Efficient Estimation of Cumulative Influence for Multiple Activation Information Diffusion Model with Continuous Time Delay

Kazumi Saito[1], Masahiro Kimura[2], Kouzou Ohara[3], and Hiroshi Motoda[4]

[1] School of Administration and Informatics, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
k-saito@u-shizuoka-ken.ac.jp

[2] Department of Electronics and Informatics, Ryukoku University
Otsu 520-2194, Japan
kimura@rins.ryukoku.ac.jp

[3] Department of Integrated Information Technology, Aoyama Gakuin University
Kanagawa 229-8558, Japan
ohara@it.aoyama.ac.jp

[4] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We show that the node cumulative influence for a particular class of information diffusion model in which a node can be activated multiple times, i.e. Susceptible/Infective/ Susceptible (SIS) Model, can be very efficiently estimated in case of independent cascade (IC) framework with asynchronous time delay. The method exploits the property of continuous time delay within a stochastic framework and analytically derives the iterative formula to estimate cumulative influence without relying on awfully lengthy simulations. We show that it can accurately estimate the cumulative influence with much less computation time (about 2 to 6 orders of magnitude less) than the naive simulation using three real world social networks and thus it can be used to rank influential nodes quite effectively. Further, we show that the SIS model with a discrete time step, i.e. fixed synchronous time delay, gives adequate results only for a small time span.

1 Introduction

The proliferation of emails, blogs and social networking services (SNS) in the World Wide Web has accelerated the creation of large social networks [1–5]. Social networks naturally mediate the spread of various information. Innovation, topics and even malicious rumors can propagate in the form of so-called “word-of-mouth” communications. Thus, it is now understood that social networks provide rich sources of information that is useful to help understand the dynamics of our society, e.g. who are the best group of people to spread the desired information, how people respond to other people’s opinion, what kind of topics propagate faster, how the public opinions are formed, how the way the information spreads differ from community to community, etc.

Several models have been proposed that simulate information diffusion through a network. The most widely-used model is the *independent cascade (IC)*. This is a fundamental probabilistic model of information diffusion [6, 7], which can be regarded as the so-called *susceptible/infective/recovered (SIR) model* for the spread of a disease [2]. This model has been used to solve such problems as the *influence maximization problem* which is to find a limited number of nodes that are influential for the spread of information [7, 8] and the *influence minimization problem* which is to suppress the spread of undesirable information by blocking a limited number of links [9]. Here, it is noted that the influence of a node is defined as the expected number of nodes that it can activate due to the stochastic nature of the information diffusion. The SIR model assumes that a node, once infected, never re-infected after it has been cured (recovered). Thus, the influence is normally defined as the expected number of recovered nodes at the end of the time span in consideration. The other class of model for the spread of a disease is the so-called *susceptible/infective/susceptible (SIS) model* [2], where a node, once infected, moves to a susceptible state and can be re-activated multiple times. A similar problem can be solved for this model, too [10, 11]. In these models, efficient methods of estimating the influence have been proposed based on bond percolation, strongly connected component decomposition, burnout and pruning [8, 11], but no analytical solutions have been found. Thus, efficiency remains that the computation time is 2 or 3 orders of magnitude faster than naive simulation.

The IC model above, whether it is used in SIR or SIS setting, cannot handle time-delays that are asynchronous and continuous for information propagation. Time step is incremented discretely and thus the node states are updated synchronously, which can be viewed that the time delay is fixed and synchronous. We call this “fixed time delay” for short. In reality, time flows continuously and thus information, too, propagates on this continuous time axis. For any node, information must be received at any time from any other nodes and must be allowed to propagate to yet other nodes at any other time with a possible delay, both in an asynchronous way. We call this “continuous time delay” for short. For example, the following scenario in case of SIS setting explains this need. Suppose a person A posted an article to a blog and a person B read it and responded a week later. Another person C posted an article on the same topic the next day A posted and B read it and responded the same day. B was activated twice, first by C and next by A although the time A was activated is earlier than C. Thus, for a realistic behavior analysis of information diffusion, we need to adopt a model that explicitly represents continuous asynchronous time delay. The continuous time delay SIR model was discussed in the machine learning problem setting in which the objective was to learn the parameters in the diffusion model from the observed time stamped node activation sequence data [3, 12]. In [12] it was shown that the parameters can be learned by maximizing the likelihood of the observed data being produced by the model. Note that there is no need to do simulation to obtain the influence degree in case of SIR setting because the final influence degree is equal to that of the model without time delay[1] since a node is not allowed to be re-activated multiple times.

