Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis

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 Univesity
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 address the problem of estimating the parameters for a continuous time delay independent cascade (CTIC) model, a more realistic model for information diffusion in complex social network, from the observed information diffusion data. For this purpose we formulate the rigorous likelihood to obtain the observed data and propose an iterative method to obtain the parameters (time-delay and diffusion) by maximizing this likelihood. We apply this method first to the problem of ranking influential nodes using the network structure taken from two real world web datasets and show that the proposed method can predict the high ranked influential nodes much more accurately than the well studied conventional four heuristic methods, and second to the problem of evaluating how different topics propagate in different ways using a real world blog data and show that there are indeed differences in the propagation speed among different topics.

1 Introduction

The rise of the Internet and the World Wide Web accelerates the creation of various large-scale social networks, and considerable attention has been brought to social networks as an important medium for the spread of information [1–5]. Innovation, topics and even malicious rumors can propagate through social networks in the form of so-called “word-of-mouth” communications. This forms a virtual society forming various kinds of communities. Just like a real world society, some community grows rapidly and some other shrinks. Likewise, some information propagates quickly and some other only slowly. Good things remain and bad things diminish as if there is a natural selection. The social network offers a nice platform to study a mechanism of society dynamics and behavior of humans, each as a member of the society. In this paper, we address the problem of how information diffuses through the social network, in particular how different topics propagate differently by inducing a diffusion model that can handle continuous time delay.

There are several models that simulate information diffusion through a network. A widely-used model is the *independent cascade (IC)*, a fundamental probabilistic model of information diffusion [6, 7], which can be regarded as the so-called *susceptible/infected/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]. The IC model requires the parameters that represent diffusion probabilities through links to be specified in advance. Since the true values of the parameters are not available in practice, this poses yet another problem of estimating them from the observed data [10].

One of the drawbacks of the IC model is that it cannot handle time-delays for information propagation, and we need a model to explicitly represent time delay. Gruhl et al. is the first to extend the IC model to include the time-delay [3]. Their model now has the parameters that represent time-delays through links as well as the parameters that represent diffusion probabilities through links. They presented a method for estimating the parameter values from the observed data using an EM-like algorithm, and experimentally showed its effectiveness using sparse Erdös-Renyi networks. However, it is not clear what they are optimizing in deriving the update formulas of the parameter values. Further, they treated the time as a discrete variable, which means that it is assumed that information propagate in a synchronized way in a sense that each node can be activated only at a specific time. 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 other nodes and must be allowed to propagate to yet other nodes at any other time, both in an asynchronous way. Thus, for a realistic behavior analyses of information diffusion, we need to adopt a model that explicitly represents continuous time delay.

In this paper, we deal with an information diffusion model that incorporates continuous time delay based on the IC model (referred to as CTIC model), and propose a novel method for estimating the values of the parameters in the model from a set of information diffusion results that are observed as time-sequences of infected (active) nodes. What makes this problem difficult is that incorporating time-delay makes the time-sequence observation data structural. There is no way of knowing from the data which node activated which other node that comes later in the sequence. We introduce an objective function that rigorously represents the likelihood of obtaining such observed data sequences under the CTIC model on a given network, and derive an iterative algorithm by which the objective function is maximized. First we test the convergence performance of the proposed method by applying it to the problem of ranking influential nodes using the network structure taken from two real world web datasets and show that the parameters converge to the correct values by the iterative procedure and can predict the high ranked influential nodes much more accurately than the well studied conventional four heuristic methods. Second we apply the method to the problem of behavioral analysis of topic propagation, i.e., evaluating how different topics propagate in different ways, using a real world blog data and show that there are indeed differences in the propagation speed among different topics.

2 Information Diffusion Model and Learning Problem

We first define the IC model according to [7], and then introduce the continuous-time IC model. After that, we formulate our learning problem.

We mathematically model the spread of information through a directed network $G = (V, E)$ without self-links, where V and E $(\subset V \times V)$ stands for the sets of all the nodes and links, respectively. We call nodes *active* if they have been influenced with the information. In the model, it is assumed that nodes can switch their states only from inactive to active, but not from active to inactive. Given an initial set S of active nodes, we assume that the nodes in S have first become active at an initial time, and all the other nodes are inactive at the time.

In this paper, node u is called a *child node* of node v if $(v, u) \in E$, and node u is called a *parent node* of node v if $(u, v) \in E$. For each node $v \in V$, let $F(v)$ and $B(v)$ denote the set of child nodes of v and the set of parent nodes of v, respectively,

$$F(v) = \{w \in V;\ (v, w) \in E\}, \quad B(v) = \{u \in V;\ (u, v) \in E\}.$$

2.1 Independent Cascade Model

Let us describe the definition of the IC model. In this model, for each link (u, v), we specify a real value $\lambda_{u,v}$ with $0 < \lambda_{u,v} < 1$ in advance. Here $\lambda_{u,v}$ is referred to as the *diffusion probability* through link (u, v).

The diffusion process unfolds in discrete time-steps $t \geq 0$, and proceeds from a given initial active set S in the following way. When a node u becomes active at time-step t, it is given a single chance to activate each currently inactive child node v, and succeeds with probability $\lambda_{u,v}$. If u succeeds, then v will become active at time-step $t+1$. If multiple parent nodes of v become active at time-step t, then their activation attempts are sequenced in an arbitrary order, but all performed at time-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.

