Acquiring Expected Influence Curve from Single Diffusion Sequence

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

[1] School of Administration and Informatics, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
{b7101,k-saito}@u-shizuoka-ken.ac.jp
[2] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp
[3] Department of Electronics and Informatics, Ryukoku University
Otsu 520-2194, Japan
kimura@rins.ryukoku.ac.jp
[4] Department of Integrated Information Technology, Aoyama Gakuin Univesity
Kanagawa 229-8558, Japan
ohara@it.aoyama.ac.jp

Abstract. We address the problem of estimating the expected influence curves with good accuracy from a single observed information diffusion sequence, for both the asynchronous independent cascade (AsIC) model and the asynchronous linear threshold (AsLT) model. We solve this problem by first learning the model parameters and then estimating the influence curve using the learned model. Since the length of the observed diffusion sequence may vary from a very long one to a very short one, we evaluate the proposed method by simulation using artificial diffusion sequence of various lengths and show that the proposed method can estimate the expected influence curve robustly from a single diffusion sequence with various lengths.

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 diffuse through social networks in the form of so-called "word-of-mouth" communications. Such social interaction processes are usually characterized by highly distributed phenomena over a social network, but the high complexity and distributed nature of these processes do not necessarily imply that these evolutions are chaotic or unpredictable. Just as natural scientists discover laws and create models for their fields, so can one, in principle, find empirical regularities and develop explanatory accounts of evolution in a social network. Especially, such predictive knowledge would be valuable for market opportunities. In this paper, as a piece of such predictive knowledge, we focus on acquiring the expected influence curve of each information source node by using information diffusion models.

Widely used information diffusion models in recent 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. 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 and LT models and their variants that incorporate asynchrnous time delay (referred to as the AsIC model and the AsLT model) [3, 12–14]. We follow the methods in [13, 14] in this paper.

Now assume that we observed a single information diffusion sequence for an information source node. How can we acquire the expected influence curve from this single instance of observation? This is the problem we want to solve. In a sence, this sequence can be regarded as a piece of crude knowledge about the expected influence curve because we can count the number of nodes that have been influenced (activated) by any time point t which we specify. However, due to its stochastic nature, such a sequence varies in a quite wide range each time we observe it, even if we know which of the two models (AsIC and AsLT) the information diffusion follows. Thus, it is undesirable to approximate the expected influence curve by a single instance of observed sequence.

In this paper, we assume that information diffuses over a network by either the AsIC model or the AsLT model, and propose a novel method for estimating the expected influence curve by first estimating parameters for the assumed models from a single observed informaition diffusion sequence and use the learned model to estimate the expected curve. In another word, our method can be viewed as a knowledge refinement method from the observed single information diffusion sequence to the expected influence curve based on the information diffusion model. We performed extensive experiments to evaluate whether the proposed method can estimate the influence curve much more accurately than the observed diffusion curve itself. The results cleary show the advantage of our method.

The paper is organized as follows. We revisit the information diffusion models and briefly explain the independent cascade model, the linear threshold model, and their asynchronous time delay versions (the models we use in this paper) : AsIC and AsLT in section 2, and revisit parameter learning algorithms for AsIC and AsLT in section 3. We then describe the estimation method of the expected influence curve in section 4, and explain the experimental results in detail in section 5, followed by some discussions in section 6. We summarize our conlcusion in section 7.

2 Information Diffusion Models

We first define the IC model according to [7], and then introduce the asynchronous IC model (AsIC). After that, we do the same for the LT model and the asynchronous LT model (AsLT). We mathematically model the spread of information over 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. 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 that time. 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

The IC model is a fundamental probabilistic model for the spread of a disease. In this model, we specify a real value $\kappa_{u,v}$ with $0 < \kappa_{u,v} < 1$ for each link (u, v) in advance. Here $\kappa_{u,v}$ 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 information source node 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 $\kappa_{u,v}$. 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 Asynchronous Independent Cascade Model

Next, we extend the IC model so as to allow continuous-time delays, and refer to the extended model as the *Asynchronous independent cascade (AsIC) model*. In the AsIC model, we specify a real value $r_{u,v}$ with $r_{u,v} > 0$ for each link $(u, v) \in E$ in advance together with diffusion probability $\kappa_{u,v}$. We refer to $r_{u,v}$ as the *time-delay parameter* through link (u, v).