In this paper, we address the problem of efficiently estimating the *cumulative influence* of a node in the network by adopting the information diffusion model that allows

[1] This is equivalent to fixed time delay in discrete time setting.

continuous time delay and multiple activation of the same node under the framework of independent cascade model, called CTSIS for short. Interestingly, although the model we considered in this paper is most complicated among the series of the models discussed above, it is possible to derive a formula analytically, under a simplified condition, that can iteratively estimate the *cumulative influence* of a node exploiting the property of continuous time delay within a stochastic framework. What makes the analysis easier is that in case of the continuous time there is only one single node that can be activated at a time, i,e, no multiple activations at different nodes at the same time, and no simultaneous activations of a node by its multiple active parents each of which has been activated at a different time in the past. Thus it does not make sense to define the node influence at a specific time and in light of SIS and continuous time delay we naturally define the influence to be an integral over a specified time span (*cumulative influence*), which is more meaningful in many practical settings.

We show that the proposed method (called *iterative method*) can accurately estimate the cumulative influence with much less computation time (about 2 to 6 orders of magnitude less) than empirical mean of the *naive simulation method* with a limited number of runs using three real world social networks with different sizes and connectivities. The method can be used to rank influential nodes quite effectively. We compare the proposed methods with two other methods, the SIS with fixed time delay and the one which is the extreme case of the propose method where the time span is set to be infinitely large (called *infinite iterative method*). We show that these are indeed less accurate and discuss under which conditions these work well, e.g. SIS with fixed time delay only works well for a small time span.

The paper is organized as follows. We revisit the information diffusion model, in particular SIS family, in section 2, and explain the proposed method of cumulative influence estimation in section 3. Then we report the experimental results in section 4, followed by discussion in section 5. We summarize our conclusion in section 6.

2 Information Diffusion Model

Let $G = (V, E)$ be a directed network, where V and E $(\subset V \times V)$ stand for the sets of all the nodes and (directed) links, respectively. For any $v \in V$, let $\Gamma(v; G)$ denote the set of the child nodes (directed neighbors) of v, that is,

$$\Gamma(v; G) = \{w \in V;\ (v, w) \in E\}.$$

We consider information diffusion models on G in the susceptible/infected/susceptible (SIS) framework. In this context, infected nodes mean that they have just adopted the information, and we call these infected nodes *active* nodes.

2.1 Basic SIS Model

We first define the basic SIS model for information diffusion on G. In the model, the diffusion process unfolds in discrete time-steps $t \geq 0$, and it is assumed that the state of a node is either active or inactive. For every link $(u, v) \in E$, we specify a real value

$\kappa_{u,v}$ with $0 < \kappa_{u,v} < 1$ in advance. Here, $\kappa_{u,v}$ is referred to as the *diffusion parameter* through link (u, v). Given an initial active node v_0 and a time span T, the diffusion process proceeds in the following way. Suppose that node u becomes active at time-step t ($< T$). Then, node u attempts to activate every $v \in \Gamma(u; G)$, and succeeds with probability $\kappa_{u,v}$. If node u succeeds, then node v will become active at time-step $t + 1$. Thus, as mentioned in 1, we can view this as synchronous fixed time delay[2]. If multiple active nodes attempt to activate node v in time-step t, then their activation attempts are sequenced in an arbitrary order. On the other hand, node u will become inactive at time-step $t + 1$ unless it is activated by an active node in time-step t. The process terminates if the current time-step reaches the final time T.

2.2 Continuous-time SIS model

Next, we extend the basic SIS model so as to allow continuous-time delays, and refer to the extended model as the *continuous-time SIS (CTSIS) model*[3]. This model can be interpreted as *susceptible/exposed/infective/susceptible (SEIS) model* in that a node does not become active (infected) instantly when activated, but wait for a while (exposed) before it gets activated (infected). Once it gets activated, it instantly turns into susceptible state. In terms of information diffusion of some topic in blog space, this activation corresponds to posting a blog article on the topic (instantaneous action).

In the CTSIS model on G, for each link $(u, v) \in E$, we specify real values $r_{u,v}$ and $\kappa_{u,v}$ with $r_{u,v} > 0$ and $0 < \kappa_{u,v} < 1$ in advance. We refer to $r_{u,v}$ and $\kappa_{u,v}$ as the *time-delay parameter* and the *diffusion parameter* through link (u, v), respectively.