2.2 Continuous-Time Independent Cascade Model

Next, we extend the IC model so as to allow continuous-time delays, and refer to the extended model as the *continuous-time independent cascade (CTIC) model*.

In the CTIC model, 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.

The diffusion process unfolds in continuous-time t, and proceeds from a given initial active set S in the following way. Suppose that a node u becomes active at time t. Then, node u is given a single chance to activate each currently inactive child node v. We choose a delay-time δ from the exponential distribution with parameter $r_{u,v}$. If node v is not active before time $t + \delta$, then node u attempts to activate node v, and succeeds with probability $\kappa_{u,v}$. If u succeeds, then v will become active at time $t + \delta$. Under the continuous time framework, it is unlikely that multiple parent nodes of v attempt to activate v for the activation at time $t + \delta$. But if they do, their activation attempts are sequenced in an arbitrary order. 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 S, let $\varphi(S)$ denote the number of active nodes at the end of the random process for the CTIC model. Note that $\varphi(S)$ is a random variable. Let $\sigma(S)$ denote the expected value of $\varphi(S)$. We call $\sigma(S)$ the *influence degree* of S for the CTIC model.

2.3 Learning problem

For the CTIC model on network G, we define the time-delay parameter vector $\boldsymbol{r}$ and the diffusion parameter vector $\boldsymbol{\kappa}$ by

$$\boldsymbol{r} = (r_{u,v})_{(u,v)\in E}, \quad \boldsymbol{\kappa} = (\kappa_{u,v})_{(u,v)\in E}.$$

In practice, the true values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ are not available. Thus, we must estimate them from past information diffusion histories observed as sets of active nodes.

We consider an observed data set of M independent information diffusion results,

$$\mathcal{D}_M = \{D_m;\ m = 1, \cdots, M\}.$$

Here, each D_m is a time-sequence of active nodes in the mth information diffusion result,

$$D_m = \langle D_m(t);\ t \in \mathcal{T}_m \rangle, \qquad \mathcal{T}_m = \langle t_m, \cdots, T_m \rangle,$$

where $D_m(t)$ is the set of all the nodes that have first become active at time t, and $\mathcal{T}_m$ is the observation-time list; t_m is the observed initial time and T_m is the observed final time. We assume that for any active node v in the mth information diffusion result, there exits some $t \in \mathcal{T}_m$ such that $v \in D_m(t)$. Let $t_{m,v}$ denote the time at which node v becomes active in the mth information diffusion result, i.e., $v \in D_m(t_{m,v})$. For any $t \in \mathcal{T}_m$, we set

$$C_m(t) = \bigcup_{\tau \in \mathcal{T}_m \cap \{s;\, s < t\}} D_m(\tau)$$

Note that $C_m(t)$ is the set of active nodes before time t in the mth information diffusion result. We also interpret D_m as referring to the set of all the active nodes in the mth information diffusion result for convenience sake. In this paper, we consider the problem of estimating the values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ from $\mathcal{D}_M$.

3 Proposed Method

We explain how we estimate the values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ from $\mathcal{D}_M$. Here, we limit ourselves to outline the derivations of the proposed method due to the lack of space. We also briefly mention how we do behavioral analysis with the method.

3.1 Likelihood function

For the learning problem described above, we strictly derive the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ to use as our objective function.

First, we consider any node $v \in D_m$ with $t_{m,v} > 0$ for the mth information diffusion result. Let $\mathcal{A}_{m,u,v}$ denote the probability density that a node $u \in B(v) \cap C_m(t_{m,v})$ activates the node v at time $t_{m,v}$, that is,

$$\mathcal{A}_{m,u,v} = \kappa_{u,v} r_{u,v} \exp(-r_{u,v}(t_{m,v} - t_{m,u})). \tag{1}$$

Let $\mathcal{B}_{m,u,v}$ denote the probability that the node v is not activated from a node $u \in B(v) \cap C_m(t_{m,v})$ within the time-period $[t_{m,u}, t_{m,v}]$, that is,

$$\begin{aligned}\mathcal{B}_{m,u,v} &= 1 - \kappa_{u,v} \int_{t_{m,u}}^{t_{m,v}} r_{u,v} \exp(-r_{u,v}(t - t_{m,u}))dt \\ &= \kappa_{u,v} \exp(-r_{u,v}(t_{m,v} - t_{m,u})) + (1 - \kappa_{u,v}). \end{aligned}\tag{2}$$

If there exist multiple active parents for the node v, i.e., $\eta = |B(v) \cap C_m(t_{m,v})| > 1$, we need to consider possibilities that each parent node succeeds in activating v at time $t_{m,v}$. However, in case of the continuous time delay model, we can ignore simultaneous activations by multiple active parents due to the continuous property. Thus, the probability density that the node v is activated at time $t_{m,v}$, denoted by $h_{m,v}$, can be expressed as

$$\begin{aligned}h_{m,v} &= \sum_{u \in B(v) \cap C_m(t_{m,v})} \mathcal{A}_{m,u,v} \left(\prod_{x \in B(v) \cap C_m(t_{m,v}) \setminus \{u\}} \mathcal{B}_{m,x,v} \right). \\ &= \prod_{x \in B(v) \cap C_m(t_{m,v})} \mathcal{B}_{m,x,v} \sum_{u \in B(v) \cap C_m(t_{m,v})} \mathcal{A}_{m,u,v} (\mathcal{B}_{m,u,v})^{-1}. \end{aligned}\tag{3}$$

Note that we are not able to know which node u actually activated the node v. This can be regarded as a hidden structure.