The diffusion process unfolds in continuous-time t, and proceeds from a given information source node 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 at exactly the same time $t + \delta$. So we ignore this possibility. 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 information source node v, let $\varphi(t; v)$ denote the number of active nodes at a specified time t, i.e. the number of nodes that have become activated by t. Note that $\varphi(t; v)$ is a random variable. Let $\sigma(t; v)$ denote the expected value of $\varphi(t; v)$. We call $\sigma(t; v)$ the *expected influence curve* of v for the AsIC model.

2.3 Linear Threshold Model

The LT model is a fundamental probabilistic model for the spread of innovation. In this model we specify a *weight* ($\omega_{u,v} > 0$) for every node $v \in V$ 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 no less than 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.

2.4 Asynchronous Linear Threshold Model

We make a similar extention to the LT model so as to allow continuous-time delays, and refer to the extended model as the *Asynchronous linear threshold (AsLT) model*. In the AsLT 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 no less than the 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. Similarly to the AsIC model, we can also define the expected influence curve $\sigma(t; v)$ of an information source node v for the AsIC model.

3 Learning Algorithms

We define the time-delay parameter vector $\boldsymbol{r}$ and the diffusion parameter vector $\boldsymbol{\kappa}$ by $\boldsymbol{r} = (r_{u,v})_{(u,v)\in E}$ and $\boldsymbol{\kappa} = (\kappa_{u,v})_{(u,v)\in E}$ for the AsIC model. Similarly, we define the parameter vectors $\boldsymbol{\omega}$ and $\boldsymbol{r}$ by $\boldsymbol{\omega} = (\omega_{u,v})_{(u,v)\in E}$ and $\boldsymbol{r} = (r_v)_{v\in V}$ for the AsLT model. In practice, the true values of these parameters are not available. Thus, we must learn them from past information diffusion histories.

We consider an observed data set of M independent information diffusion results, $\{D_m;\ m = 1, \cdots, M\}$. Here, each D_m is a set of pairs of active nodes and their activation times in the mth information diffusion result, $D_m = \{(u, t_{m,u}),\ (v, t_{m,v}),\ \cdots\}$. For each D_m, we denote the observed initial time by $t_m = \min\{t_{m,v};\ (v, t_{m,v}) \in D_m\}$, and the observed final time by $T_m \geq \max\{t_{m,v};\ (v, t_{m,v}) \in D_m\}$. Note that T_m is not necessarily equal to the final activation time. Hereafter, we express our observation data by $\mathcal{D}_M = \{(D_m, T_m);\ m = 1, \cdots, M\}$. For any $t \in [t_m, T_m]$, we set $C_m(t) = \{v;\ (v, t_{m,v}) \in D_m,\ t_{m,v} < t\}$. Namely, $C_m(t)$ is the set of active nodes before time t in the mth information diffusion result. For convenience sake, we use C_m as referring to the set of all the active nodes in the mth information diffusion result. Moreover, we define a set of non-active nodes with at least one active parent node for each by $\partial C_m = \{v;\ (u, v) \in E,\ u \in C_m,\ v \notin C_m\}$. For each node $v \in C_m \cup \partial C_m$, we define the following subset of parent nodes, each of which has a chance to activate v.

$$\mathcal{B}_{m,v} = \begin{cases} B(v) \cap C_m(t_{m,v}) & \text{if } v \in C_m(t_{m,v}), \\ B(v) \cap C_m & \text{if } v \in \partial C_m. \end{cases}$$

In oder to learn the values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ for the AsIC model, and the values of $\boldsymbol{r}$ and $\boldsymbol{\omega}$ for the AsLT model for the given $\mathcal{D}_M$, we adopt the method proposed in [13] and [14], respectively, each of which is only briefly explained here.

