2011 11th IEEE International Conference on Data Mining

Discovering Emerging Topics in Social Streams via Link Anomaly Detection

Toshimitsu Takahashi
Institute of Industrial Science
The University of Tokyo
Tokyo, Japan
Email: takahashi@tauhat.com

Ryota Tomioka
Department of Mathematical Informatics
The University of Tokyo
Tokyo, Japan
Email: tomioka@mist.i.u-tokyo.ac.jp

Kenji Yamanishi
Department of Mathematical Informatics
The University of Tokyo
Tokyo, Japan
Email: yamanishi@mist.i.u-tokyo.ac.jp

***Abstract*—Detection of emerging topics are now receiving renewed interest motivated by the rapid growth of social networks. Conventional term-frequency-based approaches may not be appropriate in this context, because the information exchanged are not only texts but also images, URLs, and videos. We focus on the social aspects of theses networks. That is, the links between users that are generated dynamically intentionally or unintentionally through replies, mentions, and retweets. We propose a probability model of the mentioning behaviour of a social network user, and propose to detect the emergence of a new topic from the anomaly measured through the model. We combine the proposed mention anomaly score with a recently proposed change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood (SDNML), or with Kleinberg's burst model. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in social network posts. We demonstrate our technique in a number of real data sets we gathered from Twitter. The experiments show that the proposed mention-anomaly-based approaches can detect new topics at least as early as the conventional term-frequency-based approach, and sometimes much earlier when the keyword is ill-defined.**

***Keywords*-Topic Detection, Anomaly Detection, Social Networks, Sequentially Discounted Maximum Likelihood Coding, Burst detection**

I. Introduction

Communication through social networks, such as Facebook and Twitter, is increasing its importance in our daily life. Since the information exchanged over social networks are not only texts but also URLs, images, and videos, they are challenging test beds for the study of data mining.

There is another type of information that is intentionally or unintentionally exchanged over social networks: mentions. Here we mean by mentions *links* to other users of the same social network in the form of message-to, reply-to, retweet-of, or explicitly in the text. One post may contain a number of mentions. Some users may include mentions in their posts rarely; other users may be mentioning their friends all the time. Some users (like celebrities) may receive mentions every minute; for others, being mentioned might be a rare occasion. In this sense, *mention is like a language* with the number of words equal to the number of users in a social network.

We are interested in detecting emerging topics from social network streams based on monitoring the mentioning behaviour of users. Our basic assumption is that a new (emerging) topic is something people feel like discussing about, commenting about, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words [1], [2]. A term frequency based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. Moreover, it cannot be applied when the contents of the messages are mostly non-textual information. On the other hands, the "words" formed by mentions are unique, requires little prepossessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

In this paper, we propose a probability model that can capture the normal mentioning behaviour of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Then this model is used to measure the *anomaly* of future user behaviour. Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behaviour of the user. We aggregate the anomaly scores obtained in this way over hundreds of users and apply a recently proposed change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding [3]. This technique can detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and pinpoint where the topic emergence is; see Figure 1. The effectiveness of the proposed approach is demonstrated on two data sets we have collected from Twitter. We show that our approach can detect the emergence of a new topic at least as fast as using the best term that was not obvious at the moment. Furthermore, we show that in two out of two data sets, the proposed link-anomaly based method can detect the emergence of the topics earlier than keyword-frequency based methods, which can be explained by the keyword ambiguity we mentioned above.

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.53

II. Related work

Detection and tracking of topics have been studied extensively in the area of topic detection and tracking (TDT) [1]. In this context, the main task is to either classify a new document into one of the known topics (tracking) or to detect that it belongs to none of the known categories. Subsequently, temporal structure of topics have been modeled and analyzed through dynamic model selection [4], temporal text mining [5], and factorial hidden Markov models [6].

Another line of research is concerned with formalizing the notion of "bursts" in a stream of documents. In his seminal paper, Kleinberg modeled bursts using time varying Poisson process with a hidden discrete process that controls the firing rate [2]. Recently, He and Parker developed a physics inspired model of bursts based on the change in the momentum of topics [7].

