Behavioral Analyses of Information Diffusion Models by Observed Data of Social Network

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 investigate how well different information diffusion models explain observation data by learning their parameters and performing behavioral analyses. We use two models (CTIC, CTLT) that incorporate continuous time delay and are extension of well known Independent Cascade (IC) and Linear Threshold (LT) models. We first focus on parameter learning of CTLT model that is not known so far, and apply it to two kinds of tasks: ranking influential nodes and behavioral analysis of topic propagation, and compare the results with CTIC model together with conventional heuristics that do not consider diffusion phenomena. We show that it is important to use models and the ranking accuracy is highly sensitive to the model used but the propagation speed of topics that are derived from the learned parameter values is rather insensitive to the model used.

1 Introduction

The growth of Internet has enabled to form various kinds of large-scale social networks, through which a variety of information including innovation, hot topics and even malicious rumors can be propagated in the form of so-called "word-of-mouth" communications. Social networks are now recognized as an important medium for the spread of information, and a considerable number of studies have been made [1–5]. Widely used information diffusion models in these studies are the *independent cascade (IC)* [6–8] and the *linear threshold (LT)* [9, 10]. They have been used to solve such problems as the *influence maximization problem* [7, 11].

These two models focus on different information diffusion aspects. The IC model is sender-centered and an active node influences its inactive neighbors *independently* with diffusion probabilities assigned to links. On the other hand, the LT model is receiver-centered and a node is influenced by its active neighbors if the sum of their weights

exceeds the threshold for the node. Which model is more appropriate depends on the situation and selecting appropriate model is not easy. In order to study this problem, first of all, we need to know how different model behaves differently and how well or badly explain the observation data. Both models have parameters that need be specified in advance: diffusion probabilities for the IC model, and weights for the LT model. However, their true values are not known in practice. This poses yet another problem of estimating them from a set of information diffusion results that are observed as time-sequences of influenced (activated) nodes. To the best of our knowledge, there are only a few methods that can estimate the parameter values for the IC models and its variant that incorporates continuous time delay (referred to as the CTIC model) [3, 12, 13], but none for the LT model.

With this background, we first propose a novel method of learning the parameter values of a variant of the LT model that incorporates continuous time delay, similar to the CTIC model. We refer to this model as the CTLT model. It is indispensable to be able to cope with continuous time delay to do realistic analyses of information diffusion because, in the real world, information propagates along the continuous time axis, and time-delays can occur during the propagation. Thus, the proposed method has to estimate not only the weight parameters but also the time-delay parameters from the observed data. Incorporating time-delay makes the time-sequence observation data structural. In order to exploit this structure, we introduce an objective function that rigorously represents the likelihood of obtaining such observed data sequences under the CTLT model on a given network, and obtain parameter values that maximize this function by deriving parameter update EM algorithm. Next, we experimentally analyze how different models affect the information diffusion results differently by applying the proposed method to two tasks and comparing the results with the method which we already developed with the CTIC model [13]. The first task is ranking influential nodes in a social network, and we show that ranking is highly sensitive to the model used. We also show that the proposed method works well and can extract influential nodes more accurately than the well studied conventional four heuristic methods that do not take diffusion phenomena explicitly. The second task is the behavioral analysis of topic propagation on a real world blog data. We show that both model well capture the propagation phenomena on different topics at this level of abstract characterization.

2 Proposed Method

2.1 Information Diffusion Model

For a given directed network (or equivalently graph) $G = (V, E)$, let V be a set of nodes (or vertices) and E a set of links (or edges), where we denote each link by $e = (v, w) \in E$ and $v \neq w$, meaning there exists a directed link from a node v to a node w. For each node v in the network G, we denote $F(v)$ as a set of child nodes of v as follows: $F(v) = \{w; (v, w) \in E\}$. Similarly, we denote $B(v)$ as a set of parent nodes of v as follows: $B(v) = \{u; (u, v) \in E\}$. We define the LT model. In this model, for every node $v \in V$, we specify a *weight* ($\omega_{u,v} > 0$) from its parent node u in advance such that $\sum_{u \in B(v)} \omega_{u,v} \leq 1$. The diffusion process from a given initial active set S proceeds according to the following randomized rule. First, for any node $v \in V$, a *threshold* θ_v is

