Learning Latent Events From Network Message Logs

Siddhartha Satpathi, Supratim Deb, R. Srikant, *Fellow, IEEE*, and He Yan

***Abstract*— We consider the problem of separating error messages generated in large distributed data center networks into error events. In such networks, each error event leads to a stream of messages generated by hardware and software components affected by the event. These messages are stored in a giant message log. We consider the unsupervised learning problem of identifying the signatures of events that generated these messages; here, the signature of an error event refers to the mixture of messages generated by the event. One of the main contributions of the paper is a novel mapping of our problem which transforms it into a problem of topic discovery in documents. Events in our problem correspond to topics and messages in our problem correspond to words in the topic discovery problem. However, there is no direct analog of documents. Therefore, we use a non-parametric change-point detection algorithm, which has linear computational complexity in the number of messages, to divide the message log into smaller subsets called episodes, which serve as the equivalents of documents. After this mapping has been done, we use a well-known algorithm for topic discovery, called LDA, to solve our problem. We theoretically analyze the change-point detection algorithm, and show that it is consistent and has low sample complexity. We also demonstrate the scalability of our algorithm on a real data set consisting of 97 million messages collected over a period of 15 days, from a distributed data center network which supports the operations of a large wireless service provider.**

***Index Terms*— Unsupervised learning, data mining, event message log, change point detection, Bayesian inference, data center networks, time series mixture.**

I. Introduction

The delivery of modern data and web-based services requires the execution of a chain of network functions at different elements in distributed data-centers. This is true for video-based services, gaming services, cellular data/voice services, etc., each of which requires processing from multiple coupled networked entities hosting different network functions. For example, modern wireless networks rely on servers and virtual machines (VM) residing in distributed data centers to establish voice calls or data sessions, authenticate users, check user compliance with monthly voice/data limits, verify if users have paid their monthly bills, add to users' bills for extra services, etc., all of which are done before completing a call. Efficient management and operations of these services is of paramount importance as networks grow increasingly complex with the advent of technologies like virtualization and 5G. An integral component of network management is the ability to identify and understand *error events*, when failures occur in the hardware and/or software components of the network. However, the complex interdependence between coupled networking functions poses a significant challenge in characterizing an error event due to the fact that error messages can be generated in network elements beyond the actual source of error. In this paper, we are interested in the problem of mining latent error event information from messages generated by servers, VMs, base stations, routers, and links in large-scale distributed data center networks. The mined events are useful for troubleshooting purposes. Also, the correlations captured through each learned event could be subsequently used for on-line detection of potential errors. While our methodology is broadly applicable to any type of data center network, we validate our algorithms by applying them to a large data set provided by a major wireless network service provider, so we will occasionally use terminology specific to this application to motivate our problem and solution methodology.

In most operational networks, all messages and alarms from distributed network elements are logged with time stamps into message logs. The logs from different network elements could be pooled together in a central database for subsequent analysis. While mining error logs have been studied extensively in different contexts, (see [1], [2] for excellent surveys; also see Section I-B), there are some fundamental differences in our setting. Modern data center and communication networks consist of components bought from different vendors, and each component is designed to generate an error message when it cannot execute a job. This poses a challenge in mining messages because there is no common model or standard that dictates the content and format of these error messages. Another challenge stems from the fact that each end-to-end service consists of multiple network functions each of which generates diverse error messages when failures happen. The following example provides an illustration.

Motivating Example: Suppose Alice makes a cellphone call to Bob. This call is first routed through a base station which is attached to a data center verifying the caller credentials. If Alice is not at her home location, a VM at this data center must contact a database at her home location to verify

Manuscript received November 6, 2018; revised April 30, 2019; accepted July 1, 2019; approved by IEEE/ACM Transactions on Networking Editor T. He. Date of publication July 31, 2019; date of current version August 16, 2019. The work of S. Satpathi and R. Srikant was supported in part by AT&T, and in part by the NSF under Grant NeTS 1718203 and Grant CPS ECCS 1739189. *(Corresponding author: Siddhartha Satpathi.)*

S. Satpathi and R. Srikant are with University of Illinois at Urbana-Champaign (UIUC), Champaign, IL 61820-5783 USA (e-mail: sidd.piku@gmail.com; rsrikant@illinois.edu).

S. Deb was with AT&T Labs, Florham Park, NJ 07932-1049 USA. He is now with Facebook, New York, NY 10003 USA.

H. Yan is with AT&T Labs, Florham Park, NJ 07932-1049 USA.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes a PDF containing Appendices A-J and proofs.

Digital Object Identifier 10.1109/TNET.2019.2930040

her credentials. Once the credentials are verified, the caller's cellular base station connects to the base station near Bob through a complicated network spanning many geographical locations. Consider two potential error scenarios: (i) an error occurs at a router in the path from Bob to Alice's base station, (ii) error at a router connecting the data centers verifying the caller's credentials. In either scenario, the call will fail to be established leading to the generation of error messages not only at the failed routers but also at network functions (implemented in a cluster of VMs) responsible for call establishment. Furthermore, if the error leads to additional call failures, then respective base-stations could send alarms indicating higher than normal call failures. Additionally, depending on vendor of a given network element, the timing and content of the error messages could be different.

Indeed, the source, timing, and message-components of the error are all latent. In this paper we are interested in extracting patterns from messages generated by common faults/errors (also referred to as events). Specifically, our goal of this paper is to mine event signatures (i.e., distribution of messages for each event) and event occurrences (i.e., the begin and the end time of each event) from the message log. Based on the motivating example, we now note the following fundamental characteristics which make our error event mining problem challenging:

- In our setting, the source of an error is usually not known. There could be error messages due to network-component level failures or due to network service-level failure. In case of a service-level failure, error-message could be generated by a component that is functional by itself. For example, when the link between an authentication server and the network core fails, this could lead to call establishment failures which are logged by network functions responsible for call establishment. Furthermore, the same type of error log-message could be generated due to many different errors. From a data modeling point of view, each latent) event can be viewed as a probabilistic-mixture of multiple log-messages at different elements and also, the set of log messages generated by different events could have non-zero intersection.
- Each error event can produce a sequence of messages, including the same type of message multiple times, and the temporal order between distinct messages from the same event could vary based on the latency between network elements, network-load, co-occurrence of other uncorrelated events, etc. Thus, the temporal pattern of messages may also contain useful information for our purpose. In our model, the message occurrence times are modeled as a stochastic process.
- These messages could correspond to multiple simultaneous events without any further information on the start-time and end-time of each event.
- An additional challenge arises from the fact that network topology information is unknown, because modern networks are very complicated and are constantly evolving due to the churn (addition or deletion) of routers and switches. Third-party vendor software and hardware have no way of providing information to localize and understand the errors. Thus, topological information cannot be used for event mining purposes.

The practical novelty of our work comes from modeling for all of the above factors and proposing scalable algorithms that learn the latent event signatures (the notion of signature will be made precise later) along with their occurrence times.

Remark 1: In different works on event mining (see Section I-B), the concept of event is different depending on the problem-context. In our work, an event *simply refers to a real-world occurrence of a fault/incident somewhere in the distributed/networking system such that each event leads to a generation of error messages at multiple network elements.*

A. Contributions

We model each event as a probabilistic mixture of messages from different sources.[1] In other words, the probability distribution over messages characterizes an event, and thus acts as the signature of the event. Each occurrence of an event also has a start/end time and several messages can be generated during the occurrence of an event. We only observe the messages and their time-stamps while the event signatures and duration window is unknown; also there could be multiple simultaneous events occurring in the network. Given this setting, we study the following unsupervised learning problem: *given collection of time-stamped log-messages, learn the latent event signatures and event start/end times.*

The main contributions of the paper are as follows:

- *Novel algorithmic framework:* We present a novel way of decomposing the problem into simpler sub-problems. Our method, which we will call CD-LDA, decomposes the problem into two parts: the first part consists of a change-point detection algorithm which identifies time instants at which either a new event starts in the network or an existing event comes to an end, and, the second part of the algorithm uses Latent Dirichlet Allocation (LDA) (see [3]) to classify messages into events. This observation that one can use change-point detection, followed by LDA, for event classification is one of the novel ideas in the paper.
- *Scalable change-point detection:* While the details of the LDA algorithm itself are standard, non-parametric change-point detection as we have used in this paper is not as well studied. We adapt an idea from [4] to design an $O(n)$ algorithm where n is the number of messages in the message log. Our change detection algorithm uses an easy to compute total-variation (TV) distance. We analyze the sample complexity of (i.e, the number of samples required to detect change points with a high-degree of accuracy) of our change-point detection algorithm using the method of types and Pinsker's inequality from information theory. To the best of our knowledge, no such sample complexity results exist for the algorithm in [4].

[1] It is more precise to use the terminology event-class to refer to a specific fault-type; each occurrence can be referred to as an instance of some event class. However, for simplicity, we simply refer to event-class as event and we just say occurrence of the event to mean instance of this class.

- *Experimental validation:* We use two different real-world data sets from a large operational network to perform the following validation of our approach. First, we compare our algorithm to two existing approaches adapted to our setting: a Bayesian inference-based algorithm and Graph-based clustering algorithm. We show the benefits of our approach compared to these methods in terms of scalability and performance, by applying it to small samples extracted from a large data set consisting of 97 million messages. Second, we validate our method against two real world events by comparing the event signature learned by our method with domain expert validated event signature for a smaller data set consisting of 700K messages.[2] Finally, we also show results to indicate scalability of our method by applying to the entire 97 million message data set.

We note that this paper is an extended version of [5].

B. Context and Related Work

Data-driven techniques have been shown to be very useful in extracting meaningful information out of system-logs and alarms for large and complex systems. The primary goal of this "knowledge" extraction is to assist in diagnosing the underlying problems responsible for log-messages and events. Two excellent resources for the large body of work done in the area are [1], [2]. Next, we outline some of the key challenges in this knowledge extraction, associated research in the area, and our problem in the context of existing work.