All the above mentioned studies make use of textual content of the documents, but not the social content of the documents. The social content (links) have been utilized in the study of citation networks [8]. However, citation networks are often analyzed in a stationary setting.

The novelty of the current paper lies in focusing on the social content of the documents (posts) and in combining this with a change-point analysis.

III. Proposed Method

The overall flow of the proposed method is shown in Figure 1. We assume that the data arrives from a social network service in a sequential manner through some API. For each new post we use samples within the past T time interval for the corresponding user for training the mention model we propose below. We assign anomaly score to each post based on the learned probability distribution. The score is then aggregated over users and further fed into a change-point analysis.

A. Probability Model

We characterize a post in a social network stream by the number of mentions k it contains, and the set V of names (IDs) of the users mentioned in the post. Formally, we consider the following joint probability distribution

$$P(k, V|\theta, \{\pi_v\}) = P(k|\theta) \prod_{v \in V} \pi_v. \tag{1}$$

Here the joint distribution consists of two parts: the probability of the number of mentions k and the probability of each mention given the number of mentions. The probability of the number of mentions $P(k|\theta)$ is defined as a geometric distribution with parameter θ as follows:

$$P(k|\theta) = (1-\theta)^k \theta. \tag{2}$$

On the other hand, the probability of mentioning users in V is defined as independent, identical multinomial distribution with parameters π_v ($\sum_v \pi_v = 1$).

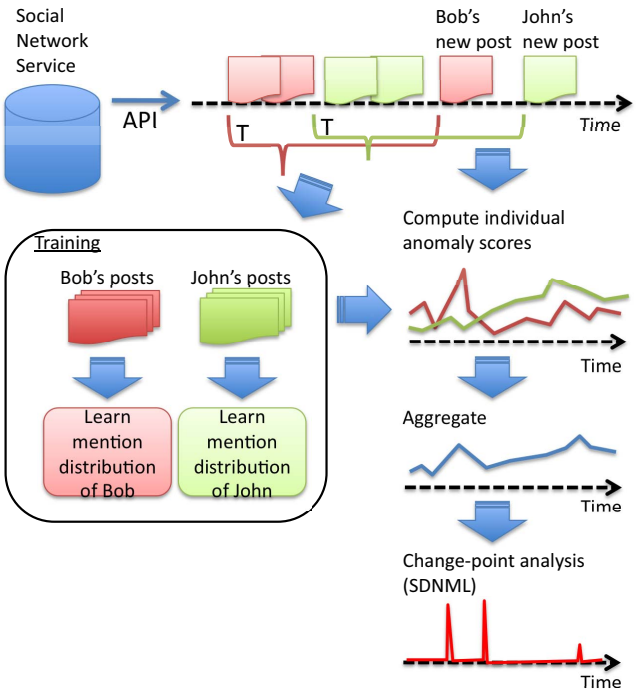


Figure 1. Overall flow of the proposed method.

Suppose that we are given n training examples $\mathcal{T} = \{(k_1, V_1), \ldots, (k_n, V_n)\}$ from which we would like to learn the predictive distribution

$$P(k, V|\mathcal{T}) = P(k|\mathcal{T}) \prod_{v \in V} P(v|\mathcal{T}). \tag{3}$$

First we compute the predictive distribution with respect to the the number of mentions $P(k|\mathcal{T})$. This can be obtained by assuming a beta distribution as a prior and integrating out the parameter θ. The density function of the beta prior distribution is written as follows:

$$p(\theta|\alpha, \beta) = \frac{(1-\theta)^{\beta-1}\theta^{\alpha-1}}{B(\alpha, \beta)},$$

where α and β are parameters of the beta distribution and $B(\alpha, \beta)$ is the beta function. By the Bayes rule, the predictive distribution can be obtained as follows:

$$\begin{aligned} P(k|\mathcal{T}, \alpha, \beta) &= \frac{P(k, k_1, \ldots, k_n|\alpha, \beta)}{P(k_1, \ldots, k_n|\alpha, \beta)} \\ &= \frac{\int_0^1 (1-\theta)^{\sum_{i=1}^n k_i + k + \beta - 1} \theta^{n+1+\alpha-1} \mathrm{d}\theta}{\int_0^1 (1-\theta)^{\sum_{i=1}^n k_i + \beta - 1} \theta^{n+\alpha-1} \mathrm{d}\theta}. \end{aligned}$$

