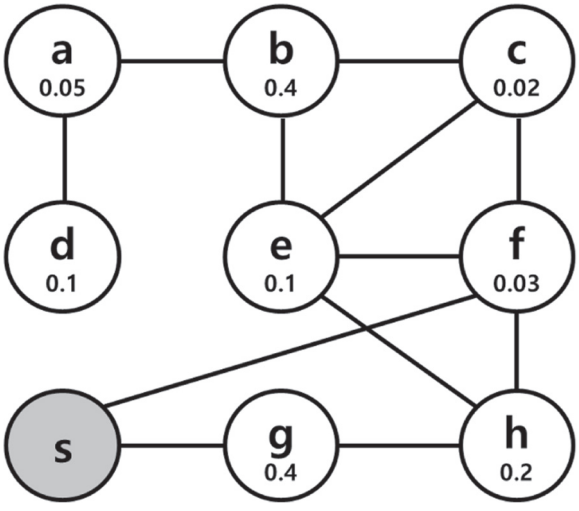

Figure 1 An example of the INS problem. Each node's propagation rate is provided as a number in the node. With the spreader node s, when $k = 1$, we should choose g as an initially activated node to maximize $|S_t|$; however, when every edge has the same propagation probability, the optimal choice might be either f rather than g.

The user propagation weight ω represents each user's average propagation rate to her neighbors. Given a user u, $\omega(u)$ is defined as $\tau(u)/(\rho(u)/d(u))$ where $\tau(u)$ and $\rho(u)$ are the number of u's posts shared by u's neighbors and the number of u's all posts, respectively. For example, if a user u with 1,000 neighbors wrote 10 posts and gets 100 shares, $\omega(u)$ is $100/(10 \cdot 1000) = 0.01$.

The content interestingness $\phi(r)$ of information r represents a measure to determine how much users want to share the information r with their neighbors. Naturally, higher content interestingness ϕ of a piece of information may facilitate higher propagation for the information through a network. Previous studies [19,20] showed that propagation probability λ can be greatly changed with the content of information (i.e., content interestingness ϕ).

The decay factor γ at hop N represents the ratio between the propagation probability at hop N and the propagation probability at hop $N-1$. In practice, the propagation probability might decay exponentially as the cascades spread away from the information source. Here, one possible explanation would be that the freshness of the information would drop as the time goes on.

With these parameters, given an edge $(u,v) \in E$, a spreader $s \in V$ and a piece of information r, $\lambda(u,v,s,r)$ is finally defined as follows [10]:

$$\lambda(u,v,s,r) = \min\left\{\omega(u) \cdot \phi(r) \cdot \gamma^{\delta(u,s,r)-1}, 1\right\} \tag{1}$$

where $\delta(u,s,r)$ is the number of times the information r is to be relayed from s to u.

For example, when $\phi(r) = 0.0136$, $\delta(u,s,r) = 3$, and $\gamma = 0.2$, a user u with $\omega(u) = 1$ would activate his (or her) neighbor v with the probability of about 0.0005 ($\approx 1 \cdot 0.0136 \cdot (0.2)^2$).

In this paper, we also use these parameters and the propagation probability equation. We particularly performed experiments with a real dataset (Twitter users and messages related to the 2010 UK election campaign) instead of using randomly generated synthetic parameters to provide more realistic simulation results than the previous study [10] which was not capable of considering the correlation between the number of neighbors and the propagation rate that might significantly affect the information diffusion process.

Neighbor selection schemes

For the `INS` problem described in Section 'Influential neighbor selection problem,' we basically use a greedy strategy to select the influential neighbors.

Assume that a spreader $s \in V$ wants to spread a piece of information r through the network $G = (V,E)$ by sharing r with its $\min(k, d(s))$ neighbors at the initial step. Node s first tries to assess the influence of information diffusion for each neighbor $v \in N(s)$, respectively, by collecting the information about v. We note that the neighbors' influence should be estimated based on s's local information only, rather than the whole network. Since online social networks, such as Facebook, typically provide APIs to obtain the neighborhood information about the user, s can automatically collect the information about her own neighbors. After estimating the neighbors' influences, s selects the top $\min(k, d(s))$ nodes with the highest influence values from $N(s)$; that is, for the `IC` model in Section 'Influential neighbor selection problem', these nodes are selected as the set of initially activated nodes $S_0 \subseteq V$.

