Graph Topic Scan Statistic for Spatial Event Detection

Yu Liu[†] Baojian Zhou[‡] Feng Chen[‡] David W. Cheung[†]
[†] The University of Hong Kong, Hong Kong [‡] University at Albany-SUNY, Albany, NY 12222
[†] {yliu4, dcheung}@cs.hku.hk [‡] {bzhou6, fchen5}@albany.edu

ABSTRACT

Spatial event detection is an important and challenging problem. Unlike traditional event detection that focuses on the timing of global urgent event, the task of spatial event detection is to detect the spatial regions (e.g. clusters of neighboring cities) where urgent events occur. In this paper, we focus on the problem of spatial event detection using textual information in social media. We observe that, when a spatial event occurs, the topics relevant to the event are often discussed more coherently in cities near the event location than those far away. In order to capture this pattern, we propose a new method called Graph Topic Scan Statistic (Graph-TSS) that corresponds to a generalized log-likelihood ratio test based on topic modeling. We first demonstrate that the detection of spatial event regions under Graph-TSS is NP-hard due to a reduction from classical node-weighted prize-collecting Steiner tree problem (NW-PCST). We then design an efficient algorithm that approximately maximizes the graph topic scan statistic over spatial regions of arbitrary form. As a case study, we consider three applications using Twitter data, including Argentina civil unrest event detection, Chile earthquake detection, and United States influenza disease outbreak detection. Empirical evidence demonstrates that the proposed Graph-TSS performs superior over state-of-the-art methods on both running time and accuracy.

Keywords

Scan Statistic; Topic Model; Spatial Event Detection; Large Graph

1. INTRODUCTION

Spatial event detection, such as detection of disease outbreaks, civil unrests, earthquake, and financial crises, is an important and challenging problem. Unlike traditional event detection that focuses on the timing of global urgent event, the task of spatial event detection is to detect spatial regions (e.g. clusters of neighboring cities) where urgent events occur. To motivate this scenario, consider the cholera outbreak problem [25] as shown in Figure 1. Suppose we have a network of counties (vertices) and each vertex has a feature referring to the number of cases of cholera in that county on a given day. Suppose further that two vertices are connected by an edge if they share a boundary. The task is to detect spatial regions (connected sub-graphs) of arbitrary form as indicators of ongoing cholera outbreaks in the noisy background data.

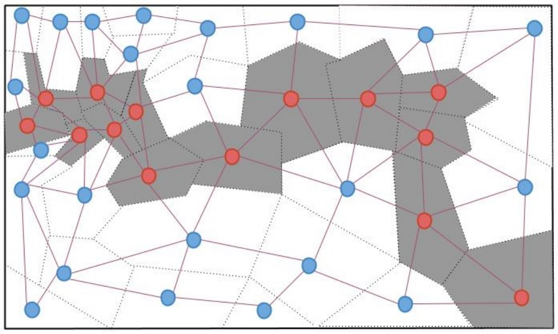


Figure 1: A potential cholera outbreak leads to elevated intensities of infected cases in counties near the river, which forms as an irregularly shaped connected sub-graph (cluster) of counties. (Redrawn from[25])

Spatial event detection usually utilizes traditional channels, where collection of information, such as patient data, crimes, and financial transaction, is hysteretic and costly. With the popularity of low-cost GPS chips and smart phones, micro-blogging services such as Twitter, Tumblr and Weibo have become important tools for online users to share breaking news, interesting stories and rich media content. Unlike traditional media or channels, social microblogs media provides a more fruitful, timely and vast amount of data available on the Internet at almost no cost. Social media also helps spread information earlier and faster than traditional media. For example, Twitter firstly leaked credible word of Osama bin Laden's death before President Obama's announcement, and there were a half million tweets (and only 800 news mentions) one hour after the event [1]. Another example is that, the information of pestilence spreads in the social media before newspaper or TV's announcement [6]. Because social media often discusses these events in advance, compared with the tradition media, it can be a sig-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983744

nificant study to detect the urgent event by using the textual information of social media.