Both the integrals on the numerator and denominator can be obtained in closed forms as beta functions and the predictive distribution can be rewritten as follows:

$$P(k|\mathcal{T}, \alpha, \beta) = \frac{B(n+1+\alpha, \sum_{i=1}^n k_i + k + \beta)}{B(n+\alpha, \sum_{i=1}^n k_i + \beta)}.$$

Using the relation between beta function and gamma function, we can further simplify the expression as follows:

$$P(k|\mathcal{T},\alpha,\beta) = \frac{n+\alpha}{m+k+\beta}\prod_{j=0}^{k}\frac{m+\beta+j}{n+m+\alpha+\beta+j}, \quad (4)$$

where $m = \sum_{i=1}^{n} k_i$ is the total number of mentions in the training set $\mathcal{T}$.

Next, we derive the predictive distribution $P(v|\mathcal{T})$ of mentioning user v. The maximum likelihood (ML) estimator is given as $P(v|\mathcal{T}) = m_v/m$, where m is the number of total mentions and m_v is the number of mentions to user v in the data set $\mathcal{T}$. The ML estimator, however, cannot handle users that did not appear in the training set $\mathcal{T}$; it would assign probability zero to all these users, which would appear infinitely anomalous in our framework. Instead we use the Chinese Restaurant Process (CRP; see [9]) based estimation. The CRP based estimator assigns probability to each user v that is proportional to the number of mentions m_v in the training set $\mathcal{T}$; in addition, it keeps probability proportional to γ for mentioning someone who was not mentioned in the training set $\mathcal{T}$. Accordingly the probability of known users is given as follows:

$$P(v|\mathcal{T}) = \frac{m_v}{m+\gamma} \quad (\text{for } v\text{: } m_v \geq 1). \quad (5)$$

On the other hand, the probability of mentioning a new user is given as follows:

$$P(\{v : m_v = 0\}|\mathcal{T}) = \frac{\gamma}{m+\gamma}. \quad (6)$$

B. Computing the link-anomaly score

In order to compute the anomaly score of a new post $\boldsymbol{x} = (t, u, k, V)$ by user u at time t containing k mentions to users V, we compute the probability (3) with the training set $\mathcal{T}_u^{(t)}$, which is the collection of posts by user u in the time period $[t-T, t]$ (we use $T = 30$ days in this paper). Accordingly the link-anomaly score is defined as follows:

$$s(\boldsymbol{x}) = -\log P(k|\mathcal{T}_u^{(t)}) - \sum_{v\in V}\log P(v|\mathcal{T}_u^{(t)}). \quad (7)$$

The two terms in the above equation can be computed via the predictive distribution of the number of mentions (4), and the predictive distribution of the mentionee (5)–(6), respectively.

C. Combining Anomaly Scores from Different Users

The anomaly score in (7) is computed for each user depending on the current post of user u and his/her past behaviour $\mathcal{T}_u^{(t)}$. In order to measure the general trend of user behaviour, we propose to aggregate the anomaly scores obtained for posts $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ using a discretization of window size $\tau > 0$ as follows:

$$s'_j = \frac{1}{\tau}\sum_{t_i\in[\tau(j-1),\tau j]} s(\boldsymbol{x}_i), \quad (8)$$

where $\boldsymbol{x}_i = (t_i, u_i, k_i, V_i)$ is the post at time t_i by user u_i including k_i mentions to users V_i.