For the purpose of influence estimation, we test the following four selection schemes based on the 'number of friends' and 'user propagation weight' each user has:

- `Random` selection: Pick $\min(k, d(s))$ nodes randomly from $N(s)$. This scheme is very simple and easy to implement – the spreader s does not need any knowledge of the network topology.
- `Degree` selection: Pick the $\min(k, d(s))$ highest degree nodes from $N(s)$. This scheme requires the degree knowledge of neighbors.
- `Propagation`-weight selection: Pick the $\min(k, d(s))$ highest user propagation weight nodes from $N(s)$. This scheme requires the user propagation weight knowledge of the nodes. To calculate $\omega(v)$ for s's neighbor $v \in N(s)$, the information about $\tau(v)$, $\rho(v)$ and $d(v)$ is required where $\tau(v)$ and $\rho(v)$ are the number of v's posts shared by v's neighbors and the number of v's all posts, respectively.
- `Hybrid` selection: Pick the $\min(k, d(s))$ nodes $v \in V$ with the highest *weighted* node degree $\omega d(v)$ which is defined as $\omega d(v) = \omega(v) \cdot d(v)$. At the first glance, this scheme requires the knowledge of both the degree and the user propagation weight of neighbors. In fact, however, this scheme can be simply implemented without the knowledge about node degree since $\omega(v) \cdot d(v)$ is calculated as $\tau(v)/\rho(v)$; $d(v)$ is automatically canceled in the calculation.

The algorithm of those schemes is commonly specified as follows:

```
1:  procedure SELECT-k-NEIGHBORS(G, s, k, SCHEME)
2:      S_0 ← ∅                                  ▷ initialize the set of initially activated nodes S_0
3:      N ← Find-Neighbor-Set(s)                 ▷ find the node s's neighbors
4:      m ← min(k, |N|)                          ▷ calculate min(k, d(s)) to determine the
                                                   number of selecting neighbors (i.e.,
                                                   |N| = d(s))
5:      switch SCHEME do
6:          case Random
7:              Q ← Random-Shuffle(N)            ▷ construct a queue Q with m elements
                                                   randomly drawn from N
8:          case Degree
9:              Q ← Construct-Max-Queue(N)       ▷ construct a max-priority queue Q where the
                                                   key is the node degree d(v) for v ∈ N
10:         case Propagation
11:             Q ← Construct-Max-Queue(N)       ▷ construct a max-priority queue Q where the
                                                   key is the node propagation weight ω(v) for
                                                   v ∈ N
12:         case Hybrid
13:             Q ← Construct-Max-Queue(N)       ▷ construct a max-priority queue Q where the
                                                   key is the weighted node degree ωd(v) for
                                                   v ∈ N
14:     for i ← 1, m do
15:         v ← Extract-Max(Q)                   ▷ get a node v with the maximum key (i.e., the
                                                   promising candidate in each greedy scheme
                                                   except for Random)
16:         Insert(S_0, v)                       ▷ insert the selected node v into the set of
                                                   initially activated nodes S_0
17:     end for
18: end procedure
```

The proposed algorithm runs in $O(|N| + m \log |N|)$ time – the maximum-priority queue Q can be constructed bottom-up in $O(|N|)$; the priority queue operations (extract-max) can be performed in $O(\log |N|)$ in each of m iterations. In practice, the term of '$m \log |N|$' can be simply ignored since m is less than or equal to a constant k. That is, these schemes can efficiently be performed.

Furthermore, we note that these schemes seem to be the most reasonable and promising for the `INS` problem since we cannot calculate network centrality metrics, such as closeness and betweenness [21], which require the knowledge of the entire network topology. Here, we do not consider the other metrics (e.g., [22]) to estimate node centrality based on localized information alone since the previous work [8] already showed that these metrics are ineffective for the `INS` problem compared with node degree.

The communication costs of all these schemes are $O(d(s))$ since the spreader s can obtain $d(v)$, $\omega(v)$, or $\omega d(v)$ through only direct communications with each neighbor $v \in N(s)$.

Experimental results

In this section, we analyze the performance of the selection schemes presented in Section 'Neighbor selection schemes.' Our goal was to find the best neighbor selection scheme to maximize information diffusion in Twitter through the experiments.

For experiments, we used the Twitter dataset [23] related to the 2010 UK general election between the 5th and 12th of May since this dataset reflects typical behavior of information diffusion in a political campaign.

To remove insignificant test cases, we filtered out users who did not either write any posts or had any followers and constructed the Twitter follower graph for those users. In this graph, each node represents a Twitter user and each edge represents a follow relation as a directed edge going from the followed user to the follower; when a user u follows another user v, we added an edge from v to u with a weight of $\omega(v)$ from the point of view of information flow. Unlike previous studies that used network topology alone [8] or synthetic datasets [10], we used real propagation rate $\omega(u)$ for each node u calculated from the Twitter messages. That is, from the collected tweets, we counted the number of tweets produced by u and the number of u's tweets shared (i.e., retweeted) by u's neighbors, respectively, to calculate the user u's propagation rate with those numbers.