Next, for the mth information diffusion result, we consider any link $(v, w) \in E$ such that $v \in C_m(T_m)$ and $w \notin D_m$. Let $g_{m,v,w}$ denote the probability that the node w is not activated by the node v within the observed time period $[t_m, T_m]$. We can easily derive the following equation:

$$g_{m,v,w} = \kappa_{v,w} \exp(-r_{v,w}(T_m - t_{m,v})) + (1 - \kappa_{v,w}). \tag{4}$$

Here we can naturally assume that each information diffusion process finished sufficiently earlier than the observed final time, i.e., $T_m \gg \max\{t; D_m(t) \neq \emptyset\}$. Thus, as $T_m \to \infty$ in equation (4), we assume

$$g_{m,v,w} \; = \; 1 - \kappa_{v,w}. \tag{5}$$

Therefore, by using equations (3), (5), and the independence properties, we can define the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ by

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M) = \prod_{m=1}^{M} \left(\prod_{t \in \mathcal{T}_m} \prod_{v \in D_m(t)} h_{m,v} \prod_{v \in D_m} \prod_{w \in F(v) \setminus D_m} g_{m,v,w} \right). \tag{6}$$

Here, we retained the product with respect to $v \in D_m(t)$ for completeness, but in practice there is only one v in $D_m(t)$.

In this paper, we focus on the above situation (i.e., equation (5)) for simplicity, but we can easily modify our method to cope with the general one (i.e., equation (4)). Thus, our problem is to obtain the values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$, which maximize equation (6). For this estimation problem, we derive a method based on an iterative algorithm in order to stably obtain its solution.

3.2 Estimation method

We describe our estimation method. Let $\bar{\boldsymbol{r}} = (\bar{r}_{u,v})$ and $\bar{\boldsymbol{\kappa}} = (\bar{\kappa}_{u,v})$ be the current estimates of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$, respectively. For each $v \in D_m$ and $u \in B(v) \cap C_m(t_{m,v})$, we define $\alpha_{m,u,v}$ by

$$\alpha_{m,u,v} = \mathcal{A}_{m,u,v}(\mathcal{B}_{m,u,v})^{-1} \Big/ \sum_{x \in B(v) \cap C_m(t_{m,v})} \mathcal{A}_{m,x,v}(\mathcal{B}_{m,x,v})^{-1}. \tag{7}$$

Let $\bar{\mathcal{A}}_{m,u,v}$, $\bar{\mathcal{B}}_{m,u,v}$, $\bar{h}_{m,v}$, and $\bar{\alpha}_{m,u,v}$ denote the values of $\mathcal{A}_{m,u,v}$, $\mathcal{B}_{m,u,v}$, $h_{m,v}$, and $\alpha_{m,u,v}$ calculated by using $\bar{\boldsymbol{r}}$ and $\bar{\boldsymbol{\kappa}}$, respectively.

From equations (3), (5), (6), we can transform our objective function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M)$ as follows:

$$\log \mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M) = Q(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}}) - H(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}}), \tag{8}$$

where $Q(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}})$ is defined by

$$Q(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}}) = \sum_{m=1}^{M} \left(\sum_{t \in \mathcal{T}_m} \sum_{v \in D_m(t)} Q_{m,v} + \sum_{v \in D_m} \sum_{w \in F(v) \setminus D_m} \log(1 - \kappa_{v,w}) \right),$$
$$Q_{m,v} = \sum_{u \in B(v) \cap C_m(t_{m,v})} \log(\mathcal{B}_{m,u,v}) + \sum_{u \in B(v) \cap C_m(t_{m,v})} \bar{\alpha}_{m,u,v} \log\left(\mathcal{A}_{m,u,v}(\mathcal{B}_{m,u,v})^{-1}\right) \tag{9}$$

and $H(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}})$ is defined by

$$H(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}}) = \sum_{m=1}^{M} \sum_{t \in \mathcal{T}_m} \sum_{v \in D_m(t)} \sum_{u \in B(v) \cap C_m(t_{m,v})} \bar{\alpha}_{m,u,v} \log \alpha_{m,u,v}. \tag{10}$$

Since $H(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}})$ is maximized at $\boldsymbol{r} = \bar{\boldsymbol{r}}$ and $\boldsymbol{\kappa} = \bar{\boldsymbol{\kappa}}$ from equation (10), we can increase the value of $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M)$ by maximizing $Q(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}})$ (see equation (8)). Note here that although $\log \mathcal{A}_{m,u,v}$ is a linear combination of $\log \kappa_{u,v}$, $\log r_{u,v}$, and $r_{u,v}$, $\log \mathcal{B}_{m,u,v}$ cannot be written as such a linear combination (see equations (1), (2)). In order to cope with this problem of $\log \mathcal{B}_{m,u,v}$, we transform $\log \mathcal{B}_{m,u,v}$ in the same way as above, and define $\beta_{m,u,v}$ by

$$\beta_{m,u,v} = \kappa_{u,v} \exp(-r_{u,v}(t_{m,v} - t_{m,u})) \,/\, \mathcal{B}_{m,u,v}$$