D. Change-point detection via Sequentially Discounting Normalized Maximum Likelihood Coding

Given an aggregated measure of anomaly (8), we apply a change-point detection technique based on the SDNML coding [3]. This technique detects a change in the statistical dependence structure of a time series by monitoring the compressibility of the new piece of data. The SDNML proposed in [3] is an approximation for normalized maximum likelihood (NML) code length that can be computed sequentially and employs discounting in the learning of the AR models; see also [11], [12].

Algorithmically, the change point detection procedure can be outlined as follows. For convenience, we denote the aggregate anomaly score as x_j instead of s'_j.

1. 1st layer learning: Let $x^{j-1} := \{x_1, \ldots, x_{j-1}\}$ be the collection of aggregate anomaly scores from discrete time 1 to $j-1$. Sequentially learn the SDNML density function $p_{\text{SDNML}}(x_j|x^{j-1})$ $(j = 1, 2, \ldots)$; see Appendix A for details.

2. 1st layer scoring: Compute the intermediate change-point score by smoothing the log loss of the SDNML density function with window size κ as follows:

$$y_j = \frac{1}{\kappa}\sum_{j'=j-\kappa+1}^{j}\left(-\log p_{\text{SDNML}}(x_j|x^{j-1})\right).$$

3. 2nd layer learning Let $y^{j-1} := \{y_1, \ldots, y_{j-1}\}$ be the collection of smoothed change-point score obtained as above. Sequentially learn the second layer SDNML density function $p_{\text{SDNML}}(y_j|y^{j-1})$ $(j = 1, 2, \ldots)$; see Appendix A.

4. 2nd layer scoring Compute the final change-point score by smoothing the log loss of the SDNML density function as follows:

$$Score(y_j) = \frac{1}{\kappa}\sum_{j'=j-\kappa+1}^{j}\left(-\log p_{\text{SDNML}}(y_j|y^{j-1})\right). \quad (9)$$

E. Dynamic Threshold Optimization (DTO)

We make an alarm if the change-point score exceeds a threshold, which was determined adaptively using the method of dynamic threshold optimization (DTO) [13].

In DTO, we use a 1-dimensional histogram for the representation of the score distribution. We learn it in a sequential and discounting way. Then, for a specified value ρ, to determine the threshold to be the largest score value such that the tail probability beyond the value does not exceed ρ. We call ρ a *threshold parameter*.

The details of DTO are summarized in Algorithm 1.

IV. Experiments

A. Experimental setup

We collected four data sets "Job hunting", "Youtube", "NASA", and "BBC"from Twitter. Each data set is associated with a list of posts in a service called Togetter[1]; Togetter

[1]http://togetter.com/

Algorithm 1 Dynamic Threshold Optimization (DTO) [13]

Given: $\{Score_j | j = 1, 2, \ldots\}$: scores, N_H: total number of cells, ρ: parameter for threshold, λ_H: estimation parameter, r_H: discounting parameter, M: data size

Initialization: Let $q_1^{(1)}(h)$be a uniform distribution.

for $j = 1, \ldots, M-1$ **do**

Threshold optimization: Let l be the least index such that $\sum_{h=1}^{l} q^{(j)}(h) \geq 1 - \rho$. The threshold at time j is given as

$$\eta(j) = a + \frac{b-a}{N_H - 2}(l+1).$$

Alarm output: Raise an alarm if $Score_j \geq \eta(j)$.

Histogram update:

$$q_1^{(j+1)}(h) = \begin{cases} (1-r_H)q_1^{(j)}(h) + r_H & \text{if } Score_j \text{ falls into the } h\text{th cell,} \\ (1-r_H)q_1^{(j)}(h) & \text{otherwise.} \end{cases}$$

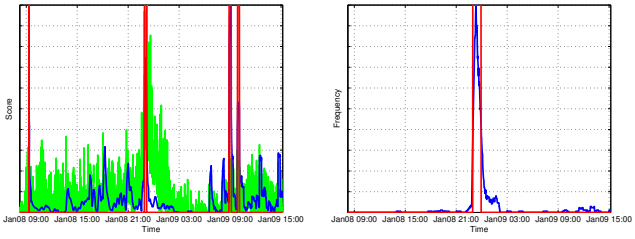
$q^{(j+1)}(h) = (q_1^{(j+1)}(h) + \lambda_H)/(\sum_h q_1^{(j+1)}(h) + N_H \lambda_H)$.