3.1 Learning Parameters of AsIC Model

To learn the values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ from $\mathcal{D}_M$ for the AsIC model, we revisit the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\omega}$ to use as the objective function [13]. First, we consider any node $v \in C_m$ with $t_{m,v} > t_m$ for the mth information diffusion result. Let $\Phi_{m,u,v}$ 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,

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

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

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

As explained in 2.2, it is not necessary to consider simultaneous activations by multiple active parents even if $\eta = |B(v) \cap C_m(t_{m,v})| > 1$. Thus, the probability density that the node v is activated at time $t_{m,v}$, denoted by $h^{(IC)}_{m,v}$, can be expressed as

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

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 $v \in C_m$ and $w \notin C_m$. Let $g^{(IC)}_{m,v,w}$ denote the probability that the node w is not activated by the node v during the observed time period $[t_m, T_m]$. We can easily derive the following equation:

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

Therefore, by using equations (3), (4), 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

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M) = \prod_{m=1}^{M} \prod_{v \in C_m} \left(h^{(IC)}_{m,v} \prod_{w \in F(v) \setminus C_m} g^{(IC)}_{m,v,w} \right), \tag{5}$$

Thus, our problem is to obtain the time-delay parameter vector $\boldsymbol{r}$ and the diffusion parameter[2]. vector $\boldsymbol{\kappa}$, which together maximize Equation (5). To obtain the values of $\boldsymbol{r}$ and $\boldsymbol{\kappa}$, we can employ a learning method based on the Expectation-Maximization algorithm in order to stably obtain its solutions [13].

3.2 Learning Parameters of AsLT Model

To learn the values of $\boldsymbol{r}$ and $\boldsymbol{\omega}$ from $\mathcal{D}_M$ for the AsLT model, we also revisit the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\omega}$ to use as the objective function [14]. For the sake of technical convenience, we introduce a slack weight $\omega_{v,v}$ for each node $v \in V$ such that $\omega_{v,v} + \sum_{u \in B(v)} \omega_{u,v} = 1$. Here note that such a slack weight $\omega_{v,v}$ never contributes to the activation of v and 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.

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 no less than the threshold θ_v at the time when one of these active parent nodes, $u \in \mathcal{B}_{m,v}$, became first active. However, in case of $|\mathcal{B}_{m,v}| > 1$, there is no way of exactly knowing the actual nodes due to the continuous time-delay. Suppose that a node v was actually activated when a node $\zeta \in \mathcal{B}_{m,v}$ 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 $h^{(LT)}_{m,v}$, can be expressed as

$$h^{(LT)}_{m,v} = \sum_{u \in \mathcal{B}_{m,v}} \omega_{u,v} r_v \exp(-r_v(t_{m,v} - t_{m,u})). \tag{6}$$

Here we define $h^{(LT)}_{m,v} = 1$ if $t_{m,v} = t_m$.

[2] We use "diffusion parameter" and "diffusion probability" interchangeably depending on the contex

Next, we consider any node $w \in V$ belonging to $\partial C_m = \{w; (v, w) \in E \wedge v \in C_m(T_m) \wedge w \notin C_m(T_m)\}$ for the mth result. Let $g_{m,v}$ denote the probability that the node v is not activated during the observed time period $[t_m, T_m]$. We can calculate $g_{m,v}$ as follows:

$$\begin{aligned} g_{m,v}^{(LT)} &= 1 - \sum_{u \in \mathcal{B}_{m,v}} \omega_{u,v} \int_{t_{m,u}}^{T_m} r_v \exp(-r_v(t - t_{m,u}))dt \\ &= 1 - \sum_{u \in \mathcal{B}_{m,v}} \omega_{u,v}(1 - \exp(-r_v(T_m - t_{m,v}))) \\ &= \omega_{v,v} + \sum_{u \in B(v) \setminus \mathcal{B}_{m,v}} \omega_{u,v} + \sum_{u \in \mathcal{B}_{m,v}} \omega_{u,v} \exp(-r_v(T_m - t_{m,v})). \end{aligned} \tag{7}$$

Therefore, by using Equations (6) and (7), 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

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M) = \prod_{m=1}^{M} \left(\prod_{v \in C_m} h_{m,v}^{(LT)} \right) \left(\prod_{v \in \partial C_m} g_{m,v}^{(LT)} \right). \tag{8}$$

Thus, our problem is to obtain the time-delay parameter vector $\boldsymbol{r}$ and the diffusion parameter vector $\boldsymbol{\omega}$, which together maximize Equation (8). To obtain the values of $\boldsymbol{r}$ and $\boldsymbol{\omega}$, we can also employ a learning method based on the Expectation-Maximization algorithm in order to stably obtain its solutions [14].

