Tractable Models for Information Diffusion in Social Networks

Masahiro Kimura[1] and Kazumi Saito[2]

[1] Department of Electronics and Informatics, Ryukoku University
Otsu, Shiga 520-2194, Japan
[2] NTT Communication Science Laboratories, NTT Corporation
Seika-cho, Kyoto 619-0237, Japan

Abstract. When we consider the problem of finding influential nodes for information diffusion in a large-scale social network based on the *Independent Cascade Model (ICM)*, we need to compute the expected number of nodes influenced by a given set of nodes. However, a good estimate of this quantity needs a large amount of computation in the ICM. In this paper, we propose two natural special cases of the ICM such that a good estimate of this quantity can be efficiently computed. Using real large-scale social networks, we experimentally demonstrate that for extracting influential nodes, the proposed models can provide novel ranking methods that are different from the ICM, typical methods of social network analysis, and "PageRank" method. Moreover, we experimentally demonstrate that when the propagation probabilities through links are small, they can give good approximations to the ICM for finding sets of influential nodes.

1 Introduction

Recently, considerable attention has been devoted to investigating social networks [9,5,4,7,11,6], since the progress of the Internet, the World Wide Web, and blogs has enabled us to collect real large-scale social networks. Here, a social network is a network of relationships and interactions among social entities such as individuals, organizations and groups. Examples include email networks, hyperlink networks of web sites, trackback networks of blogs, and scientific collaboration networks. Since information, ideas, and influence can propagate through a social network in the form of "word-of-mouth" communications, it is an important research issue to find influential nodes for information diffusion in the underlying network in terms of sociology and marketing. Namely, it is significant to investigate the problem of finding nodes that generate a large spread of information. For example, Domingos and Richardson [2,12], and Kempe *et al.* [5] in particluar studied the *influence maximization problem*, that is, the problem of choosing a set of k nodes to target for initial activation such that it yields the largest expected spread of information, where k is a given integer.

In order to investigate these problems, we need a model of information diffusion in a social network. Although models for diffusion processes in a network

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 259–271, 2006.

have been studied in various fields including epidemiology, sociology, marketing and physics [5,4], one of the conceptually simplest models is the *Independent Cascade Model (ICM)* used by Goldenberg *et al.* [3], Kempe *et al.* [5], and Gruhl *et al.* [4]. The ICM is a stochastic process model in which information propagates from a node to its neighboring nodes at each time-step according to some probabilistic rule. Therefore, when we consider the problem of finding sets of influential nodes in a social network based on the ICM, we need to compute the expected number $\sigma(A)$ of nodes influenced by a given set A of nodes. It is an open question to compute $\sigma(A)$ exactly by an efficient method, and so good estimates were obtained by simulating the random process many times [5]. However, such computations become very heavy for a large-scale social network.

In this paper, as natural special cases of the ICM, we propose two novel information diffusion models such that a good estimate of $\sigma(A)$ can be efficiently computed. Using large real data from a blog network and a scientific collaboration network, we experimentally explore properties of the proposed models. First, we experimentally compare the proposed models with the ICM, typical methods of social network analysis [13], and "PageRank" method [1] in terms of ranking methods to extract influential nodes, and show that the proposed models provide novel scalable ranking methods that can in general extract nontrivial nodes as influential nodes. We also demonstrate that when the propagation probabilities through links are small, the proposed models can provide good approximations to the ICM for finding sets of influential nodes in a social network. On the other hand, if we consider the influence maximization problem in the ICM, a provable performance guarantee for a natural greedy alogrithm was obtained by Kempe *et al.* [5]. We extend this result to the proposed models.

2 Independent Cascade Model

Based on the work of Kempe *et al.* [5], we recall the definition of the ICM, and an approximation theory for the influence maximization problem in the ICM.

2.1 Definition

We consider the ICM for the spread of a certain information through a social network represented by a directed graph. First, we call nodes *active* if they have accepted the information. We assume that nodes can switch from being inactive to being active, but cannot switch from being active to being inactive. When node u first becomes active at step t, it is given a single chance to activate each currently inactive *child* v, and succeeds with probablity $p_{u,v}$. Here, $p_{u,v}$ is a constant that is independent of the history of the process, and node v is called a *child* of node u and node u is called a *parent* of node v if there is a directed link (u,v) from u to v. If u succeeds, then v will become active at step $t+1$. If multiple parents of v first become active at step t, then their activation attempts are sequenced in an arbitrary order, but performed at step t. Whether or not u succeeds, it cannot make any further attempts to activate v in subsequent rounds. The process terminates if no more activations are possible.