Let T be the time span. The diffusion process unfolds in continuous-time t, and proceeds from a given initial active node v_0 in the following way. Suppose that a node u becomes active at time t ($< T$). Then a delay-time δ is chosen for u's every child node $v \in \Gamma(u; G)$ from the exponential distribution with parameter $r_{u,v}$. If $t + \delta \leq T$, v is activated by u with success probability $\kappa_{u,v}$ at $t + \delta \leq T$. Under the continuous time framework, there is no possibility that multiple parent nodes of v simultaneously activate v exactly at the same time $t + \delta$. The process terminates if the current time reaches the final time T.

2.3 Influence Function

Let T be the time span for the CTSIS model on G. We consider a time-interval $[T_0, T_1]$ with $0 \leq T_0 < T_1 \leq T$. For any node $v \in V$, let $S(v; T_0, T_1)$ denote the total number of nodes activated within time-interval $[T_0, T_1]$ for the probabilistic diffusion process from an initial active node v under the CTSIS model. Note that $S(v; T_0, T_1)$ is a random variable. Let $\sigma(v; T_0, T_1)$ denote the expected value of $S(v; T_0, T_1)$. We call $\sigma(v; T_0, T_1)$ the *cumulative influence degree* of node v within time-interval $[T_0, T_1]$. Note that σ is a function defined on V. We call the function $\sigma(\cdot; T_0, T_1) : V \rightarrow \mathbf{R}$ the *cumulative influence function* for the CTSIS model within time-interval $[T_0, T_1]$ on network G.

[2] This may well be called as “no time delay” because time delay is not explicitly represented in the formulation.

[3] Note that the information propagates at a certain time point, but its delay can be continuous.

It is important to estimate the cumulative influence function $\sigma(\cdot; T_0, T_1)$ efficiently. In theory we can simply estimate it by simulating the CTSIS model in the following way. First, a sufficiently large positive integer M is specified. For each $v \in V$, the diffusion process of the CTSIS model is simulated from initial active node v, and the total number of nodes activated within time-interval $[T_0, T_1]$, $S(v; T_0, T_1)$, is calculated. Then, $\sigma(v; T_0, T_1)$ is estimated as the empirical mean of $S(v; T_0, T_1)$ that are obtained from M such simulations. We refer to this estimation method as the *naive simulation method*. However, as shown in the experiments, this is extremely inefficient, and cannot be practical (out of question). In this paper, we deal with the case "$T_0 = 0, T_1 = T$" for simplicity, and we denote $\sigma(v; 0, T)$ by $\sigma(v; T)$.

3 Estimation Methods

For a given directed graph $G = (V, E)$, we identify each node with a unique integer from 1 to $|V|$. Then we can define the adjacency matrix $\boldsymbol{A} \in \{0, 1\}^{|V|\times|V|}$ by setting $a_{u,v} = 1$ if $(u, v) \in E$; otherwise $a_{u,v} = 0$. We also define the probability matrix $\boldsymbol{P} \in [0, 1)^{|V|\times|V|}$ by replacing each element $a_{u,v}$ to the corresponding diffusion probability $\kappa_{u,v}$ if $(u, v) \in E$. Let $\boldsymbol{f}_v \in \{0, 1\}^{|V|}$ be a vector whose v-th element is 1 and other elements are 0, and $\mathbf{1} \in \{1\}^{|V|}$ be a vector whose elements are all 1.

3.1 Infinite Iterative Method

We can calculate the number of nodes that are reachable with J-steps starting from a node v by $\boldsymbol{f}_v^T \boldsymbol{A}^J \mathbf{1}$. Thus, when considering the diffusion probabilities, we can calculate the vector of the expected number of reachable nodes starting from each node within J steps by $\boldsymbol{P}\mathbf{1} + \cdots + \boldsymbol{P}^J\mathbf{1}$. Therefor, in case that the time-interval is $[0, \infty]$, according to the definition of the CTSIS model, we obtain the cumulative influence degree $\boldsymbol{\sigma}_\infty$ as follows:

$$\boldsymbol{\sigma}_\infty = \sum_{J=1}^{\infty} \boldsymbol{P}^J \mathbf{1}, \tag{1}$$

Note that the vector $\boldsymbol{\sigma}_\infty$ consists of values of the cumulative influence functions, i.e., $\sigma(\cdot; \infty)$. We refer to this estimation method as the *infinite iterative method*.