The constructed graph consists of 45,179 nodes and 1,938,734 edges representing a sub-network of Twitter. This graph has the following properties: (1) its average degree is 42.91; (2) its number of strongly connected components is 4,559; and (3) the number of weakly connected components is 18 (i.e., this graph is divided into 18 disconnected components).

We used the `IC` model described in Section 'Influential neighbor selection problem' to evaluate the performance of the schemes presented in Section 'Neighbor selection schemes', with varying the number of initially activated neighbors k. The propagation probability $\lambda(u, v, s, r)$ on an edge $(u, v) \in E$ was defined with the spreader $s \in V$ and a piece of information r described in Section 'Influential neighbor selection problem'.

In each simulation run, we randomly picked a spreader with a piece of information r and then selected its k neighbors according to a selection criterion presented in Section 'Neighbor selection schemes', where the content interestingness $\phi(r)$ was randomly drawn from the normal distribution with the mean of 0.0136 and the standard

deviation of 0.0501, according to real data [19]. We also set the decay factor $\gamma = 0.2$ according to the mean of decay factors observed in the same dataset.

For evaluation, we observed the changes in the number of activated nodes during the 200th time steps. With a fixed k, we repeated this 500 times to minimize the bias of the test samples (randomly selected spreaders); we measured the ratio of the average number of activated nodes per test sample to the total number of nodes in the network. To establish a fair comparison, the parameter values were the same for all selection schemes in the ith run. Figure 2 shows how these values are changed over time t with $k = 1, 3, 5$, or 7 under the `IC` model.

From this figure, we can see that the `hybrid` selection scheme outperformed the other selection schemes: When $k = 1$, in the `hybrid` selection scheme, the ratio of the average number of activated nodes to the total number of nodes were over 0.002 (i.e., 2%) while the ratios were below 0.002 in `degree` and `propagation`-weight selection schemes. As k increased to 7, the gap between `hybrid` and other selection schemes was rather reduced but still seemed significant. This shows that we can effectively spread information using the `hybrid` scheme, even without consideration of the 'number of friends' information since the node degree is not needed to use the `hybrid` scheme. Interestingly, the `degree` selection was slightly better than the `propagation`-weight selection when $k = 3$, 5, or 7, while these schemes produced almost the same results when $k = 1$.