However, the language used in social media is highly informal, ungrammatical, and dynamic, and thus traditional natural language processing (NLP) techniques cannot be directly applied. Different events tend to have different contexts, and their relevant keywords are often unpredictable. It is hence difficult to develop an event detection model directly based on keyword frequencies. Instead, we focus on topic-level analysis, as topics are considered high-level semantic summarizations of a corpus from different aspects. We observe that, when a spatial event occurs, the topics relevant to the event are often discussed more coherently in cities near the event location than those far away. For example, people lived in “civil unrest” event region may discuss more related topics, such as strike, work, teacher, while people lived outside the region may discuss other popular topics. We thus propose a new method called Graph Topic Scan Statistic (Graph-TSS) that corresponds to a generalized log-likelihood ratio test (GLRT) to decide between the **null hypothesis** (H_0): the topic of each tweet posted at any city follows the same multinomial distribution $\mathbf{Multinomial}(\pi)$, and the **alternative hypothesis** ($H_1(S)$): the topic of each tweet posted within the event region S follows a different multinomial distribution $\mathbf{Multinomial}(\theta)$, and the topic of each tweet posted outside the event region S follows the multinomial distribution $\mathbf{Multinomial}(\pi)$, where $\theta \neq \pi$. The problem of spatial event detection is then formalized as the maximization of the corresponding generalized log-likelihood ratio function over all possible spatial regions. Then, empirical calibration approach is employed to deal with the multiple testing issue. Specifically, to ensure the identified anomalous sub-graph is statistically significant at the $1 - \alpha$ (e.g $\alpha = 0.05$) confidence level, we estimate a threshold such that the probability of the current test statistic Graph-TSS being above the threshold is α under the null hypothesis. If the test statistic Graph-TSS of the identified sub-graph is above the empirically calibrated threshold, we will reject the null hypothesis of no sub-graph, i.e. an alert will be raised.

We note that, although a number of geographical topic models have been proposed in recent years [19, 9, 33, 17, 4, 11, 16], these models cannot identify subtle signals of coherent regional topics in noisy social media data that are important for the detection of spatial events in their early stages, as these models are all generative probabilistic models, and do not have sufficient discriminative power for this specific task. Our proposed Graph-TSS has strong connections to traditional spatial scan statistic models that have been shown effective in a variety of applications related to spatial event detection, such as detection of road traffic congestion [24], water pollution [32], crime hotspots [20, 30], and disease outbreaks [14]. These spatial scan statistic models are designed for analyzing non-textual data (e.g., counts of infected cases) collected from traditional sources such as hospitals, emergency departments, and drug stores. Our proposed Graph-TSS can be considered a new variant of spatial scan statistic for analyzing noisy textual data. The main contributions of this paper are as follows:

- **Formulation of a new Graph topic scan statistic.** We formulate a new Graph-TSS based on generalized log-likelihood ratio test and topic models. To the best of our knowledge, this is the first kind of spatial scan statistic for spatial event detection using noisy textual data.
- **Development of an approximate algorithm for graph scanning.** We prove that the spatial event detection problem under Graph-TSS is NP-hard via a reduction from classical node-weighted prize-collecting Steiner tree problem (NW-PCST), and develop an efficient algorithm that approximately maximizes the Graph-TSS over all possible spatial regions, with time complexity around $O(|\mathbb{E}| \log^3 |\mathbb{V}|)$, where $|\mathbb{V}|$ and $\mathbb{E}$ refer to the total number of vertices and the number of edges, respectively.
- **Comprehensive experiments to validate the effectiveness and efficiency of the proposed algorithm.** We conduct the experiments on the real Twitter data and practical applications. According to the results, the proposed method outperforms existing methods for the three applications, including civil unrest detection, earthquake detection, and influenza disease outbreak detection.

The rest of this paper is organized as follows: Section 2 introduces the related work. Section 3 presents a detailed analysis of civil unrest event patterns on real Twitter data. We present the proposed Graph-TSS and approximation algorithm for sub-graph detection in Section 4. Experiments on real Twitter datasets and the three practical applications are presented in Section 5. Section 6 concludes our work and describes the future work.

2. RELATED WORK

We briefly review three lines of related work: Spatial scan statistic-based methods, Geographical topic models, and Burst detection-based approaches.

Spatial scan statistic based methods detect connected or correlated subgraphs which are unexpected given the typical data distribution (e.g. Gaussian, Poisson, or mixture of Gaussians). Existing methods can be categorized into two groups, namely parametric and nonparametric methods. **Parametric methods** assume specific forms of distribution for features of normal and abnormal vertices, respectively, and formalize the anomaly detection as a hypothesis testing problem. Depending on the specific forms of distributions assumed, a number of methods have been proposed, including Kulldorff’s spatial scan statistic [13], expectation based scan statistic [23, 21], the elevated mean scan statistic [26], and various other variant scan statistic methods. Here, Kulldorff’s spatial scan statistic and expectation based scan statistic utilize the observed count (number of cases), e.g. the number of patients, to detect the area of outbreak event. On the other hand, **nonparametric methods** estimate a p-value for each vertex of spatial graph by comparing the current features of this vertex with its features in the historical data [5, 18, 27], rather than associating specific forms of distributions with normal and abnormal vertices. These approaches usually maximize a score function $F(S)$ of p-value in a sub-graph S, and typically nonparametric methods measure the significance of the collection of p-values in sub-graph. Recent studies [5, 32] show that these approaches perform better than the burst detection-based methods in the problem of spatial event detection. However, nonparametric methods do not capture the latent semantics in textual information.

