Burst Detection in a Sequence of Tweets based on Information Diffusion Model

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

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

Abstract. We propose a method of detecting the period in which a burst of information diffusion took place from an observed diffusion sequence data over a social network and report the results obtained by applying it to the real Twitter data. We assume a generic information diffusion model in which time delay associated with the diffusion follows the exponential distribution and the burst is directly reflected to the changes in the time delay parameter of the distribution (inverse of the average time delay). The shape of the parameter change is approximated by a series of step functions and the problem of detecting the change points and finding the values of the parameter is formulated as an optimization problem of maximizing the likelihood of generating the observed diffusion sequence. Time complexity of the search is almost proportional to the number of observed data points (possible change points) and very efficient. We apply the method to the real Twitter data of the 2011 To-hoku earthquake and tsunami, and show that the proposed method is by far efficient than a naive method that adopts exhaustive search, and more accurate than a simple greedy method. Two interesting discoveries are that a burst period between two change points detected by the proposed method tends to contain massive homogeneous tweets on a specific topic even if the observed diffusion sequence consists of heterogeneous tweets on various topics, and that assuming the information diffusion path is a line shape tree can give a good approximation of the maximum likelihood estimator when the actual diffusion path is not known.

1 Introduction

Recent technological innovation and popularization of high performance mobile/smart phones has changed our communication style drastically and the use of various social media such as Twitter and Facebook has been affecting our daily lives substantially. In

these social media, information propagates through the social network formed based on friendship relations. Especially, Twitter, micro-blog in which the number of characters is limited to 140, is now very popular among the young generation due to its handiness and easiness of usage, and it is fresh to our memory that Twitter played a very important role as the information infrastructure during the recent natural disaster, both domestic and abroad, including the 2011 To-hoku earthquake and tsunami in Japan.

In these social networks, there have been proposed several measures, called centrality, that characterize nodes in the network based on the structure of the network [11, 1, 3]. While such centrality measures can be used to identify those nodes that play an important role in diffusing information over the network, it has also been shown that measures based solely on the network structure are not good enough to a such problem of influence maximization [4, 5] in which the task is to identify a limited number of nodes which together maximize the information spread and that explicit use of information diffusion mechanism is essential [5]. In general, the mechanism is represented by a probabilistic diffusion model. Most representative and basic ones are the Independent Cascade (IC) model [2, 4] and the Linear Threshold (LT) model [12, 13] including their extended versions that explicitly handle asynchronous time delay, Asynchronous time delay Independent Cascade (AsIC) model [8] and Asynchronous time delay Linear Threshold (AsLT) model [9]. In fact, the nodes and links that are identified to be influential using these models are substantially different from those identified by the existing centrality measures.

In reality, we observe that the information on a certain topic propagates explosively for a very short period of time. Because such information affects our behaviour strongly, it is important to understand the observed event in a timely manner. This brings in an important and interesting problem which is to accurately and efficiently detect the burst from the observed information diffusion data and to identify what caused this burst and how long it persisted. Any of the above mentioned probabilistic models cannot handle this kind of problem because they assume that information diffuses in a stationary environment, i.e. model parameters are stationary. Zhu and Shasha [14] approched this problem without relying on a diffusion model. They detected a burst period for a target event by counting the number of its occurences in a given time window and cheking whether it exceeds a predetermined threshold or not. Kleinberg [6] challenged this problem using a hidden Markov model in which bursts appear naturally as state transitions, and successfully identified the hierarchical structure of e-mail messages. Sun et al. [10] extended Kleinberg's method so as to detect correlated burst patterns from multiple data streams that co-evolve over time.

We handle this problem by assuming that parameters in the diffusion model changed due to unknown external environmental factors and devise an efficient algorithm that accurately detects the changes in the parameter values from a single observed diffusion data sequence. In particular we note that the parameter related to the time delay is most crucial in the burst detection and focus on detecting the changes in the time delay parameter that defines the delay distribution. We modeled the time delay in AsIC and AsLT models by the exponential distribution, thus we do the same in this paper. This corresponds to associating the burst with the information diffusion with a shorter time

delay. By focusing only on this time delay, we can devise a generic algorithm that does not depend on a specific information diffusion model, e.g. be it either AsIC or AsLT.

More precisely, we assume that time delay parameter changes are approximated by a series of step functions and propose an optimization algorithm that maximizes the likelihood ratio that is the ratio of the likelihood of observing the data assuming the time delay parameter changes (change points and parameter values between the successive change points) to the likelihood of observing the data assuming that there is no changes in the time delay parameter. The algorithm is based on iterative search based on recursive splitting with delayed backtracking, and requires no predetermined threshold. The time complexity is almost proportional to the number of observed data points (candidates of possible change points). We apply the method to the Twitter data observed during the 2011 To-hoku earthquake and tsunami and confirm that the proposed method can efficiently and accurately detect the change points. We further analyze the content of the tweets and report the discovery that even use of the diffusion sequence data of the same user ID (not necessarily the data on a specific topic) allows us to identify that a specific topic is talked intensively around the beginning of the period where the burst is detected, and the assumption we made that the information diffusion path is a line shape tree gives a good approximation of the maximum likelihood estimator in this problem setting. Finally, we discuss that although the detected change points do not correspond exactly to nodes in a social network that caused the burst period, the detected change points are useful to find such nodes because we can limit nodes to be considered by focusing on those around them.