For an initial active set A, let $\sigma(A)$ denote the expected number of active nodes at the end of the process. We call $\sigma(A)$ the *influence* of target set A.

2.2 Approximation Theory

We consider the influence maximization problem in the ICM. Namely, for a given positive integer k, we consider finding a set A_k^* of k nodes to target for initial activation such that $\sigma(A_k^*) \geq \sigma(B)$ for any set B of k nodes based on the ICM. For this problem, we analyze the following natural greedy algorithm:

```
1. Start with $B = \emptyset$.
2. for $i = 1$ to $k$ do
3.     Choose a node $v_i$ maximizing $\sigma(B \cup \{v_i\}) - \sigma(B)$.
4.     Set $B \leftarrow B \cup \{v_i\}$.
5. end for
```

Let B_k denote the set of k nodes obtained by this algorithm. Then, Kempe *et al.* [5] proved that $\sigma(B_k) \geq (1 - 1/e)\ \sigma(A_k^*)$, that is, they presented an approximation guarantee for this algorithm. Their proof relies on the theory of *submodular functions* [8]. Here, for a function f that maps a subset of a finite ground set U to a nonnegative real number, f is called *submodular* if $f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$ for any $u \in U$ and any pair $\{S, T\}$ of subsets of U with $S \subset T$. They proved the result of the approximation guarantee by showing that the function σ is submodular for the ICM.

However, for a naive implementation of this greedy algorithm, we need to compute the influence $\sigma(A)$ for each target set A. Since it is not clear how to evaluate $\sigma(A)$ exactly by an efficient method, Kempe *et al.* [5] obtained a good estimate by simulating the random process $10,000$ times for each target set. They argued that the quality of the approximation after $10,000$ iterations is comparable to that after $300,000$ or more iterations.

3 Proposed Models

We propose two novel information diffusion models as natural special cases of the ICM, and describe an approximate computation of influence $\sigma(A)$ for them. Moreover, we extend the approximation theory for the influence maximization problem by Kempe *et al.* [5] to the proposed models.

3.1 Definitions

We define two natural special cases of the ICM. Let A be the initial active set in the network, that is, the set of nodes that first become active at step 0. For nodes u and v in the network, let $d(u, v)$ denote the graph distance from u to v, and let $d(A, v)$ denote the graph distance from A to v, that is, $d(A, v) = \min_{u \in A} d(u, v)$. When there is no path from u to v, we set $d(u, v) = \infty$. Note that the value of $d(A, v)$ can be efficiently computed by graph theory [9].

First, we define the *Shortest-Path Model (SPM)*. The SPM is a special case of the ICM such that each node v has the chance to become active only at step $t = d(A, v)$. In other words, each node is activated only through the shortest paths from the initial active set. Namely, the SPM is a special type of the ICM where only the most efficient information spread can occur.

Next, we slightly generalize the SPM within the ICM, and define the *SP1 Model (SP1M)*. In the SP1M, each node v has the chance to become active only at steps $t = d(A, v)$ and $t = d(A, v) + 1$. In other words, node v cannot be activated excluding the paths from A to v whose length are equal to $d(A, v)$ or $d(A, v) + 1$.

We define the influence $\sigma(A)$ of target set A for the SPM and SP1M in the same way as the ICM.

3.2 Approximate Computation of Influence

We consider computing efficiently an approximate value of $\sigma(A)$ for the SPM and SP1M. Let V be the set of all the nodes in the network, N the number of elements of V, and V_A the set of nodes v such that $d(A, v) < \infty$. For any $v \in V$, let $P_t(v; A)$ denote the probability that v first becomes active at step t, and let $PA(v)$ denote the set of all the parent nodes of v. Here, note that $P_t(v; A) = 0$ for any $t \geq 0$ if $v \notin V_A$. Note also that for each $v \in A$, $P_t(v; A) = 1$ if $t = 0$, and $P_t(v; A) = 0$ if $t > 0$.