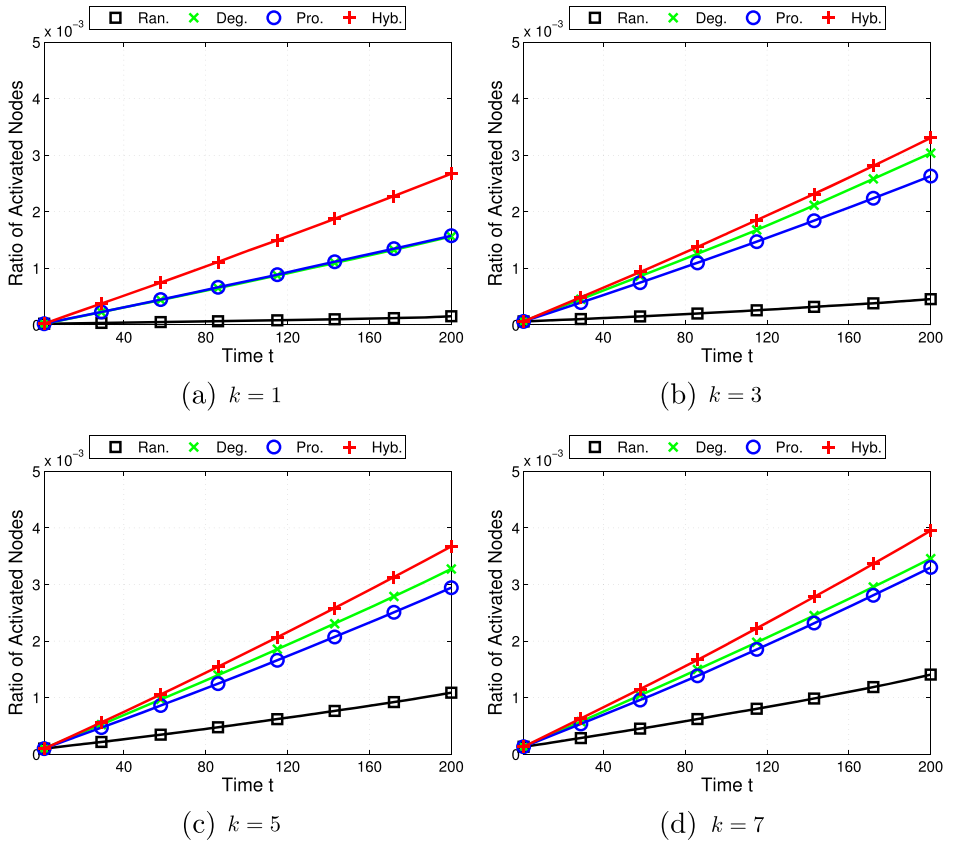

Figure 2 Changes in the ratio for real propagation rates. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network over time t for the *real* propagation rates.

Paired one-tailed t-tests with $\alpha = 0.01$ were used to compare the performance of the neighbor selection schemes in a statistically significant manner. We tested whether the distributions of the numbers of the activated nodes between schemes after the final time step (i.e., the 200th) were statistically different. The test results show that the performance of all the schemes appeared to be significantly different, except for the comparison of `propagation`-weight and `degree` when $k = 1$ ($p = 0.5202$).

To examine the influence of real propagation rate, we performed additional experiments using the above Twitter datasets, except the use of synthetic propagation rate parameters. As for synthetic propagation rates, the distribution of nodes' propagation rates were obtained by randomly shuffling the associations between nodes and propagation rates (i.e., we randomly permute user propagation rates and sequentially assign them to users) while keeping the network topology. Figure 3 shows how the numbers of activated nodes were changed over time t with $k = 1, 3, 5$, or 7 for the cases of using synthetic propagation rates. Since the resulting numbers with *synthetic* propagation rates are quite different from those with *real* propagation rates, we used a different y-axis scale for clarity.

Unlike the cases of using *real* propagation rates, we can see that the `propagation`-weight scheme is significantly better than the `degree` scheme. Moreover, the performance of the `propagation`-weight scheme is almost similar to that of the `hybrid` scheme – this trend appears to be totally different from the cases of using *real* propagation rates but similar to the results presented in Kim et al. [10] which also used

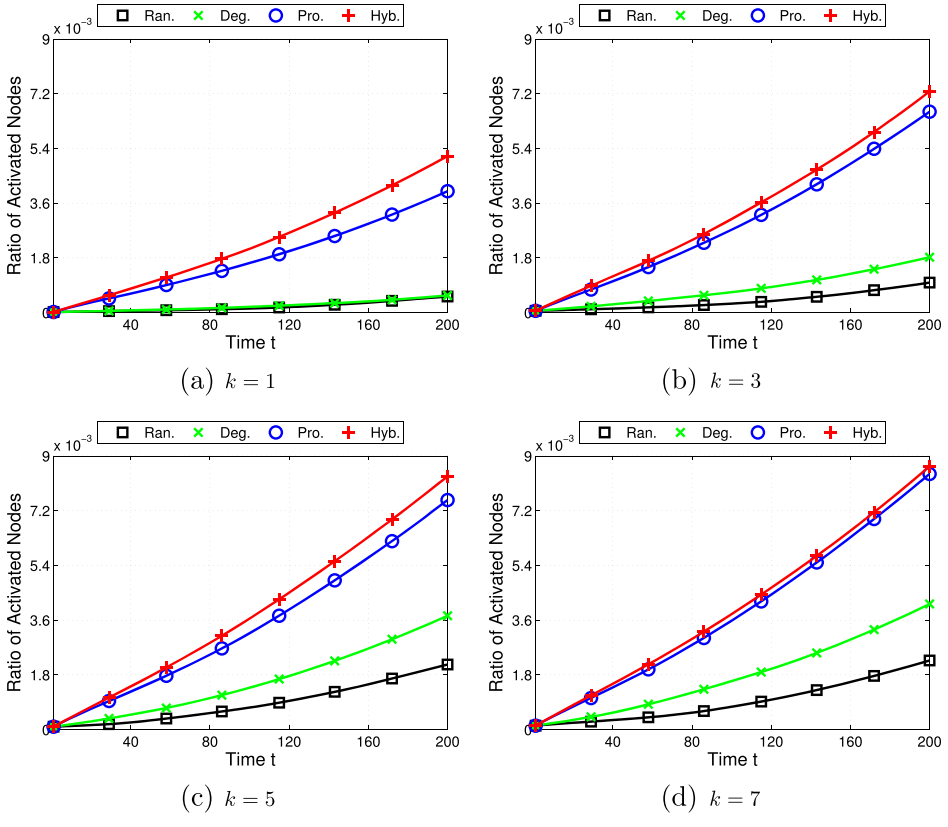

Figure 3 Changes in the ratio for synthetic propagation rates. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network over time t for the *synthetic* propagation rates.

synthetic parameters for experiments. This implies that there exists the reciprocal relationship between the node degree and propagation rate in a real-world social network. To show this, we calculate the Spearman correlation coefficient between the degree ranking of nodes and their propagation rate ranking. The computed rank correlation is 0.1936 ($P < 0.0001$), which indicates there exists a weak correlation between them.