Finally, as the solution which maximizes $Q(\boldsymbol{r}, \boldsymbol{\kappa}; \bar{\boldsymbol{r}}, \bar{\boldsymbol{\kappa}})$, we obtain the following update formulas of our estimation method:

$$r_{u,v} = \frac{\sum_{m \in \mathcal{M}_{u,v}^{+}} \bar{\alpha}_{m,u,v}}{\sum_{m \in \mathcal{M}_{u,v}^{+}} (\bar{\alpha}_{m,u,v} + (1 - \bar{\alpha}_{m,u,v})\bar{\beta}_{m,u,v})(t_{m,v} - t_{m,u})},$$
$$\kappa_{u,v} = \frac{1}{|\mathcal{M}_{u,v}^{+}| + |\mathcal{M}_{u,v}^{-}|} \sum_{m \in \mathcal{M}_{u,v}^{+}} (\bar{\alpha}_{m,u,v} + (1 - \bar{\alpha}_{m,u,v})\bar{\beta}_{m,u,v}),$$

where $\mathcal{M}^+_{u,v}$ and $\mathcal{M}^-_{u,v}$ are defined by

$$\mathcal{M}^+_{u,v} = \{m \in \{1, \cdots, M\};\ u, v \in D_m,\ v \in F(u),\ t_{m,u} < t_{m,v}\},$$
$$\mathcal{M}^-_{u,v} = \{m \in \{1, \cdots, M\};\ u \in D_m,\ v \notin D_m,\ v \in F(u)\}.$$

Note that we can regard our estimation method as a kind of the EM algorithm.

3.3 Behavioral analysis

Thus far, we assumed that the parameters (time-delay and diffusion) can vary with respect to links but remain the same irrespective of the topic of information diffused, following Gruhl et al. [3]. However, they may be sensitive to the topic.

Our method can cope with this by assigning m to a topic, and placing a constraint that the parameters depends only on topics but not on links throughout the network G, that is $r_{m,u,v} = r_m$ and $\kappa_{m,u,v} = \kappa_m$ for any link $(u, v) \in E$. This constraint is required because, without this, we have only one piece of observation for each (m, u, v) and there is no way to learn the parameters. Noting that we can naturally assume that people behave quite similarly for the same topic, this constraint should be acceptable. Under this setting, we can easily obtain the parameter update formulas. Using each pair of the estimated parameters, (r_m, κ_m), we can analyze the behavior of people with respect to the topics of information, by simply plotting (r_m, κ_m) as a point of 2-dimensional space.

3.4 Simple case analysis

We analyze a few properties of our proposed estimation method by using simple cases. Assume that a node v became active at time t on some information result. We denote the active parent nodes of v by $u_1, \cdots, u_N$. First, we consider a simple case that diffusion parameter κ is 1 for any link, time-delay parameter r is the same for all links, and the activation times of $u_1, \cdots, u_N$ are all zeros. Then, as shown in equation (3), the probability density that the node v is activated at time t by one of the parent nodes, can be expressed as follows

$$h_v = \sum_{n=1}^{N} r \exp(-rt) \left(1 - \int_0^t r \exp(-r\tau) d\tau\right)^{N-1} = Nr \exp(-Nrt).$$

Similarly, for a case that the parent nodes $u_1, \cdots, u_N$ became active at times $t_1, \cdots t_N$ $(< t)$, respectively, we can easily obtain the following probabilty.

$$h_v = Nr \exp\left(-Nr\left(t - \frac{1}{N}\sum_{n=1}^{N} t_n\right)\right).$$

For the information diffusion result, by solving the maximum likelihood problem which maximizes $\log h_v$ with respect to r, the estimation of the average delay time of our model can be obtained as follows:

$$r^{-1} = N\left(t - \frac{1}{N}\sum_{n=1}^{N} t_n\right).$$

Thus, we can see that this estimation is N times larger than the simple average of time differences. Namely the information diffuses more quickly when there exist multiple active parents, i.e., r^{-1}/N, and this fact coincide with our intuition. Thus for information diffusion phenomena, some simple statistics such as average delay time may fail to obtain the intrinsic property, and this suggest that adequate information diffusion models are vital.

Next, we consider another simple case that diffusion parameter κ is the same for all links, time-delay parameter r is the same for all links, and the activation times of $u_1, \cdots, u_N$ are all zeros. Here diffusion parameter is also a variable. Then the probability density that the node v is activated at time t can be expressed as follows

$$h_v = N\kappa r \exp(-rt)(\kappa \exp(-rt) + (1-\kappa))^{N-1}.$$

Now, by setting $f(\kappa, r) = \log h_v$, we consider maximizing $f(\kappa, r)$ with respect to κ and r. Here we obtain the first- and second-order derivatives of $f(\kappa, r)$ with respect to κ as follows:

$$\frac{\partial f(\kappa, r)}{\partial \kappa} = \frac{1}{\kappa} + (N-1)\frac{\exp(-rt) - 1}{\kappa \exp(-rt) + (1-\kappa)}$$

$$\frac{\partial^2 f(\kappa, r)}{\partial \kappa \partial \kappa} = -\frac{1}{\kappa^2} - (N-1)(\frac{\exp(-rt) - 1}{\kappa \exp(-rt) + (1-\kappa)})^2$$