chosen uniformly at random from the interval [0, 1]. At time-step t, an inactive node v is influenced by each of its active parent nodes, u, according to weight $\omega_{u,v}$. If the total weight from active parent nodes of v is at least threshold θ_v, that is, $\sum_{u \in B_t(v)} \omega_{u,v} \geq \theta_v$, then v will become active at time-step $t+1$. Here, $B_t(v)$ stands for the set of all the parent nodes of v that are active at time-step t. The process terminates if no more activations are possible. Next, we extend the LT model so as to allow continuous-time delays, and refer to the extended model as the *continuous-time linear threshold (CTLT) model*. In the CTLT model, in addition to the weight set $\{\omega_{u,v}\}$, we specify real values r_v with $r_v > 0$ in advance for each node $v \in V$. We refer to r_v as the *time-delay parameter* on node v. Note that r_v depends only on v, which means that it is the node v's decision when to receive the information once the activation condition has been satisfied. The diffusion process unfolds in continuous-time t, and proceeds from a given initial active set S in the following way. Suppose that the total weight from active parent nodes of v became at least threshold θ_v at time t for the first time. Then, v will become active at time $t + \delta$, where we choose a delay-time δ from the exponential distribution with parameter r_v. Further, note that even though some other non-active parent nodes of v become active during the time period between t and $t + \delta$, the activation time of v, $t + \delta$, still remains the same. The other diffusion mechanisms are the same as the LT model.

For an initial active node v, let $\varphi(v)$ denote the number of active nodes at the end of the random process for the CTLT model. Note that $\varphi(v)$ is a random variable. Let $\sigma(v)$ denote the expected value of $\varphi(v)$. We call $\sigma(v)$ the *influence degree* of v for the CTLT model.

2.2 Learning problem

For the sake of technical convenience, we introduce a slack weight $\omega_{v,v}$ for each node $v \in V$ so as to be $\omega_{v,v} + \sum_{u \in B(v)} \omega_{u,v} = 1$. Here note that such a slack weight $\omega_{v,v}$ never contributes to the activation of v. We define the parameter vectors $\boldsymbol{r}$ and $\boldsymbol{\omega}$ by $\boldsymbol{r} = (r_v)_{v \in V}$ and $\boldsymbol{\omega} = (\omega_{u,v})_{(u,v) \in E}$. In practice, their true values are not available. Thus, we must estimate them from past information diffusion histories.

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 (called mth result, hereafter for simplicity),

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

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 initial observed time and T_m is the final observed time. We assume that for any active node v in the mth 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 has become active in the mth result, i.e., $v \in D_m(t_{m,v})$. For any $t \in \mathcal{T}_m$, we set

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

Note that $C_m(t)$ is the set of nodes that had become active before time t in the mth result. We also interpret D_m as referring to the set of all the active nodes in the mth result for convenience sake. The problem is to estimate the values of $\boldsymbol{r}$ and $\boldsymbol{\omega}$ from $\mathcal{D}_M$.

2.3 Likelihood function

For the learning problem described above, we derive the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\omega}$ in a rigorous way to use as our objective function. Here note that for each node v, since a threshold θ_v is chosen uniformly at random from the interval $[0, 1]$, we can regard each weight $\omega_{*,v}$ as a multinomial probability, namely, $\omega_{v,v} + \sum_{u \in B(v)} \omega_{u,v} = 1$.