We begin with the SPM. We consider calculating $\sigma(A)$ from a computation of $P_t(v; A)$ for any $t \geq 0$ and $v \in V$. Note first that for any $v \in V_A$, $P_t(v; A) = 0$ if $t \neq d(A, v)$. Thus, we focus on $t = d(A, v)$ for any $v \in V_A$. Then, it is easily shown that $P_t(v; A)$ is computed by

$$P_t(v; A) = \sum_{W \subset PA(v)} P_{t-1}(W | PA(v); A)\, P_t(W \to v), \tag{1}$$

where the summation is taken over all subsets of $PA(v)$, $P_{t-1}(W|PA(v); A)$ denotes the probability that subset W first becomes active at step $t - 1$ in $PA(v)$, and $P_t(W \to v)$ denotes the probability that v is activated from W at step t when W is infectious. Here, we put $PA(v) = \{u_1, \cdots, u_K\}$, and use the following one-to-one correspondence between a subset W of $PA(v)$ and a binary K-vector $\boldsymbol{h} = (h_1, \cdots, h_K)$; for each k, $h_k = 1$ if $u_k \in W$, and $h_k = 0$ if $u_k \notin W$. Following Domingos and Richardson [2], we approximate the joint probabilities $\{P_{t-1}(W|PA(v); A);\ W \subset PA(v)\}$ by their maximal entropy estimates given the mariginals $\{P_{t-1}(u_k; A);\ k = 1, \cdots, K\}$. This yields

$$P_{t-1}(W|PA(v); A) = \prod_{k=1}^{K} P_{t-1}(u_k; A)^{h_k} \left(1 - P_{t-1}(u_k; A)\right)^{1-h_k}_{.} \tag{2}$$

Note here that we can also obtain the same result by assuming that events $E_{1,t-1}, \cdots, E_{K,t-1}$ are independent, where each $E_{k,t-1}$ is the event that node u_k first becomes active at step $t - 1$. Thus, by (1) and (2), we have

$$P_t(v;A)$$
$$= \sum_{\boldsymbol{h}} \left[\left\{ \prod_{k=1}^{K} P_{t-1}(u_k;A)^{h_k} \left(1 - P_{t-1}(u_k;A)\right)^{1-h_k} \right\} \left\{ 1 - \prod_{k=1}^{K} (1 - p_{u_k,v})^{h_k} \right\} \right],$$

$$\begin{aligned} P_t(v;A) &= \sum_{\boldsymbol{h}} \prod_{k=1}^{K} P_{t-1}(u_k;A)^{h_k} (1 - P_{t-1}(u_k;A))^{1-h_k} \\ &\quad - \sum_{\boldsymbol{h}} \prod_{k=1}^{K} \left\{ P_{t-1}(u_k;A)\ (1 - p_{u_k,v}) \right\}^{h_k} \left(1 - P_{t-1}(u_k;A)\right)^{1-h_k} \\ &= \prod_{k=1}^{K} \left\{ P_{t-1}(u_k;A) + (1 - P_{t-1}(u_k;A)) \right\} \\ &\quad - \prod_{k=1}^{K} \left\{ P_{t-1}(u_k;A)\ (1 - p_{u_k,v}) \ + \ (1 - P_{t-1}(u_k;A)) \right\} \\ &= 1 - \prod_{k=1}^{K} \left(1 \ - \ p_{u_k,v}\, P_{t-1}(u_k;A)\right). \end{aligned}$$

Under this approximation, we estimate $\sigma(A)$ as $\sigma(A) = \sum_{v \in V_A} P_{d(A,v)}(v;A)$.

Next, we consider the SP1M. In this case, for any $v \in V_A$, $P_t(v;A) = 0$ if $t \neq d(A,v)$, $d(A,v) + 1$. Thus, we focus on $t = d(A,v)$ and $t = d(A,v) + 1$ for any $v \in V_A$. In the same way as the case of the SPM, we approximate $P_t(v;A)$ by

$$P_t(v;A) = (1 - P_{t-1}(v;A)) \left\{ 1 - \prod_{u \in PA(v)} \left(1 \ - \ p_{u,v}\, P_{t-1}(u;A)\right) \right\}.$$

Under this approximation, we estimate $\sigma(A)$ as $\sigma(A) = \sum_{v \in V_A} (P_{d(A,v)}(v;A) + P_{d(A,v)+1}(v;A))$.