4 Expected Influence Curve Acquisition

Thus far, we assumed that the time-delay and diffusion parameters can vary with respect to nodes and links. However, as mentioned earlier, we address the problem of estimating the influence curves from single observed diffusion sequences. Thus, in order to avoid overfitting to the observed data, we place a constraint that the parameters are uniform on nodes and links throughout the network G. Therefore, we set $r_{u,v} = r$ and $\kappa_{u,v} = \kappa$ for any link $(u, v) \in E$ in case of the AsIC model and $r_v = r$ and $\omega_{u,v} = \kappa |B(v)|^{-1}$ for any node $v \in V$ and link $(u, v) \in E$ in case of the AsLT model, where note that $0 < \kappa < 1$ and $\omega_{v,v} = 1 - \kappa$. Namely, since parameter κ of the AsLT model can be interpreted as a kind of diffusion probability, we employ the same symbol as used in the AsIC model. Without this constraint there is no way to learn the parameters since we only have one sequence of observation that covers only a small part of existing links.

We describe our method for acquiring an expected influence curve under the AsIC and AsLT model. Assume that we have observed the following single information diffusion sequence from the information source node v_0 at time t_0.

$$d = \{(v_0, t_0),\ (v_1, t_1),\ \cdots,\ (v_T, t_T)\}$$

First, by using the method described in Section 3.1 or 3.2, we can learn a pair of model parameters, κ and r, from the observed diffusion sequence d. Next, by using the method

described in Section 2.2 or 2.4, we obtain the following K sets of simulated diffusion sequences

$$s_k = \{(v_0, t_0),\ (v_{k,1}, t_{k,1}),\ \cdots,\ (v_{k,T}, t_{k,T})\},\ k = 1,\ \cdots,\ K.$$

Here note that the information source node v_0 at time t_0 is the same for all sequences, but their final activation times $\{t_{k,T}\}$ as well as their numbers of activated nodes $\{|s_k|\}$ vary in quite wide range, as shown later in our experiments. Finally, by using the generated sequences $S = \{s_1,\ \cdots,\ s_K\}$, we can estimate the expected influence curve $\sigma(t, v_0)$ as follows:

$$\sigma(t; v_0, d) = \frac{1}{K} \sum_{k=1}^{K} |\{(v, \tau) \in s_k\ ; \tau \leq t\}| \tag{9}$$

This method needs three kinds of input information, i.e., the single observed diffusion sequence d, the topology of observed social network G, and the number of diffusion simulation trials K; then it outputs the expected influence curve $\sigma(t, v_0)$. Below we summarize the estimation algorithm.

step 1 Learn a pair of parameters κ and r from d.
step 2 Generate $S = \{s_1, ..., s_K\}$ by simulating information diffusion K times with the learned parameters κ and r.
step 3 Calculate the expected influence curve $\sigma(t; v_0, d)$ as the average of S.

In our experiments, the number of diffusion simulation trials is set to $K = 100$.

5 Experiments

We evaluate the feasibility of the proposed estimation method using the topologies of two large real network data.

5.1 Evaluation Procedure

Below we describe a procedure to evaluate our proposed method.

proc. 1 Decide information diffusion model: AsIC or AsLT, and choose its true parameters κ^* and r^*, and an information source node v_0 at time t_0.
proc. 2 Generate a set of N diffusion sequences D under the setting of proc. 1.
proc. 3 Calculate the expected influence curve $\sigma(t; v_0)$ from D (by Equation (9) with S replaced by D) and the empirical influence curve $\varphi(t; v_0, d_n)$ from each $d_n \in D$.
proc. 4 Estimate the expected influence curve $\sigma(t; v_0, d_n)$ from each $d_n \in D$ by the proposed method in Section 4.
proc. 5 Calculate the RMSE curves E_C and E_D for evaluation.

In reality it is almost impossible to obtain the actual expected influence curve from observation. Thus our evaluation resorts to experiments based on synthetic data by assuming an information diffusion model, AsIC or AsLT, with a pair of model parameters,