However, there exist some intrinsic limitations to the simple iterative method, i.e., we cannot specify arbitrary time-interval $[T_0, T_1]$ and diffusion probabilities for this method. As for the diffusion probabilities, when the largest eigenvalue of the probability matrix $\boldsymbol{P}$ is less than 1, we can guarantee to obtain finite value of $\boldsymbol{\sigma}_\infty$. In a simple case that the diffusion parameters are uniform for any link, i.e., $\kappa_{u,v} = \kappa$ for any $(u, v) \in E$, since the probability matrix $\boldsymbol{P}$ is equivalent to $\kappa\boldsymbol{A}$, the diffusion parameter κ must be less than the reciprocal of the the largest eigenvalue of the adjacency matrix $\boldsymbol{A}$. Incidentally, the calculation formula for this simple case is quite similar to that of Bonacich's centrality [13] and identical to that of Katz's measure [14].

3.2 Proposed Method

We want to estimate the cumulative influence degree within time-interval $[T_0, T_1]$ for arbitrary diffusion probabilities. To this end, we introduce the probability $R(J; T_0, T_1)$ that diffusion takes J-steps within this time-interval according to the CTSIS model. Here, in order to simplify our derivation, we focus on the simplest case that the time-delay parameters are uniform for any link, i.e., $r_{u,v} = r$ for any $(u, v) \in E$, although our approach can be naturally extended to more complex settings. In a special case where $T_0 = 0$ and $T_1 = T$, we denote this probability by $R(J; T)$. Here we note that $R(J; T_0, T_1) = R(J; T_1) - R(J; T_0)$. Thus we focus on calculation of $R(J; T)$.

Let δ_j be a random variable of a time-delay for the j-th step $(1 \leq j \leq J)$. In order to meet the condition that the diffusion takes J-steps within time-interval $[0, T]$, the total sum of the time-delays must be less than T, i.e., $0 \leq \delta_1 + \cdots + \delta_J \leq T$. In case of $J = 1$, we can easily obtain the following formula.

$$R(1; T) = \int_0^T r \exp(-r\delta_1) d\delta_1 = 1 - \exp(-rT). \tag{2}$$

In case of $J \geq 2$, due to the independence of time-delay trials, we can calculate the probability $R(J; T)$ as follows:

$$R(J; T) = \int_0^T \int_0^{T-\delta_1} \cdots \int_0^{T-(\delta_1+\cdots+\delta_{J-1})} \prod_{j=1}^{J} r \exp(-r\delta_j) d\delta_1 \cdots d\delta_J \tag{3}$$

Here by noting the following two formulas,

$$\int_0^{T-(\delta_1+\cdots+\delta_{J-1})} r \exp(-r\delta_J) d\delta_J = 1 - \exp(-rT) \prod_{j=1}^{J-1} \exp(r\delta_j),$$

$$\int_0^T \cdots \int_0^{T-(\delta_1+\cdots+\delta_{J-2})} r^{J-1} \exp(-rT) d\delta_1 \cdots d\delta_{J-1} = \exp(-rT) \frac{(rT)^{J-1}}{(J-1)!},$$

we can calculate Eq. 3 as follows:

$$R(J; T) = R(J-1; T) - \exp(-rT) \frac{(rT)^{J-1}}{(J-1)!} \tag{4}$$

Therefore, from Eqs. 2 and 4, we can derive the following explicit formula:

$$R(J; T) = 1 - \exp(-rT) \sum_{j=1}^{J} \frac{(rT)^{j-1}}{(j-1)!}. \tag{5}$$

Here, we can easily see that $R(J; T)$ is a monotonic decreasing function approaching to zero as J increases.

Now, by combining Eqs. 1 and 5, we can derive a new method for estimating the cumulative influence degree within time-interval $[T_0, T_1]$ for arbitrary diffusion probabilities. We can formulate the key formula as follows:

$$\boldsymbol{\sigma}_{[T_0,T_1]} = \sum_{J=1}^{\infty} R(J; T_0, T_1) \boldsymbol{P}^J \mathbf{1}. \tag{6}$$

Below we can summarize the algorithm of the proposed method.