We discuss how the performance of the proposed neighbor selection schemes may change with the number of initially activated nodes k. To accelerate the speed of information diffusion, a possible straightforward approach is to increase the number of initially activated neighbors k. Probably, we can imagine that even the naive `random` selection scheme can also be used to efficiently disseminate a piece of information if k increases sufficiently.

To demonstrate the effects of k, we analyzed the ratio of the average number of activated nodes after the 50th, 100th, 150th, and 200th time steps, respectively, with k ranging from 1 to 7. The experimental results are shown in Figure 4.

Unsurprisingly, the effects of k were rather limited for the process of early stage – in all selection strategies, the number of activated nodes were not greatly increased until the 50th step. After around that time, however, the performance of all selection schemes overall was improved as k increased. The ratios of activated nodes, except the `random`

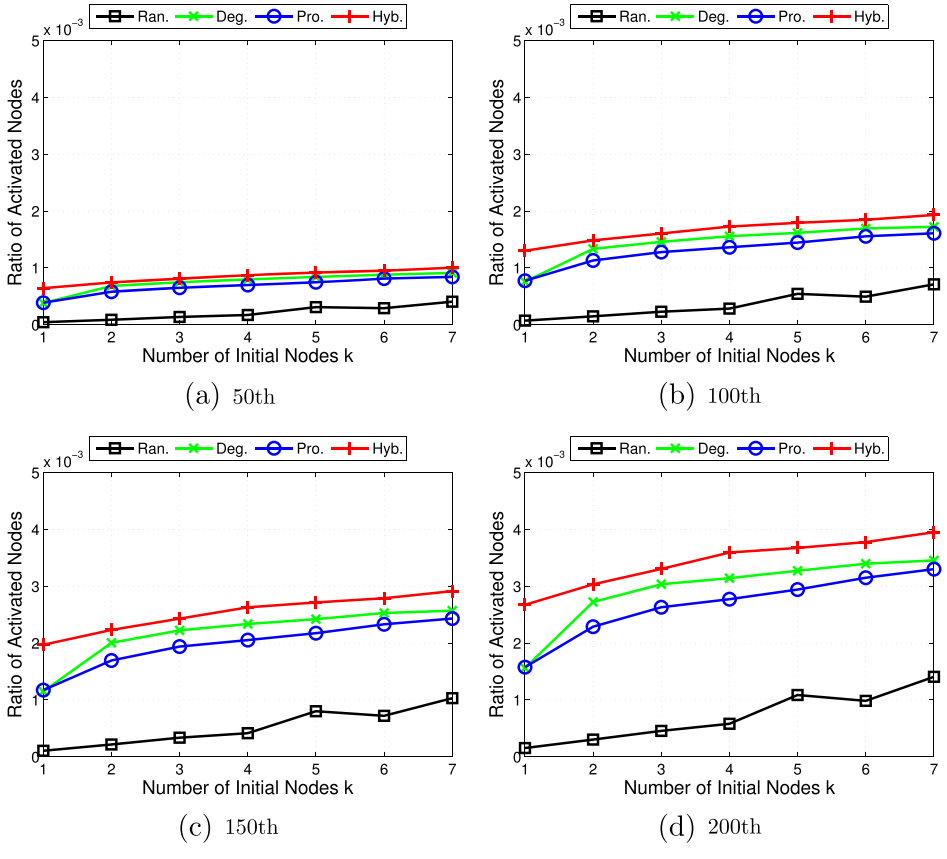

Figure 4 Changes in the ratio with the number of initially activated neighbors *k*. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network with the number of initially activated neighbors k.

selection scheme, show almost a similar pattern – the curves commonly had gentle slopes. Although the `random` selection scheme was relatively highly affected by k, the average number of activated nodes in the `random` selection scheme was still below 0.002 after the 200th time step even for $k = 7$. We can also see that the performance gaps between the schemes still existed with k. Even when the time step is 50th, the `hybrid`, `degree`, and `propagation`-weight selection schemes were significantly better than the `random` selection scheme. Moreover, after the 200th time steps, the performance gaps between selection schemes are clearly shown. We note that although the number of the initial activated nodes is really important, selection scheme is also important to accelerate the diffusion of information through the network. The performance of the `hybrid` selection scheme, when $k = 4$, was better than the other schemes even when $k = 7$.