As investigated by Leskovec *et al.* [6], it seems that large cascades of information diffusion happen rarely. We believe that this kind of real situations can be reasonably simulated by using SPM or SP1M with relatively small $p_{u,v}$. Using real social networks, we experimentally confirmed that the proposed estimation methods can be effective for the SPM and SP1M especially with relatively small $p_{u,v}$ (see Appendix). These results imply that for the SPM and SP1M, $\sigma(A)$ can be efficiently estimated in a reasonable situation.

3.3 Extension of Approximation Theory

For the SPM and SP1M, we consider the influence maximization problem, and investigate an approximation guarantee for the greedy algorithm defined in Sect. 2.2. We fix an integer k $(1 \leq k < N)$. Let A_k^* be a set that maximizes the value of σ over all k-element subsets of V, and let B_k be the k-element set obtained by the greedy algorithm. Then, we can obtain the same result as that for the ICM.

Theorem 1. *In the SPM and SP1M, we have the following approximation guarantee for the greedy algorithm:* $\sigma(B_k) \geq (1 - 1/e)\sigma(A_k^*)$.

Proof. We prove this inequality in the same way as [5]. By the theory of submodular functions (see Theoerm 2.1 in [5]), it is sufficient to prove that σ is submodular in the SPM and SP1M. According to the proof of Theorem 2.2 in [5], we view the ICM in terms of *live* and *blocked* links. We first consider the SPM. Let X denote one sample set of outcomes for all the coin flips on the directed links in the network. Let $P(X)$ denote the probability of sample X. For any $u \in V$, let $S(u)$ be the set of shortest paths from u to each node in V, and let $L(u; X)$ be the set of live link paths from u with respect to X. We define $R(u; X)$ to be the set of nodes that belong to the paths in $S(u) \cap L(u; X)$. For any $A \subset V$, we define $\sigma_X(A)$ to be the number of nodes in $\cup_{u \in A} R(u; X)$. Then, we have $\sigma(A) = \sum_X P(X)\sigma_X(A)$, where the summation is taken over all samples. We can easily prove that σ_X is submodular and a nonnegative linear combination of submodular functions is also submodular. Hence, σ is submodular in the SPM. Similarly, we can also prove that σ is submodular in the SP1M.

4 Experimental Evaluation

Using real large-scale social networks, we experimentally explore properties of the proposed models.

4.1 Data Sets

We used two different data sets of large-scale social networks. The details of these data sets are given below.

Blog Network Data. First, we used a tackback network of blogs as an example of a social network. By tracing ten steps ahead the trackbacks from the blog of the theme “JR Fukuchiyama Line Derailment Collision” in the site “Theme salon of blogs” (`http://blog.goo.ne.jp/usertheme/`), we collected a large connected trackback network in May, 2005. Here, the total numbers of blogs and trackbacks were $12,047$ and $39,960$, respectively. Since bloggers discuss various topics and establish mutual communications by putting trackbacks on each other’s blogs, we regarded a link created by a trackback as a bidirectional link for simplicity. We call this data set the BN data.

Collaboration Network Data. Next, we employ a collaboration network obtained from co-authorships of physics papers as an example of a social network, where each undirected link is regarded as a bidirectional link. We used the coauthorship network of the Los Alamos Condensed Matter e-print archives investigated by Palla *et al.* [11]. Here, the total numbers of nodes and undirected links were $30,561$ and $125,959$, respectively. The network consisted of 668 connected components, and the total number of nodes in the maximal connected component was $28,502$. We call this data set the CN data.

4.2 Experimental Settings and Fundamental Statistics

In our experiments, we assigned a uniform probability of p to each directd link in the network for the ICM, SPM, and SP1M, that is, $p_{u,v} = p$ for any directed link (u, v). As regards large and small propagation probabilities, we investigated $p = 10\%$ and $p = 1\%$, respectively.