Thus, for a given parameter r, since the above second-order derivative is negative definite, we can see that there exists a unique global solution with respect to κ. As for r, we obtain the following derivatives:

$$\frac{\partial f(\kappa, r)}{\partial r} = \frac{1}{r} - t - (N-1)\frac{t\kappa \exp(-rt)}{\kappa \exp(-rt) + (1-\kappa)}$$

$$\frac{\partial^2 f(\kappa, r)}{\partial r \partial r} = -\frac{1}{r^2} + (N-1)\frac{t^2\kappa(1-\kappa)\exp(-rt)}{(\kappa \exp(-rt) + (1-\kappa))^2}$$

Clearly, we cannot guarantee that the above second-order derivative is negative definite. More specifically, this value is likely to be negative one when $r \ll 1$, but the configuration of the objective function can be relatively complex when $r \gg 1$. In our experiments, we empirically evaluate this point by using models with $r = 2$ and $r = 0.5$. However, in case that the diffusion parameters are sufficiently, the second term of right-hand-side of the above equation becomes positive but substantially small one, Here, such a case has been widely explored by most exiting studies including our own experiments. This means that we can expect to have a desirable property for estimating parameter r, as described for parameter κ. Clearly, we need to perform further theoretical and empirical studies because our estimation problem is slightly more complex than the simple case, especially we need to simultaneously estimate both diffusion and time-delay parameters, κ and r. Whereas, we consider that our estimation problem still have a desirable property due to the above facts.

4 Experiments with Artificial data

We evaluated the effectiveness of the proposed learning method using the topologies of two large real network data. First, we evaluated how accurately it can estimate the parameters of the CTIC model from $\mathcal{D}_M$. Next, we considered applying our learning method to the problem of extracting influential nodes, and evaluated how well our learned model can predict the high ranked influential nodes with respect to influence degree $\sigma(v)$, $(v \in V)$ for the true CTIC model.

4.1 Experimental Settings

In our experiments, we employed two datasets of large real networks used in [9], which exhibit many of the key features of social networks. The first one is a trackback network of Japanese blogs. The network data was collected by tracing the trackbacks from one blog in the site *goo*[2] in May, 2005. We refer to this network data as the blog network. The blog network was a strongly-connected bidirectional network, where a link created by a trackback was regarded as a bidirectional link since blog authors establish mutual communications by putting trackbacks on each other's blogs. The blog network had $12,047$ nodes and $79,920$ directed links. The second one is a network of people that was derived from the "list of people" within Japanese Wikipedia. We refer to this network data as the Wikipedia network. The Wikipedia network was also a strongly-connected bidirectional network, and had $9,481$ nodes and $245,044$ directed links.

Here, we assumed the simplest case where $r_{u,v}$ and $\kappa_{u,v}$ are uniform throughout the network G, that is, $r_{u,v} = r$, $\kappa_{u,v} = \kappa$ for any link $(u, v) \in E$. Then, our task is to estimate the values of r and κ. According to [7], we set the value of κ relatively small. In particular, we set the value of κ to a value smaller than $1/\bar{d}$, where $\bar{d}$ is the mean out-degree of a network. Since the values of $\bar{d}$ were about 6.63 and 25.85 for the blog and the Wikipedia networks, respectively, the corresponding values of $1/\bar{d}$ were about 0.15 and 0.03. Thus, as for the true value of the diffusion parameter κ, we decided to set $\kappa = 0.1$ for the blog network and $\kappa = 0.01$ for the Wikipedia network. As for the true value of the time-delay parameter r, we decided to investigate two cases: one with a relatively high value $r = 2$ (a short time-delay case) and the other with a relatively low value $r = 0.5$ (a long time-delay case) in both networks. We used the training data $\mathcal{D}_M$ in the learning stage, which is constructed by generating each D_m from a randomly selected initial active node $D_m(0)$ using the true CTIC model. T_m was chosen to be effectively ∞.

We note that the influence degree $\sigma(v)$ of a node v is invariant with respect to the values of the delay-parameter $\boldsymbol{r}$. In fact, the effect of $\boldsymbol{r}$ is to delay the times when nodes become active, that is, parameter $r_{u,v}$ only controls how soon node v actually becomes active when node u activates node v. Therefore, the set of active nodes under the CTIC model coincides with that under the IC model after a sufficiently long time-period. As described in Section 3.1, we assume that the observed time-period is sufficiently long. Thus, we can evaluate the $\sigma(v)$ of the CTIC model by the influence degree of v for the corresponding IC model. We estimated the influence degrees $\{\sigma(v);\ v \in V\}$ using the method of [8] with the parameter value $10,000$, where the parameter represents the number of bond percolation processes (we do not describe the method here due to the page limit). The average value and the standard deviation of the influence degrees was 87.5 and 131 for the blog network, and 8.14 and 18.4 for the Wikipedia network.

[2] `http://blog.goo.ne.jp/`

Table 1: Learning performance by the proposed method.