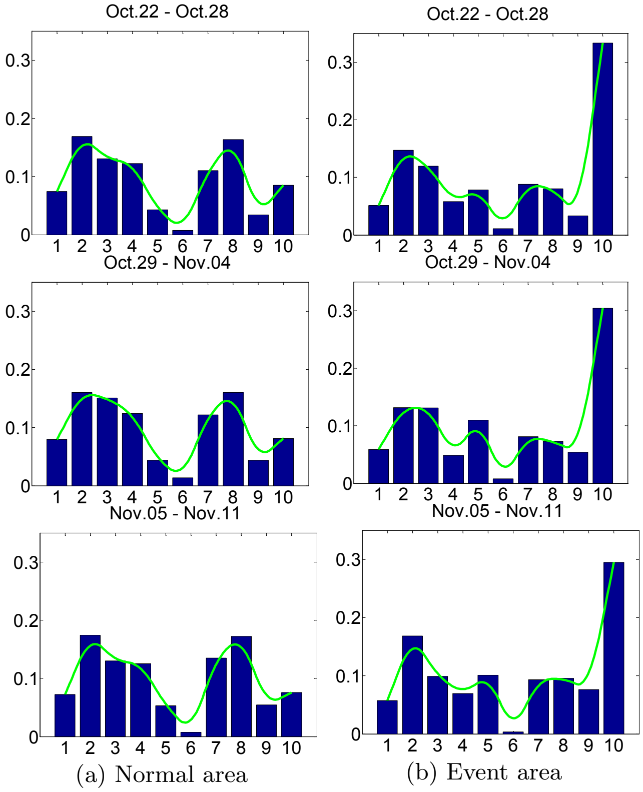
Geographical topic models [19, 9, 33, 4, 11] estimate language distributions (over a predefined vocabulary) that are distinct in some geographic regions. For example, several normal distributions are assigned to regions which have a distribution over the set of topics in Yin's model or Hong's model [9, 33, 4]. Clearly, there now can be several Gaussian regions sharing the same topic. Therefore, these models can discover some geographical regions / sub-graphs whose topics are distinct from others. A more general approach for modelling arbitrary, complex features such as geolocations was introduced by Agovic and Banerjee [3]. Given that the similarity between topic distributions of documents directly depends on their respective position in the feature space, topic distributions of documents can be sampled from a Gaussian process (GP) prior which encodes the similarity of couments in the feature space. Essentially, geographical topic models aim to partition the global region into several sub-regions, and assign all sub-regions with distinct labels, such as coastline or mountain. Therefore, the specific purpose leads to the weak performance in the problem of spatial event detection. The latter contains many new phenomena which we have not seen in the geographical topic modeling regime, to discover which, we need new methods and new theoretic frameworks. In addition, geographical topic model-based approaches can only detect event in geographical graph. Thus, they cannot be directly applied to document streams that are embedded in a general graph, e.g. social network relationship graph.

Burst detection-based methods search for space-time regions / sub-graphs where the aggregated counts of some predefined terms are abnormally high compared with the counts outside the regions / sub-graphs. Here, some extension methods like, ST Burst Detection [15], Feature-pivot Clustering [7] and the burst detection technique proposed by Kleinberg [10] are popular. Sakaki et al.[28] consider spatial-temporal Kalman filtering, which is similar to space-time burst detection to track the geographical trajectory of hot spots of tweets related to earthquakes. However, as a traditional event detection methods, burst detection-based methods focus on the detection of events influencing the global region. For example, it can discover the burst timing and regions when the large-scale disease outbreak appears, because the signal becomes significant in the large-scale event. In addition, different events tend to have different contexts, and their relevant keywords are often unpredictable, which constantly result in loss of detection.

In summary, the proposed Graph-TSS utilizes the latent semantics of textual information for spatial event detection, while both burst detection-based methods and nonparametric methods cannot capture the latent semantics of textual information. In addition, Graph-TSS can detect the local events (i.e. only significant in local area) by using the proposed efficient scan algorithm, while geographical topic models-based approaches only aim to partition the global region into several distinct sub-regions.

3. DATA ANALYSIS

The motivation for this work is firmly built on the observations of social media data and real-world events. We opted to use Twitter dataset because of its ready accessibility through APIs. In this section, we provide some analysis on twitter data and ground truth civil unrest events as a base for our method construction, which makes our assumption more reliable.