Mining and Clustering Unstructured Logs: Log-messages are unstructured textual data without any annotation for the underlying fault. A significant amount of research has focused on converting unstructured logs to common *semantic events* [2]. Note that the notion of *semantic events* is different from the actual real-world events responsible for generating the messages, nevertheless, such a conversion helps in providing a canonical description of the log-messages that enables subsequent correlation analysis. These works exploit the structural similarity among different messages to either compute an intelligent log-parser or cluster the messages based on message texts [2], [6]–[8]. Each cluster can be viewed as an semantic event which can help in diagnosing the underlying root-cause. One work closely related to ours is [9], in which the authors mine network log messages to first extract templates and then learn pairwise *implication* rules between template-pairs. Our setting and objective are somewhat different, we model events as message-distributions from different elements with each event occurrence having certain start and end times; the messages belonging to an event and the associated occurrence time-windows are hidden (to be learned). A more recent work [10] develops algorithms to mine underlying structural-event as a work-flow graph. The main differences are that, each transaction is a fixed sequence of messages unlike our setting where each message could be generated multiple times based on some hidden stochastic process, and furthermore, in our setting, there could be multiple events manifested in the centralized log-server.

Mining Temporal Patterns: Log-messages are time-series data and thus the temporal patterns contain useful information. Considerable amount of research has gone into learning latent patterns, trends and relationship between events based on timing information in the messages [11]–[13]. We refer to [2], [14], [15] for survey of these approaches. Extracted event-patterns could be used to construct event correlation graphs that could be mined using techniques such as graph-clustering. Specifically, these approaches are useful when event-streams are available as time-series. We are interested in scenarios where each event is manifested in terms of time-series of unstructured messages and furthermore, same message could arise from multiple events. Nevertheless, certain techniques developed for temporal event mining could be adapted to our setting as we describe in Section IV-A2; our results indicate that such an adaptation works well under certain conditions. Note that, our goal is to also learn the event-occurrence times.

Event-Summarization: In large dynamic systems, messages could be generated from multiple components due to reasons ranging from software bugs, system faults, operational activities, security alerts etc. Thus it is very useful to have a global summarized snapshot of messages based on logs. Most works in this area exploit the inter-arrival distribution and co-occurrence of events [2], [16]–[19] to produce summarized correlation between events. These methods are useful when the event-stream is available and possible event-types are known in advance. This limits the applicability to large-systems like ours where event types are unknown along with their generation time-window.

The body of work closest to out work are the works on event-summarization. However, there are some fundamental differences in our system: (i) we do not have a readily available event-stream, instead, our observables are log-messages, (ii) the event-types are latent variables not known in advance and all we observe are message streams, (iii) the time-boundaries of different latent-events is a learning objective, and (iv) since we are dealing with large system with multiple components where different fault-types are correlated, the same message could be generated for different root-causes (real-world events).

Apart from the above, a recent paper [20] which uses deep learning models for anomaly detection in message logs by modelling logs as a natural language sequence is also worth a mention.

II. Problem Statement and Preliminaries

Before we describe our problem statement, we first explain the notion of messages in the context of our work.

Message: In our work, messages generated by different network elements are one of two types: *syslog texts* in the form of raw-texts, and *alarms*.

1) *Syslog texts:* These are raw-textual messages sent by software components from different elements to a logging server. Raw syslog data fields include

[2] Note that manual inference of event signatures is not scalable; we did this for the purpose of validation.

timestamp, source, and message text. Since the number of distinct messages are very large and many of them have common patterns, it is often useful [2], [6]–[8] to decompose the message text into two parts: an *invariant* part called template, and *parameters* associated with template. For example, a syslog message "`service wqffv failed due to connection failure to IP address a.b.c.d using port 8231`" would reduce to template "`service wqffv failed due to connection failure to IP address * using port *`." There are many existing methods to extract such templates [1], [2], ranging from tree-based methods to NLP based methods. In our work, we have a template-extraction pre-processing step before applying our methods. We also say *message* to simply mean the extracted templates.

2) *Alarms:* Network alarms are indication of faults and each alarm type refers to the specific fault condition in a network element. Each alarm has a unique name and the occurrences are also tagged with timestamps. In this work, we view each alarm as a message. Note that, since each alarm has a unique name/id associated with it, we do not pre-process alarms before applying our methods. Example of alarms are `mmscRunTImeError`, `mmscEAIFUnavailable` sent from a network service named MMSC.

Problem Statement: We are given a data set $\mathcal{D}$ consisting of messages generated by error events in a large distributed data-center network. We assume that the messages are generated in the time interval $[0, T]$. The set of messages in the data set come from discrete and finite set $\mathcal{M}$.

We use the term message to mean either a template extracted from a message or an alarm-id. Each message has a timestamp associated with it, which indicates when the message was generated. Suppose that an event e started occurring at time S_e and finished at time F_e. In the interval of time $[S_e, F_e]$, event e will generate a mixture of messages from a subset of $\mathcal{M}$, which we will denote by $\mathcal{M}_e$. In general, an event can occur multiple times in the data set. If an event e occurs multiple times in the data set, then each occurrence of the event will have start and finish times associated with it.

As noted before, for simplicity, we will say event to mean an event-class and occurrence of an event to mean an instance from the class. An event e is characterized by its message set $\mathcal{M}_e$ and the probability distribution with which messages are chosen from $\mathcal{M}_e$, which we will denote by $p^{(e)}$, i.e., $p_m^{(e)}$ denotes the probability that event e will generate a message $m \in \mathcal{M}_e$. For compactness of notation, one can simply think of $p^{(e)}$ as being defined over the entire set of messages $\mathcal{M}$, with $p_m^{(e)} = 0$ if $m \notin \mathcal{M}_e$. Thus, $p^{(e)}$ fully characterizes the event e and can be viewed as the signature of the event. We assume that the support set of messages for two different events are not identical.

It is important to note that the data set simply consists of messages from the set $\mathcal{M}$; there is no explicit information about the events in the data set, i.e., the event information is latent. The goal of the paper is to solve the following inference problem: from the given data set $\mathcal{D}$, identify the set of events that generated the messages in the data set, and for each instance of event, identify when it started and finished. In other words, the output of the inference algorithm should contain the following information:

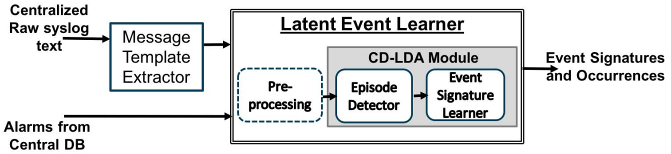


Fig. 1. Figure showing the machine-learning pipeline. Our main contribution is in "Latent Event Learner" module, specifically proposing the CD-LDA algorithm.

- The number of E events which generated the data set.
- The signatures of these events: $p^{(1)}, p^{(2)}, \ldots, p^{(E)}$.
- For each event $e \in \{1, 2, \ldots, E\}$, the number of times it occurred in the data set and, for each occurrence, its start and finish times.

Notations: We use the notation $X_i \in \mathcal{M}$, for the i^{th} message. Also, let t_i be the timestamp associated with the i^{th} message. Thus the data set $\mathcal{D}$ can be characterized by tuples $(X_1, t_1), (X_2, t_2), \ldots (X_n, t_n)$ of n data points.

Machine-Learning Pipeline: In Figure 1, we show the machine-learning pipeline for completeness. This paper focuses on the module "Latent Event Learner" which has data-processing step followed by the key proposed algorithm in the paper, namely CD-LDA algorithm which we describe in Section III. Syslog texts require more pre-processing while alarms do not. We have shown the two types of messages in the pipeline figure, but for the purposes for developing an algorithm, in the rest of the paper, we only refer to messages without distinguishing between them.

III. Algorithm CD-LDA

We now present our solution to this problem which we call CD-LDA (Change-point Detection-Latent Dirichlet Allocation). The key novelty in the paper is the connection that we identify between event identification in our problem and topic modeling in large document data sets, a problem that has been widely studied in the natural language processing literature. In particular, we process our data set into a form that allows us to use a widely-used algorithm called LDA to solve our problem. In standard LDA, we are given multiple documents, with many words in each document. The goal is to identify the mixture of latent topics that generated the documents, where each topic is identified with a collection of words and a probability distribution over the words. Our data set has similar features: we have events (which are the equivalents of topics) and messages (which are the equivalents of words) which are generated by the events. However, we do not have a concept of documents. A key idea in our paper is to divide the data set into smaller data sets, each of which will be called an episode. The episodes will be the equivalents of documents in our problem. We do this using a technique called non-parametric change-point detection.

Now we describe the concept of an episode. An episode is an interval of time over which the same set of events occur i.e. there is no event-churn, and at time instants on either side of the interval, the set of events that occur are different from the set of events in the episode. Thus, we can divide our data set of events such that no two consecutive episodes have the same set of events. We present an example to clarify the concept of an episode. Suppose the duration of the message data set $T = 10$. Suppose event 1 occurred from time 0 to time 5, event 2 occurred from time 4 to time 8, and event 3 occurred from time 5 to time 10. Then there are four episodes in this data set: one in the time interval $[0, 4]$ where only one event occurs, one in the time interval $[4, 5]$ where events $1, 2$ occur, one in the time interval $[5, 8]$ where events $2, 3$ occur and finally one in $[8, 10]$ where only event 3 occurs. We assume that between successive episodes, at most one new event starts or one existing event ends.

We use change-point detection to identify episodes. To understand how the change-point detection algorithm works, we first summarize the characteristics of an episode:

- An episode consists of a mixture of events, and each event consists of a mixture of messages.
- Since neighboring episodes consist of different mixtures of events, neighboring episodes also contain different mixtures of messages (due to our assumption that different events do not generate the same set of messages).
- Thus, successive episodes contain different message distributions and therefore, the time instances where these distributions change are the episode boundaries, which we will call *change points*.
- In our data set, the messages contain time stamps. In general, the inter-arrival time distributions of messages are different in successive episodes, due to the fact that the episodes represent different mixtures of events. This fact can be further exploited to improve the identification of change points.