We now move to the discussion on the performance of neighbor selection schemes when the content interestingness ϕ changes by fixing $k = 1$. We analyzed the ratio of the average number of activated nodes after the 50th, 100th, 150th, and 200th time steps, respectively, with mean of $\phi(r)$ ranging from 0.01 to 0.05 (and standard deviation 0.0501). The experimental results are shown in Figure 5.

Overall, the performance of all selection schemes except the `random` selection was improved and that of the `hybrid` selection scheme was particularly increased among those schemes with $\phi(r)$ compared with the other schemes. Therefore, the `hybrid`

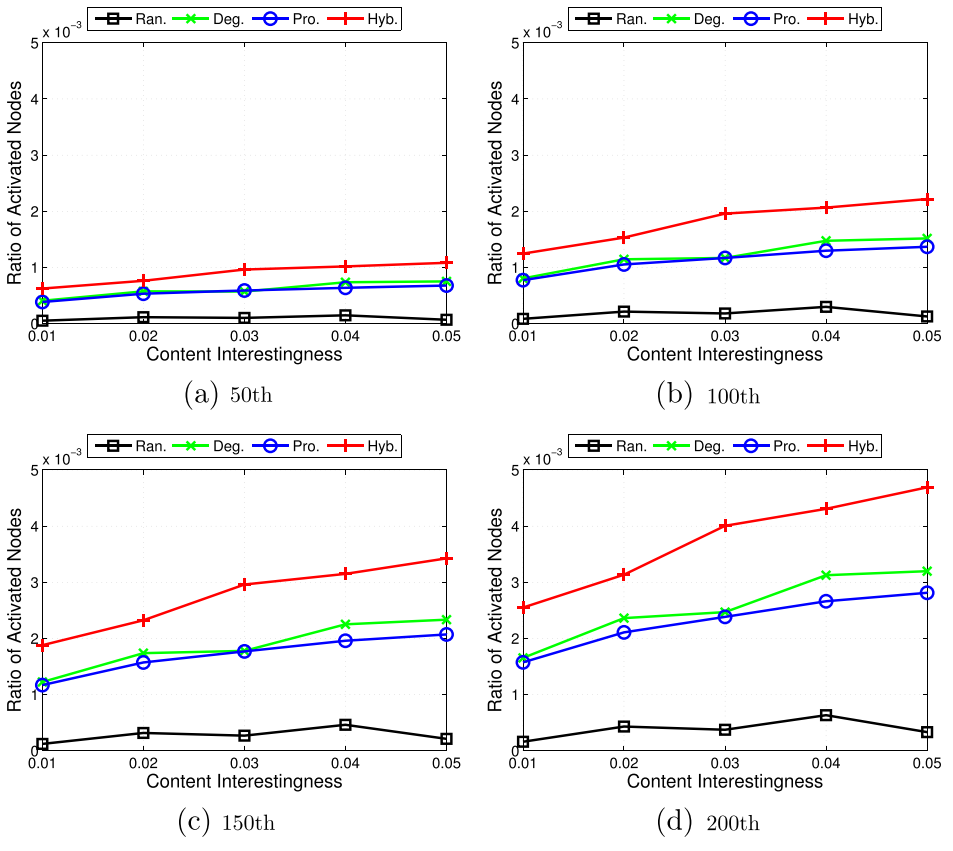

Figure 5 Changes in the ratio with the content interestingness $\phi(r)$. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network with the content interestingness $\phi(r)$.

selection scheme is still recommendable even for contents with a high $\phi(r)$. We note that the performance of the `random` selection was not significantly affected by $\phi(r)$.

Finally, we discuss the effects of the decay factor γ presented in Section 'Influential neighbor selection problem'. To demonstrate the effects of γ, we analyzed the ratio of the average number of activated nodes after the 50th, 100th, 150th, and 200th time steps, respectively, with $k = 1$ and γ ranging from 0.2 to 1. The experimental results are shown in Figure 6. We use a different y-axis scale on this figure since the numbers of activated nodes were greatly increased with γ.

Although the effects of γ were rather limited for the process of early stage within the 100th time step, after the 150th time step, the performance of all selection schemes except the `random` selection generally improved and the gaps between the schemes grew with γ.

Thus our suggestion is to use the `hybrid` selection scheme even with a large decay factor γ. Interestingly, we can also observe two different patterns in Figure 6: one is for the `hybrid` and `degree` selection schemes, which tends to increase quickly when $\gamma = 0.6$, and the other one is for the `propagation-weight` and `random` selection schemes, which tends to increase relatively slowly.