(a) Normal area

(b) Event area

Figure 2: Example topic distributions with respect to civil unrest event from Oct 1st, 2013 to Dec 31st, 2013. The horizontal axis refers to topic index, while the vertical axis refers to probability of the corresponding topic. Green curves are used to fit the topic distributions.

We collected three months Twitter data from Argentina, and the civil unrest event labels, called Golden Standard Report (GSR), were collected and confirmed from the local newspapers that are accessible from Internet. Specifically, there are 12 events and 0.58 million tweets in each week on average (0.02 million tweets in event area and 0.56 million tweets in normal area). Topic modeling was used to extract the topics of tweets and analyze the topic distributions of event region and non-event (normal) region.

Figure 2 shows three example weekly topic distributions in normal area and event area. We observe that the topic distributions of these three weeks are very similar in normal area, and they are significantly different from the topic distributions in event area, especially the probability of topic 10.

In order to demonstrate the comprehensive observation, we present all weekly topic distributions (totally 13 weeks). Topic distribution (multinomial distribution) is fitted by curve. Figure 3 shows the comparison of topic distributions of normal area and event area during the consecutive 13 weeks, and result demonstrates topic distributions in event area are different from the topic distributions in normal area. As expected, we observe that the topic distribution are very similar in normal area.

4. GRAPH TOPIC SCAN STATISTIC

In this section, we propose a novel method called Graph Topic Scan Statistic (Graph-TSS), which combines scan statistic, sub-graph detection and topic modeling together. Specifically, we start with an overview of the basic idea in Section 4.1; section 4.2 presents the proposed Graph-TSS; and section 4.3 presents an efficient approximate algorithm that maximizes the Graph-TSS over connected sub-graphs to identify the most anomalous graph clusters.

4.1 General Idea

People like to post information occurred nearby, e.g. tweets, on the social media, and social media also helps spread information earlier and faster. For example, users who live in the "civil unrest" event region are more likely to post tweets related to this "civil unrest" event on the early time, while other users are less likely to post tweets about this event. That means that the topic popularity may be different between inside and outside the event region. For instance, people lived in "civil unrest" event region may discuss more related topics, such as strike, work, teacher, and people lived outside the region may discuss other popular topics. The results of data analysis described above also quantitatively confirm this intuition. Therefore, we aim to utilize this observation to find the most anomalous sub-graph region S, where the topic popularity is different.

The proposed method, Graph-TSS, will return a sub-graph (i.e. abnormal area) of the spatial graph that approximately maximizes the proposed graph-TSS, which is formulated based on the log-likelihood ratio. The log-likelihood ratio can be regarded as the degree of difference between the sub-graph (event) region and normal region. Therefore, the detected sub-graph S is the region where topic distribution is significantly distinct. Here, the returned sub-graph consists of one or several connected component(s), because most spatial events, e.g. civil unrest, earthquake or infectious disease, occurred in one or several connected regions. In real world, the connected sub-graph can be regarded as a cluster of neighboring cities, which share boundaries, and isolated city never exist. Note that the present work is the first work to propose the concept of Graph topic scan statistic.

4.2 Formulation of Graph Topic Scan Statistic

We assume that the collection of geocoded documents D has an embedded graph. Suppose we have a network of cities of a country, and the network can be regarded as a undirected graph. Given the Graph, each vertex refers to a location (e.g. city) l in the country, which has many geocoded documents D_l that consist of a collection of keywords $\overrightarrow{w_d}$. Suppose further that two vertices are connected by an edge if they share a boundary. We wish to identify possible event outbreaks sub-graph S at an early stage in the noisy background data.

In order to determine which connected sub-graph S is the most anomalous, we generalize the Graph-TSS, which extends Kulldorff's spatial scan statistic and was originally proposed for modeling spatial-temporal count data. The Graph-TSS detects an anomalous sub-graph by searching over a large number of sub-graphs, where each sub-graph S consists of some subset of the locations l, and finding the sub-graph S which maximizes the Graph-TSS. We formally define a topic to be a distribution over a fixed vocabulary, and topics represent the semantic summarizations of a corpus. We firstly give the definitions of null hypothesis $\boldsymbol{H_0}$ and alternative hypotheses $\boldsymbol{H_1(S)}$.