Based on our discussion so far in this section, CD-LDA has two-phases as follows:

Change-point

1) *detection:* In this phase, we detect the start and end time of each episode. In other words, we identify the time-points where a new event started or an existing event ended. This phase is described in detail in Section III-A.
2) *Applying LDA:* In this phase, we show that, once episodes are known, LDA based techniques can be used to solve the problem of computing message distribution for each event. Subsequently, we can also infer the occurrence times for each event. This phase along with the complete algorithm is described in Section III-B.

A. Change-Point Detection

Suppose we have n data points and a known number of change points k. The data points between two consecutive change points are drawn i.i.d from the same distribution.[3] In the inference problem, each data point could be a possible change point. A naive exhaustive search to find the k best locations would have a computational complexity of $O(n^k)$. Nonparametric approaches to change-point detection aim to solve this problem with much lower complexity even when the number of change points is unknown and there are few assumptions on the family of distributions, [4], [21], [22].

The change point detection algorithm we use is hierarchical in nature. This is inspired by the work in [4]. Nevertheless our algorithm has certain key differences as discussed in section III-C1. It is easier to understand the algorithm in the setting of only one change point, i.e., two episodes. Suppose that τ is a candidate change point among the n points. The idea is to measure the change in distribution between the points to the left and right of τ. We use the TV distance between the empirical distributions estimated from the points to the left and right of the candidate change point τ. In our context the TV distance between two probability mas functions p and q is given by one half the $L1$ distance $0.5||p - q||_1$. This is maximized over all values of τ to estimate the location of the change point. If the distributions are sufficiently different in the two episodes the TV distance between the empirical distributions is expected to be highest for the correct location of the change point in comparison to any other candidate point τ (we rigorously prove this in the proof Theorem 1, 2).

Further, we also have different inter-arrival times for messages in different episodes. Hence we use a combination of TV distance and mean inter-arrival time as the metric to differentiate the two distributions[4] We denote this metric by $\widehat{D}(l)$.

$$\widehat{D}(l) = \|\widehat{p}_L(l) - \widehat{p}_R(l)\|_1 + |\widehat{\mathbb{E}}S_L(l) - \widehat{\mathbb{E}}S_R(l)|, \tag{1}$$

where $\widehat{p}_L(l)$, $\widehat{p}_R(l)$ are empirical estimates of message distributions to the left and right l and $\widehat{\mathbb{E}}S_L(l)$, $\widehat{\mathbb{E}}S_R(l)$ are empirical estimates of the mean inter-arrival time to the left and right of l, respectively. The empirical distributions $\widehat{p}_L(l)$, $\widehat{p}_R(l)$ have M components. For each $m \in \mathcal{M}$, we can write

$$\widehat{p}_{L,m}(l) = \frac{\sum_{i=1}^{l-1} \mathbb{1}\{X_i = m\}}{l} \tag{2}$$

$$\widehat{p}_{R,m}(l) = \frac{\sum_{i=l}^{n} \mathbb{1}\{X_i = m\}}{n-l}. \tag{3}$$

The mean inter-arrival time $\widehat{\mathbb{E}}S_L(l)$ and $\widehat{\mathbb{E}}S_L(l)$ are defined as

$$\widehat{\mathbb{E}}S_L(l) = \frac{\sum_{i=1}^{l-1} \Delta t_i}{l} \tag{4}$$

$$\widehat{\mathbb{E}}S_R(l) = \frac{\sum_{i=l}^{n} \Delta t_i}{n-l}. \tag{5}$$

[3] The i.i.d. assumption is not always true in practice as messages could be sparser in time in the beginning of an event. Indeed, the algorithms developed in this work does not rely on the i.i.d. assumption, however, the assumption allows us to prove useful theoretical guarantees

[4] One can potentially use a weighted combination of the TV distance and mean inter-arrival time as a metric with the weight being an hyper parameter. While the unweighted metric performs well in out real-life datasets, it is an interesting future direction of research to understand how to optimally choose a weighted combination in general.

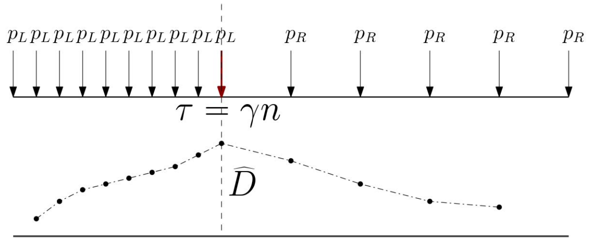

p_L p_L p_L p_L p_L p_L p_L p_L p_L p_L
p_R p_R p_R p_R p_R
$\tau = \gamma n$
$\widehat{D}$

Fig. 2. Example change point with two episodes.

We sometimes write $\widehat{D}(l)$ as $\widehat{D}(\widetilde{\gamma}n)$, where the argument $l = \widetilde{\gamma}n$. Symbol $\widetilde{\gamma}$ denotes the index l as a fraction of n and it can take n discrete values between 0 to 1. $\mathbb{1}\{A\}$ takes value 1 only when event A occurs and 0 otherwise.

Algorithm 1 describes the algorithm in the one change point case. To make the algorithm more robust, we declare a change point only when the episode length is at least αn and the maximum value of the metric (1) is at least δ.

Let us consider a simple example to illustrate the idea of change-point detection with one change-point. Suppose we have a sequence of messages with unequal inter-arrival times as shown in Fig. 2. All the messages are the same, but the first half of the messages arrive at a rate higher than the second half of the messages. In this scenario, our metric reduces to the difference in the mean inter-arrival times between the two episodes. So, $\widehat{D}(l) = |\widehat{\mathbb{E}}S_L(l) - \widehat{\mathbb{E}}S_R(l)|$. The function $\widehat{D}$ in terms of data point l for this example is shown in Fig 2. As we show later in section III-C, the shape of $\widehat{D}$ will be close to the following when the number of samples is large: $\widehat{D}$ will be increasing to the left of change point $\tau = \gamma n$, attain its maximum at the change point and decrease to the right.

Algorithm 1 Change Point Detection With One Change Point

```
Input: parameter $\delta > 0, \alpha > 0$.
Output: changept denoting whether a change point exists
    and the location of the change point $\tau$.
Find $\tau \in \arg\max_l \widehat{D}(l)$
if $\widehat{D}(\tau) > \delta$ and $\alpha n < \tau < 1 - \alpha n$ then
    return $changept = 1$, $\tau$.
else
    return $changept = 0$.
```

Next, we consider the case of multiple change points. When we have multiple change points, we apply Algorithm 1 hierarchically until we cannot find a change point. Algorithm 2 CD$(\mathcal{D}, \alpha, \delta)$ is presented below.

Algorithm 2 CD$(\mathcal{D}, \alpha, \delta)$

```
Input: data points $\mathcal{D}$, minimum value of TV distance $\delta$,
    minimum episode length $\alpha$.
Output: Change points $\tau_1, \ldots, \tau_k$.
Run FINDCHANGEPT$(1, n)$.
procedure FINDCHANGEPT$(L, H)$
    changept, $\tau$ $\leftarrow$ ALGORITHM 1
        $(X_L, X_{L+1}, \ldots, X_H, \alpha, \delta)$.
    if changept exists then
        $\tau_l \leftarrow$ FINDCHANGEPT$(L, \tau)$,
        $\tau_h \leftarrow$ FINDCHANGEPT$(\tau, H)$.
        return $\{\tau_l, \tau, \tau_h\}$
    else
        return
```

The above algorithm tries to detect a single change point first, and if such a change point is found, it divides the data set into two parts, one consisting of messages to the left of the change point and the other consisting of messages to the right of the change point. The single change-point detection algorithm is now applied to each of the two smaller datasets. This is repeated recursively till no more change points are detected.

1) Discussion: What Metric for Change Point Detection?: We have used the TV distance between two distributions to estimate the change point in metric 1. One can also use other distance measures like the l_2 distance, the Jensen-Shannon (J-S) distance, the Hellinger distance, or the metric used in [4]. The metric used in [4] is shown to be an unbiased estimator of the l_2 distance for categorical data in Appendix J of the supplementary material. We argue that for our data set, all of the above distances give similar performance. Our data set has 97m points and 39330 types of messages. In the region where the number of data points is much more than the dimension of the distribution, estimating a change point through all of the above metrics give order wise similar error rate. We show this through synthetic data experiments since we do not know the ground truth to compute the error in estimating the change point in the real dataset.

TABLE I
COMPARISON BETWEEN DIFFERENT METRICS FOR CHANGE POINT

$\|\|p - q\|\|_1 = 0.1$					
Metric	TV	l_2	Unbiased l_2, [4]	J-S	Hellinger
$\|\widehat{\tau}/n - 0.5\|$	0.021	0.030	0.025	0.030	0.030

We present one such experiment with a synthetic dataset here. Consider two distributions p and q whose support set consists of 10 points. We assume that p is the uniform distribution, while $q[1] = q[2] = \ldots = q[5] = 0.09$, and $q[6] = q[7] = \ldots = q[10] = 0.11$. There are $n = 25000$ data points. The first half of the data points are independently drawn from p and the second half of the data points are drawn from q. Table I shows the absolute error in estimating the change point at $0.5n$ to be of the order of 10^{-2} for all the distance metrics.

We test the l_1 distance metric on real data and we show in section IV-B that it is satisfactory. Since we do not know the ground truth, we take a small part of the real data set where we can can visually identify the approximate location of the major change points. The change point algorithm with l_1 metric correctly estimates these locations.

A graph based change point detection algorithm in [23] can be adapted to our problem such that the metric computation is linear in the number of messages. We can do this if we consider a graph with nodes as the messages and edges connecting message of the same type. But, one can show that the metric in [23] is not consistent for this adaptation.

B. Latent Dirichlet Allocation

In the problem considered in this paper, each episode can be thought of as a document and each message can be thought of as a word. Like in the LDA model where each topic is latent, in our problem, each event is latent and can be thought of as a distribution over messages. Unlike LDA-based document modeling, we have time-stamps associated with messages, which we have already used to extract episodes from our data set. Additionally, this temporal information can also be used in a Bayesian inference formulation to extract events and their signatures. However, to make the algorithm simple and computationally tractable, as in the original LDA model, we assume that there is no temporal ordering to the episodes or messages within the episodes. Our experiments suggest that this choice is reasonable and leads to very good practical results. However, one can potentially use the temporal information too as in [24], [25], and this is left for future work.