Conclusions

Given the increasing popularity of online social networks, it is of growing interest to investigate the characteristics of epidemic spreading, in order to accelerate or mitigate it. Kim and Yoneki [8] introduced the optimization problem to find *influential neighbors* for

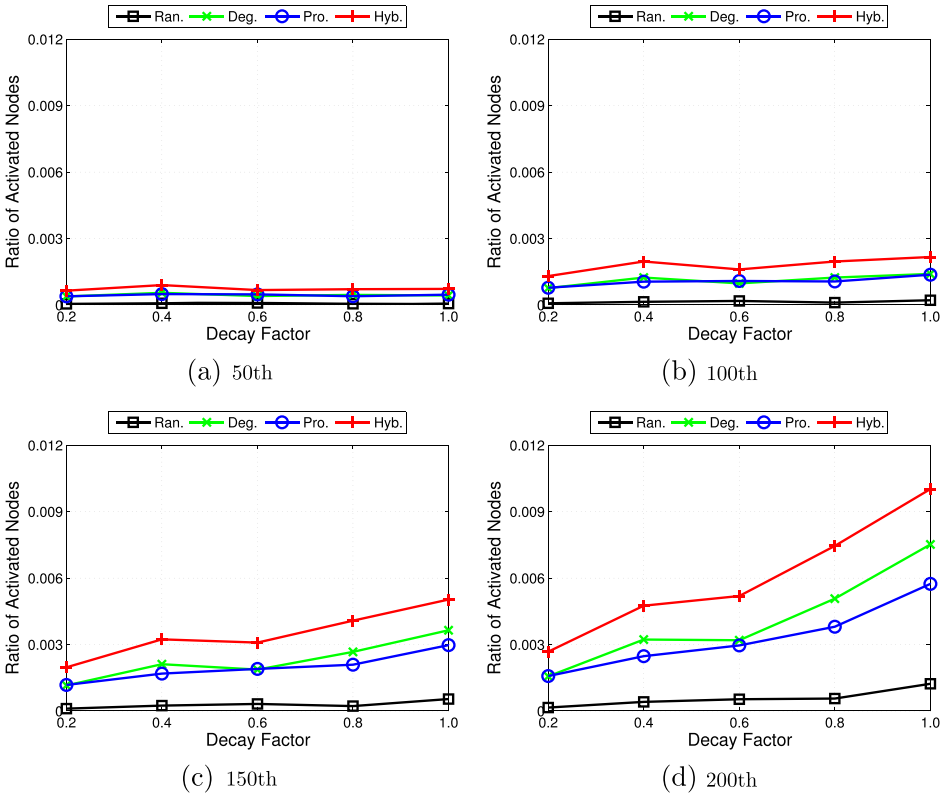

Figure 6 Changes in the ratio with the decay factor γ. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network with the decay factor γ.

maximizing information diffusion. We have extended their work by introducing several parameters (*user propagation weight*, *decay factor*, and *content interestingness*) to provide a more general and practical information diffusion model.

We presented four neighbor selection schemes (`random`, `degree`, `propagation-weight`, and `hybrid` selection) and explored their feasibility. We compared these selection schemes by computing the ratio of the average number of activated nodes to the total number of nodes in the network. We discussed which selection methods are generally recommended under which conditions.

Our experimental results showed that the `hybrid` selection scheme produced the best results of maximizing information diffusion through intensive simulation. Even with a small k, the `hybrid` selection scheme outperformed the other selection schemes with a relatively large k. Since the `hybrid` selection scheme can use the information about the users' posts alone, we can efficiently spread the information without the information about the 'number of friends' each user has. Unlike the results based on synthetic parameters, the `degree` scheme is significantly better than the `propagation-weight` scheme.

As an extension to this work, we are considering a theoretical study to formally generalize and verify our results in order to consider a wide range of application environments (e.g., each of which will have different levels of content interestingness). We will also develop a more extended framework for information diffusion. We may consider not only a spreader with the knowledge about the user's neighbors but also a spreader with a partial knowledge of the network topology (e.g., a subset of users or neighbors of neighbors). For example, we will extend the concept of the `INS` problem by expanding the set of the initially activated nodes with the distance from the information spreader.