Suppose that a node v became active at time $t_{m,v}$ for the mth result. Then, we know that the total weight from active parent nodes of v became at least threshold θ_v at the time when one of these active parent nodes, $u \in B(v) \cap C_m(t_{m,v})$, became first active. However, in case of $|B(v) \cap C_m(t_{m,v})| > 1$, there is no way of exactly knowing the actual node due to the continuous time-delay. Suppose that a node v was actually activated when a node $\zeta \in B(v) \cap C_m(t_{m,v})$ became activated. Then θ_v is between $\sum_{u \in B(v) \cap C_m(t_{m,\zeta})} \omega_{u,v}$ and $\omega_{\zeta,v} + \sum_{u \in B(v) \cap C_m(t_{m,\zeta})} \omega_{u,v}$. Namely, the probability that θ_v is chosen from this range is $\omega_{\zeta,v}$. Here note that such events with respect to different active parent nodes are mutually disjoint. Thus, the probability density that the node v is activated at time $t_{m,v}$, denoted by $h_{m,v}$, can be expressed as

$$h_{m,v} = \sum_{u \in B(v) \cap C_m(t_{m,v})} \omega_{u,v} r_v \exp(-r_v(t_{m,v} - t_{m,u})). \tag{1}$$

Next, we consider any node $w \in V$ belonging to $\partial D_m = \{w; (v, w) \in E \wedge v \in C_m(T_m) \wedge w \notin D_m\}$ for the mth result. Let $g_{m,w}$ denote the probability that the node w is not activated by the node v within the observed time period $[t_m, T_m]$. 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 $T_m \to \infty$, we obtain

$$g_{m,w} = 1 - \sum_{v \in B(w) \cap C_m(T_m)} \omega_{v,w}. \tag{2}$$

Therefore, by using Equations (1) and (2), and the independence properties, we can define the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\omega}$ by

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M) = \prod_{m=1}^{M} \left(\prod_{t \in \mathcal{T}_m} \prod_{v \in D_m(t)} h_{m,v} \prod_{w \in \partial D_m} g_{m,w} \right). \tag{3}$$

Thus, our problem is to obtain the time-delay parameter vector $\boldsymbol{r}$ and the diffusion parameter vector $\boldsymbol{\omega}$, which maximizes Equation (3). For this estimation problem, we can derive an estimation method based on the Expectation-Maximization algorithm in order to stably obtain its solutions, although we skip its derivation due to a space limitation.

2.4 Behavioral analysis

Thus far, we assumed that the time-delay and diffusion parameters can vary with respect to nodes and links but independent of the topic of information diffused. However, they may be sensitive to the topic.

Our method can cope with this by assigning a different m to a different topic, and placing a constraint that the parameters depends only on topics but not on nodes and links throughout the network G, that is $r_{m,v} = r_m$ and $\omega_{m,u,v} = q_m|B(v)|^{-1}$ for any node $v \in V$ or link $(u, v) \in E$. Here note that $0 < q_m < 1$ and $\omega_{v,v} = 1 - q_m$. 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, q_m), we can analyze the behavior of people with respect to the topics of information, by simply plotting (r_m, q_m) as a point of 2-dimensional space (See Fig. 2 in Section 3.2).

3 Experiments

We applied the proposed learning method to two tasks to analyze how different models affect the information diffusion results differently and compared the results with the method which we already developed with the CTIC model [13]. First, we applied it to the problem of extracting influential nodes, and evaluated the performance of the CTLT model, i.e. parameter learning and influential node prediction, using the topologies of four large real network data. Next, we applied our method to behavioral analysis using a real world blog data based on the method described in section 2.4 and investigated how each topic spreads throughout the network.

3.1 Ranking Influential Nodes

Experimental Settings We employed four datasets of large real networks, which are all bidirectional connected networks. The first one is a trackback network of Japanese blogs used in [14] and had 12,047 nodes and 79,920 directed links (the blog network). The second one is a network of people that was derived from the "list of people" within Japanese Wikipedia, also used in [14], and had 9,481 nodes and 245,044 directed links (the Wikipedia network). The third one is a network derived from the Enron Email Dataset [15] by extracting the senders and the recipients and linking those that had bidirectional communications and there were 4,254 nodes and 44,314 directed links (the Enron network). The fourth one is a co-authorship network used in [16] and had 12,357 nodes and 38,896 directed links (the coauthorship network).