The paper is organized as follows. Section 2 briefly describes the framework of information diffusion model on which our problem setting is based. Section 3 elucidates the problem setting, and Section 4 describes the change point detection method including two other methods that are used for comparison. Section 5 reports experimental results using real Twitter data. Section 6 summarizes what has been achieved in this work and addresses the future work.

2 Information Diffusion Model Framework

We consider information diffusion over a social network whose structure is defined as a directed graph $G = (V, E)$, where V and E ($\subset V \times V$) represent a set of all nodes and a set of all links, respectively. Suppose that we observe a sequence of information diffusion $C = \{(v_0, t_0), (v_1, t_1), \cdots, (v_N, t_N)\}$ that arose from the information released at the source node v_0 at time t_0. Here, v_n is a node where the information has been propagated and t_n is its time. We assume that the time points are ordered such that $t_{n-1} < t_n$ for any $n \in \{1, \cdots N\}$. We further assume, as a standard setting, that the actual information diffusion paths of a sequence C correspond to a tree that is embedded in the directed graph G representing the social network[7], i.e., the parent node which passed the information to a node v_n is uniquely identified to be $v_{p(n)}$. Here, $p(n)$ is a function that returns the node identification number of the parent of the node v_n in the range of $\{0, \cdots, n-1\}$.

The information diffusion model we consider here is any model that explicitly incorporates the concept of asynchronous time delay such as AsIC model [8] and AsLT

model [9] in contrast to the traditional IC model [2, 4] and LT model [12, 13] that do not consider the time delay. Said differently, it is a model that allows any real value for the time t_n at which the information has been propagated to a node v_n and assumes a certain probability distribution for the time delay $t_n - t_{p(n)}$. In this paper, we use the exponential distribution for the time delay, but any other distribution such as power law is feasible exactly in the same way.

3 Problem Settings

In this section we formally define the change point detection problem. As mentioned in Section 1, we assume that some unknown change took place in the course of information diffusion and what we observe is a sequence of information diffusion of some topic in which the change is encapsulated. Thus, our goal is to detect each change point and how long the change persisted from there. Note that we basically pay attention to a diffusion sequence of a certain topic. From our previous result that people's behaviors are quite similar when talking the same topic [8, 9], we can assume that the time delay parameter $r_{u,v}$ which is in principle defined for each link $(u, v) \in E$ takes a uniform value regardless of the link it passes through. In other word, we set $r_{u,v} = r\ (\forall (u, v) \in E)$ and thus, the time delay of information diffusion is represented by the following simple exponential distribution $p(t_n - t_{p(n)}; r) = r \exp(-r(t_n - t_{p(n)}))$.

With this preparation, we mathematically define the change point detection problem. Let's assume that we observe a set of time points of information diffusion sequence $\mathcal{D} = \{t_0, t_1, \cdots, t_N\}$. Let the time of the j-th change point be T_j $(t_0 < T_j < t_N)$. The delay parameter that the distribution follows switches from r_j to r_{j+1} at the j-th change point T_j. Namely, we are assuming a series of step functions as a shape of parameter changes. Let the set comprising J change points be $\mathcal{S}_J = \{T_1, \cdots, T_J\}$, and we set $T_0 = t_0$ and $T_{J+1} = t_N$ for the sake of convenience $(T_{j-1} < T_j)$. Let the division of $\mathcal{D}$ by $\mathcal{S}_J$ be $\mathcal{D}_j = \{t_n; T_{j-1} < t_n \leq T_j\}$, i.e., $\mathcal{D} = \{t_0\} \cup \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{J+1}$, and $|\mathcal{D}_j|$ represent the number of observed points in $(T_{j-1}, T_j]$. Here, we request that $|\mathcal{D}_j| \neq 0$ for any $j \in \{1, \cdots, J+1\}$ and there exists at least one t_n and $t_n \in \mathcal{D}_j$ is satisfied.

The log-likelihood for the $\mathcal{D}$, given a set of change points $\mathcal{S}_J$, is calculated, by defining the parameter vector $\mathbf{r}_{J+1} = (r_1, \cdots, r_{J+1})$, as follows.

$$\begin{aligned} L(\mathcal{D}; \mathbf{r}_{J+1}, \mathcal{S}_J) &= \log \prod_{j=1}^{J+1} \prod_{t_n \in \mathcal{D}_j} r_j \exp(-r_j(t_n - t_{p(n)})) \\ &= \sum_{j=1}^{J+1} |\mathcal{D}_j| \log r_j - \sum_{j=1}^{J+1} r_j \sum_{t_n \in \mathcal{D}_j} (t_n - t_{p(n)}). \end{aligned} \tag{1}$$

Thus, the maximum likelihood estimate of the parameter of Equation (1) is given by

$$\hat{r}_j^{-1} = \frac{1}{|\mathcal{D}_j|} \sum_{t_n \in \mathcal{D}_j} (t_n - t_{p(n)}), \quad j = 1, \cdots, J+1. \tag{2}$$