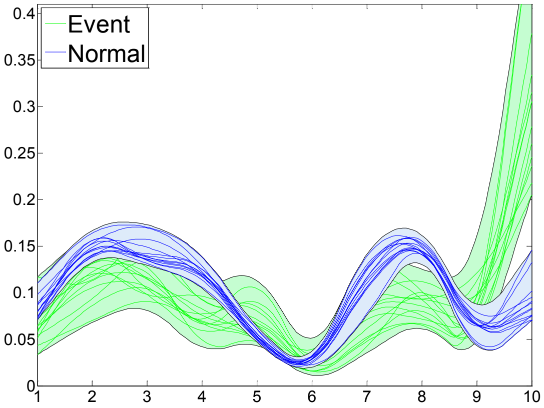


Figure 3: Topic distribution per week with respect to civil unrest event from Oct 1, 2013 to Dec 31, 2013. Blue curves refer to the topic distribution of each week in normal area, while green curves refer to the topic distribution of each week in event area.

$\boldsymbol{H_0}$: refers that the topic of each document d posted at any vertex follows the same multinomial distribution $Multinomial(\pi)$, i.e. $z_d \sim Multinomial(\pi)$, $\forall d$.

$\boldsymbol{H_1(S)}$: assumes that the topic of each document d posted outside S follows the multinomial distribution $Multinomial(\pi)$, and the topic of each document d posted within S follows a different multinomial distribution $Multinomial(\theta)$, i.e. $z_d \sim Multinomial(\theta)$, $\forall d$ posted in S; $z_d \sim Multinomial(\pi)$, otherwise; where $\theta \neq \pi$.

Given a set of alternative hypotheses $H_1(S)$ and a null hypothesis H_0, the **Graph-TSS** $\mathscr{F}(S)$ for a given sub-graph S is the ratio of the data log-likelihood under the alternative and null hypotheses:

$$\mathscr{F}(S) = \log(\frac{\max_\theta Pr(Data|H_1(S), \theta)}{Pr(Data|H_0, \pi)}) \quad (1)$$

where θ and π refer to the topic distributions of the collection of geocoded documents in sub-graph S and normal region, respectively.

The assumption that we fix topic distribution under null hypothesis $\boldsymbol{H_0}$ lays on two reasons. First, we observe this pattern in data analysis. Second, in order to detect changes of topic distributions as indicators of spatial events, we need to fix the background distribution.

The sub-graph S with the highest values of the Graph-TSS is that which are most likely to have been generated under the alternative hypothesis instead of the null hypothesis of no sub-graph.

To ensure the identified anomalous sub-graph is statistically significant at the $1 - \alpha$ (e.g $\alpha = 0.05$) confidence level, we need to estimate a significant threshold such that the probability of the current test statistic Graph-TSS being above the threshold is α under the null hypothesis. Traditional spatial scan statistic approaches [13] deal with this

Table 1: Notations used in Graph-TSS

Symbol	Description
L	Collection of locations
D	Collection of geocoded documents
D_l	Collection of geocoded documents posted at location l
d	Document identity
w	Word identity
$\overrightarrow{w_d}$	Word vector of document d
z	Topic assignment of a document
l	Location (e.g. city) identity
S	Abnormal sub-graph region
B	Normal region
x	The vector form of S, i.e. S =**supp**(x), $x \in \{0,1\}^{\vert L\vert}$
ϕ	Topic word distribution
θ	Topic distribution in sub-graph region
π	Topic distribution in normal region
α	The statistical significant level

multiple testing issue by "randomization testing", generating a large number of replica datasets under the null hypothesis and finding the maximum sub-graph statistic score for each replica dataset. A sub-graph S must score higher $\mathscr{F}(S)$ than approximately 95% of the replica datasets to be significant at $\alpha = 0.05$. However, randomization testing is computationally expensive, multiplying the computation time by $R + 1$, where R is the number of Monte Carlo replications performed. Moreover, Neill [22] indicates that empirical calibration performed better than randomization testing. Therefore, we use historical training data of normal area to empirically calibrate the significant threshold. Finally, if the test statistic Graph-TSS of the identified sub-graph $\mathscr{F}(S)$ is above the empirically calibrated threshold, we will reject the null hypothesis of no sub-graph, i.e. an alert will be raised.

The notations used in the Graph-TSS are listed in Table 1.

It can be proved that finding this sub-graph, consist of several connected components, is a NP-hard. We analyze the hardness of this problem below.