κ^* and r^* which we assume to be true (proc. 1). Then, by performing simulation based on the model with the true parameters, we can prepare a set of N synthetic diffusion sequences denoted by $D = \{d_1, \cdots, d_N\}$ (proc. 2). Next, by applying Equation (9) with respect to D (instead of S), we can obtain a reasonably accurate expected influence curve $\sigma(t; v_0)$ (proc. 3). Here we can also obtain an empirical influence curve for each of the generated sequence d_n defined by $\varphi(t; v_0, d_n) = |\{(v, \tau) \in d_n \,; \tau \leq t\}|$ (proc. 3)[3]. On the other hand, by regarding each of the generated sequence d_n as a single observed diffusion sequence, we can estimate the expected influence curve $\sigma(t; v_0, d_n)$ by our method proposed in Section 4 (proc. 4). Finally, we evaluate the average accuracy of the expected influence curves estimated by our method by means of the RMSE (Root Mean Squered Error) curve $E_C(t)$ and compare it with that of the empirical influence curves denoted by $E_D(t)$. Here these RMSE curves, $E_C(t)$ and $E_D(t)$, are defined as follows.

$$E_C(t) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\sigma(t; v_0, d_n) - \sigma(t; v_0))^2}, \quad E_D(t) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\varphi(t; v_0, d_n) - \sigma(t; v_0))^2}.$$

We can consider that the RMSE curve for $E_D(t)$ corresponds to the average accuracy of the single observed diffusion sequence when we interprete it as a piece of crude knowledge.

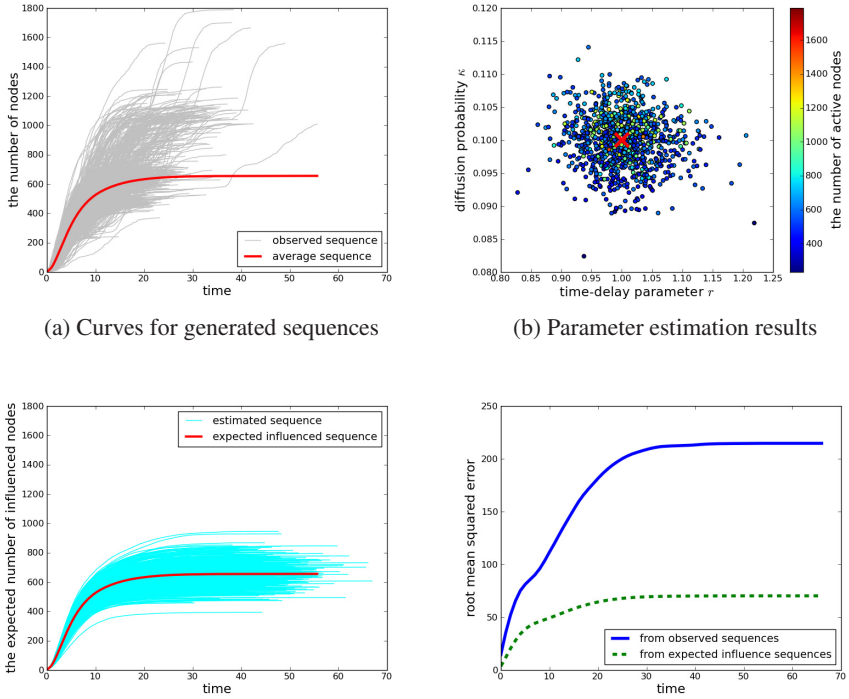
5.2 Experimental Settings

In our experiments, we employed two datasets of large real networks used in [8], which exhibit many of the key features of social networks. The first one is a trackback network of Japanese blogs. The network data were collected by tracing the trackbacks from one blog in the site *goo*[4] 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.

We determined the values of r and κ of the two models which we assumed to be true in the following way. In the AsIC model, we calculated the mean out-degree $\bar{d}$ and set two different values of κ in reference to $1/\bar{d}$, one smaller than $1/\bar{d}$ according to [7] and the other larger than $1/\bar{d}$ to see how a different value affects the result. 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, we decided to set $\kappa = 0.1$ and 0.3 for the blog network and $\kappa = 0.03$ and 0.09 for the Wikipedia network as the true values. As for the time-delay parameter r, we simply decided to set it to 1.0 because changing r is equivalent to changing the time scale accordingly. In the AsLT

[3] Note that d_n is not continuous but $\varphi(t; v_0, d_n)$ is continuous with respect to t.