1. Set each element of $\boldsymbol{\sigma}_{[T_0,T_1]}$ to 0, and set $J \leftarrow 1$ and $\boldsymbol{x} \leftarrow \mathbf{1}$.
2. Calculate $\boldsymbol{x} \leftarrow \boldsymbol{P}\boldsymbol{x}$ and if $R(J; T_0, T_1)\|\boldsymbol{x}\| < \eta$, then output $\boldsymbol{\sigma}_{[T_0,T_1]}$ and terminate.
3. Set $\boldsymbol{\sigma}_{[T_0,T_1]} \leftarrow \boldsymbol{\sigma}_{[T_0,T_1]} + R(J; T_0, T_1)\boldsymbol{x}$ and $J \leftarrow J + 1$ and return to 2.

In this algorithm, $\boldsymbol{x} \in \mathbb{R}^{|V|}$ is a vector to calculate the expected number of the J-step reachable nodes, and η is a parameter for the termination condition. In our experiments, η is set to a sufficiently small number, i.e., 10^{-12}.

4 Experiments

We first evaluate the performance (accuracy) of the proposed method (*iterative method*) by comparing with the *naive simulation method* with different number of runs to estimate the empirical mean using three large real social networks. We then compare the *iterative method* with two other methods, the *infinite iterative method* and the *SIS with fixed time delay method* in terms of the estimated *cumulative influence degree* for the CTSIS model using the same networks. Finally we compare the efficiency (computation time) of the *iterative method* with the *naive simulation method*. In all the experiments, we consider the simplest case where the both diffusion and time-delay parameters of the CTSIS model are uniform for any link.

4.1 Datasets

We employed three datasets of large real networks. These are all bidirectionally connected networks. The first one is a network of people that was derived from the "list of people" within Japanese Wikipedia, also used in [15], and has $9,481$ nodes and $245,044$ directed links (the Wikipedia network). The second one is a network derived from the Enron Email Dataset [16] by extracting the senders and the recipients and linking those that had bidirectional communications, and has $4,254$ nodes and $44,314$ directed links (the Enron network). The third one is a Coauthorship network used in [17] and has $12,357$ nodes and $38,896$ directed links (the coauthorship network).

4.2 Accuracy Evaluation

We evaluated the accuracy of the proposed method by comparing it with the *naive simulation method* mentioned in section 2.3. We speculate that the *cumulative influence degree* estimated by taking the empirical mean of the results of the *naive simulation method* converges asymptotically to the true value as the number of simulations M increases. Thus, we first examined how the difference of the estimated cumulative influence degree between the *iterative method* and the *naive simulation method* changes as M changes for the three networks.

The difference was evaluated by

$$\epsilon_M = \sum_{v \in V} |\sigma(v; T) - s_M(v; T)|/|V|, \tag{7}$$

where $\sigma(v;T)$ and $s_M(v;T)$ are the *cumulative influence degree* of node v estimated by the *iterative method* and the *naive simulation method*, respectively. We used $T = 10^4$ and varied M from 100, 1,000, and 10,000.

In these experiments we determined the values for the diffusion and time-delay parameters as follows. As noted in 3.1, it is required that the diffusion parameter κ must be less than $eig(\boldsymbol{A})^{-1}$, the reciprocal of the largest eigenvalue of the adjacency matrix $\boldsymbol{A}$ of the network for the *infinite iterative method* to obtain a finite value of $\boldsymbol{\sigma}_\infty$. The values of $eig(\boldsymbol{A})^{-1}$ for the Wikipedia, Enron, and Coauthorship networks were 0.00674, 0.0205, and 0.105, respectively. Thus, we adopted 0.0067, 0.02, and 0.1 as the values of κ for these networks, respectively. These are the largest values that the *infinite iterative method* can take. We set $r = 1$ for the time-delay parameter. This is equivalent to setting the average time delay to be a unit time which is consistent to the discrete time step of the *SIS with fixed time delay* method.

Table 1 summarizes the results, from which we can see that the estimation difference decreases as M increases and it becomes reasonably small at $M = 10,000$ for all the three networks. We are able to verify our speculation and conclude that the proposed *iterative method* can indeed estimate the *cumulative influence* accurately.