Further, substituting Equation (2) to Equation (1) leads to

$$L(\mathcal{D}; \hat{\mathbf{r}}_{J+1}, \mathcal{S}_J) = -N - \sum_{j=1}^{J+1} |\mathcal{D}_j| \log\left(\frac{1}{|\mathcal{D}_j|} \sum_{t_n \in \mathcal{D}_j} (t_n - t_{p(n)})\right). \tag{3}$$

Therefore, the change point detection problem is reduced to the problem of finding the change point set $\mathcal{S}_J$ that maximizes Equation (3). However, Equation (3) alone does not allow us to directly evaluate the effect of introducing $\mathcal{S}_j$. We, thus, reformulate the problem as the maximization problem of log-likelihood ratio. If we do not assume any change point, i.e., $\mathcal{S}_0 = \emptyset$, Equation (3) is reduced to

$$L(\mathcal{D}; \hat{r}_1, \mathcal{S}_0) = -N - N \log\left(\frac{1}{N} \sum_{n=1}^{N} (t_n - t_{p(n)})\right). \tag{4}$$

Thus, the log-likelihood ratio of the case where we assume J change points and the case where we assume no change points is given by

$$\begin{aligned} LR(\mathcal{S}_J) &= L(\mathcal{D}; \hat{\mathbf{r}}_{J+1}, \mathcal{S}_J) - L(\mathcal{D}; \hat{r}_1, \mathcal{S}_0) \\ &= N \log\left(\frac{1}{N} \sum_{n=1}^{N} (t_n - t_{p(n)})\right) - \sum_{j=1}^{J+1} |\mathcal{D}_j| \log\left(\frac{1}{|\mathcal{D}_j|} \sum_{t_n \in \mathcal{D}_j} (t_n - t_{p(n)})\right). \end{aligned} \tag{5}$$

We consider the problem of finding the set of change points $\mathcal{S}_J$ that maximizes $LR(\mathcal{S}_J)$ defined by Equation (5).

We note that, in general, it is conceivable that we are not able to acquire the complete tree structure of the diffusion sequence data. Thus, here, we consider two extreme cases, one in which the information spreads fastest (star shape tree) and the other in which the information spread slowest (line shape tree). The function which defines the parent node becomes $p(n) = 0$ for the former and $p(n) = n - 1$ for the latter. In case where there is no change point, the maximum likelihood estimator is $r^{-1} = (t_1 + \cdots + t_N)/N - t_0$ for the former and $r^{-1} = (t_N - t_0)/N$ for the latter. While we conjecture that in reality the optimal value lies in between these two extreme values, under the assumption that the actual tree structure of the diffusion data is unknown, we consider to approximate the optimal value by using either one of them. Here, note that in the former case, the maximum likelihood estimator represents the average diffusion delay time between the source node v_0 and each node v_i which is assumed to be connected to v_0 by a direct link, while in the latter case, it represents the average time interval between successive observation time points. Considering that the burst period we want to detect is much shorter than the other non burst periods, the latter case (line shape tree) seems to be more suitable for our aim. Therefore, $LR(\mathcal{S}_J)$ defined by Equation (5) becomes

$$LR(\mathcal{S}_J) = N \log\left(\frac{t_n - t_0}{N}\right) - \sum_{j=1}^{J+1} |\mathcal{D}_j| \log\left(\frac{T_j - T_{j-1}}{|\mathcal{D}_j|}\right). \tag{6}$$

We compared the bursts detected by using the two extreme values, and found that the use of line shape tree gave a better results and decided to use Equation (6) in our experiments.

4 Change Points Detection Method

We consider the problem of detecting change points as a problem of finding a subset $\mathcal{S}_J \subset \mathcal{D}$ when the set of time points of information diffusion result $\mathcal{D} = \{t_0, t_1, \cdots, t_N\}$ and the number of change points J are given. In other words, we search for J time points that are most likely to be the change points from a sequence of N observation points. In what follows, we explain each of the three methods, naive method (an exhaustive search), simple method (a greedy search), and the proposed method that is a combination of a greedy search and a local search.

4.1 Naive Method

The simplest method is to exhaustively search for the best set of J change points $\mathcal{S}_J$. Clearly the time complexity of this naive approach is $O(N^J)$. Thus, the number of change points detectable would be limited to $J = 2$ in order for the solution to be obtained in a reasonable amount of computation time when N is large enough.

4.2 Simple Method

We describe the simple method which is applicable when the number of change points J is large. This is a progressive binary splitting without backtracking. We fix the already selected set of $(j-1)$ change points $\mathcal{S}_{j-1}$ and search for the optimal j-th change point T_j and add it to $\mathcal{S}_{j-1}$. We repeat this procedure from $j = 1$ to J.

The algorithm is given below.

Step1. Initialize $j = 1$, $\mathcal{S}_0 = \emptyset$.
Step2. Search for $T_j = \arg\max_{t_n \in \mathcal{D}} \{LR(\mathcal{S}_{j-1} \cup \{t_n\})\}$.
Step3. Update $\mathcal{S}_j = \mathcal{S}_{j-1} \cup \{T_j\}$.
Step4. If $j = J$, output $\mathcal{S}_J$ and stop.
Step5. $j = j + 1$, and return to Step2.