THEOREM 1. ***The problem 1****, finding a sub-graph (unions of individual trees) S that maximized Graph-TSS $\mathscr{F}(S)$, is NP-hard*

PROOF 1. *Consider an instance of the NP-hard* node-weighted prize-collecting Steiner tree *problem (NW-PCST) [12], defined by an n-node, undirected graph $G = (V, E)$, a non-negative cost $c(v)$ and a non-negative penalty value $\sigma(v)$ for each vectex $v \in V$, we wish to find a tree T that minimizes $\sum_{v\in T} c(v) + \sum_{v\in V\setminus T} \sigma(v)$. We show that this can be viewed as a special case of problem 1.*

Given an arbitrary instance of NW-PCST problem, the task is equivalent to find a tree T that maximize $\sum_{v\in T} \sigma(v) - \sum_{v\in T} c(v)$. Denoted $p_v = \sigma(v) - c(v)$, p_v will be positive or negative. Therefore, the NW-PCST problem is equivalent to find a tree that maximized $\sum_{v\in T} p(v)$. Note that for an instance of problem 1, we only find 1 connected tree, rather than g connected components. For an instance of problem 1, there are only two topics and the topics inside the event region and outside the event region do not overlap, i.e. $\theta = I - \pi, I = [1, 1]^T$. Therefore, Graph-TSS $\mathscr{F}(S)$ can be presented as:

$$\mathscr{F}(S) = \sum_{l\in S} \sum_{d\in D_l} \log \frac{Pr(d|H_1(S), I - \pi)}{Pr(d|H_0, \pi)}$$

Thus, this instance is equivalent to NW-PCST problem, if we set the log-likelihood ratio $\sum_{d\in D_l} \log \frac{Pr(d|H_1(S),I-\pi)}{Pr(d|H_0,\pi)}$ in each location is p_v, where topic distribution of normal region π can be obtained from historical data. Since the NW-PCST problem is NP-hard , the above implies that problem 1 is also NP-hard.

When we compute the Graph-TSS $\mathscr{F}(S)$, the latent parameter θ need to be estimated before. However, the latent parameter θ cannot be estimated by using the maximum likelihood estimation directed, and θ depends on the given sub-graph S. To solve this problem, we employ the *mixture of unigrams* model to model the geocoded documents in each vertex (location). In the mixture of unigrams model, each document only has one specific topic. Topic is influenced by the property of region, i.e. whether this region is event region or normal region. When one document is posted in abnormal area, the corresponding topic z is generated from a topic distribution $Multinomial(\theta) = p(z|S)$. When the document is posted in normal region B, its corresponding topic z is generated from another topic distribution $Multinomial(\pi) = p(z|B)$.

To generate the above documents dataset, we employ the following generative process:

- For each document $d \in S$, which is written in the abnormal area S
  - Draw a topic $z \sim p(z|S)$
  - For each word w in document d, draw $w \sim p(w|z)$
- For each document $d \in B$, which is written in the normal region B
  - Draw a topic $z \sim p(z|B)$
  - For each word w in document d, draw $w \sim p(w|z)$

Let $x \in \{0,1\}^{|L|}$ be the vector form of S. That is, if document is posted in location l belonging to sub-graph S, the value of x_l is 1, others are 0.

Here we introduce a few notations used later:

- **supp**(x): the support of vector x refers to a set containing the indices corresponding to nonzero entries in x, i.e., **supp**$(x) = \{i|x_i \neq 0\}$.
- b_Ω: a vector that has $(b_\Omega)_i = b_i$ for $i \in \Omega$ and $(b_\Omega)_i = 0$ otherwise, given a subset Ω.

In order to estimate parameters, we use Expectation Maximization(EM) algorithm, which iteratively computes a local maximum of likelihood. Given the word vector $\overrightarrow{w_d}$ of document d, the joint probability over d and its corresponding topic z can be :

$$\begin{aligned} & p(d, z) = p(\overrightarrow{w_d}, z, S, B) && (2) \\ = \quad & p(z|S)^{x_l} p(z|B)^{1-x_l} \prod_{w\in\overrightarrow{w_d}} p(w|z) \end{aligned}$$

- In E-step, the hidden variable $p(z|d)$ is updated according to Bayes formulas as in Equation 3.

$$p(z|d) = \frac{p(d,z)}{p(d)} = \frac{p(d,z)}{\sum_z p(d,z)}$$
$$= \frac{p(z|S)^{x_l} p(z|B)^{1-x_l} \prod_{w \in \overrightarrow{w_d}} p(w|z)}{\sum_z p(z|S)^{x_l} p(z|B)^{1-x_l} \prod_{w \in \overrightarrow{w_d}} p(w|z)} \quad (3)$$

- In M-step, we find the estimation parameter that maximizes the expectation of the complete likelihood, i.e. $\text{argmax}_\theta \sum_z p(z|d) \log \prod_{d \in D} p(d,z)$, using the following updating formulas:

$$\theta = p(z|S) = \frac{\sum_{l \in L} \sum_{d \in D_l} x_l p(z|d)}{\sum_{l \in L} x_l |D_l|} \quad (4)$$

The topic distribution in normal region $p(z|B) = \frac{\sum_{d \in D} p(z|d)}{D}$ and word distributions $p(w|z) = \frac{\sum_{d \in D} c(d,w) p(z|d)}{\sum_{w \in V} \sum_{d \in D} c(d,w) p(z|d)}$ are learnt from historical normal region data. In addition, we assume the word distributions $\phi = p(w|z)$ are the same both inside and outside the sub-graph S. An emerging spatial event is characterized as a new topic distribution, instead of new topics.