If we apply the LDA algorithm to our episodes, the output will be the event signatures $p^{(e)}$ and episode signatures $\theta^{(\mathcal{E})}$, where an episode signature is a probability distribution of the events in the episode. In other words, LDA assumes that each message in an episode is generated by first picking an event within an episode from the episode signature and then picking a message from the event based on the event signature.

For our event mining problem, we are interested in event signatures and finding the start and finish times of each occurrence of an event. Therefore, the final step (which we describe next) is to extract the start and finish times from the episode signatures.

Putting it All Together: In order to detect all the episodes in which the event e occurs prominently, we proceed as follows. We collect all episodes $\mathcal{E}$ for which the event occurrence probability $\theta_e^{(\mathcal{E})}$ is greater than a certain threshold $\eta > 0$. We declare the start and finish times of the collected episodes as the start and finish times of the various occurrences of the event e. If an event spans many contiguous episodes, then the start time of the first episode and the end time of the last contiguous episode can be used as the start and finish time of this occurrence of the event. However, for simplicity, this straightforward step in not presented in the detailed description of the algorithm in ALGORITHM 3.

Algorithm 3 CD-LDA($\mathcal{D}, \alpha, \delta, \eta$)

```
Input: data points $\mathcal{D}$, threshold of occurrence of an event in an episode $\eta$, the minimum value of TV distance $\delta$, minimum episode length $\alpha$.
Output: Event signatures $p^{(1)}, p^{(2)}, \dots, p^{(E)}$, Start and finish time $S_e, F_e$ for each event $e$.
Change points $\tau_1, \dots, \tau_k \leftarrow$ CD($\mathcal{D}, \alpha, \delta$). Episode $\mathcal{E}_i \leftarrow \{X_{\tau_{i-1}}, \dots, X_{\tau_i}\}$ for $i = 1$ to $k+1$.
$p^{(1)}, \dots, p^{(E)}; \theta^{(\mathcal{E}_1)}, \dots, \theta^{(\mathcal{E}_{k+1})} \leftarrow$LDA($\mathcal{E}_1, \dots, \mathcal{E}_{k+1}$)
Consider event $e$. $\mathcal{G}_e \leftarrow$ Set of all episodes $\mathcal{E}$ such that $\theta_e^{(\mathcal{E})} > \eta$. $S_e, F_e \leftarrow$ start and finish times of all episodes in set $\mathcal{G}_e$.
```

Remark 2: We use the Gibbs sampling based inference from [26] on the LDA model. For a discussion on the comparison between different inference methods ([3], [26]–[30]) for the LDA model, see Appendix A in the supplementary material.

Note that the LDA algorithm requires an input for the number of events E. However, one can run LDA for different values of E and choose the one with maximum likelihood [3]. Hence E need not be assumed to be an input to CD-LDA. One can also use the Hierarchical Dirichlet Process (HDP) algorithm [31] which is an extension of LDA and figure out the number of topics from the data. In our experiments, we use the maximum likelihood approach to estimate the number of events. This is explained in section IV-C1.

C. Analysis of CD

As mentioned earlier, the novelty in the CD-LDA algorithm lies in the connection we make to topic modeling in document analysis. In this context, our key contribution is an efficient algorithm to divide the data set of messages into episodes (documents). Once this is done, the application of the LDA of episodes (documents), consisting of messages (words) generated by events (topics) is standard. Therefore, the correctness and efficiency of the CD part of the algorithm will determine the correctness and efficiency of CD-LDA as a whole. We focus on analyzing the CD part of the algorithm in this section. Due to space limitations, we only present the main results here, and the proofs can be found in the supplementary material.

Section III-C1 shows that the computational complexity of CD algorithm is linear in the number of data points. Section III-C2 contains the asymptotic analysis of the CD algorithm while section III-C3 has the finite sample results.

1) Computational Complexity of CD: In this section we discuss the computational complexities of Algorithm 1 and Algorithm 2. We will first discuss the computational complexity of detecting a change point in case of one change point. Algorithm 1 requires us to compute $\arg\max_l \widehat{D}(l)$ for $1 \leq l \leq n$. From the definition of $\widehat{D}(l)$ in (1), we only need to compute the empirical probability estimates $\widehat{p}_L(l)$, $\widehat{p}_R(l)$, and the empirical mean of the inter arrival time $\widehat{\mathbb{E}}S_L(l)$, $\widehat{\mathbb{E}}S_R(l)$ for every value of l between 1 to n.

We focus on the computation of $\widehat{p}_L(l)$, $\widehat{p}_R(l)$. Consider any message m in the distribution. For each m, we can compute $\widehat{p}_{L,m}(l)$, $\widehat{p}_{R,m}(l)$ in $O(n)$ for every value of l by using neighbouring values of $\widehat{p}_{L,m}(l-1)$, $\widehat{p}_{R,m}(l-1)$.

$$\begin{aligned} \widehat{p}_{L,m}(l) &= \frac{(l-1)\widehat{p}_{L,m}(l-1) + \mathbb{1}\{X_{l-1} = m\}}{l}, \\ \widehat{p}_{R,m}(l) &= \frac{(n-l+1)\widehat{p}_{R,m}(l-1) - \mathbb{1}\{X_{l-1} = m\}}{n-l} \end{aligned} \tag{6}$$

The computation of $\widehat{\mathbb{E}}S_L(l)$, $\widehat{\mathbb{E}}S_R(l)$ for every value of l from 1 to n is similar.

Performing the above computations for all M messages, results in a computational complexity of $O(nM)$. In the case of k change points, it is straightforward to see that we require $O(nMk)$ computations. In much of our discussion, we assume M and k are constants and therefore, we present the computational complexity results in terms of n only.

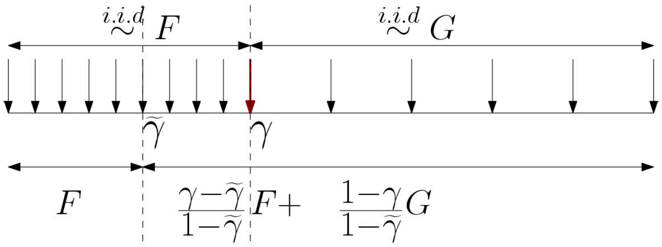

$\overset{i.i.d}{\sim} F$
$\overset{i.i.d}{\sim} G$
$\widetilde{\gamma}$
γ
F
$\frac{\gamma-\widetilde{\gamma}}{1-\widetilde{\gamma}}F+ \; \frac{1-\gamma}{1-\widetilde{\gamma}}G$

Fig. 3. Consistency with two change points.

Related Work: Algorithm 2 executes the process of determining change points hierarchically. This idea was inspired by the work in [4]. However, the metric $\widehat{D}$ we use to detect change points is different from that of [4]. The change in metric necessitates a new analysis of the consistency of the CD algorithm which we present in the next subsection. Further, for our metric, we are also able to derive sample complexity results which are presented in a later subsection.

2) The Consistency of Change-Point Detection: In this section we discuss the consistency of the change-point detection algorithm, i.e., when the number of data points n goes to infinity one can accurately detect the location of the change points. In both this subsection and the next, we assume that the inter-arrival times of messages within each episode are i.i.d., and are independent (with possibly different distributions) across episodes.

Theorem 1: For $\widetilde{\gamma} \in (0,1)$, $D(\widetilde{\gamma}) = \lim_{n\to\infty} \widehat{D}(\widetilde{\gamma}n)$ is well-defined and $D(\widetilde{\gamma})$ attains its maximum at one of the change points if there is at least one change point.

Proof: We only provide a proof for the single change point case here and refer the interested reader to Appendix B in the supplemental material for the proof of the multiple change point case. Let the change point be at index τ. The location of the change point is determined by the point where $\widehat{D}(l)$ maximizes over $1 < l < n$. We will show that when n is large the argument where $\widehat{D}(l)$ maximizes converges to the change point τ.

Suppose all the points X to the left of the change point τ are chosen i.i.d from a distribution F and all the points from the right of τ are chosen from a distribution G, where $F \neq G$. Also, say the inter-arrival times Δt_i's are chosen i.i.d from distribution F_t and G_t to the left and right of change point τ, respectively. Let $l = \widetilde{\gamma}n$, $0 < \widetilde{\gamma} < 1$ be the index of any data point and $\tau = \gamma n$, the index of the change point.

Case 1 $\widetilde{\gamma} \leq \gamma$: Suppose we consider the value of $\widehat{D}(l) = \widehat{D}(\gamma n)$ to the left of the actual change point, i.e, $l < \tau$ or $\widetilde{\gamma} < \gamma$. The distribution to the left of $\widetilde{\gamma}n$, $\widehat{p}_L(\widetilde{\gamma}n)$, has all the data points chosen from the distribution F. So $\widehat{p}_L(\widetilde{\gamma}n)$ is the empirical estimate for F. On the other hand, the data points to the right of $\widetilde{\gamma}n$ come from a mixture of distribution F and G. $\widehat{p}_R(\widetilde{\gamma}n)$ has $\frac{\gamma-\widetilde{\gamma}}{1-\widetilde{\gamma}}$ fraction of samples from F and $\frac{1-\gamma}{1-\widetilde{\gamma}}$ fraction of samples from G. Figure 3 below explains it pictorially.