Here note that in Step3 elements of the change point set $\mathcal{S}_j$ are reindexed to satisfy $T_{i-1} < T_i$ for $i = 2, \cdots, j$. Clearly, the time complexity of the simple method is $O(NJ)$ which is fast. Thus, it is possible to obtain the result within a allowable computation time for a large N. However, since this is a greedy algorithm, it can be trapped easily to a poor local optimal.

4.3 Proposed Method

We propose a method which is computationally almost equivalent to the simple method but gives a solution of much better quality. We start with the solution obtained by the simple method $\mathcal{S}_J$, pick up a change point T_j from the already selected points, fix the rest $\mathcal{S}_J \setminus \{T_j\}$ and search for the better value T'_j of T_j, where $\cdot \setminus \cdot$ represents set difference. We repeat this from $j = 1$ to J. If no replacement is possible for all j ($j = 1, \cdots J$), i.e. $T'_j = T_j$ for all j, no better solution is expected and the iteration stops.

The algorithm is given below.

Step1. Find $\mathcal{S}_J$ by the simple method and initialize $j = 1$, $k = 0$.
Step2. Search for $T'_j = \arg\max_{t_n \in \mathcal{D}} \{LR(\mathcal{S}_J \setminus \{T_j\} \cup \{t_n\})\}$.
Step3. If $T'_j = T_j$, set $k = k + 1$, otherwise set $k = 0$, and update $\mathcal{S}_J = \mathcal{S}_J \setminus \{T_j\} \cup \{T'_j\}$.
Step4. If $k = J$, output $\mathcal{S}_J$ and stop.
Step5. If $j = J$, set $j = 1$, otherwise set $j = j + 1$, and return to Step2.

It is evident that the proposed method requires computation time several times larger than that of the simple method, but it is much less than that of the naive method. How much the computation time increases compared to the simple method and how much the solution quality increases await for the experimental evaluation, which we will report in Section 5.

5 Experimental Evaluation

We experimentally evaluate the computation time and the accuracy of the change point detection using the real world Twitter information diffusion sequence data based on the methods we described in the previous section. We, then, analyze in depth the top 6 diffusion sequences in terms of the log-likelihood ratio based on the detected change points and burst periods, show that the line shape tree approximation is much better than the star shape tree approximation, and investigate whether or not we are able to identify which node in a social network caused the burst from the detected change points.

5.1 Experimental Settings

The information diffusion data we used for evaluation are extracted from 201,297,161 tweets of 1,088,040 Twitter users who tweeted at least 200 times during the three weeks from March 5 to 24, 2011 that includes March 11, the day of 2011 To-hoku earthquake and tsunami. It is conceivable to use a retweet sequence in which a user sends out other user's tweet without any modification. But there exist multiple styles of retweeting (official retweet and unofficial retweet), and it is very difficult to accurately extract a sequence of tweets in an automatic manner considering all of these different styles. Therefore, in our experiments, noting that each retweet includes the ID of the user who sent out the original tweet in the form of "@ID", we extracted tweets that include @ID format of each user ID and constructed a sequence data for each user. More precisely, we used information diffusion sequences of 798 users for which the length of sequences are more than 5,000 (number of tweets). Note that each diffusion sequence includes retweet sequences on multiple topics. Since we do not know the ground truth of the change points for each sequence if there are changes in it, we used the naive method which exhaustively search for all the possible combinations of the change points as giving the ground truth. We had to limit the number of change points to 2 ($J = 2$) in order for the naive method to return the solution in a reasonable amount of computation time. The experimental results explained in the next subsection is obtained by using a machine with Intel(R) Xeon(R) CPU W5590 @3.33GHz and 32GB memory.

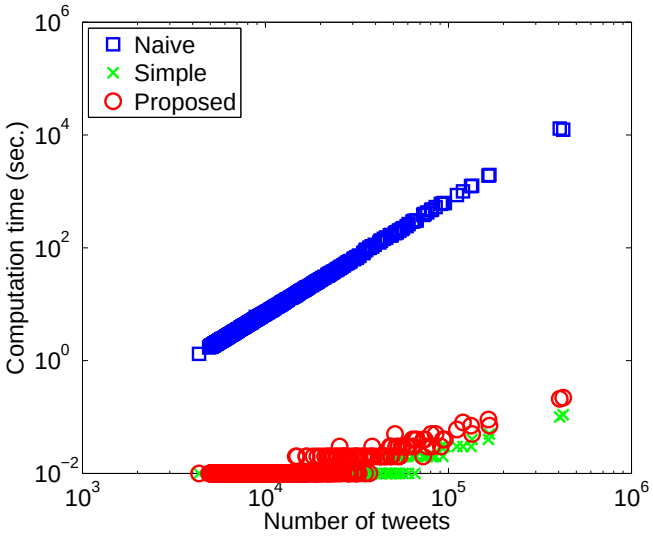


Fig. 1: Comparison of computation time among the three (naive, simple, and proposed) methods.