Combining the above equations, the Graph-TSS can be described as following formulas:

$$f(x) = \log\left(\frac{\prod_{l \in L} (\prod_{d \in D_l} \sum_z p(z,d|\theta))^{x_l}}{\prod_{l \in L} (\prod_{d \in D_l} \sum_z p(z,d|\pi))^{x_l}}\right)$$
$$= \sum_{l \in L} \left(x_l \sum_{d \in D_l} \log\left(\frac{\sum_z p(z,d|\theta)}{\sum_z p(z,d|\pi)}\right)\right) \quad (5)$$

where $f(x) \equiv \mathscr{F}(S)$, S = **supp**(x). And $f(x)$ can be expressed by the linear combination of the Graph-TSS in each location l.

4.3 Approximation Algorithm for Sub-graph Detection

Based on the proposed Graph-TSS, the task of Graph-TSS is to return a sub-graph S that maximized the proposed Graph-TSS $\mathscr{F}(S)$ defined in Equation 5. In order to make our method more powerful, the sub-graph S that Graph-TSS returns consists of g connected components:

$$S = \underset{S \subseteq L, \gamma(S) = g, |S| \le s}{\text{argmax}} \mathscr{F}(S) \quad (6)$$

where $\gamma(S)$ refers to the number of connected components in S, and s refers to the upper bound on the number of vertices.

According to the hardness analysis, it is necessary to develop approximate solutions. We now refine our anomalous spatial event detection problem as the general sub-graph detection problem, which make our algorithm can be used in both spatial graph and general graph. The notations used in the Sub-graph Detection are listed in Table 2.

Table 2: Notations used in Sub-graph Detection

Symbol	Description
G	The undirected, weighted graph derived from the spatial graph
$\mathbb{V}$	The set of nodes
$\mathbb{E}$	The set of edges
w	node weights
N	the number of nodes, is equal to the $\lvert L \rvert$ defined before.
g	the connected components of S
s	the upper bound on the number of nodes
$supp(x)$	a set contains the indices corresponding to nonzero entries in vector x
b_Ω	a vector that has $(b_\Omega)_i = b_i$ for $i \in \Omega$ and $(b_\Omega)_i = 0$ otherwise, given a subset Ω

Let $G = (\mathbb{V}, \mathbb{E}, w)$ be an undirected, weighted graph with $\mathbb{V} = \{v_1, \cdots, v_N\}$, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$, and $w : \mathbb{V} \to \mathbb{R}$ denote node weights, which can be calculated from (5). Denote $\mathbb{M} = \{S \subseteq V | \gamma(S) = g, |S| \le s\}$, and $\mathscr{F}(S)$ as the objective function. Then, the problem is then formulated to maximize:

$$\max_{S \in \mathbb{M}} \mathscr{F}(S) \quad (7)$$

As S can be represented as the vector form $x \in [0,1]^N$, the above function can be reformulated as the following form:

$$\max_{x \in \{0,1\}^N \cap \mathcal{M}} f(x), \quad (8)$$

where $f(x)$ refers to the vector function of $\mathscr{F}(S)$, and $\mathcal{M} = \{x \in \mathbb{R}^N | \, \mathbf{supp}(x) \in \mathbb{M}\}$. Here, the number of vertices N is equivalent to the $|L|$ in spatial event detection problem.

The above function is difficult to optimize as the domain of elements of x is discrete. We use the following convex relaxation:

$$\max_{x \in \mathcal{M}} f(x) - \frac{1}{2}\|x\|^2, \quad (9)$$

The solution can be found as: $x_i \leftarrow 1$ if $x_i > 0$, $x_i \leftarrow 0$ otherwise. We apply a variant of Iterative Hard Thresholding (IHT), namely, Graph-IHT [34] (see in Algorithm 1), for solving Equation 9.