end for

is a collaborative service where people can tag Twitter posts that are related to each other and organize a list of posts that belong to a certain topic. Our goal is to evaluate whether the proposed approach can detect the emergence of the topics recognized and collected by people. We have selected two data sets, "Job hunting", and "BBC" each corresponding to a user organized list in Togetter. For each data set we collected posts from users that appeared in each list (participants). The number of participants were 200 for "Job hunting" data set and 47 for "BBC" data set, respectively. Please find the results for the other two data sets and combination of the proposed link-anomaly-based method and Kleinberg's burst model in our technical report [14].

We compared our proposed approach with a keyword-based change-point detection method. In the keyword-based method, we looked at a sequence of occurrence frequencies (observed within one minute) of a keyword related to the topic; the keyword was manually selected to best capture the topic. Then we applied DTO described in Section III-E to the sequence of keyword frequency. In our experience, the sparsity of the keyword frequency seems to be a bad combination with the SDNML method; therefore we did not use SDNML in the keyword-based method. We use the smoothing parameter $\kappa = 15$, and the order of the AR model 30 in the experiments; the parameters in DTO was set as $\rho = 0.05$, $N_H = 20$, $\lambda_H = 0.01$, $r_H = 0.005$.

A drawback of the keyword-based dynamic thresholding is that the keyword related to the topic must be known in advance, although this is not always the case in practice. The change-point detected by the keyword-based methods can be thought of as the time when the topic really emerges. Hence our goal is to detect emerging topics as early as the keyword based methods.



(a) Link-anomaly-based change-point analysis. Green: Aggregated anomaly score (8) at $\tau = 1$ minute. Blue: Change-point score (9). Red: Alarm time.

(b) Keyword-frequency-based method. Blue: Frequency of keyword "Job hunting" per one minute. Red: Alarm time.

Figure 2. Result of "Job hunting" data set. The initial controversial post was posted on 22:50, Jan 08.

Table I
DETECTION TIME AND THE NUMBER OF DETECTIONS. THE FIRST DETECTION TIME IS DEFINED AS THE TIME OF THE FIRST ALERT AFTER THE EVENT/POST THAT INITIATED EACH TOPIC; SEE CAPTIONS FOR FIGURES 2–3 FOR THE DETAILS.

Method		"Job hunting"	"BBC"
Link-anomaly based method	♯ of detections	4	3
	1st detect time	22:55, Jan 08	**19:52, Jan 21**
Keyword-frequency based method	♯ of detections	1	1
	1st detect time	22:57, Jan 08	22:41, Jan 21

B. *"Job hunting" data set*

This data set is related to a controversial post by a famous person in Japan that "the reason students having difficulty finding jobs is, because they are stupid" and various replies.

The keyword used in the keyword-based methods was "Job hunting." Figure 2(a) shows the result of the proposed link-anomaly-based change detection. Figure 2(b) shows the result of the keyword-frequency-based change detection.

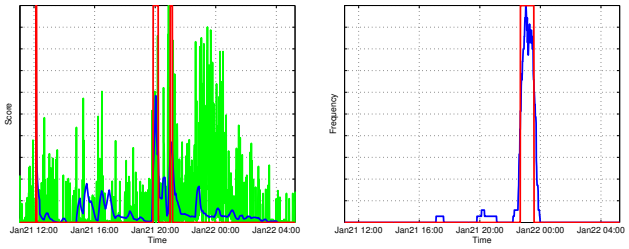
The first alarm time of the proposed link-anomaly-based change-point analysis was 22:55, whereas that for the keyword-frequency-based counterpart was 22:57; see also Table I. From Figure 2, we can observe that the proposed link-anomaly-based methods were able to detect the emerging topic as early as keyword-frequency-based methods.