5.2 Main Results

Performance Evaluation Figure 1 shows the computation time that each method needed to produce the results. The horizontal axis is the length of the information diffusion data sequences, and the vertical axis is the computation time in second. The results clearly indicate that the naive method requires the largest computation time. The computation time is quadratic to the sequence length as predicted. In contrast, the computation time for the simple and the proposed methods is much shorter and it increases almost linearly to the increase of the sequence length for both. The proposed method requires more computation time due to the extra iteration needed for delayed backtracking. In fact, the number of extra iteration is 2.2 on the average and 7 at most.

Figure 2 shows the accuracy of the detected change points. We regarded that the solution obtained by the naive method is the ground truth. The horizontal axis is the sequence ranking of the log-likelihood ratio for the naive method (ranked from the top to the last), and the vertical axis is the logarithm of the likelihood ratio of the solution of each method. The results indicate that the simple method has lower likelihood ratio for all the range, meaning that it detects change points which are different from the optimal ones, but the proposed method can detect the correct optimal change points except for the low ranked sequences for which the likelihood ratio is small as is evident from the result in that the red curve representing the proposed method is indistinguishable from the blue curve representing the naive method. The reason why the accuracy of the proposed method for sequences with low likelihood decreases may be because the burst period is not clear for these sequences. In summary, out of the 798 sequences in total, the proposed method gave the correct results for 713 sequences (98.4%), whereas the simple method gave the correct results for only 171 sequences (21.4%). The average ratio of the likelihood ratio of the proposed method to that of the naive method (optimal

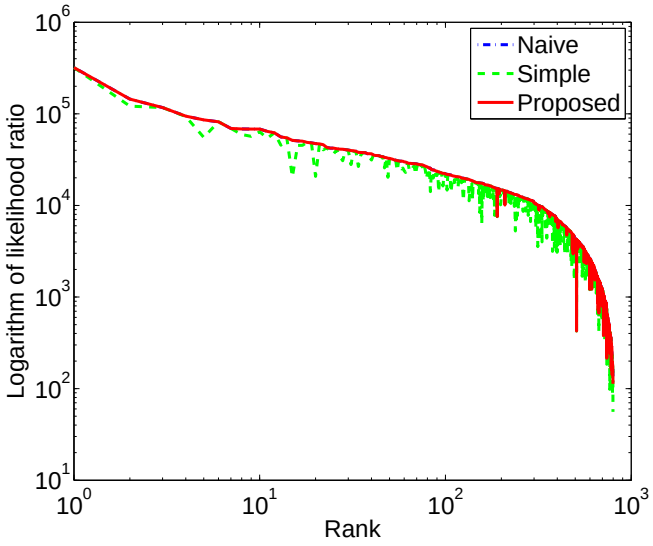


Fig. 2: Comparison of accuracy among the three (naive, simple, and proposed) methods.

solution) is 0.976, whereas the corresponding ratio for the simple method is 0.881, revealing that the proposed method gives much closer ratio to the optimal likelihood ratio. These results confirm that the proposed method can increase the change point detection accuracy to a great extent compared to the simple method with only a small penalty for the increased computation time.

In Depth Analyses on Detected Change Points and Burst Periods Next, we had a closer look at the top 6 diffusion sequences in terms of the log-likelihood ratios. Table 1 shows the total number of tweets included in the sequence, the starting and the ending time of the burst period, and the main topics that appeared near the beginning of the burst. Figure 3 shows how the cumulative number of tweets increases as time goes for each diffusion sequence. The horizontal axis is time and the vertical axis is the cumulative number of tweets. The two red vertical lines in the graph are the change (starting and ending) points detected by the proposed method, and the interval between them is the burst period.

As is understood from Table 1, explosive retweeting of the information of urgent need about the earthquake for a short period of time triggered the start of the burst (with the exception of the 4th ranked sequence). The 4th ranked sequence is for the account called "ordinary timeline" which was set up for allowing to tweet everyday topics by adding "@itsumonoTL" at the beginning of the tweet when people are in voluntary restraint mood after the disastrous earthquake. We can say, with the exception

[1] NHK is the government operated broadcaster.