According to the work of Kempe *et al.* [5], we estimated the influence $\sigma(A)$ of target set A in the ICM as follows: We started the process by initially activating A, and counted the number of active nodes at the end of the process. We then used the empirical mean obtained by simulating the stochastic process $10,000$ times as the estimate. However, these estimates needed very heavy computations. For example, for the CN data, the estimates of $\{\sigma(v);\, v \in V\}$ for the ICM with $p = 1\%$ needed about 3 hours, and those with $p = 10\%$ needed about 115 hours. Incidentally, in both cases of $p = 1\%$ and $p = 10\%$, it took about 5 and 20 minutes, respectively, to compute the estimates of $\{\sigma(v);\, v \in V\}$ for the SPM and SP1M based on the proposed estimation methods. Here, all our experimentation was undertaken on a single Dell PC with an Intel 3GHz Pentium D processor, with 2GB of memory. From these facts, we also confirm that as p increases, the processing time for estimating $\{\sigma(v);\, v \in V\}$ for the ICM much increases, while the processing times for the SPM and SP1M hardly change. Therefore, we can deduce that unlike the proposed models, the ICM needs a very large amount of computation for solving the influence maximization problem with $p = 10\%$ in a large-scale network based on the natural greedy algorithm. In particular, sophisticated techniques such as parallel computing must be needed to practically solve this problem for the ICM with $p = 10\%$ in our data sets.

When we estimated $\{\sigma(v);\, v \in V\}$ through $10,000$ simulations for the ICM, we also computed the standard deviation for each node. For example, for $p = 10\%$, the average standard deviations in the BN data and the CN data were 139.12 and $2,092.60$, respectively. Here, the average of $\{\sigma(v);\, v \in V\}$ in the BN data and that in the CN data were 87.80 and $1,586.59$, respectively. We see from these facts that the number of finally influenced nodes can greatly vary every simulation in the ICM.

From the above observations we deduce that a large amount of computation can be generally needed to obtain good estimates of $\{\sigma(v);\, v \in V\}$ for the ICM in a large-scale network, and so the ICM can be a computationally expensive model. Thus, for reference purposes, we also investigated the special case where the influence $\sigma(A)$ of target set A is estimated through 100 simulations in the ICM. We refer to this special model as the *ICM100*.

4.3 Ranking Problem

First, we consider extracting influential nodes from the network by ranking nodes based on influence measure. The ICM, ICM100, SPM, and SP1M can measure the influence of node v by $\sigma(v)$. On the other hand, "degree centrality", "closeness centrality", and "betweenness centrality" are commonly used as influence measure in sociology [13], where the degree of node v is defined as the number

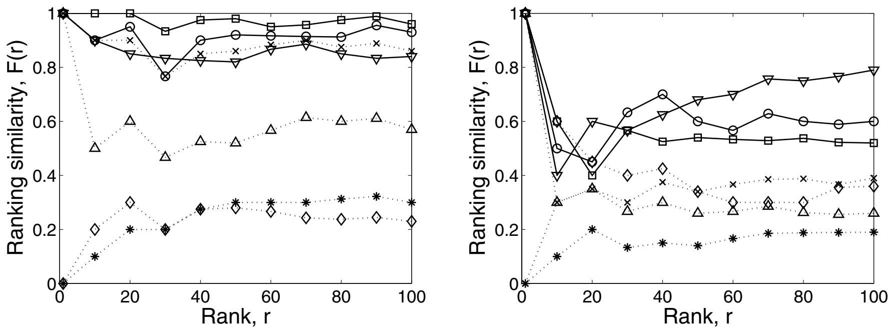

Fig. 1. Results for ranking similarities in the BN data. Left: $p = 1\%$. Right: $p = 10\%$. ("▽": ICM100. "○": SPM. "□": SP1M. "×": degree centrality. "◊": closeness centrality. "∗": betweenness centrality. "△": PageRank.)

of links attached to v, the closeness of node v is defined as the reciprocal of the average distance between v and other nodes in the network, and the betweenness of node v is defined as the total number of shortest paths between pairs of nodes that pass through v. We also consider measuring the influence of each node by its "authoritativeness" obtained by the "PageRank" method [1], since this is a well known method for identifying authoritative or influential pages in a hyperlink network of web pages. This method has a parameter ε; when we view it as a model of a random web surfer, ε corresponds to the probability with which a surfer jumps to a page picked uniformly at random [10]. In our experiments, we used a typical setting of $\varepsilon = 0.2$.

In terms of ranking methods for extracting influential nodes from the network, we compare the proposed models with the others for each value of p, so we introduce the *ranking similarity* $F(r)$ at rank r that quantifies the degree of similarity between two ranking methods. Based on F-measure, $F(r)$ is defined as follows: Let $L(r)$ and $L'(r)$ be the respective sets of top r nodes for the two ranking methods that we compare. Then, $F(r) = |L(r) \cap L'(r)| \ / \ r$, where $|S|$ indicates the number of elements in a set S. We focus on ranking similarities at high ranks since we are interested in extracting influential nodes.