Blog network (r = 2)			Wikipedia network (r = 2)			Blog network (r = 0.5)			Wikipedia network (r = 0.5)		
M	$\mathcal{E}_r$	$\mathcal{E}_\kappa$	M	$\mathcal{E}_r$	$\mathcal{E}_\kappa$	M	$\mathcal{E}_r$	$\mathcal{E}_\kappa$	M	$\mathcal{E}_r$	$\mathcal{E}_\kappa$
20	0.013	0.015	20	0.036	0.034	20	0.011	0.012	20	0.026	0.028
40	0.010	0.010	40	0.024	0.016	40	0.010	0.007	40	0.021	0.023
60	0.008	0.008	60	0.013	0.015	60	0.009	0.005	60	0.018	0.021
80	0.007	0.007	80	0.012	0.013	80	0.004	0.004	80	0.014	0.012
100	0.005	0.005	100	0.006	0.011	100	0.004	0.004	100	0.007	0.006

4.2 Comparison Methods

We compared the predicted result of the high ranked influential nodes for the true CTIC model by the proposed method with four heuristics widely used in social network analysis.

The first three of these heuristics are "degree centrality", "closeness centrality", and "betweenness centrality". These are commonly used as influence measure in sociology [11], where the out-degree of node v is defined as the number of links going out from 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. The fourth is "authoritativeness" obtained by the "PageRank" method [12]. We considered this measure 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 [13]. In our experiments, we used a typical setting of $\varepsilon = 0.15$.

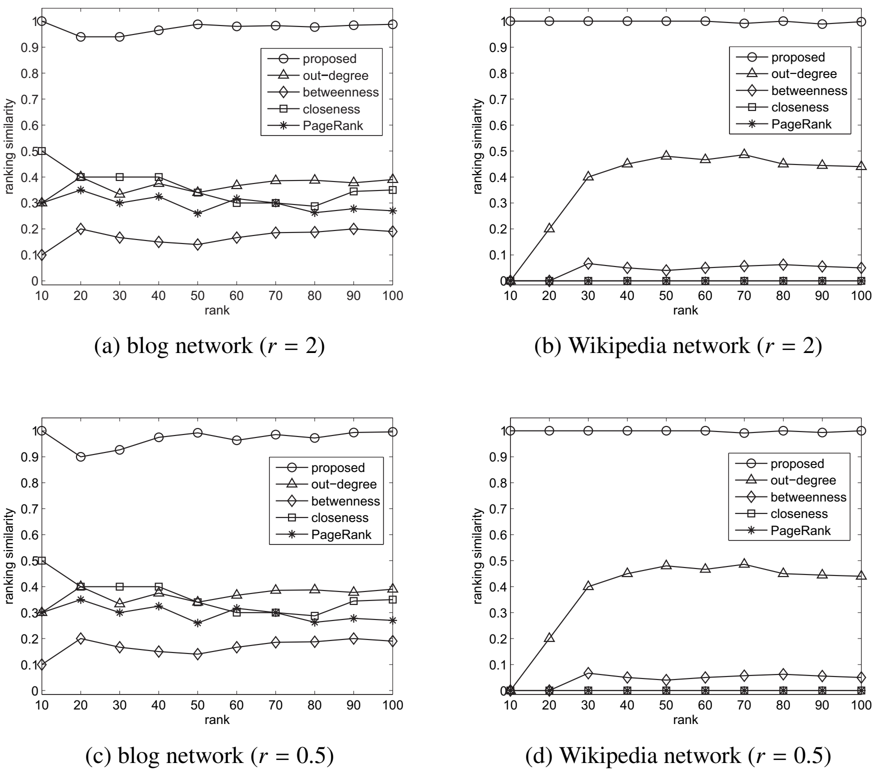
4.3 Experimental Results

First, we examined the parameter estimation accuracy by the proposed method. Let r_0 and κ_0 be the true values of the parameters r and κ, respectively, and let $\hat{r}$ and $\hat{\kappa}$ be the values of r and κ estimated by the proposed method, respectively. We evaluated the learning performance in terms of the error rates,

$$\mathcal{E}_r = \frac{|r_0 - \hat{r}|}{r_0}, \quad \mathcal{E}_\kappa = \frac{|\kappa_0 - \hat{\kappa}|}{\kappa_0}.$$

Table 1 shows the average values of $\mathcal{E}_r$ and $\mathcal{E}_\kappa$ for different numbers of training samples, M, where we performed the same experiment ten times independently. Here, the true value of r is $r = 2$ and $r = 0.5$. Our algorithm can converge to the true values efficiently when there is a reasonable amount of training data. The results demonstrate the effectiveness of the proposed method.