C. *"BBC" data set*

This data set is related to angry reactions among Japanese Twitter users against a BBC comedy show that asked "who is the unluckiest person in the world" (the answer is a Japanese man who got hit by nuclear bombs in both Hiroshima and Nagasaki but survived).

The keyword used in the keyword-based models is "British" (or "Britain"). Figure 3(a) shows the result of link-anomaly-based change detection. Figure 3(b) shows the same result for the keyword-frequency-based method.

The first alarm time of the link-anomaly-based method was 19:52, which is earlier than the keyword-frequency-based method at 22:41. See Table I.



(a) Link-anomaly-based change-point analysis. Green: Aggregated anomaly score (8) at $\tau = 1$ minute. Blue: Change-point score (9). Red: Alarm time.

(b) Keyword-frequency-based method. Blue: Frequency of keyword "British" per one minute. Red: Alarm time.

Figure 3. Result of "BBC" data set. The first post about BBC's comedy show was posted on 17:08, Jan 21.

Table II
NUMBER OF ALARMS FOR THE PROPOSED CHANGE-POINT DETECTION METHOD BASED ON THE LINK-ANOMALY SCORE (8) FOR VARIOUS SIGNIFICANCE LEVEL PARAMETER VALUES ρ.

ρ	"Job hunting"	"BBC"
0.01	4	3
0.05	4	3
0.1	8	3

D. Discussion

The proposed link-anomaly based method detects an emerging topic significantly earlier than the keyword-frequency based method on "BBC" data set, whereas for "Job hunting" data set the detection times are almost the same; see Table I.

The above observation is natural, because for "Job hunting" data set, the keyword seemed to have been unambiguously defined from the beginning of the emergence of the topic, whereas for "BBC" data set, the keywords are more ambiguous. For "BBC" data set, interestingly the "bursty" areas found by the link-anomaly-based change-point analysis and the keyword-frequency-based method seem to be disjoint (Figures 3(a) and 3(b)). This is probably because there was an initial stage where people reacted individually using different words and later there was another stage in which the keywords are more unified.

In our approach, the alarm was raised if the change-point score exceeded a dynamically optimized threshold based on the significance level parameter ρ. Table II shows results for a number of threshold parameter values. We see that as ρ increased, the number of false alarms also increased. Meanwhile, even when it was so small, our approach was still able to detect the emerging topics as early as the keyword-based methods. We set $\rho = 0.05$ as a default parameter value in our experiment.

V. CONCLUSION

In this paper, we have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behaviour of users instead of the textual contents. We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee. Furthermore, we have combined the proposed mention model with recently proposed SDNML change-point detection algorithm [3] to pin-point the emergence of a topic.

We have applied the proposed approach to two real data sets we have collected from Twitter. In all the data sets our proposed approach showed promising performance; the detection by the proposed approach was as early as term-frequency based approaches in the hindsight of the keywords that best describes the topic that we have manually chosen afterwards. Furthermore, for "BBC" data set, in which the keyword that defines the topic is more ambiguous than the other data set, the proposed link-anomaly based approach has detected the emergence of the topics much earlier than the keyword-based approach.

ACKNOWLEDGMENTS

This work was partially supported by MEXT KAKENHI 23240019, 22700138, Aihara Project, the FIRST program from JSPS, initiated by CSTP, Hakuhodo Corporation, NTT Corporation, and Microsoft Corporation (CORE Project).

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang *et al.*, "Topic detection and tracking pilot study: Final report," in *Proceedings of the DARPA broadcast news transcription and understanding workshop*, 1998.

[2] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Min. Knowl. Disc.*, vol. 7, no. 4, pp. 373–397, 2003.

[3] Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, "Real-time change-point detection using sequentially discounting normalized maximum likelihood coding," in *Proceedings. of the 15th PAKDD*, 2011.

[4] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in *Proceedings of the 10th ACM SIGKDD*, 2004, pp. 811–816.

[5] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proceedings of the 11th ACM SIGKDD*, 2005, pp. 198–207.