Table 1: Estimation difference between the *iterative method* (proposed) and the *naive simulation method*

network	M		
	100	1,000	10,000
Wikipedia	0.196	0.062	0.020
Enron	0.552	0.190	0.062
Coauthorship	0.298	0.096	0.031

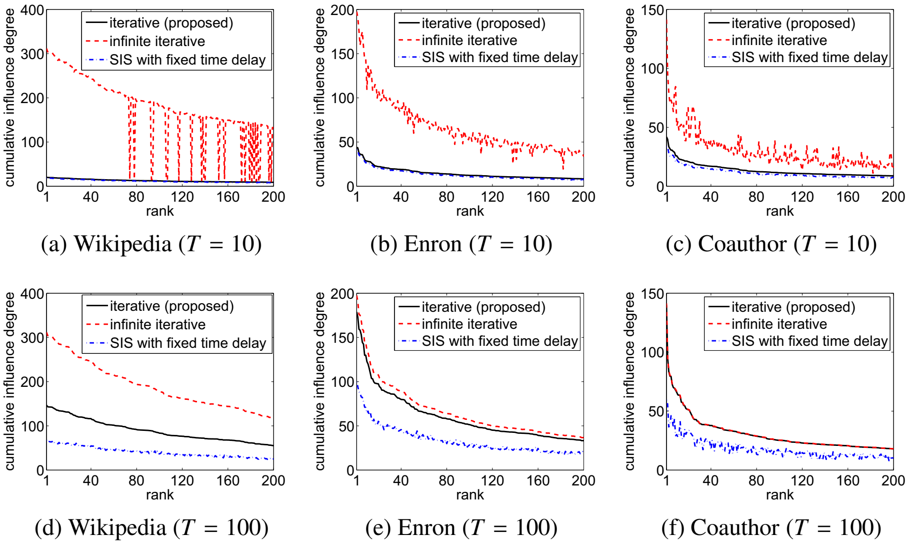
4.3 Cumulative Influence Degree Comparison

Next, we investigated how well the other approaches can approximate the *cumulative influence degree*. We compared two approaches. One is the *infinite iterative method* described in 3.1. The other is the *SIS with fixed time delay* method [11][5]. The *SIS with fixed time delay* method uses bond percolation on the layered graph which is constructed from the original social network with each layer added on top as the time proceeds[10] and much more efficiently estimates the *cumulative influence degree* than the *naive simulation method*. We used the same M (= 10,000) from the result in 4.2. For each network, we investigated two cases, one with a short time span $T = 10$ and the other with a long time span $T = 100$. Note that we set r=1 and thus, the average time delay $\overline{\delta} = 1$. We selected the top 200 most influential nodes that the *iterative method* identified and compared their *cumulative influence degree* with the values that the other two methods estimated for the same 200 nodes.

Figure 1 illustrates the results of comparison. We can see that the *infinite iterative method* estimate the *cumulative influence degree* fairly well for a long time span $T = 100$ except for the Wikipedia network, but it tends to overestimate it for a short time span $T = 10$. In contrast, the *SIS with fixed time delay method* tends to underestimate

[4] We had to set the value to be small so that the naive simulation returns the result within a day.

[5] Note that in [11] the influence degree was defined to be the expected number of active nodes at the end of observation time T, but here the algorithm in [11] is modified to calculated the *cumulative influence degree*.



(a) Wikipedia ($T = 10$) (b) Enron ($T = 10$) (c) Coauthor ($T = 10$)

(d) Wikipedia ($T = 100$) (e) Enron ($T = 100$) (f) Coauthor ($T = 100$)