Next, we compared the proposed method with the out-degree, the betweenness, the closeness, and the PageRank methods in terms of the capability of ranking the influential nodes. For any positive integer k ($\leq |V|$), let $L_0(k)$ be the true set of top k nodes,



(a) blog network ($r = 2$)

(b) Wikipedia network ($r = 2$)

(c) blog network ($r = 0.5$)

(d) Wikipedia network ($r = 0.5$)

Fig. 1: Performance comparison in extracting influential nodes.

and let $L(k)$ be the set of top k nodes for a given ranking method. We evaluated the performance of the ranking method by the *ranking similarity* $F(k)$ at rank k, where $F(k)$ is defined by

$$F(k) = \frac{|L_0(k) \cap L(k)|}{k}.$$

We focused on ranking similarities only at high ranks since we are interested in extracting influential nodes. Figures 1a and 1c show the results for the blog network, and Figures 1b and 1d show the results for the Wikipedia network, where the true value of r is $r = 2$ and $r = 0.5$ for Figures 1a and 1b, and Figures 1c and 1d, respectively. In these figures, circles, triangles, diamonds, squares, and asterisks indicate ranking similarity $F(k)$ as a function of rank k for the proposed, the out-degree, the betweenness, the closeness, and the PageRank methods, respectively. For the proposed method, we plotted the average value of $F(k)$ at k for five experimental results in the case of $M = 100$. The proposed method gives far better results than the other heuristic based methods for the both networks, demonstrating the effectiveness of our proposed learning method.

5 Behavioral Analysis of Real World Blog Data

We applied our method to behavioral analysis using a real world blog data based on the method described in 3.3 and investigated how each topic spreads throughout the network.

5.1 Experimental Settings

The network we used is a real blogroll network in which bloggers are connected to each other. We note that when there is a blogroll link from blogger y to another blogger x, this means that y is a reader of the blog of x. Thus, we can assume that topics propagate from blogger x to blogger y. According to [14], we suppose that a topic is represented as a URL which can be tracked down from blog to blog. We used the database of a blog-hosting service in Japan called *Doblog* [3]. The database is constructed by all the Doblog data from October 2003 to June 2005, and contains $52,525$ bloggers and $115,552$ blogroll links.

We identified all the URLs mentioned in blog posts in the Doblog database, and constructed the following list for each URL from all the blog posts that contain the URL:

$$\langle (v_1, t_1), \cdots, (v_k, t_k) \rangle, \quad (t_1 < \cdots < t_k),$$

where v_i is a blogger who mentioned the URL in her/his blog post published at time t_i. By taking into account the blogroll relations for the list, we estimated such paths that the URL might propagate through the blogroll network. We extracted $7,356$ URL propagation paths from the Doblog dataset, where we ignored the URLs that only one blogger mentioned. Out of these, only those that are longer than 10 time steps are chosen for analyses, resulting into 172 sequences. Each sequence data represents a topic, and a topic can be distributed in multiple URLs. The same URL can appear in different sequences. Here note that the time stamp of each blog article is different from each other and thus, the time intervals in the sequence $< t_1, t_2, ..., t_k >$ are not a fixed constant.

5.2 Experimental Results

We ran the experiments for each identified URL and obtained the corresponding parameters κ and r. Figure 2 is a plot of the results for the major URLs. The horizontal axis is the diffusion parameter κ and the vertical axis is the delay parameter r. The latter is normalized such that $r = 1$ corresponds to a delay of one day, meaning $r = 0.1$ corresponds delay of 10 days. We only explain three URLs that exhibit some interesting propagation properties. The circle is a ULR that corresponds to the musical baton which is a kind of telephone game on the Internet. It has the following rules. First, a blogger is requested to respond to five questions about music by some other bologger (receive the baton) and the requested blogger replies to the questions and designate the next five bloggers with the same questions (pass the baton). It is shown that this kind of message propagates quickly (less than one day on the average) with a good chance (one out of

[3] Doblog(`http://www.doblog.com/`), provided by NTT Data Corp. and Hotto Link, Inc.

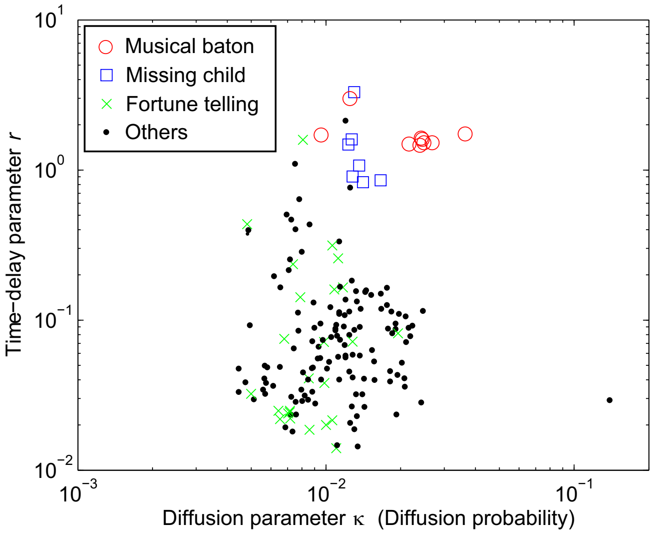


Fig. 2: Results for the Doblog database.

25 to 100 persons responds). This is probably because people are interested in this kind of message passing. The square is a URL that corresponds to articles about a missing child. This also propagates quickly with a meaningful probability (one out of 80 persons responds). This is understandable considering the urgency of the message. The cross is a ULR that corresponds to articles about fortune telling. Peoples responses are diverse. Some responds quickly (less than one day) and some late (more than one month after), and they are more or less uniformly distributed. The diffusion probability is also nearly uniformly distributed. This reflects that each individual's interest is different on this topic. The dot is a URL that corresponds to one of the other topics. Interestingly, the one in the bottom right which is isolated from the rest is a post of an invitation to a rock music festival. This one has a very large probability of being propagated but with very large time delay. In general, it can be said that the proposed method can extract characteristic properties of a certain topics reasonably well only from the observation data.