So $\widehat{p}_L(l)$ and $\widehat{p}_R$ defined in (3) converges to

$$\widehat{p}_L(l) \to F, \quad \widehat{p}_R(l) \to \frac{\gamma-\widetilde{\gamma}}{1-\widetilde{\gamma}}F + \frac{1-\gamma}{1-\widetilde{\gamma}}G. \tag{7}$$

Similarly, we can say that the empirical mean estimates $\widehat{ES}_L(l)$ and $\widehat{ES}_R(l)$ converge to

$$\widehat{ES}_L(l) \to \mathbb{E}F_t, \quad \widehat{\mathbb{E}S}_R(l) \to \frac{\gamma-\widetilde{\gamma}}{1-\widetilde{\gamma}}\mathbb{E}F_t + \frac{1-\gamma}{1-\widetilde{\gamma}}\mathbb{E}G_t. \tag{8}$$

We can combine (7) and (8) to say that $\widehat{D}(\widetilde{\gamma}n) \to D(\widetilde{\gamma})$ where

$$\begin{aligned}\widehat{D}(\widetilde{\gamma}n) &= \|\widehat{p}_L(\widetilde{\gamma}n) - \widehat{p}_R(\widetilde{\gamma}n)\| + |\mathbb{E}S_L(\widetilde{\gamma}n) - \mathbb{E}S_R(\widetilde{\gamma}n)| \\ &\to D(\widetilde{\gamma}) := \frac{1-\gamma}{1-\widetilde{\gamma}}(\|F-G\|_1 + |\mathbb{E}F_t - \mathbb{E}G_t|). \end{aligned} \tag{9}$$

Note that from the definition of D, $D(\gamma) = \|F-G\|_1 + |\mathbb{E}F_t - \mathbb{E}G_t|$.

Case 2 $\widetilde{\gamma} > \gamma$: Proceeding in a similar way to Case 1, we can show

$$\widehat{D}(\widetilde{\gamma}n) \to D(\widetilde{\gamma}) := \frac{\gamma}{\widetilde{\gamma}}(\|F-G\|_1 + |\mathbb{E}F_t - \mathbb{E}G_t|). \tag{10}$$

From Case 1 and Case 2, we have

$$\begin{aligned}\widetilde{\gamma} \leq \gamma, \quad \widehat{D}(\widetilde{\gamma}n) \to D(\widetilde{\gamma}) &= \frac{1-\gamma}{1-\widetilde{\gamma}}D(\gamma) \\ \widetilde{\gamma} > \gamma, \quad \widehat{D}(\widetilde{\gamma}n) \to D(\widetilde{\gamma}) &= \frac{\gamma}{\widetilde{\gamma}}D(\gamma).\end{aligned} \tag{11}$$

Equation (11) shows that the maximum of $D(\widetilde{\gamma})$ is obtained at $\widetilde{\gamma} = \gamma$. □

3) The Sample Complexity of Change-Point Detection: In the previous subsection, we studied the CD algorithm in the limit as $n \to \infty$. In this section, we analyze the algorithm when there are only a finite number of samples. For this purpose, we assume that the inter-arrival distribution of messages have sub-Gaussian tails.

We say that Algorithm CD is correct if the following conditions are satisfied. Let $\epsilon > 0$ be a desired accuracy in estimation of the change point.

Definition 1: Given $\epsilon > 0$, Algorithm CD is correct if

- *there are change points $0 < \frac{\tau_1}{n} = \gamma_1, \ldots, \frac{\tau_k}{n} = \gamma_k < 1$ and the algorithm gives $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_k$ such that $\max_i |\widehat{\gamma}_i - \gamma_i| < \epsilon$.*
- *there is no change point and $\widehat{D}(\gamma n) < \delta, \forall \gamma \in \{\gamma_1, \ldots, \gamma_k\}$.*

Now we can state the correctness theorem for Algorithm 2. The sample complexity is shown to scale logarithmically with the number of change points.

Theorem 2: Algorithm 2 is correct in the sense of Definition 1 with probability $(1-\beta)$ if

$$n = \Omega\left(\max\left(\frac{\log\left(\frac{2k+1}{\beta}\right)}{\epsilon^2}, \frac{M^{1+c}}{\epsilon^{2(1+c)}}\right)\right),$$

for sufficiently small α, δ, ϵ and for any $c > 0$.

Remark 3: The proof of this theorem uses the method of types and Pinsker's inequality. We present here the proof for the single change point case. Due to space constraints, we move the proof for multiple change points to Appendix D in the supplementary material.

Proof: We first characterize the single change point case in finite sample setting. In order to get the sample complexity, we prove the correctness for Algorithm 1 as per Definition 1 with high probability. Before we go into the proof, we state the assumptions on α, δ, ϵ under which the proof is valid.

- Suppose a change point exists at index γn and the metric $\widehat{D}(\gamma n)$ converges to $D(\gamma)$ at the change point. Then ϵ can only be chosen in following region: ϵ has to be less than the value of the metric at the change point, $\epsilon < D(\gamma)$; ϵ has to be less than the minimum episode length, $\epsilon < \min(\gamma, 1-\gamma)$.
- If a change point exists at index γn, α has to chosen less than the minimum episode length minus ϵ, $\alpha < \min(\gamma, 1-\gamma) - \epsilon$.
- The threshold $\delta < D(\gamma) - \epsilon$.

Under the above assumptions we show that Algorithm 1 is correct as per the Definition 1 with probability at least

$$1 - (6n+4)\exp\left(-\frac{\min(\delta,1)^2\epsilon^2\alpha^2}{512\max(\sigma^2,1)}n + M\log(n)\right).$$

Suppose

$$\widehat{\gamma}n = \arg\max_{\widetilde{\gamma}n} \widehat{D}(\widetilde{\gamma}n).$$

The idea is to upper bound the probability when Algorithm 1 is not correct. From Definition 1 this happens when,

- Given a change point exists at $\gamma \in (0,1)$,

  $$(\widehat{D}(\widehat{\gamma}n) > \delta, |\gamma - \widehat{\gamma}| < \epsilon, \alpha < \widehat{\gamma} < 1-\alpha)^c$$

  occurs. Say the event E_1 denotes $E_1 = \{\widehat{D}(\widehat{\gamma}) > \delta, |\gamma - \widehat{\gamma}| < \epsilon, \alpha < \widehat{\gamma} < 1-\alpha\}$.
- Given a change point does not exist,

  $$\widehat{D}(\widehat{\gamma}) > \delta, \alpha < \widehat{\gamma} < 1-\alpha.$$

  When a change point does not exist we write $\gamma = 0$. Say the event E_2 denotes $E_2 = \{\gamma = 0, \alpha < \widehat{\gamma} < 1-\alpha\}$

So

$$\begin{aligned} &P(\text{Algorithm 1 is NOT correct}) \\ &\qquad \le P(E_1^c|0<\gamma<1) + P(\widehat{D}(\widehat{\gamma}) > \delta|E_2). \end{aligned} \tag{12}$$

We analyze each part in (12) separately.

Case 1: Suppose no change point exists and say all the data points are drawn from the same mutinomial distribution F and all inter-arrival times are generated i.i.d from a distribution F_t. Given event E_2, if $\|\widehat{p}_L(\widehat{\gamma}n) - F\|, \|\widehat{p}_R(\widehat{\gamma}n) - F\|, |\widehat{\mathbb{E}}S_L(\widehat{\gamma}n) - \mathbb{E}F_t|, |\widehat{\mathbb{E}}S_R(\widehat{\gamma}n) - \mathbb{E}F_t|$ are all less than $\delta/4$, then $\widehat{D}(\widehat{\gamma}) < \delta$. So $P(\widehat{D}(\widehat{\gamma}) > \delta|E_2) \le P(\|\widehat{p}_L(\widehat{\gamma}n) - F\| > \delta/4|E_2) + P(\|\widehat{p}_R(\widehat{\gamma}n) - F\| > \delta/4|E_2) + P(|\widehat{\mathbb{E}}S_L(\widehat{\gamma}n) - \mathbb{E}F_t| > \delta/4|E_2) + P(|\widehat{\mathbb{E}}S_R(\widehat{\gamma}n) - \mathbb{E}F_t| > \delta/4|E_2)$. Now, we can use Sanov's theorem followed by Pinsker's inequality to upper bound each of the above terms as

$$\begin{aligned} &P(\widehat{D}(\widehat{\gamma}) > \delta|E_2) \\ &\le (n\widehat{\gamma}+1)^M\exp(-n\delta^2/16) \\ &\quad + ((1-\widehat{\gamma})n+1)^M\exp(-n\delta^2/16) + 2\exp(-\alpha n\delta^2/32\sigma^2) \\ &\quad + 2\exp(-\alpha n\delta^2/32\sigma^2) \\ &\le 4(n+2)^M\exp\left(-n\frac{\alpha\delta^2}{32\max(\sigma^2,1)}\right). \end{aligned} \tag{13}$$

Case 2: Next, we look at the case when a change point exists at γn. Say the messages are drawn from a distribution F to the left of the change point and G to the right of the change point. Also, suppose the inter-arrival time distribution to the left of the change point is F_t and the inter-arrival time distribution to the right is G_t. According to our assumptions, α is chosen such that $\alpha + \epsilon < \gamma < 1 - (\alpha+\epsilon)$. Hence

$$\begin{aligned} &P(E_1^c|0<\gamma<1) \\ &\le P(\widehat{D}(\widehat{\gamma}n) < \delta|0<\gamma<1) \\ &\quad + P(|\widehat{\gamma}-\gamma| > \epsilon|\widehat{D}(\gamma) > \delta, 0<\gamma<1) \\ &\quad + P(\alpha < \widehat{\gamma} < 1-\alpha|\widehat{D}(\gamma) > \delta, |\widehat{\gamma}-\gamma| < \epsilon, 0<\gamma<1). \end{aligned} \tag{14}$$

Given the assumption on α, $P(\alpha < \widehat{\gamma} < 1-\alpha|\widehat{D}(\gamma) > \delta, |\widehat{\gamma}-\gamma| < \epsilon, 0<\gamma<1) = 0$. The rest of the proof deals with upper bounding $P(\widehat{D}(\widehat{\gamma}n) < \delta|0<\gamma<1)$ and $P(|\widehat{\gamma}-\gamma| > \epsilon|\widehat{D}(\gamma) > \delta, 0<\gamma<1)$.

In lemma 1-3 we develop the characteristics of $\widehat{\gamma}$ and $D(\widehat{\gamma})$ when a change point exists at γn. Lemma 1-3 are proved in Appendix E, F of the supplementary material. First, we analyze the concentration of $\widehat{D}(\widetilde{\gamma}n)$ for any value of $\widetilde{\gamma}$ in the Lemma 1.