Here, we assumed the simplest case where $\omega_{u,v} = q|B(v)|^{-1}$ and $r_v = r$ for any $u, v \in V$. One reason behind this assumption is that there is no need that the observation sequence data have to pass through every link at least once. This drastically reduces the amount of data necessary to learn the parameters. Then, our task is to estimate the values of q and r. The true value of q was decided to be set to 0.9 in order to achieve reasonably high influence degrees of nodes, and the true value of r was decided to be chosen from two values, one with a relatively high value $r = 2$ (a short time-delay case) and the other with a relatively low value $r = 1/2$ (a long time-delay case). The training data $\mathcal{D}_M$ in the learning stage was constructed by generating each D_m from a randomly selected initial active node $D_m(0)$ using the true CTLT model. We chose

Table 1: Parameter estimation accuracy by the proposed method.

Blog network			Wikipedia network			Enron network			Coauthorship network		
r^*	$\mathcal{E}_q$	$\mathcal{E}_r$	r^*	$\mathcal{E}_q$	$\mathcal{E}_r$	r^*	$\mathcal{E}_q$	$\mathcal{E}_r$	r^*	$\mathcal{E}_q$	$\mathcal{E}_r$
2	0.024	0.060	2	0.015	0.028	2	0.013	0.031	2	0.023	0.043
1/2	0.017	0.012	1/2	0.016	0.007	1/2	0.011	0.004	1/2	0.024	0.011

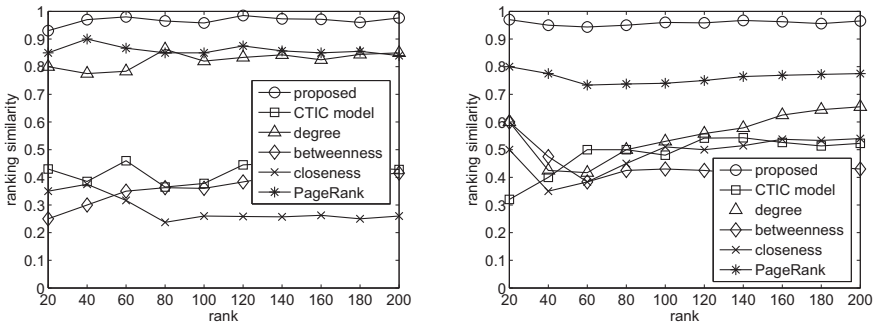
$T_m = \infty$ and used $M = 100$. We repeated the same experiment for each network five times independently.

We measure the influence of node v by the influence degree $\sigma(v)$ for the CTLT model that has generated $\mathcal{D}_M$. We compared the result of the high ranked influential nodes for the true CTLT model predicted by the proposed method with four heuristics widely used in social network analysis and the CTIC model based method [13]. The four heuristics are the same as those used in [13], "degree centrality", "closeness centrality", "betweenness centrality", and "authoritativeness" . The first three heuristics are commonly used as influence measure in sociology [17]. The authoritativeness is obtained by the "PageRank" method [18] which is a well known method for identifying authoritative or influential pages in a hyperlink network of web pages [5]. The CTIC model based method employs the CTIC model as the information diffusion model[13], where we learn the parameters of the CTIC model from the observed data $\mathcal{D}_M$, and rank nodes according to the influence degrees based on the learned model.

Experimental Results First, we examined the performance of estimating parameters by the proposed method. Let q^* and r^* denote the true values of q and r, respectively. Let $\hat{q}$ and $\hat{r}$ be the values of q and r estimated by the proposed method, respectively. We evaluated the parameter estimation accuracy by the errors $\mathcal{E}_q = |q^* - \hat{q}|$ and $\mathcal{E}_r = |r^* - \hat{r}|$. Table 1 shows the average values of $\mathcal{E}_q$ and $\mathcal{E}_r$ of five trials. We observe that the estimated values were close to the true values. The results demonstrate the effectiveness of the proposed method.