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



(a) Curves for generated sequences

(b) Parameter estimation results

(c) Curves from estimated parameters

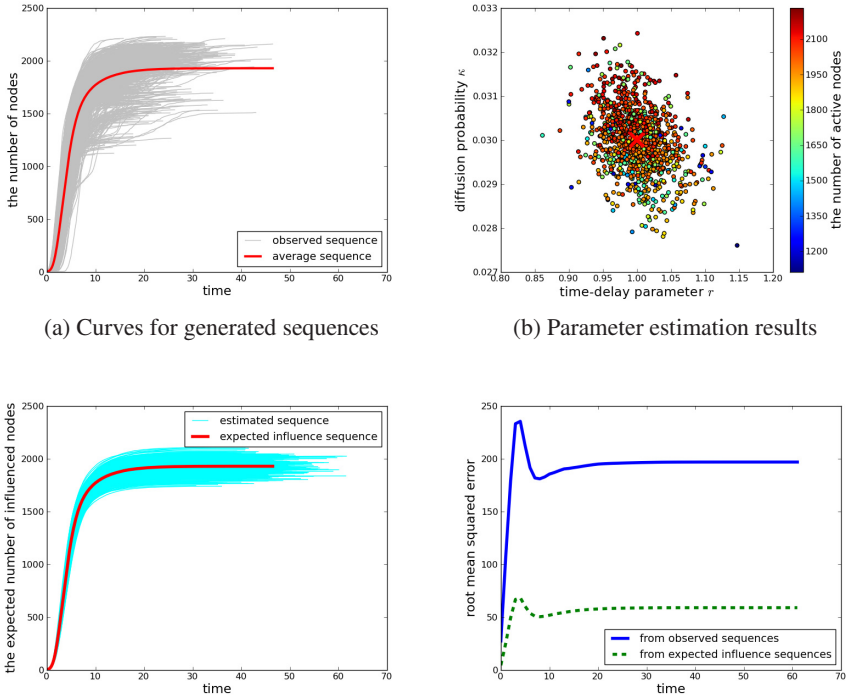
(d) Accuracy comparison

Fig. 1: The result set of blog network under the AsIC model ($\kappa^* = 0.1$)

model, we only chose one value for κ. This is because we found that the information does not reach out far in the AsLT model and we needed to set a large value for κ to realize a decent diffusion. A value of 0.9 was a proper choice for κ. The time-delay parameter was set to $r = 1.0$, same as for the AsIC model.

5.3 Experimental results

blog network under the AsIC model Figure 1 is the results of blog network under the AsIC model for the parameters $\kappa = 0.1$ and $r = 1.0$ (proc. 1). Figure 1(a) plots individual sequence data when the diffusion simulation was repeated $N = 1000$ times starting from the same initial source node (proc. 2). The horizontal axis is the time and the vertical axis is the number of active nodes. As shown in the figure, we observe a wide variety of influence curves with respect to time (dipicted in grey) due to the stochastic nature of the AsIC model. Here our task is to estimate the expected influence curve (depicted in red (black)), which is approximated by the empirical mean of the 1000 gray curves (proc. 3). Figure 1(b) is to show that it is possible to estimate the parameters of the AsIC model, i.e. time-delay parameter r and diffusion probability κ even from a single diffusion sequence. There are 1000 dots and each dot is the esitmated results (r, κ) from the corresponding sequence (proc. 4). We observe that the parameter estimation results



(a) Curves for generated sequences

(b) Parameter estimation results

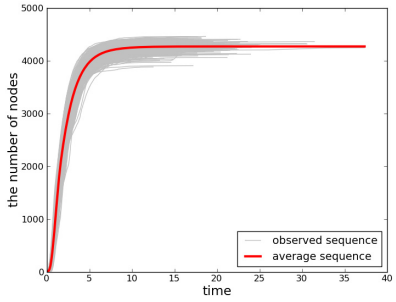
(c) Curves from estimated parameters

(d) Accuracy comparison

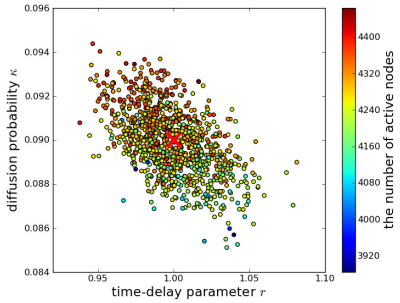
Fig. 2: The result set of Wikipedia network under the AsIC model ($\kappa^* = 0.03$)