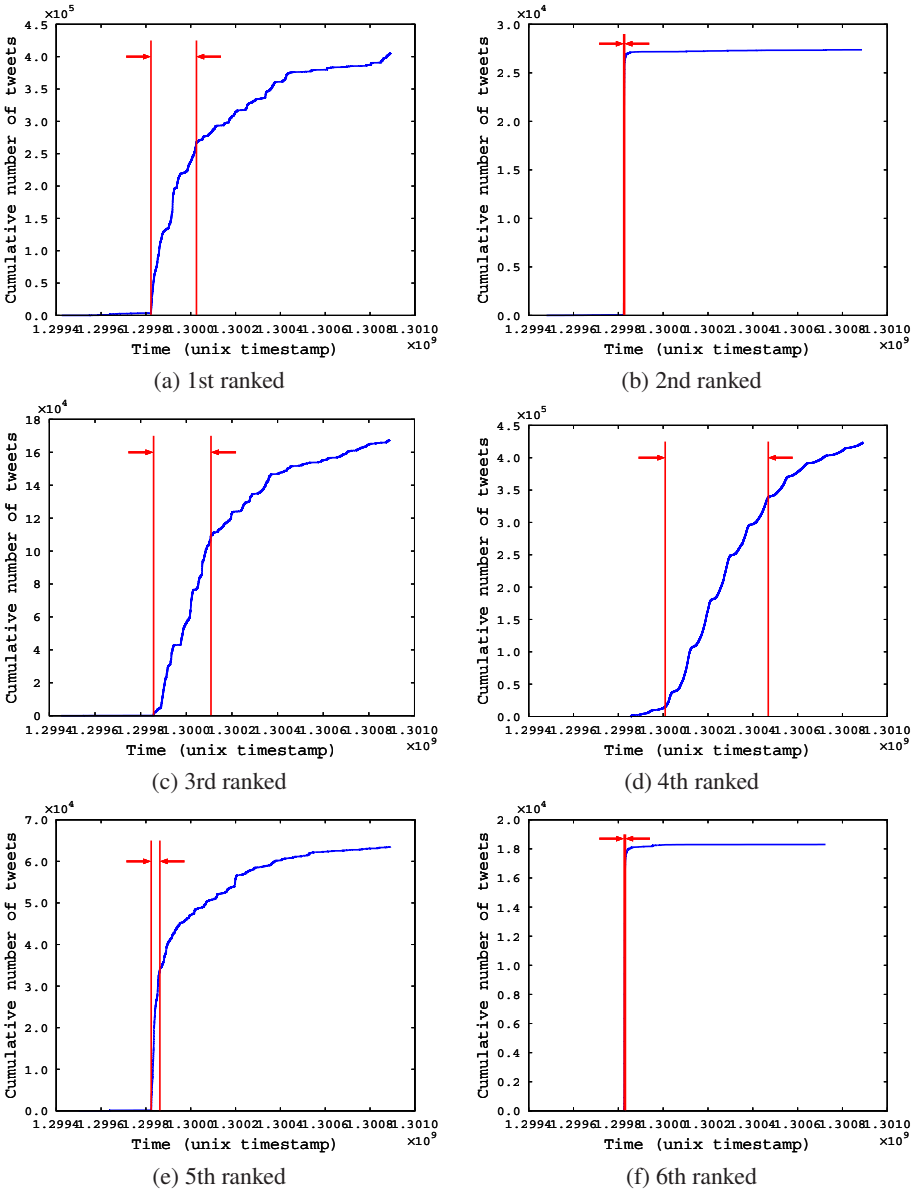
[2] Great Hanshin-Awaji Earthquake occurred on January 17, 1995 in Kobe area and 6,434 people lost their lives.

Table 1: Major topics appearing at the beginning of the burst periods of the top 6 diffusion results in terms of log-likelihood ratio

Ranking	Length	Detected burst period		Major topics at the beginning of the burst period
		Start	End	
1	450,739	2011/3/11 14:48:13	2011/3/13 23:13:04	Retweets of the earthquake bulletin posted by the PR department of Japan Broadcasting Corporation, NHK (@NHK_PR).[1]
2	27,372	2011/3/11 15:13:57	2011/3/11 16:19:26	Retweets of the article on to-do list at the time of earthquake onset posted by a victim of the Great Hanshin-Awaji Earthquake. [2]
3	167,528	2011/3/12 00:18:19	2011/3/14 22:08:20	Retweets of the article on measures against cold at an evacuation site posted by the news department of NHK (@nhk_seikatsu).
4	423,594	2011/3/13 18:38:50	2011/3/19 02:20:58	Ordinary tweets irrelevant to the earthquake posted to a special account "@itsumonoTL".
5	63,485	2011/03/11 15:05:08	2011/03/12 01:52:13	Retweets of the earthquake bulletin posted by the Fire and Disaster Management Agency (@FDMA_JAPAN).
6	18,299	2011/3/11 15:45:17	2011/3/11 17:19:02	Retweets of a call for help posted by a user who seemed to be buried under a server rack (later found to be a false rumor).

of such a special case of "ordinary timeline", that we are able to detect efficiently a time period where tweets on a specific topic (of urgent need in this example) are intensively retweeted by looking at the change points detected by the proposed method even from the diffusion sequence that contains multiple topics.

We note that the cumulative number of the tweets for the 2nd and 6th ranked diffusion sequences is smaller than the other 4 sequences from Table 1, and the burst period of these 2 sequences are much shorter than others and there is little changes in the number of tweets before and after the burst from Figure 3. This difference is considered to come from whether the account is private or public. Among these 4 sequences, except for the exceptional 4th one, the remaining 3 are all from the public organization accounts (1st and 3rd are NHK and 5th is FDMA). Information posted by these accounts tends to disseminate widely everyday. Thus, considering this situation, it is natural to observe that the cumulative number of tweets shows a relatively smooth increase as seen in Figure 3 by adding multiple bursts of short periods about the earthquake-related information of urgent need as shown in Table 1. Figure 3(e) has only one smooth change during the burst period, which indicates that the earthquake bulletin in Table 1 is the only source of the burst. On the other hand, we see multiple smooth changes with discontinuity of the gradient at each boundary during the burst period in Figures 3(a) and (c). This implies that there can be other sources of the burst than shown in Table 1. Indeed, it is possible to identify these change points by increasing the value of J (an example explained later). On the other hand, Figures 3(b) and (f) shows that the information posted by an individual that is rarely retweeted in ordinary situations can be propagated explosively if it is of urgent need, e.g. timely information about earthquake.