Figs. 1 and 2 show the experimental results, where the ranking similarities $F(r)$ between the ICM method and the other methods are displayed at rank r $(1 \leq r \leq 100)$. Here, downward-pointing triangles, circles, squares, crosses, diamonds, asterisks, and upward-pointing triangles indicate the results for the ICM100, SPM, SP1M, degree centrality, closeness centrality, betweenness centrality, PageRank, respectively. We can observe that as ranking methods to extract influential nodes, the proposed models in general yield ranking results that are different from the ICM, typical methods of social network analysis, and PageRank method. This implies that the SPM and SP1M can provide novel ranking methods that in general extract nontrivial nodes as influential nodes. We can also observe that for $p = 1\%$, the ranking similarities of the proposed

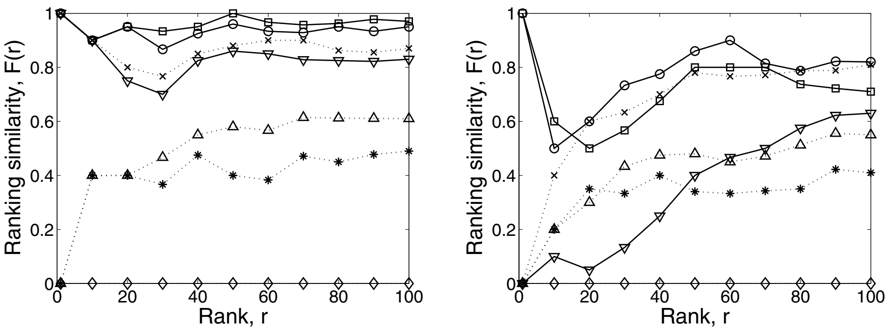

Fig. 2. Results for ranking similarities in the CN data. Left: $p = 1\%$. Right: $p = 10\%$. (“$\triangledown$”: ICM100. “$\circ$”: SPM. “$\Box$”: SP1M. “$\times$”: degree centrality. “$\Diamond$”: closeness centrality. “$*$”: betweenness centrality. “$\triangle$”: PageRank.)

models and ICM were very high, and higher than those of the ICM100 and ICM. These results imply the following remarkable result: When p is small, the SPM and SP1M can give good approximations to the ICM in terms of ranking methods for extracting influential nodes in a social network. However, we note that the SPM and SP1M do not necessarily provide good estimates of $\{\sigma(v);\ v \in V\}$ for the ICM even if p is small. On the other hand, we can see that the ranking similarities of the ICM100 and ICM were not high for $p = 10\%$ in particular. These results imply that good estimates of $\{\sigma(v);\ v \in V\}$ for the ICM cannot necessarily be obtained by using 100 simulations.

Moreover, we note that unlike the ICM, the proposed models can provide scalable ranking methods such as the typical methods of social network analysis. Namely, the ranking methods based on the proposed models can practically be applied even to a large-scale social network with $p = 10\%$. In fact, their computational complexities are almost comparable to those of the betweenness centrality and closeness centrality methods. We believe that this property is important for a practical ranking method based on information diffusion in a social network.

4.4 Influence Maximization Problem

We further investigate whether or not the proposed models can approximate the ICM for extracting sets of influential nodes in a social network, when propagation probabilities through links are small. For this purpose, we employ the task of approximately solving the influence maximization problem in the ICM with $p = 1\%$. To perform this task, we apply the ICM, ICM100, SPM, and SP1M with $p = 1\%$ in the following way: As an approximate solution for a target set size k, we use the optimal k-element set obtained by the natural greedy algorithm based on each model. Let B_k^0, B_k^1, B_k^2, and B_k^3 denote the optimal k-element sets based on the ICM, ICM100, SPM, and SP1M, respectively. To simplify our explanation, let $\sigma^0(A)$ denote the influence $\sigma(A)$ of targe set A for the ICM with $p = 1\%$. We evaluate the

Table 1. Performance of approximate solutions for the influence maximization problem in the ICM with $p = 1\%$ in the BN data