[6] A. Krause, J. Leskovec, and C. Guestrin, "Data association for topic intensity tracking," in *Proceedings of the 23rd ICML*, 2006, pp. 497–504.

[7] D. He and D. S. Parker, "Topic dynamics: an alternative model of bursts in streams of topics," in *Proceedings of the 16th ACM SIGKDD*, 2010, pp. 443–452.

[8] H. Small, "Visualizing science by citation mapping," *Journal of the American society for Information Science*, vol. 50, no. 9, pp. 799–813, 1999.

[9] D. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII—1983*. Springer, 1985, pp. 1–198.

[10] J. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from time series," *IEEE T. Knowl. Data En.*, vol. 18, no. 44, pp. 482–492, 2006.
[11] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE T. Inform. Theory*, vol. 47, no. 5, pp. 1712–1717, 2002.
[12] J. Rissanen, T. Roos, and P. Myllymäki, "Model selection by sequentially normalized least squares," *Journal of Multivariate Analysis*, vol. 101, no. 4, pp. 839–849, 2010.
[13] K. Yamanishi and Y. Maruyama, "Dynamic syslog mining for network failure monitoring," *Proceeding of the 11th ACM SIGKDD*, p. 499, 2005.
[14] T. Takahashi, R. Tomioka, and K. Yamanishi, "Discovering emerging topics in social streams via link anomaly detection," arXiv:1110.2899v1 [stat.ML], Tech. Rep., 2011.

APPENDIX

Suppose that we observe a discrete time series x_t ($t = 1, 2, \ldots,$); we denote the data sequence by $x^t := x_1 \cdots x_t$. Consider the parametric class of conditional probability densities $\mathcal{F} = \{p(x_t|x^{t-1} : \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbf{R}^p\}$, where $\boldsymbol{\theta}$ is the p-dimensional parameter vector and we assume x^0 to be an empty set. We denote the maximum likelihood (ML) estimator given the data sequence x^t by $\hat{\boldsymbol{\theta}}(x^t)$; i.e., $\hat{\boldsymbol{\theta}}(x^t) := \mathrm{argmax}_{\boldsymbol{\theta}\in\mathbf{R}^p} \prod_{j=1}^t p(x_j|x^{j-1} : \boldsymbol{\theta})$. The sequential normalized maximum likelihood (SNML) distribution is defined as follows:

$$p_{\text{SNML}}(x_t|x^{t-1}) := \frac{p(x^t : \hat{\boldsymbol{\theta}}(x^t))}{K_t(x^{t-1})}, \tag{10}$$

where the normalization constant $K_t(x^{t-1}) := \int p(x^t|\hat{\boldsymbol{\theta}}(x^t))\mathrm{d}x_t$ is necessary because the new sample x_t is used in the estimation of parameter vector $\hat{\boldsymbol{\theta}}(x^t)$ and the numerator in (10) is not a proper density function. We call the quantity $-\log p_{\text{SNML}}(x_t|x^{t-1})$ the *SNML code-length*. It is known from [12] that the cumulative SNML code-length, which is the sum of SNML code-length over the sequence, is optimal in the sense that it asymptotically achieves the shortest code-length.

The sequentially discounting normalized maximum likelihood (SDNML) is obtained by applying the above SNML to the class of autoregressive (AR) model and replacing the ML estimation in (10) with a *discounted* ML estimation, which makes the SDNML-based change-point detection algorithm more flexible than an SNML-based one. Let $x_t \in \mathbf{R}$ for each t. We define the pth order AR model as follows:

$$p(x_t|x_{t-k}^{t-1} : \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\Big(x_t - \sum_{i=1}^p a^{(i)} x_{t-i}\Big)^2\right),$$

where $\boldsymbol{\theta}^\top = (\boldsymbol{a}^\top, \sigma^2) = ((a^{(1)}, \ldots, a^{(p)}), \sigma^2)$ is the parameter vector.