6 Discussion

Being able to handle the time more precisely brings a merit to the analysis of such information diffusion as in a blog data because the time stamp is available in the unit of second. There are subtle cases where it is not self evident to which value to assign the time when the discretization has to be made. We have solved this problem.

There are many pieces of work in which time sequence data is analyzed assuming a certain model behind. Ours also falls in this category. The proposed approach brings in a new perspective in which it allows to use the structure of a complex network as a kind of background knowledge in a more refined way. There are also many pieces of work on topic propagation analyses, but they focus mostly on the analyses of average propagation speed (propagation speed distribution) and average life time. Our method is new and different in that we explicitly address the diffusion phenomena incorporating diffusion probability and time delay as well as the structure of the network.

The proposed method derives the learning algorithm in a principled way. The objective function has a clear meaning of the likelihood by which to obtain the observed data, and the parameter is iteratively updated in such a way to maximize the likelihood, guaranteeing the convergence. Due to the property of continuous time, we excluded the possibility that a node is activated simultaneously by multiple parent nodes. It is also straightforward to formulate the likelihood taking the possibility of the simultaneous activation into account. However, the numerical experiments revealed that the results are not as accurate as the current model. Having to explore millions of paths with very small probability does harm numerical computation. This is, in a sense, similar to the problem of feature selection in building a classifier. It is known that the existence of irrelevant features is harmful even though the classification algorithm can in theory ignore those irrelevant features.

The CTIC model is a continuous-time information diffusion model that extends the discrete-time model by Gruhl et al [15]. We note that their model is based on the popular IC model and models a time-delay by a geometic distribition. In the CTIC model, we models a time-delay by an exponential distribution as a natural extension. Song et al [16] also modeled time-delays for information flow by exponential distributions when they formulated an information flow model as a continuous-time Markov chain (i.e., a random-surfer model). Thus, we can regard the CTIC model as a natural continuous-time information diffusion model based on the IC model, and investigating the CTIC model can be an important research issue. As shown in Section 2.2, the CTIC model is rather complicated, and developing a learning algorithm of the CTIC model is challenging. In this paper, we presented an effective method for estimating the parameters of the CTIC model from observed data, and applied it to node-ranking and social behavioral data analysis. However, the time-delay distribution for real information diffusion must be more complex, and a power-law distribution and others might be more suitable. Our future work includes incorpolating various more realistic distributions as the time-delay distribution.

We consider that our proposed ranking method presents a novel concept of centrality based on the information diffusion model, i.e., *the CTIC model*. Actually, nodes identified as higher ranked by our method are substantially different from those by each of the conventional methods. This means that our method enables a new type of social network analysis if past information diffusion data are available. Note that this is not to claim to replace them with the proposed method, but simply to propose that it is an addition to them which has a different merit in terms of information diffusion.

We note that the analysis we showed in this paper is the simplest case where κ and r take a single value each for all the links in E. However, the method is very general. In a more realistic setting we can divide E into subsets $E_1, E_2, ..., E_N$ and assign a different value κ_n and r_n for all the links in each E_n. For example, we may divide the nodes into two groups: those that strongly influence others and those not, or we may divide the nodes into another two groups: those that are easily influenced by others and those not. We can further divide the nodes into multiple groups. If there is some background knowledge about the node grouping, our method can make the best use of it.

7 Conclusion

We emphasized the importance of incorporating continuous time delay for the behavioral analysis of information diffusion through a social network, and addressed the problem of estimating the parameters for a continuous time delay independent cascade (CTIC) model from the observed data by rigorously formulating the likelihood of obtaining these data and maximizing the likelihood iteratively with respect to the parameters (time-delay and diffusion). We tested the convergence performance of the proposed method by applying it to the problem of ranking influential nodes using the network structure from two real world web datasets and showed that the parameters converge to the correct values efficiently by the iterative procedure and can predict the high ranked influential nodes much more accurately than the well studied four heuristic methods. We further applied the method to the problem of behavioral analysis of topic propagation using a real world blog data and showed that there are indeed sensible differences in the propagation patterns in terms of delay and diffusion among different topics

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research, U.S. Air Force Research Laboratory 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. Saito, K., Kimura, M., Nakano, R., Motoda, H.: Finding influential nodes in a social network from information diffusion data. In: Proceedings of the International Workshop on Social Computing and Behavioral Modeling (SBP09). (2009) 138–145
11. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press, Cambridge, UK (1994)
12. Brin, S., L.Page: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
13. Ng, A.Y., Zheng, A.X., Jordan, M.I.: Link analysis, eigenvectors and stability. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01). (2001) 903–910
14. Adar, E., Adamic, L.A.: Tracking information epidemics in blogspace. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). (2005) 207–214
15. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proceedings of the 13th International World Wide Web Conference (WWW 2004). (2004) 107–117
16. Song, X., Chi, Y., Hino, K., Tseng, B.L.: Information flow modeling based on diffusion rate for prediction and ranking. In: Proceedings of the 16th International World Wide Web Conference (WWW 2007). (2007) 191–200