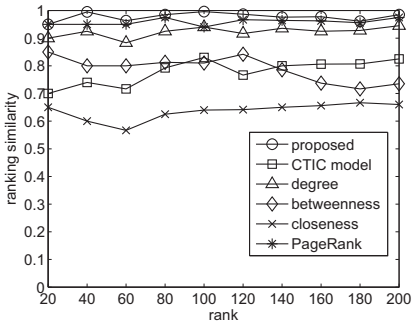
Next, in terms of extracting influential nodes from the network $G = (V, E)$, we evaluated the performance of the ranking methods mentioned above by the *ranking similarity* $\mathcal{F}(k) = |L^*(k) \cap L(k)|/k$ within the rank $k(> 0)$, where $L^*(k)$ and $L(k)$ are the true set of top k nodes and the set of top k nodes for a given ranking method, respectively. We focused on the performance for high ranked nodes since we are interested in extracting influential nodes. Figure 1 shows the results in the case of $r^* = 2$ for the blog, the Wikipedia, the Enron, and the coauthorship networks, respectively. For the proposed and the CTIC model methods, we plotted the average value of $\mathcal{F}(k)$ at k for five experimental results stated earlier. The results in the case of $r^* = 1/2$ for the proposed and the CTIC model methods were very similar to those in the case of $r^* = 2$. We see that the proposed method gives better results than the other methods for these networks, demonstrating the effectiveness of our proposed learning method. We also observe that the CTIC model method does not work well for predicting the high ranked influential nodes for the CTLT model for the problem setting we employed.

[5] As for the jump parameter ε of PageRank, we used a typical setting of $\varepsilon = 0.15$.

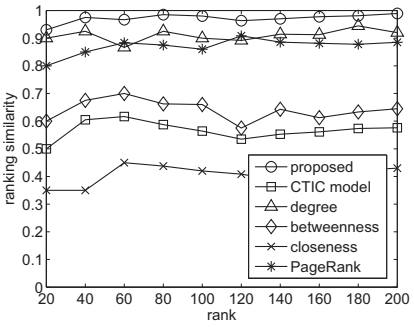


(a) Blog network

(b) Wikipedia network



(c) Enron network



(d) Coauthorship network

Fig. 1: Performance comparison in extracting influential nodes in the case of $r^* = 2$.

3.2 Behavioral Analysis of Real World Blog Data

Experimental Settings To compare the result by the proposed method with that by the CTIC model based method [13], we used the same real blogroll network as [13], which was generated from the database of a blog-hosting service in Japan called *Doblog* [6]. In the network, bloggers are connected to each other and we assume that topics propagate from blogger x to another blogger y when there is a blogroll link from y to x because this means that y is a reader of the blog of x. In addition, according to [19], it is supposed that a topic is represented as a URL which can be tracked down from blog to blog. We used the same propagation sequences of 172 URLs as [13] for this analysis, each of which is longer than 10 time steps. Please refer to [13] for more detailed description of the network generation and URL sequences.

Experimental Results We ran the experiments for each identified URL and obtained the corresponding parameters q and r. Figure 2 is a plot of the results for the major URLs. The horizontal axis is the diffusion parameter q and the vertical axis is the delay

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

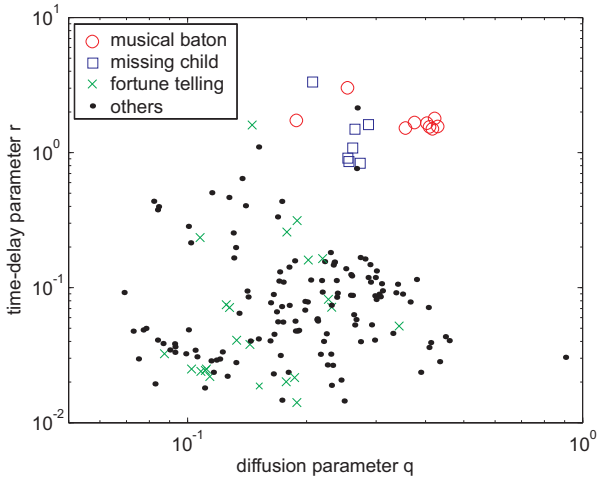


Fig. 2: Results for the Doblog database.