are scatterd around the true values $(r^*, \kappa^*) = (1.0, 0.1)$, which were used to generate each sequence. The color (greyness) in the bar on the right indicates the lengh of the sequence, and the results are not very sensitive to the length unless it is very short. Figure 1(c) shows the estimated influence curves (depicted in cyan (grey)), each of which is obtained by performing simulation $K = 100$ times from the correspondng initial source node using the AsIC model with the same parameters learned from the corresponding original diffusion sequence. The target expected influence curve is the same as in Figure 1(a). Figure 1(d) shows the RMSE (Root Mean Squared Error) curves for both the original influence $\varphi(t; v_0, d_n)$ (Figure 1(a)) and the estimated influence $\sigma(t; v_0, d_n)$ (Figure 1(c)) with respect to the target influence (proc. 5). As shown, we observe that the RMSE for the estimated curve is much smaller (less than 1/3) than the one for the original one. Thus, we can say that the estimated influence curve is much closer to the expected influence curve than the original curve. Similar result is obtained for the case of $\kappa^* = 0.3$.

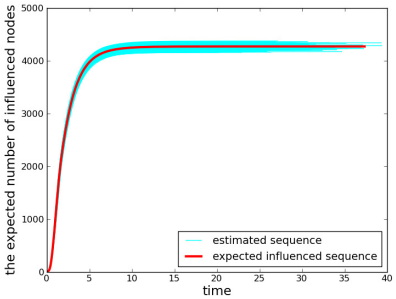
Wikipedia network under the AsIC model Figures 2 and 3 are the results of Wikipedia network under the AsIC model for $\kappa^* = 0.03$ and $\kappa^* = 0.09$, respectively. In both cases, the RMSE for the estimated curve is much smaller (about 1/4 for $\kappa^* = 0.03$ and about



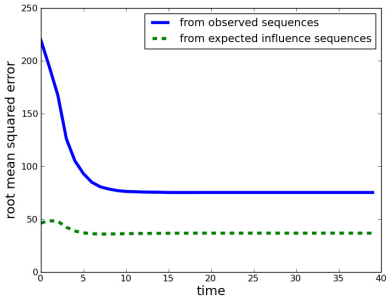
(a) Curves for generated sequences



(b) Parameter estimation results



(c) Curves from estimated parameters

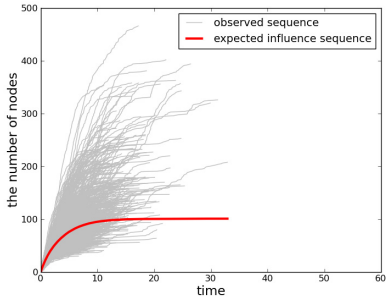


(d) Accuracy comparison

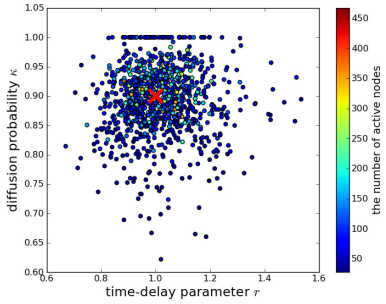
Fig. 3: The result set of Wikipedia network under the AsIC model ($\kappa^* = 0.09$)

1/2 for $\kappa^* = 0.09$) in the proposed method. The results for $\kappa^* = 0.03$ is similar to the results of blog network except that the shape of the RMSE curve is different. However, the results for $\kappa^* = 0.09$ reveal different behaviors. When the diffusion probability is large, the information propagates far enough and individual sequence becomes similar to each other. Note that the number of nodes is almost doubled. The accuracy becomes better accordingly, especially for the original influence $\varphi(t; v_0, d_n)$. In general the proposed method is more effective when the diffusion probability is small and the observation sequences are diversified.

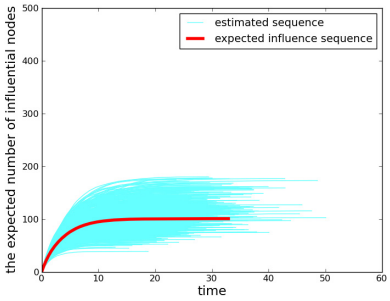
blog network under the AsLT model Figure 4 shows the results of blog network under the AsLT model for $\kappa^* = 0.9$. Unlike the AsIC model, the information does not spread far and wide and the sequences are short. Accordingly the number of active nodes are much smaller (less than 500) and the errors in the parameter estimation are larger than the AsIC model. But still, we can say that the parameters are estimated reasonably well and the RMSE is much smaller (about 1/3) in the proposed method. Similar results are obtained for Wikipedia network.