In order to compute the SDNML density function we need the discounted ML estimators of the parameters in $\boldsymbol{\theta}$. We define the discounted ML estimator of the regression coefficient $\hat{\boldsymbol{a}}_t$ as follows:

$$\hat{\boldsymbol{a}}_t = \mathop{\mathrm{argmin}}_{\boldsymbol{a}\in\mathbf{R}^p} \sum_{j=t_0+1}^t w_{t-j} \left(x_j - \boldsymbol{a}^\top \bar{\boldsymbol{x}}_j\right)^2, \tag{11}$$

where $w_{j'} = r(1-r)^{j'}$ is a sequence of sample weights with the discounting coefficient r ($0 < r < 1$); t_0 is the smallest number of samples such that the minimizer (11) is unique; $\bar{\boldsymbol{x}}_j := (x_{j-1}, x_{j-2}, \ldots, x_{j-k})^\top$. Note that the error terms from older samples receive geometrically decreasing weights in (11). The larger the discounting coefficient r is, the smaller the weights of the older samples become; thus we have stronger discounting effect. Moreover, we obtain the discounted ML estimator of the variance $\hat{\tau}_t$ as follows:

$$\hat{\tau}_t := \sum_{j=t_0+1}^t r(1-r)^{t-j} \hat{e}_j^2$$

where we define $\hat{e}_j^2 = \left(x_j - \hat{\boldsymbol{a}}_j^\top \bar{\boldsymbol{x}}_j\right)^2$. Clearly when the discounted estimator of the AR coefficient $\hat{\boldsymbol{a}}_j$ is available, $\hat{\tau}_t$ can be computed in a sequential manner.

In the sequel, we first describe how to efficiently compute the AR estimator $\hat{\boldsymbol{a}}_j$. Finally we derive the SDNML density function using the discounted ML estimators $(\hat{\boldsymbol{a}}_t, \hat{\tau}_t)$.

The AR coefficient $\hat{\boldsymbol{a}}_j$ can simply be computed by solving the least-squares problem (11). It can, however, be obtained more efficiently using the iterative formula described in [12]. Here we repeat the formula for the discounted version presented in [3]. First define the sufficient statistics $V_t \in \mathbf{R}^{p\times p}$ and $\boldsymbol{\chi}_t \in \mathbf{R}^p$ as follows:

$$V_t := \sum_{j=t_0+1}^t w_j \bar{\boldsymbol{x}}_j \bar{\boldsymbol{x}}_j^\top, \quad \boldsymbol{\chi}_t := \sum_{j=t_0+1}^t w_j \bar{\boldsymbol{x}}_j x_j.$$

Using the sufficient statistics, the discounted AR coefficient $\hat{\boldsymbol{a}}_j$ from (11) can be written as follows:

$$\hat{\boldsymbol{a}}_t = V_t^{-1} \boldsymbol{\chi}_t.$$

Note that $\boldsymbol{\chi}_t$ can be computed in a sequential manner. The inverse matrix V_t^{-1} can also be computed sequentially using the Sherman-Morrison-Woodbury formula as follows:

$$V_t^{-1} = \frac{1}{1-r} V_t^{-1} - \frac{r}{1-r} \frac{V_t^{-1} \bar{\boldsymbol{x}}_t \bar{\boldsymbol{x}}_t^\top V_t^{-1}}{1-r+c_t},$$

where $c_t = r\bar{\boldsymbol{x}}_t^\top V_t^{-1} \bar{\boldsymbol{x}}_t$.

Finally the SDNML density function is written as follows:

$$p_{\text{SDNML}}(x_t|x^{t-1}) = \frac{1}{K_t(x^{t-1})} \frac{s_t^{-(t-t_0)/2}}{s_{t-1}^{-(t-t_0-1)/2}},$$

where the normalization factor $K_t(x^{t-1})$ is calculated as follows:

$$K_t(x^{t-1}) = \frac{\sqrt{\pi}}{1-d_t} \sqrt{\frac{1-r}{r}} (1-r)^{-\frac{t-m}{2}} \frac{\Gamma((t-t_0-1)/2)}{\Gamma((t-t_0)/2)},$$

with $d_t = c_t/(1-r+c_t)$.