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. In general, from this result, it can be said that the proposed method can extract characteristic properties of certain topics reasonably well only from the observation data. We only explain three URLs that exhibit some interesting propagation properties. The circle is a URL that corresponds to the musical baton which is a kind of telephone game on the Internet. It is shown that this kind of message propagates quickly (less than one day on the average) with a good chance (one out of 25 to 100 persons responds). This is probably because people are easily interested in and influenced by 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 URL 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 (not necessarily the same).

4 Discussion

With the addition of the proposed method, we now have ways to compare the diffusion process with respect to two models (the CTIC model and the CTLT model) for the same observed dataset. Being able to learn the parameters of these models enable us to analyze the diffusion process more precisely. Comparing the results bring us deeper insights into the relation between models and information diffusion processes. Hence, we consider the contribution of the proposed method is significant.

Indeed, we obtained two interesting insights through the comparative experiments in the previous section. The first one comes from the results of ranking influential nodes, in which the ranking accuracy by the proposed method was better than those by the conventional heuristics, which was sort of expected, but the accuracy by the CTIC method was not, which is rather surprising. This means that the ranking results that involve detailed probabilistic simulation is very sensitive to the underlying model assumed to generate the observed data. In fact, the similar results were obtained when the role of the two models are switched, i.e. data generated by CTIC and the model assumed to be CTLT (results not shown due to the space limitation). In other words, it is very important to select an appropriate model for the analysis of information diffusion from which the data has been generated. However, this is a very hard problem in reality. The second one comes from the results of the behavior analysis of topic propagation. The pattern shown in Fig.2 was very similar to that by the CTIC method shown in [13]. Regardless of the model used, in both results, the parameters for the topics that actually propagated quickly/slowly in observation converged to the values that enable them to propagate quickly/slowly on the model. Namely, we can say that the difference of models used has little influence on the relative difference of topic propagation property which indeed strongly depends on topic itself. Both models are well defined and can explain this property at this level of abstraction. However, we have to carefully choose a model at least when solving such problems as the influence maximization problem [7, 11], a problem at a more detailed level.

5 Conclusion

We considered the problem of analyzing information diffusion process in a social network using two kinds of information diffusion models, incorporating continuous time delay, the CTIC model and the CTLT model, and investigated how the results differ according to the model used. To this end, we proposed a novel method of learning the parameters of the CTLT model from the observed data, and experimentally confirmed that it works well on real world datasets. We also obtained the following two important observations through the experiments for the two tasks. One is that in learning the information diffusion parameters of nodes and links, the learning results are highly sensitive to the model used. The other is that in analyzing the topic-oriented characteristics such as the propagation speed of each topic, using different models has little influence on the analysis results. These two contrasting observations may hold only for well-defined diffusion models such as the CTIC and CTLT models. These findings would help us consider whether we should select a model carefully, or not. In practice, as there are numerous factors that affects the information diffusion process, it is difficult to select an appropriate model in a more realistic setting. This model selection is our future work.

Acknowledgment

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., 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
9. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of National Academy of Science, USA **99** (2002) 5766–5771
10. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research **34** (2007) 441–458
11. 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
12. 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
13. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Proceedings of the 1st Asian Conference on Machine Learning (ACML2009). (2009) 322–337
14. 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
15. 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
16. 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
17. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press, Cambridge, UK (1994)
18. Brin, S., L.Page: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
19. 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