Lemma 1: $|\widehat{D}(\widetilde{\gamma}n) - D(\widetilde{\gamma})| \le \epsilon$ w.p. at least $1 - 3n\exp\left(-\frac{\epsilon^2\alpha^2}{128\sigma^2}n + M\log(n)\right)$ for all values of $\widetilde{\gamma}$ when $\widehat{D}(\widetilde{\gamma}n)$ is defined.

Lemma 1 shows that the empirical estimate $\widehat{D}(\widetilde{\gamma}n)$ is very close to the asymptotic value $D(\widetilde{\gamma})$ with high probability. Recall that the argument at which $\widehat{D}$ maximizes is $\widehat{\gamma}n$. we next show in Lemma 3 that the value of metric D at $\widehat{\gamma}$ is very close to the value of the D at the change point γ.

Lemma 2: $|D(\gamma) - D(\widehat{\gamma})| < 2\epsilon$ w.p. $1 - 3n\exp\left(-\frac{\epsilon^2\alpha^2}{128\sigma^2}n + M\log(n)\right)$

Finally, in Lemma 3 we show that $\widehat{\gamma}$ is close to the change point γ with high probability.

Lemma 3: $|\widehat{\gamma} - \gamma| < \epsilon$ w.p. $1 - 3n\exp\left(-\frac{\epsilon^2D^2(\gamma)\alpha^2}{512\sigma^2}n + M\log(n)\right)$.

Also, using lemma 2 and assuming that δ is chosen such that $\delta < D(\gamma) - \epsilon$,

$$\begin{aligned} &P(\widehat{D}(\widehat{\gamma}n) < \delta|0<\gamma<1) \\ &\qquad \le P(\widehat{D}(\widehat{\gamma}n) < \delta|0<\gamma<1, |\widehat{D}(\widehat{\gamma}n) - D(\gamma)| < \epsilon) \\ &\qquad\quad + P(|\widehat{D}(\widehat{\gamma}n) - D(\gamma)| > \epsilon) \\ &\qquad \le 0 + 3n\exp\left(-\frac{\epsilon^2\alpha^2}{128\sigma^2}n + M\log(n)\right) \end{aligned} \tag{15}$$

Lemma 3 gives a bound on $P(|\widehat{\gamma}-\gamma| > \epsilon|\widehat{D}(\gamma) > \delta, 0<\gamma<1)$. Using this along with (15) in (14) we have

$$\begin{aligned} &P(E_1^c|0<\gamma<1) \\ &\quad \le 3n\exp\left(-\frac{\epsilon^2\alpha^2}{128\sigma^2}n + M\log(n)\right) \\ &\qquad\qquad + 3n\exp\left(-\frac{\epsilon^2D^2(\gamma)\alpha^2}{512\sigma^2}n + M\log(n)\right). \end{aligned} \tag{16}$$

Finally, putting together (13) and (16) into (12), we have

$$P(\text{Algorithm 1 is NOT correct}) \leq (6n+4)\exp(-\frac{\min(\delta,1)^2\epsilon^2\alpha^2}{512\max(\sigma^2,1)} + M\log(n+2)) \quad (17)$$

Ignoring the constants in (17), we can derive the sample complexity result for the one change point case. □

IV. Evaluation With Real Datasets

We now present our experimental results with real data sets from large operational network. The purpose of experiments is three-fold. First, we wish to compare our proposed CD-LDA algorithm with other techniques proposed (adapted to our setting) in the literature. Second, we want to validate our results against manual expert-derived event signature for a prominent event. Third, we want to understand the scalability of our method with respect to very large data sets.

Datasets Used: We use two data sets: one from a legacy network of physical elements like routers, switches etc., and another from a recently deployed virtual network function (VNF). The VNF dataset is used to validate our algorithm by comparing with expert knowledge. The other one is used to show that our algorithm is scalable, i.e., it can handle large data sets and it is less sensitive to the hyper parameters.

- **Dataset-1:** This data set consists of around 97 million raw syslog messages collected from 3500 distinct physical network elements (mostly routers) from a nationwide operational network over a 15-day period in 2017. There are 39330 types of messages.
- **Dataset-2:** The second data set consists of around $728,000$ messages collected from 285 distinct physical/virtual network elements over a 3 month period from a newly deployed *virtual network function* (VNF) which is implemented on a data-center using multiple VMs.

We implemented the machine-learning pipeline as shown in Figure 1. The main algorithmic component in the figure shows CD-LDA algrothm; however, for the purpose of comparison, we also implemented two additional algorithms described shortly. Before the data is applied to any of the algorithms, there are two-steps, namely, Template-extraction (in case of textual syslog data) and pre-processing (for both syslog and alarms). These steps are described in Appendix I in the supplementary material.

A. Benchmark Algorithms

We compare CD-LDA with the following algorithms.

1) Algorithm B: A Bayesian Inference Based Algorithm: We consider a fully Bayesian inference algorithm to solve the problem. A Bayesian inference algorithm requires some assumptions on the statistical generative model by which the messages are generated. Our model here is inspired by topic modelling across documents generated over multiple eras [24]. Suppose that there are E events which generated our data set, and event e has a signature $p^{(e)}$ as mentioned earlier. The generative model for generating each message is assumed to be as follows.

- To generate a message, we first assume that an event $e \in [1, 2, \ldots, E]$ is chosen with probability P_e.
- Next, a message m is chosen with probability $p_m^{(e)}$.
- Finally, a timestamp is associated with the message which is chosen according to a beta distribution $\beta(a_e, b_e)$, where the parameters of the beta distribution are distinct for different events.

The parameters of the generative model $P_e, p_m^{(e)}, a_e, b_e$ are unknown. As in standard in such models, we assume a prior on some of these parameters. Here, as in [24], we assume that there is a prior distribution on q over the space of all possible P and a prior r over the space of all possible $p^{(e)}$. The prior r is assumed to be independent of e. Given these priors, the Bayesian inference problem becomes a maximum likelihood estimation problem, i.e.,

$$\max_{a_e, b_e, {p^{(e)}}_e, P} \mathbb{P}_{q,r}(\mathcal{D}|P, \{p^{(e)}\}_e).$$

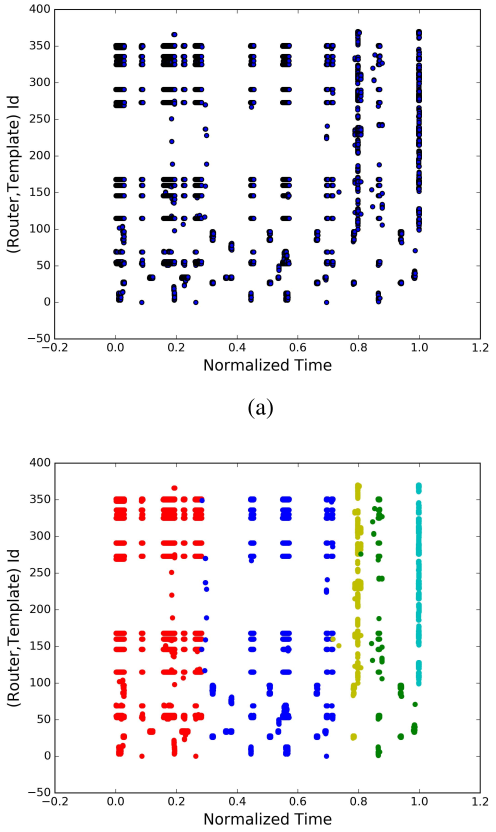
We use Gibbs sampling to solve the above maximization problem. There are two key differences between Algorithm B and proposed CD-LDA. CD-LDA first breaks up the datasets into smaller episodes whereas Algrothm-B uses prior distributions (the beta distributions) to model the fact that different events happen at different times. We show that, such an algorithm works, but the inference procedure is dramatically slow due to additional parameters to infer $\{a_e, b_e\}_e$.

2) Algorithm C: A Graph-Clustering Based Algorithm: For the purposes of comparison, we will also consider a very simple graph-based clustering-based algorithm to identify events. This algorithm is inspired from graph based clustering used in event log data in [32]. The basic idea behind the algorithm is as follows: we construct a graph whose nodes are the messages in the set $\mathcal{M}$. We divide the continuous time interval $[0, T]$ into T/w timeslots, where each timeslot is of duration w. For simplicity, we will assume that T is divisible by w. We draw an edge between a pair of nodes (messages) and label the edge by a distance metric between the messages, which roughly indicates the likelihood with which two messages are generated by the same event. Then, any standard distance-based clustering algorithm on the graphs will cluster the messages into clusters, and one can interpret each cluster as an event. Clearly, the algorithm has the following major limitation: it can detect $\mathcal{M}_e$ for an event e and not $p^{(e)}$. In some applications, this may be sufficient. Therefore, we consider this simple algorithm as a possible candidate algorithm for our real data set.

We now describe how the similarity metric is computed for two nodes i and j. Let N_i be the number of timeslots during which a message i occurs and let N_{ij} be the number of timeslots during which both i and j appear in the same timeslot. Then, the distance metric between nodes i and j is defined as

$$\rho_{ij} = 1 - \frac{N_{ij}}{N_i + N_j - N_{ij}}.$$

Thus, a smaller ρ_{ij} indicates that i and j co-occur frequently. The idea behind choosing this metric is as follows: messages generated by the same event are likely to occur closer together



(a)

(b)

Fig. 4. (a) Top panel shows scatter plot of different message-ids over the period of comparison and (b) bottom panel shows the episodes detected by CD phase of Algorithm CD-LDA.

in time. Thus, ρ_{ij} being small indicates that the messages are more likely to have been generated by the same event, and thus are closer together in distance.

B. Results: Comparison With Benchmark Algorithms

For the purposes of this section only, we consider a smaller slice of data from Dataset-1. Instead of considering all the 97 million messages, we take a small slice of 10,000 messages over a 3 hour duration from 135 distinct routers. Let us call this data set $\mathcal{D}_s$. There are two reasons for considering this smaller slice. Firstly, it is easier to visually observe the ground truth in this small data set and verify visually if CD-LDA is giving us the ground truth. We can also compare the results from different methods with this smaller data set. Secondly, as we show later in this section, the Bayesian inference Algorithm-B is dramatically slow and so running it over the full dataset is not feasible. Nevertheless, the smaller dataset allows us to validate the key premise behind our main algorithm, i.e., the decomposition of the algorithm into the CD and LDA parts.