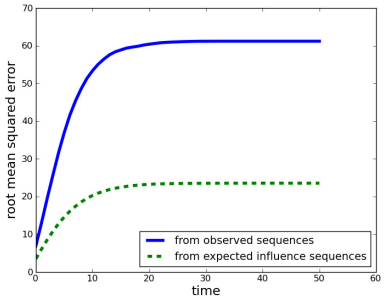
(a) Curves for generated sequences



(b) Parameter estimation results



(c) Curves from estimated parameters



(d) Accuracy comparison

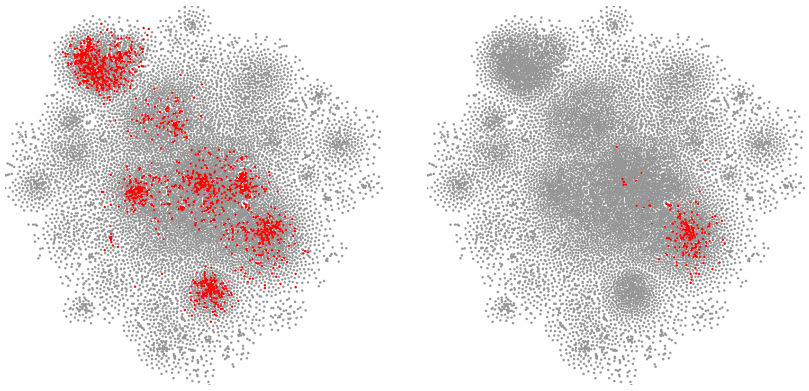
Fig. 4: The result set of blog network under the AsLT model ($\kappa^* = 0.9$)

5.4 Visual Analyses

We saw that observation sequences are diverse in general due to the stochastic nature of the diffusion process. The differences in diffusion patterns are best understood by visualizing the active nodes. Figure 5 visualizes two extreme diffusion patterns for blog network of Figure 2 by using Cross-entropy method [15]. The red dots indicate active nodes and the gray dots non-active nodes. Figure 5(a) is the pattern for the longest sequence and Figure 5(b) is the one for the shortest sequence. We observe that dots are not uniformly distributed but have some dense reagions forming communities. In Figure 5(a) the information diffuses across many communities and spread widely, whereas in Figure 5(b) it is trapped within the same community of the initial source node and does not spread. Consequently, the number of active nodes in Figure 5(a) is 1,789 and that in Figure 5(b) is only 220. Simialr result is also observed in Wikipedia network.

6 Discussion

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



(a) Visualization of high diffusion result (b) Visualization of low diffusion result

Fig. 5: Visualization of blog network

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. In this setting we learn κ_n and r_n for $n = 1, 2, ..., N$ from a single observation sequence.

We aimed to estimate the expected influence curve assuming two different information diffusion models in this paper but the framework of the proposed method can be applied to other models as well as other measures. For example, if we are interested in how different opinions spread [16], we can use the Voter model and estimate the expected opiniton share curve under this framework. Which measure and model to use depends on the problem we want to solve and the evaluation must be based on a task-specific performance measure.

7 Conclusion

One of the challenges of social network analysis is to estimate the expected influence degree with respect to time (expected influence curve). Because of the stochastic nature of information diffusion, a single observation sequence is not reliable to use as an approximation of this curve. We proposed a novel method to estimate the expected influence curve with good accuracy from a single observed information diffusion sequence assuming two types of information diffusion models: the asynchronous independent cascade (AsIC) model and the asynchronous linear threshold (AsLT). The method first learns the model parameters from a single observation sequence and next use the learned model to estimate the expected influence curve. We showed that parameter learning from a single sequence is feasible and practical, and the estimated influence curve is much more accurate than using the observed sequence as its approximation by extensive experiments using two real world networks.

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-10-4053, 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. Kimura, M., Saito, K., 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. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Behavioral analyses of information diffusion models by observed data of social network. In: Proceedings of the the 2010 International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP 2010). (2010) 149–158
15. Yamada, T., Saito, K., Ueda, N.: Cross-entropy directed embedding of network data. In: Proceedings of the 20th International Conference on Machine Learning (ICML03). (2003) 832–839
16. Kimura, M., Saito, K., Motoda, H., Ohara, K.: Learning to predict opinion share in social networks. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-10). (2010)