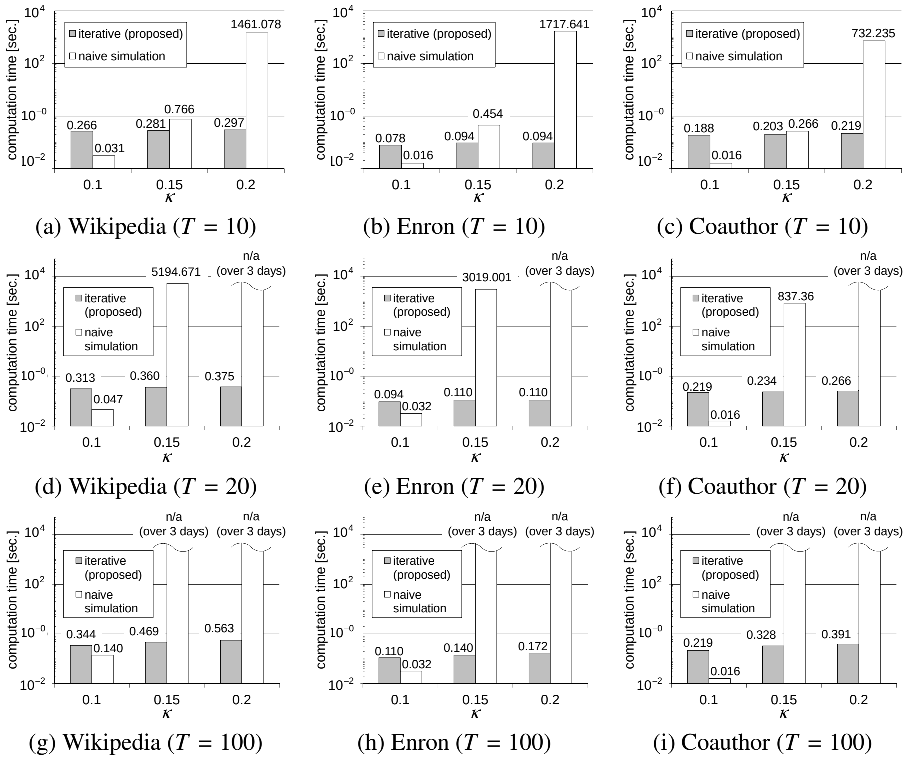
Fig. 1: Comparison in cumulative influence degrees of top 200 influential nodes

the *cumulative influence degree* for a large time span $T = 100$ but it does well for a short time span $T = 10$. These results show that these two methods cannot correctly estimate the *cumulative influence degree* for an arbitrary time span.

It is noted that there are many bumps in the graphs for the cases where the estimation of the other two methods is very poor, i.e. $T = 10$ for the *infinite iterative method* and $T = 100$ for the *SIS with fixed time delay method*. This implies that the ranking results by these methods are different from the true ranking by the *iterative method*. The curves becomes smoother when the estimation becomes better.

4.4 Efficiency Evaluation

We see in 4.3 that both *infinite iterative method* and *SIS with fixed time delay method* do not accurately estimate the *cumulative influence degree*, and we compare the computation time of the *iterative method* with the *naive simulation method* for $M = 1$. The results are shown in Fig. 2 for three values of the time span T= 10, 20, 100 and for each of the three networks. Three values are chosen for κ. The minimum values are the same as the ones used in 4.2 and 4.3, and the other values are obtained by multiplying 1.5 in sequence. The *iterative method* returns the values in less than 0.5 sec. for all cases and very insensitive to the parameter values. The *native simulation method* is only efficient when the κ is very small and requires exponentially increasing time as κ increase. In deed it did not return the values within 3 days in many cases. Considering that this is for a single simulation, use of the *naive simulation method* is not practical and out of question.



(a) Wikipedia ($T = 10$) (b) Enron ($T = 10$) (c) Coauthor ($T = 10$)

(d) Wikipedia ($T = 20$) (e) Enron ($T = 20$) (f) Coauthor ($T = 20$)

(g) Wikipedia ($T = 100$) (h) Enron ($T = 100$) (i) Coauthor ($T = 100$)

Fig. 2: Comparison in computation time

5 Discussions

We mentioned in 3.1 that the *cumulative influence degree* derived by the *infinite iterative method* is similar to the centrality proposed by Bonacich [13] and identical to the Katz' measure [14]. In [13] the standard centrality e_u of node u is defined by

$$\lambda e_u = \sum_{v \in V} a_{u,v} e_v, \tag{8}$$

where λ is a constant introduced to ensure a non-zero solution, and $\boldsymbol{A}$ is the adjacency matrix ($a_{u,v}$ is its element) as before. Bonacich generalized Eq. 8 by introducing the strength of relationship β, which is equivalent to κ in this paper, and derived the generalized centrality $c_u(\alpha, \beta)$ as

$$c_u(\alpha, \beta) = \sum_{v \in V} (\alpha + \beta c_v(\alpha, \beta)) a_{u,v}, \tag{9}$$

where α is a normalization constant. It is easily shown that $c_i(\alpha, \beta)$ is written in a matrix notation as

$$\boldsymbol{c}(\alpha, \beta) = \alpha \sum_{J=0}^{\infty} \beta^J \boldsymbol{A}^{J+1} \mathbf{1} = \alpha(\boldsymbol{A}\mathbf{1} + \beta \boldsymbol{A}^2 \mathbf{1} + \beta^2 \boldsymbol{A}^3 \mathbf{1} + \cdots). \tag{10}$$