Target set size	ICM	ICM100	SPM	SP1M
$k = 1$	3.87	3.87	3.87	3.87
$k = 10$	30.06	27.67	30.07	30.06
$k = 20$	51.84	44.40	51.84	51.87
$k = 30$	71.83	57.79	71.96	72.01

Table 2. Performance of approximate solutions for the influence maximization problem in the ICM with $p = 1\%$ in the CN data

Target set size	ICM	ICM100	SPM	SP1M
$k = 1$	3.78	3.78	3.78	3.78
$k = 10$	33.35	30.61	33.44	33.42
$k = 20$	59.40	51.80	59.39	59.53
$k = 25$	71.59	59.76	71.33	71.69

performance of an approximate solution B_k^i by the value of $\sigma^0(B_k^i)$. Of course, we estimated $\sigma^0(B_k^i)$ through $10,000$ simulations.

In our experiments, we examined such approximate solutions from $k = 1$ to $k = 30$ in the BN data, and from $k = 1$ to $k = 25$ in the CN data. Tables 1 and 2 show the experimental results, where the value of $\sigma^0(B_k^i)$ $(i = 0, 1, 2, 3)$ for each target set size k is displayed. We observe that the evaluation values for the proposed models were almost the same as those for the ICM, and better than those for the ICM100. These results imply that when p is small, the proposed models can provide good approximations to the ICM for finding sets of influential nodes in a social network.

We also examined the processing times for computing the approximate solutions. Let $t_+(k)$ be the processing time for computing B_k^i given B_{k-1}^i. Fig. 3 shows the processing time $t_+(k)$ at target set size k for each model, where left-pointing triangles, downward-pointing triangles, circles, and squares indicate the ICM, ICM100, SPM, and SP1M, respectively. We can see that as k increases, $t_+(k)$ does not increase so much for the SPM and SP1M, but it substantially increases for the ICM and ICM100. This implies that the methods based on the proposed models can be practically performed even for a large target set size k. Namely, we can see that the proposed models are also scalable to solve the influence maximization problem based on the greedy algorithm. On the other hand, the total processing time for obtaining $\{B_k^i;\ k = 1, \cdots, 30\}$ in the BN data and that for obtaining $\{B_k^i;\ k = 1, \cdots, 25\}$ in the CN data were as follows: In the BN data, the total processing times for the ICM, ICM100, SPM, and SP1M were about 5 days, 1 hour, 19 minutes, and 1 hour,

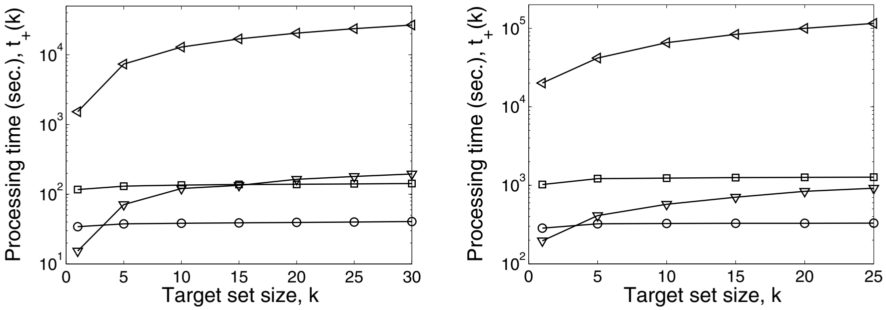

Fig. 3. Processing time $t_+(k)$ for target set size k. Left: BN data. Right: CN data. (“◁”: ICM. “▽”: ICM100. “◦”: SPM. “□”: SP1M.)

respectively. In the CN data, the total processing times for the ICM, ICM100, SPM, and SP1M were about 21 days, 4 hours, 2 hours, and 8 hours, respectively. Note here that the total processing times for the ICM were more than 100 times those for the ICM100 since $\sigma^0(B_k^0) > \sigma^0(B_k^1)$ for $k > 1$. These results show that a very large amount of computation is needed to solve the influence maximization problem for the ICM in a large-scale social network by using the greedy algorithm. Moreover, the following interesting observation is made: For the influence maximization problem in the ICM with small p, the methods based on the proposed models can be much faster than the ICM-based method, and can provide as good approximate solutions as the ICM-based method.