(a) 1st ranked (b) 2nd ranked

(c) 3rd ranked (d) 4th ranked

(e) 5th ranked (f) 6th ranked

Fig. 3: Temporal change of cumulative number of tweets in the top 6 diffusion results in terms of the highest log-likelihood ratio

Here, we report the result when we increase the number of change points. Figure 4 shows the result for the 3rd ranked sequence in Figure 3(c) when J is set to 9. There are 9 vertical lines corresponding to each change point, but the first two change points

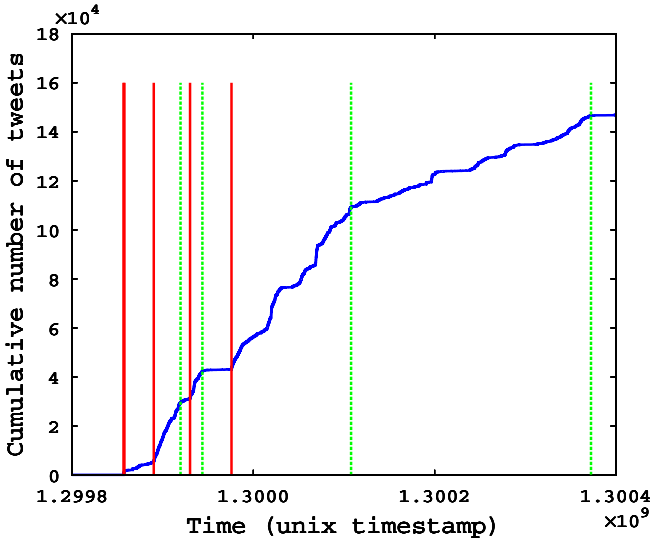
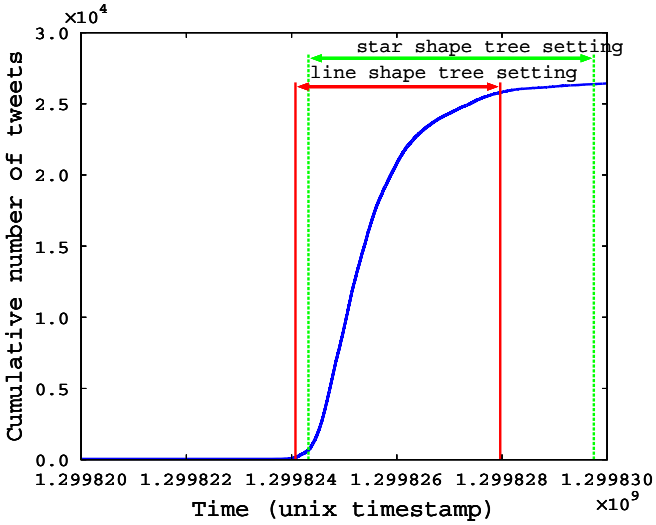


Fig. 4: Finer burst detection for the 3rd ranked sequence in Figure 4(c) when J is set to 9

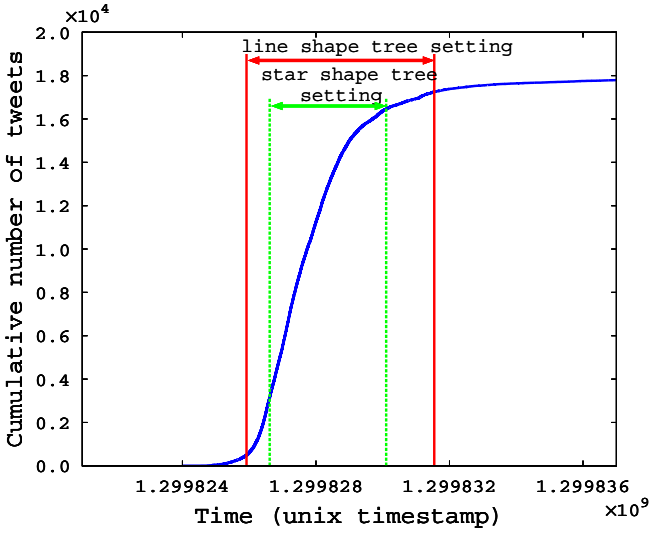
are too close and indistinguishable. Note that horizontal axis is enlarged and the range shown is different from that in Figure 3(c). We see that the detected change points are located at the boundary points where the gradients of the curves change discontinuously. Those 4 broken lines in green are considered to indicate the end of the burst because the gradient change across each boundary is rather smaller. In fact, we investigated the most recent 10 tweets for these 4 change points and confirmed that no more than half of the retweets is talking about the same topic except the one second from the last in which 7 of them are on the same topic. The remaining 5 change points (red lines) all contain at least 7 retweets (10, 8, 7, 7, 9) that are on the same topic. From this fact, we can reconfirm that there appear many tweets on the same topic during the burst period.