Comparing Eq. 1 with Eq. 10, we note that they are the same except that the generalized centrality assumes that the strength of relationship with the directed connected nodes is 1. Further, we note that the following equality holds.

$$\boldsymbol{\sigma}_\infty = \frac{\beta}{\alpha}\boldsymbol{c}(\alpha,\beta), \tag{11}$$

which is exactly the same as Katz's measure. Thus, the *cumulative influence degree* $\boldsymbol{\sigma}_\infty$ defined by the *infinite iterative method* is interpreted as a centrality measure.

We showed in 4.3 that the *infinite iterative method* well approximates the *cumulative influence degree* when the time span is large. This is evident because the *infinite iterative method* assumes an infinite time span. In the extreme limit of $T = \infty$, the *iterative method* converges to the infinite iterative method. How large T should be in order for it to be large depends on the delay time parameter r. When r gets smaller, a smaller T can be called large, e.g. $T = 10$ is large when $r = 0.1$. Similar argument can be made for the *SIS with fixed time delay method*. The *SIS with fixed time delay method* advances the time in a discrete step. Thus, it happens that multiple parents attempt to activate the same node simultaneously at the same time. If this happens, the activation count is only incremented by one. When the time span T is small, the diffusion propagation does not go far and there is not much chance that this simultaneous activation happens. This is why the *SIS with fixed time delay method* gives good results for a small time span T. However, how good the *SIS with fixed time delay method* approximates the *cumulative influence degree* depends on how close the time step is to the average delay-time $\overline{\delta}$. It overestimates the true *cumulative influence degree* for $T = 10$ when $r = 0.1$ and underestimate it when $r = 10$. We confirmed this by additional experiments but due to the space limit we do not show the figures.

6 Conclusion

In this paper we addressed the problem of efficiently estimating the *cumulative influence degree* of a node in social networks when the information diffusion follows the Susceptible/Infective/Susceptible (SIS) model with asynchronous continuous time delay based on the independent cascade framework. It is possible to analytically derive a formula by which to iteratively calculate the *cumulative influence degree* to a desired accuracy. The simplified version which corresponds to assuming an infinitely large time span is closely related to the generalized centrality measure. We showed by applying the method to three large real world social networks that the method can accurately estimate the *cumulative influence degree* with 2 to 6 orders of magnitude less computation time than the *naive simulation method*. Thus, it can be used to rank the influential nodes very efficiently. We also compared the proposed *iterative method* to the *SIS with fixed time delay model* and the *infinite iterative method* and confirmed that they generally produce poor estimates and only give good results when a specific condition holds for each.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-08-4027, and JSPS Grant-in-Aid for Scientific Research (C) (No. 20500147).

References

1. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E **66** (2002) 035101
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
3. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations **6** (2004) 43–52
4. Domingos, P.: Mining social networks for viral marketing. IEEE Intelligent Systems **20** (2005) 80–82
5. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. In: Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06). (2006) 228–237
6. 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** (2001) 211–223
7. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003). (2003) 137–146
8. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-07). (2007) 1371–1376
9. Kimura, M., Saito, K., Motoda, H.: Blocking links to minimize contamination spread in a social network. ACM Transactions on Knowledge Discovery from Data **3** (2009) 9:1–9:23
10. Kimura, M., Saito, K., Motoda, H.: Efficient estimation of influence functions fot sis model on social networks. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09). (2009) 2046–2051
11. Saito, K., Kimura, M., Motoda, H.: Discovering influential nodes for SIS models in social networks. In: Proc. of the Twelfth International Conference of Discovery Science (DS2009), LNAI 5808. (2009) 302–316
12. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Proc. of the First Asian Conference on Machine Learning, LNAI 5828. (2009) 322–337
13. Bonacichi, P.: Power and centrality: A family of measures. American Journal of Sociology **92** (1987) 1170–1182
14. Katz, L.: A new status index derived from sociometric analysis. Sociometry **18** (1953) 39–43
15. Kimura, M., Saito, K., Motoda, H.: Minimizing the spread of contamination by blocking links in a network. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-08). (2008) 1175–1180
16. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Proceedings of the 2004 European Conference on Machine Learning (ECML'04). (2004) 217–226
17. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435** (2005) 814–818