Applying CD-LDA on This Dataset Slice: Figure 4a shows the data points in x-axis and the message-ids on y-axis. Figure 4b shows the 5 episodes after the CD part of CD-LDA, where we chose $\alpha = 0.1$ and $\delta = 0.5$. For the LDA part, instead of specifying the number of events, we use maximum likelihood to find the optimal number of events and based on this, the number of events was found to be 2.

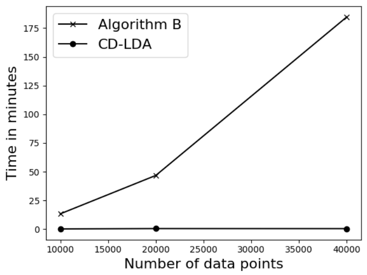


Fig. 5. Time performance: CD-LDA vs Algrithm B.

We next compare event signatures produced by CD-LDA with Algorithm B and Algorithm C.

CD-LDA Versus Algorithm B: For all unknown distributions, we assume a uniform prior in Algorithm B. Algorithm B is run with input number of events as $2, 3, 4, 5$. It turns out that, with 3 events the algorithm converges to a solution which has maximum likelihood. However, upon clustering the event signatures $p^{(e)}$ based on TV-distance between the event signatures, we find only two events. *The maximum TV-distance between the events signatures found from the two algorithms is* 0.068. Hence, we can conclude that the event signatures found by both the algorithms are very similar.

Despite the fact that Algorithm B using fewer hyper-parameters, it is not fast enough to run on large data sets. Figure 5a shows the time taken by CD-LDA and Algorithm B as we increase the size of the data set from $10,000$ to $40,000$ points. With $40,000$ data points and 12 events as input Algorithm B takes 3 hours whereas CD-LDA only takes 26.57 seconds. Clearly, we cannot practically run Algorithm B on large data sets with millions of points.

CD-LDA Versus Algorithm C: In this section we compare CD-LDA versus algorithm C on data set $\mathcal{D}_s$. Algorithm C can produce the major event clusters as CD-LDA, but does not provide the start and end time for the events. We form the co-occurrence graph for Algorithm C with edge weight as described in section IV-A2 and nodes as messages which occur more than at least 5 times in the data set $\mathcal{D}_s$. All the edges with weight more than 0.6 are discarded and we run a clique detection algorithm in the resulting graph.

We quantitatively compare the event signature $\mathcal{M}_e$ of the top two cliques found by Algorithm C with those found by CD-LDA. Suppose that message sets identified by Algorithm C for the two events are $\mathcal{M}_{e1}$ and $\mathcal{M}_{e2}$ respectively. Message sets (messages with probability more than 0.007) identified by CD-LDA for the two events are denoted by $\mathcal{S}_{e1}$ and $\mathcal{S}_{e2}$. We can now compute the Jaccard Index between the two sets.

$$\frac{|\mathcal{M}_{e1} \cap \mathcal{S}_{e1}|}{|\mathcal{M}_{e1} \cup \mathcal{S}_{e1}|} = 0.73 \quad \frac{|\mathcal{M}_{e2} \cap \mathcal{S}_{e2}|}{|\mathcal{M}_{e2} \cup \mathcal{S}_{e2}|} = 0.68.$$

Since the full Bayesian inference (Algorithm B) agrees with CD-LDA closely, we can conclude that Algorithm C gets a large fraction of the messages associated with the event correctly. However, it also misses a significant fraction of the

TABLE II

EVENTS GENERATED BY CD-LDA AND THE CONSTITUENT MESSAGES IN DECREASING ORDER OF PROBABILITY. EVENT 8 MATCHES WITH EXPERT PROVIDED EVENT SIGNATURE

Event 1	Event 2	…	Event 8
mmscRuntimeError	ISCSI_multipath		SNMP_sshd
SUDBConnectionDown	Logmon_contrail		SNMP_crond
SocketConnectionDown	VRouter-Vrouter		SNMP_AgentCheck
SUDBConnectionUp	LogFile_nova		SNMP_ntpd
SocketConnectionUp	SUDBConnectionDown		SNMP_CPU
mmscEAIFUnavailable	IPMI		SNMP_Swap
bigipServiceUp	bigipServiceDown		SNMP_Mem
bigipServiceDown	bigipServiceUp		SNMP_Filespace
SNMP_Mem	HW_IPMI		Ping_vm

messages, and additionally Algorithm C does not provide any information about start and end times of the events. Also, the events found are sensitive to the threshold for choosing the graph edges, something we have carefully chosen for this small data set.

C. Results: Comparison With Expert Knowledge and Scalability

Validation by Comparing With Manual Event Signature: The intended use-case of our methodology is for learning events where the scale of data and system does not allow for manual identification of event signatures. However, we still wanted to validate our output against a handful of event signatures inferred manually by domain experts. For the purpose of this section, we ran CD-LDA for Dataset-2 which is for an operational VNF. For this data set, an expert had identified that a known service issue had occurred on two dates: 11-Oct and 26-Nov, 2017. This event generated messages with Ids `Ping_vm`, `SNMP_AgentCheck`, `SNMP_ntpd`, `SNMP_sshd`, `SNMP_crond`, `SNMP_Swap`, `SNMP_CPU`, `SNMP_Mem`, `SNMP_Filespace`.

We ran CD-LDA on this data set with parameters $\alpha = 0.01$ and $\delta = 0.1$. We chose 10 events for the LDA phase by looking at the likelihood computed using cross validation for different number of topics. See section IV-C1 for details of the maximium likelihood approach. Table II shows the events detected by CD-LDA in decreasing order of probability. Also, top 9 messages are listed for each event. Indeed, we note that *Event* 8 *resembles the expert provided event. CD-LDA detected this event as having occurred from 2017-10-08 17:35 to 2017-10-17 15:55 and 2017-11-25 13:45 to 2017-11-26 03:10.* The longer than usual detection window for 11-Oct is due to the fact that there were other events occurring simultaneously in the network and the Event 8 contributed to small fraction of messages generated during this time window. Finally, as shown in Table II, our method also discovered several event signatures not previously known.

Scalability and Sensitivity: To understand the scalability of CD-LDA with data size, we ran it on Dataset-1 with about 97 million data points. CD-LDA was run with the following input: $\alpha = 0.01$, $\delta = 0.1$, and the number of events equal to 20. The CD part of the algorithm detects 57 change points. The sensitivity of this output with respect to α, δ is discussed next. The event signatures are quite robust to these parameter choice, but as expected, the accuracy of the start and finish time estimates of the events will be poorer for large values of α and δ. Overall, CD-LDA takes about 6 hours to run, which is quite reasonable for a dataset of this size. Reducing the running time by using other methods for implementing LDA, such as variational inference, is a topic for future work.

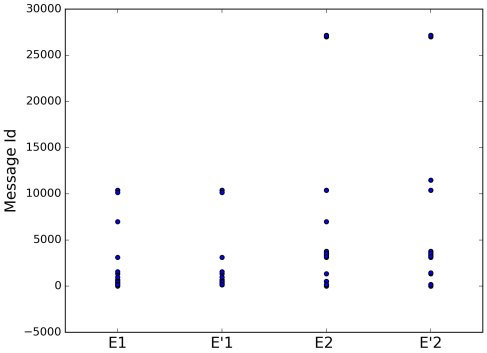


Fig. 6. Comparison of Event signatures for first two events with α_1, δ_1(E1,E2) vs α_2, δ_2(E'1,E'2).

TABLE III

COMPARING RESULTS OF CD-LDA FOR DIFFERENT VALUES OF α, δ

$\alpha_1 = 1\%, \delta_1 = 0.1$ vs $\alpha_2 = 10\%, \delta_2 = 0.5$			
$\frac{\lvert\mathcal{M}_1 \Delta \mathcal{M}'_1\rvert}{\lvert\mathcal{M}_1 \cup \mathcal{M}'_1\rvert}$	$\frac{\lvert\mathcal{M}_2 \Delta \mathcal{M}'_2\rvert}{\lvert\mathcal{M}_2 \cup \mathcal{M}'_2\rvert}$	TV dist in $p^{(1)}$	TV dist in $p^{(2)}$
0.046	0.077	0.036	0.08

TABLE IV

RESULTS OF CD-LDA ON DATASET-2 WITH $\alpha_2 = 10\%, \delta_2 = 0.5$

Event 1	Event 2
2017-02-14 00:00 to 2017-02-15 23:59	2017-02-06 19:29 to 2017-02-07 16:42
	2017-02-08 00:00 to 2017-02-08 06:25
	2017-02-08 23:59 to 2017-02-10 04:07
	2017-02-10 05:00 to 2017-02-14 00:00

Parameter α specifies the minimum duration of episode that can be detected in the change detection. By increasing δ we can control to detect the more sharp change points (change points across which the change in distribution is large), and decreasing δ helps us detect the soft change points as well. So α, δ control the granularity of the change point detection algorithm. Parameter η is a user defined parameter to detect the episodes in which a particular event occurs. We demonstrate the sensitivity of CDLDA to α and δ. We run CDLDA with $\alpha_2 = 10\%, \delta_2 = 0.5$ on Dataset-2 and compare it with results when run with parameters $\alpha_1 = 1\%, \delta_1 = 0.1$. Table IV and V shows the first two events for parameters α_1, δ_1 when compared to first two events for parameter α_2, δ_2. CDLDA detects 57 change points with α_1, δ_1 whereas it only detects 19 change points with α_2, δ_2. Despite this, Figure 6 and Table III shows that the event signatures for the first two events are almost the same. But, since the episodes are larger in duration with α_2, δ_2, the start and end times of the first two events are less accurate than α_1, δ_1. In particular, event 2 is shown to occur from 2-10 05:00 to 2-14 00:00 with α_2, δ_2 in Table V whereas it broken into two episodes, 2-10 05:00 to 2-10 13:33 and 2-10 15:27 to 2-14 00:00, with α_1, δ_1 in Table IV.