Line Shape Tree vs. Star Shape Tree Note that all of these results were obtained by assuming that the information diffuses along the line shape tree as discussed in Section 3. Here, we show that use of line shape tree gives better results than use of star shaped tree. To this end, we compared the bursts detected for the 2nd and 6th ranked information diffusion sequences which include only one burst.

The results are illustrated in Figure 5, where red solid and green broken vertical lines denote the change points detected by the naive method with the line shape and star shape settings, respectively. Only the time range of interest is extracted and shown in the horizontal axis. From these figures, we observe that use of line shape tree detects the change points more precisely as expected, which means that line shape tree gives a better approximation of the maximum likelihood estimator than star shape tree even if the actual tree shape of the diffusion path is not known to us.



(a) 2nd ranked



(b) 6th ranked

Fig. 5: Comparison of bursts detected by use of line shape tree and star shape tree for the 2nd and 6th ranked information diffusion sequences in Table 1.

Change Points in a Time Line and Nodes in a Network Remember that each observed time point corresponds to a node in a social network. In this sense, it can be said that the proposed method detects not only the change points in a time line, but also the change points in a network. However, unfortunately, those nodes do not necessarily correspond to those which actually caused the burst period. For example, in the second

ranked sequence in Table 1, we observed at least 1 retweet of the article described in Table 1 per second after the start of the burst, 2011/3/11 15:13:57, while we observed at most 20 per minute before the burst started. This shows the accuracy of the detected change point, but it also means that the node that actually influenced nodes within the burst period could exist in the period before the change point. Indeed, we observed the first retweet at 2011/3/11 15:07:05 and 69 retweets thereafter before the change point. It is natural to think that some of them played an important role on the explosive diffusion of the article. We need to know the actual information diffusion path to find such important nodes, but detecting change points in a time line would significantly reduce the effort needed to do so because the search can be focused on the limited sub-sequences around the change points. Devising a method to find such important nodes is one of our future work.

6 Conclusion

We addressed the problem of detecting the period in which information diffusion burst occurs from a single observed diffusion sequence under the assumption that the delay of the information propagation over a social network follows the exponential distribution. To be more precise, we formulated the problem of detecting the change points and finding the values of the time delay parameter in the exponential distribution as an optimization problem of maximizing the likelihood of generating the observed diffusion sequence. We devised an efficient iterative search algorithm for the change point detection whose time complexity is almost linear to the number of data points. We tested the algorithm against the real Twitter data of the 2011 To-hoku earthquake and tsunami, and experimentally confirmed that the algorithm is much more efficient than the exhaustive naive search and is much more accurate than the simple greedy search. By analyzing the real information diffusion data, we revealed that even if the data contains tweets talking about plural topics, the detected burst period tends to contain tweets on a specific topic intensively. In addition, we experimentally confirmed that assuming the information diffusion path to be the line shape tree results in much better approximation of the maximum likelihood estimator than assuming it to be the star shape tree. This is a good heuristic to accurately estimate the change points when the acutal diffusion path is not known to us. These results indicate that it is possible to detect and identify both the burst period and the topic diffused without extracting the tweet sequence for each topic and identifying the diffusion paths for each sequence, and the proposed method can be a useful tool to analyze a huge amount of information diffusion data. Our immediate future work is to compare the proposed method with existing burst detection methods that are designed for data stream. We also plan to devise a method of finding nodes that caused the burst based on the change points detected.

Acknowledgments

The Tweeter data we used in this paper were provided by Prof. Fujio Toriumi of Tokyo University and Dr. Kazuhiro Kazama of Nippon Telegraph and Telephone Corporation.

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-114111, and JSPS Grant-in-Aid for Scientific Research (C) (No. 23500194).

References

1. Bonacichi, P.: Power and centrality: A family of measures. Amer. J. Sociol. 92, 1170–1182 (1987)
2. 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, 211–223 (2001)
3. Katz, L.: A new status index derived from sociometric analysis. Sociometry 18, 39–43 (1953)
4. 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). pp. 137–146 (2003)
5. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Min. Knowl. Disc. 20, 70–97 (2010)
6. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002). pp. 91–101 (2002)
7. Sadikov, E., Medina, M., Leskovec, J., Garcia-Molina, H.: Correcting for missing data in information cascades. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM 2011). pp. 55–64 (2011)
8. 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). pp. 322–337. LNAI 5828 (2009)
9. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Selecting information diffusion models over social networks for behavioral analysis. In: Proceedings of the 2010 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010). pp. 180–195. LNAI 6323 (2010)
10. Sun, A., Zeng, D., Chen, H.: Burst detection from multiple data streams: A network-based approach. IEEE Transactions on Systems, Man, & Cybernetics Society, Part C pp. 258–267 (2010)
11. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press, Cambridge, UK (1994)
12. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of National Academy of Science, USA 99, 5766–5771 (2002)
13. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research 34, 441–458 (2007)
14. Zhu, Y., Shasha, D.: Efficient elastic burst detection in data streams. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003). pp. 336–345 (2003)