5 Conclusions

We have proposed two natural models for information diffusion in a social network, called the SPM and SP1M, such that the influence $\sigma(A)$ of a target set A can be efficiently estimated in a reasonable situation. For the influence maximization problems in the proposed models, we have provided a provable performance guarantee for the natural greedy algorithm. Using real large-scale social networks, we have experimentally explored properties of the SPM and SP1M. First, we have demonstrated that the proposed models can provide novel scalable ranking methods for extracting influential nodes in a social network. Next, we have demonstrated that when the propagation probabilities through links are small, they can give good approximations to the ICM for finding sets of influential nodes in a social network. Moreover, we have demonstrated that for solving the influence maximization problem based on the greedy algorithm, the proposed models can be scalable, and also be much faster than the ICM. Hence, we consider that the SPM and SP1M can be important models for social network analysis based on information diffusion.

References

1. Brin, S., and Page, L., The anatomy of a large-scale hypertextual Web search engine, In *Proc. WWW'98* (1998), 107–117.
2. Domingos, P., and Richardson, M., Mining the network value of customers. In *Proc. KDD'01* (2001), 57–66
3. Goldenberg, K. J., Libai, B., and Muller, E., Talk of the network: A complex systems look at the underlying process of word-of-mouth, *Marketing Letters* **12** (2001), 211–223.
4. Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A., Information diffusion through blogspace, In *Proc. WWW'04* (2004), 491–501.
5. Kempe, D., Kleinberg, J., and Tardos, E., Maximizing the spread of influence through a social network, In *Proc. KDD'03* (2003), 137–146.
6. Leskovec, J., Singh, A., and Kleinberg, J., Patterns of influence in a recommendation network, In *Proc. PAKDD'06* (2006), 380–389.
7. McCallum, A., Corrada-Emmanuel, A., and Wang, X., Topic and role discovery in social networks, In *Proc. IJCAI'05* (2005), 786–791.
8. Nemhauser, G. L., and Wolsey, L. A., *Integer and Combinatorial Optimization.* Wiley, New York, 1988.
9. Newman, M. E. J., Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, *Physical Review E* **64** (2001), 016132.
10. Ng, A. Y., Zheng, A. X., and Jordan, M. I., Link analysis, eigenvectors and stability, In *Proc. IJCAI'01* (2001), 903–901.
11. Palla, G., Derényi, I., Farkas I., and Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society, *Nature* **435** (2005), 814–818.
12. Richardson, M., and Domingos, P., Mining knowledge-sharing sites for viral marketing, In *Proc. KDD'02* (2002), 61–70.
13. Wasserman, S., and Faust, K., *Social Network Analysis.* Cambridge University Press, Cambridge, 1994.

Appendix

Performance Evaluation of Influence Estimation. In Sect. 3.2, we have proposed methods to estimate the influence $\sigma(A)$ of a target set A for the SPM and SP1M. Using several real social networks, we experimentally confirmed that the methods can be effective for the SPM and SP1M with relatively small propagation probabilities through links. Here, we describe the experimental results for the estimates of $\{\sigma(v);\ v \in V\}$ in the BN data.

In the experiments, we examined both cases of $p = 1\%$ and $p = 10\%$. First, we estimated the values of $\sigma(v)$ for the SPM and SP1M through simulating the stochastic processes $10,000$ times like the case of the ICM, and adopted them as the true values of $\sigma(v)$. Then, the average m and standard deviation s of $\{\sigma(v);\ v \in V\}$ were as follows:

SPM: ($p = 1\%$: $m = 1.081$, $s = 0.126$), ($p = 10\%$: $m = 4.212$, $s = 4.061$).
SP1M: ($p = 1\%$: $m = 1.085$, $s = 0.138$), ($p = 10\%$: $m = 8.322$, $s = 10.668$).

Let $\hat{\sigma}(v)$ denote the estimate of $\sigma(v)$ by the proposed methods for the SPM and SP1M. We measured the approximation performance by error $\mathcal{E} = \sum_{v \in V} |\sigma(v) - \hat{\sigma(v)}| \ / \ N$. The results were as follows:

SPM: ($p = 1\%$: $\mathcal{E} = 0.002$), ($p = 10\%$: $\mathcal{E} = 0.045$).
SP1M: ($p = 1\%$: $\mathcal{E} = 0.003$), ($p = 10\%$: $\mathcal{E} = 0.479$).

These results show that the proposed estimation methods can be effective in a reasonable situation.