TABLE V

RESULTS OF CD-LDA ON DATASET-2 WITH $\alpha_1 = 1\%, \delta_1 = 0.1$

Event 1	Event 2
2017-02-14 00:00 to 2017-02-15 23:59	2017-02-05 06:21 to 2017-02-07 16:42
	2017-02-08 00:00 to 2017-02-10 00:00
	2017-02-10 03:07 to 2017-02-10 04:07
	2017-02-10 05:00 to 2017-02-10 13:33
	2017-02-10 15:27 to 2017-02-14 00:00

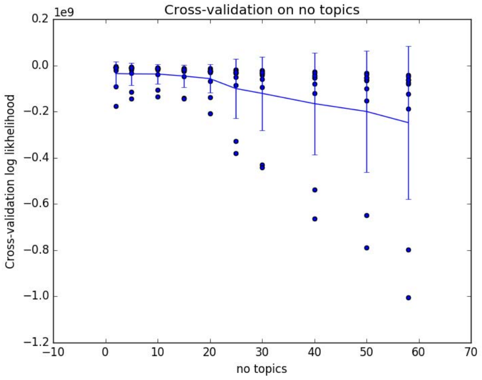


Fig. 7. Likelihood vs number of topics in Dataset-1

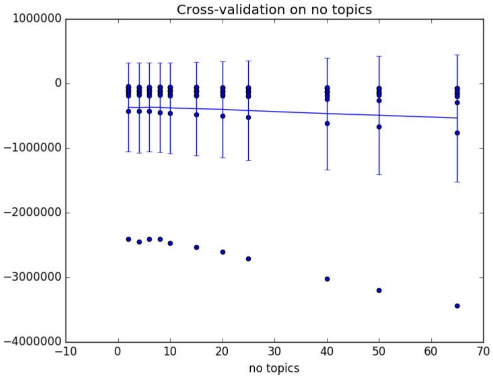


Fig. 8. Likelihood vs number of topics in Dataset-2

1) Selection of the Number of Topics in LDA: For Dataset-1, we do 10-fold cross validation. We group the 58 documents found by change detection into 10 sets randomly. We compute the likelihood on one group with a model trained using documents in the remaining 9 groups. We plot the average likelihood in Figure 7 vs the number of topics. There is a decrease in likelihood around 20 and hence, we choose the number of topics as 20.

For Dataset-2, we do 10-fold cross validation and choose the number of topics as 10 from the Figure 8 below. In this case, we create the 10 groups of documents in the following way. Out of 58 documents, group 1 has document number $1, 11, 21 \ldots$, group 2 has documents $2, 22, 32, \ldots$, etc. Sub sampling in this fashion respects the ordering in the documents.

V. CONCLUSIONS AND FUTURE WORK

In this paper we look at the problem of detecting events in an error log generated by a distributed data center network. The error log consists of error messages with time stamps. Our goal is to detect latent events which generate these messages and find the distribution of messages for each event. We solve this problem by relating it to the topic modelling problem in documents. We introduce a notion of episodes in the time series data which serves as the equivalent of documents. Also we propose a linear time change detection algorithm to detect these episodes. We present consistency and sample complexity results for this change detection algorithm. Further we demonstrate the performance of our algorithm on a real dataset by comparing it with two benchmark algorithms existing in the literature. We believe, our approach is generic enough to be applied to other problem settings where the data has similar characteristics as network logs.

REFERENCES

[1] T. Li, L. Shwartz, and G. Y. Grabarnik, "System event mining: Algorithms and applications," in *Proc. 23rd ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Halifax, NS, Canada, 2017. [Online]. Available: https://users.cs.fiu.edu/~taoli/event-mining/

[2] T. Li *et al.*, "Data-driven techniques in computing system management," *ACM Comput. Surv.*, vol. 50, no. 3, Jul. 2015, Art. no. 45. doi: 10.1145/3092697.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[4] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data," *J. Amer. Stat. Assoc.*, vol. 109, no. 505, pp. 334–345, 2013.

[5] S. Satpathi, S. Deb, R. Srikant, and H. Yan, "Learning latent events from network message logs," in *Proc. Workshop Mining Learn. Time Series*, 2018. [Online]. Available: https://milets18.github.io/papers/milets18_paper_13.pdf

[6] A. A. Makanju, A. N. Zincir-Heywood, and E. E. Milios, "Clustering event logs using iterative partitioning," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2009, pp. 1255–1264. doi: 10.1145/1557019.1557154.

[7] T. Li, F. Liang, S. Ma, and W. Peng, "An integrated framework on mining logs files for computing system management," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2005, pp. 776–781. doi: 10.1145/1081870.1081972.

[8] L. Tang and T. Li, "LogTree: A framework for generating system events from raw textual logs," in *Proc. IEEE 10th Int. Conf. Data Mining*, Dec. 2010, pp. 491–500.

[9] T. Qiu, Z. Ge, D. Pei, J. Wang, and J. Xu, "What happened in my network: Mining network events from router syslogs," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas. (IMC)*, New York, NY, USA, 2010, pp. 472–484. doi: 10.1145/1879141.1879202.

[10] F. Wu, P. Anchuri, and Z. Li, "Structural event detection from log messages," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2017, pp. 1175–1184. doi: 10.1145/3097983.3098124.

[11] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. Data Eng. (ICDE)*, Washington, DC, USA, 1995, pp. 3–14. [Online]. Available: http://dl.acm.org/citation.cfm?id=645480.655281

[12] D. Cheng, M. T. Bahadori, and Y. Liu, "FBLG: A simple and effective approach for temporal dependence discovery from time series data," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2014, pp. 382–391. doi: 10.1145/2623330.2623709.

[13] C. Zeng, Q. Wang, W. Wang, T. Li, and L. Shwartz, "Online inference for time-varying temporal dependency discovery from time series," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 1281–1290.

[14] C. H. Mooney and J. F. Roddick, "Sequential pattern mining—Approaches and algorithms," *ACM Comput. Surv.*, vol. 45, no. 2, Mar. 2013, Art. no. 19. doi: 10.1145/2431211.2431218.

[15] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. De Carvalho, and J. Gama, "Data stream clustering: A survey," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013. doi: 10.1145/2522968.2522981.

[16] Y. Jiang, C.-S. Perng, and T. Li, "Natural event summarization," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2011, pp. 765–774. doi: 10.1145/2063576.2063688.

[17] P. Wang, H. Wang, M. Liu, and W. Wang, "An algorithmic approach to event summarization," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, New York, NY, USA, 2010, pp. 183–194. doi: 10.1145/1807167.1807189.

[18] W. Peng, C. Perng, T. Li, and H. Wang, "Event summarization for system management," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2007, pp. 1028–1032. doi: 10.1145/1281192.1281305.

[19] N. Tatti and J. Vreeken, "The long and the short of it: Summarising event sequences with serial episodes," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2012, pp. 462–470. doi: 10.1145/2339530.2339606.

[20] M. Du, F. Li, G. Zheng, and V. Srikumar, "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, New York, NY, USA, 2017, pp. 1285–1298. doi: 10.1145/3133956.3134015.

[21] Y. Kawahara and M. Sugiyama, "Sequential change-point detection based on direct density-ratio estimation," *Statist. Anal. Data Mining*, vol. 5, no. 2, pp. 114–127, 2012.

[22] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, "Homogeneity and change-point detection tests for multivariate data using rank statistics," 2011, *arXiv:1107.1971*. [Online]. Available: https://arxiv.org/abs/1107.1971

[23] H. Chen and N. Zhang, "Graph-based change-point detection," *Ann. Statist.*, vol. 43, no. 1, pp. 139–176, Feb. 2015. doi: 10.1214/14-AOS1269.

[24] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2006, pp. 424–433. doi: 10.1145/1150402.1150450.

[25] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," 2012, *arXiv:1206.3298*. [Online]. Available: https://arxiv.org/abs/1206.3298

[26] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.

[27] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 856–864.

[28] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.

[29] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014. [Online]. Available: http://jmlr.org/papers/v15/anandkumar14b.html

[30] T. Bansal, C. Bhattacharyya, and R. Kannan, "A provable svd-based algorithm for learning topics in dominant admixture corpus," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1997–2005.

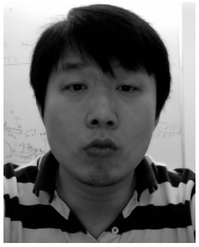
[31] C. Wang, J. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 752–760.

[32] E. Sy, S. A. Jacobs, A. Dagnino, and Y. Ding, "Graph-based clustering for detecting frequent patterns in event log data," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2016, pp. 972–977.

Siddhartha Satpathi is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Illinois at Urbana Champaign. His research interests include optimization and machine learning.

Supratim Deb received the Ph.D. degree from the University of Illinois at Urbana-Champaign in the area of communication networks in 2003. Following his Ph.D. degree, he had a post-doctoral stint at MIT. He was with Bell Labs (Lucent/Nokia) from 2005 to 2014, and was with AT&T Labs from 2015 to 2018. He has been an Engineer at Facebook since August 2018. His research interests are in the broad areas of data-driven intelligent systems, networked systems, and applied machine learning.

R. Srikant (S'90–M'91–SM'01–F'06) is currently the Fredric G. and Elizabeth H. Nearing Endowed Professor with the Department of Electrical and Computer Engineering, and a Professor with the Coordinated Science Lab, University of Illinois at Urbana-Champaign. His research interests include communication networks, machine learning, and applied probability. He has received a number of awards, including the 2015 IEEE INFOCOM Achievement Award, the 2015 IEEE INFOCOM Best Paper Award, the 2017 Applied Probability Society Best Publication Award, and the 2019 IEEE Koji Kobayashi Computers and Communications Award. He was the Editor-in-Chief of the IEEE/ACM TRANSACTIONS ON NETWORKING from 2013 to 2017.



He Yan received the M.S. and Ph.D. degrees in computer science from Colorado State University, Fort Collins, CO, USA, in 2009 and 2012, respectively. He is currently a Principal Inventive Scientist with AT&T Laboratories–Research, Bedminster, NJ, USA. His research interests include network management, anomaly detection, network data mining, network measurement, and end-to-end service monitoring.