Another interesting problem is to consider a new problem in the opposite direction to prevent (or reduce) the spread of information (e.g., rumor) by carefully monitoring the (important) users with a high 'user propagation weight' and/or 'number of friends'.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
HK leaded the research project about the influential neighbors selection problem, conducted the experiments and wrote the manuscript; KB discussed the validity of the proposed model and participated in writing the manuscript; EY provided the Twitter dataset used in our experiments and participated in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was partly supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1010) supervised by the NIPA (National IT Industry Promotion Agency) and was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2014R1A1A1003707).

Author details
[1]Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Korea. [2]Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada. [3]Computer Laboratory, University of Cambridge, Cambridge, UK.

Received: 19 November 2014 Accepted: 31 January 2015
Published online: 27 February 2015

References

1. Brown, JJ, Reingen, PH: Social ties and word-of-mouth referral behavior. J. Consum. Res. **14**(3), 350–362 (1987)
2. Goldenberg, J, Libai, B, Muller, E: Talk of the network: a complex systems look at the underlying process of word-of-mouth. Marketing Letters. **12**(3), 211–223 (2001). http://link.springer.com/article/10.1023%2FA%3A1011122126881

3. Metaxas, PT, Mustafaraj, E, Gayo-Avello, D: How (not) to predict elections. In: Proceedings of the 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT) and the 3rd International Conference on Social Computing (SocialCom), (2011)
4. Cao, Q, Sirivianos, M, Yang, X, Pregueiro, T: Aiding the detection of fake accounts in large scale social online services. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI), (2012)
5. Thomas, K, Grier, C, Paxson, V: Adapting social spam infrastructure for political censorship. In: Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats (LEET), (2012)
6. Domingos, P, Richardson, M: Mining the network value of customers. In: Proceedings of the 7th ACM Conference on Knowledge Discovery and Data Mining (KDD), (2001)
7. Kempe, D, Kleinberg, J, Tardos, E: Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM Conference on Knowledge Discovery and Data Mining (KDD), (2003)
8. Kim, H, Yoneki, E: Influential neighbours selection for information diffusion in online social networks. In: Proceedings of the 21th International Conference on Computer Communication Networks (ICCCN), (2012)
9. Cha, M, Haddadi, H, Benevenuto, F, Gummadi, KP: Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM), (2010)
10. Kim, H: Don't count the number of friends when you are spreading information in social networks. In: Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication (ICUIMC), (2014)
11. Gruhl, D, Guha, R, Liben-Nowell, D, Tomkins, A: Information diffusion through blogspace. In: Proceedings of the 13th International Conference on World Wide Web (WWW), (2004)
12. Leskovec, J, Krause, A, Guestrin, C, Faloutsos, C, VanBriesen, J, Glance, N: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM Conference on Knowledge Discovery and Data Mining (KDD), (2007)
13. Chen, W, Wang, Y, Yang, S: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM Conference on Knowledge Discovery and Data Mining (KDD), (2009)
14. Kimura, M, Saito, K: Tractable models for information diffusion in social networks. In: Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases (PKDD), (2006)
15. Chen, W, Wang, C, Wang, Y: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM Conference on Knowledge Discovery and Data Mining (KDD), (2010)
16. Wang, Y, Cong, G, Song, G, Xie, K: Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In: Proceedings of the 16th ACM Conference on Knowledge Discovery and Data Mining (KDD), (2010)
17. Kim, H, Beznosov, K, Yoneki, E: Finding influential neighbors to maximize information diffusion in twitter. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 701–706. WWW Companion '14, (2014)
18. Romero, DM, Galuba, W, Asur, S, Huberman, BA: Influence and passivity in social media. In: Proceedings of the 20th International Conference on World Wide Web (WWW), (2011)
19. Zhou, Z, Bandari, R, Kong, J, Qian, H, Roychowdhury, V: Information resonance on Twitter: watching Iran. In: Proceedings of the First ACM Workshop on Social Media Analytics (SOMA), (2010)
20. Hansen, LK, Arvidsson, A, Nielsen, FA, Colleoni, E, Etter, M: Good friends, bad news - affect and virality in Twitter. In: Future Information Technology. Communications in Computer and Information Science, pp. 34–43, (2011). http://link.springer.com/chapter/10.1007%2F978-3-642-22309-9_5
21. Kim, H, Tang, J, Anderson, R, Mascolo, C: Centrality prediction in dynamic human contact networks. Comput. Netw. **56**(3), 983–996 (2012)
22. Wehmuth, K, Ziviani, A: Daccer: Distributed assessment of the closeness centrality ranking in complex networks. Comput. Netw. **57**(13), 2536–2548 (2013)
23. Boutet, A, Kim, H, Yoneki, E: What's in Twitter, I know what parties are popular and who you are supporting now! Soc. Netw. Anal. Min. **3**(4), 1379–1391 (2013)