Algorithm 1: Graph-IHT

```
Input: Graph $G$, upper bound number of nodes $s$, the
       connected components $g$, iteration number $t$,
       log-likelihood ratio of each city $f$
Result: $\mathbf{supp}(x^{j+1})$
$x^0 \leftarrow 0$;
for $j \leftarrow 0, \ldots, t-1$ do
    $b \leftarrow -x^j + \nabla f(x^j)$ ;
    $\Gamma \leftarrow$ HEADAPPROX$(b, G, s, g)$ ;
    $z \leftarrow b_\Gamma + x^j$ ;
    $\Omega \leftarrow$ TAILAPPROX$(z, G, s, g)$ ;
    $x^{j+1} \leftarrow z_\Omega$ ;
end
if $(x_i)^{j+1} > 0$ then $(x_i)^{j+1} \leftarrow 1$ else $(x_i)^{j+1} \leftarrow 0$
```

At first, we define two approximations algorithms, namely tail approximation and head approximation.

The tail approximation is defined as follows:

$$\Omega = \mathrm{TAILAPPROX}(b, G, s, g) \tag{10}$$

And the head approximation is defined as follows:

$$\Gamma = \mathrm{HEADAPPROX}(b, G, s, g) \tag{11}$$

The tail approximation returns a sub-graph $\Omega \in \mathcal{M}$ such that $\|b - b_\Omega\| \leq c_T \cdot \min_{\Omega' \in \mathcal{M}} \|b - b_{\Omega'}\|$, while the head approximation returns a sub-graph $\Gamma \in \mathcal{M}$ such that $\|b_\Gamma\| \geq c_H \cdot \max_{\Gamma' \in \mathcal{M}} \|b_{\Gamma'}\|$. Here $c_T > 1$ and $c_H < 1$ are arbitrary, fixed constants. We note that, if $c_T = c_H = 1$, then these two approximation algorithms provides the exact solution of the model projection. These two nearly-linear time approximation algorithms with the above complementary approximation guarantees are proposed by Hedge et. al [8].

Theorem 2. *[34] Let $x \in \mathbb{R}^N$ be an optimum solution of Problem 9. Graph-IHT returns an estimate $\hat{x}$ such that:*

$$\|x - \hat{x}\|_2 \leq c\|\nabla f(x)\|_2 \tag{12}$$

where $c = 1 + \beta/(1 - \alpha)$ is a fixed constant, $\alpha = (1 + c_T)(\delta + \sqrt{1 - \alpha_0^2})$, $\beta = (1 + c_T)(\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2+\sqrt{1+\delta}}})$, $\alpha_0 = c_H(1 - \delta) - \delta, \beta_0 = (1 - \delta)(1 + c_H), \forall 0 < \delta < 1$.

Moreover, Graph-IHT runs in time:

$$O\Big(|\mathbb{E}| \cdot \log^3 N \cdot \log(\|x\|_2/\|\nabla f(x)\|_2)\Big) \tag{13}$$

Proof 2. *A sketch of the proof is presented to interpret this theorem. Applying a number of algebraic manipulations, it can be proved that the i-th iteration of Graph-IHT satisfies:*

$$\|x - x^i\| \leq \alpha\|x\|_2 + \frac{\beta}{1 - \alpha}\|\nabla f(x)\| \tag{14}$$

Then, after $t = \left\lceil \log\left(\frac{\|x\|_2}{\|\nabla f(x)\|_2}\right) / \log\frac{1}{\alpha} \right\rceil$ iterations, Graph-IHT will return an estimate $\hat{x}$ satisfying $\|x - \hat{x}\|_2 \leq (1 + \frac{\beta}{1-\alpha})\|\nabla f(x)\|_2$.

The time complexities of both head approximation and tail approximation are $O(|\mathbb{E}| \log^3 N)$, where $|\mathbb{E}|$ is the number of edges and N is the number of nodes. And the total number of iterations is $\left\lceil \log\left(\frac{\|x\|_2}{\|\nabla f(x)\|_2}\right) / \log\frac{1}{\alpha} \right\rceil$. Therefore, we can derive the overall time complexity is $O(|\mathbb{E}| \cdot \log^3 N \cdot \log(\|x\|_2/\|\nabla f(x)\|_2))$.

The whole Graph-TSS is shown in Algorithm 2, and the implementation of Graph-TSS can be divided into two stages. On the first stage, we calculate Graph-TSS of each vertex, given the sub-graph S, after learning the topic distribution of sub-graph by employing the *mixture of unigrams* model and EM algorithm. On the second stage, we use the Graph-IHT that maximizes the Graph-TSS over all possible subgraphs to identify the most anomalous sub-graph. We repeat the progress described above until the results converge.

Algorithm 2: Graph-TSS

```
Input: Collection of geocoded documents $D$, $\phi, \pi$
       learnt from historical data,$G, s, g$,number of
       iterations $t$
Result: The most anomalous sub-graph $S$
while $S$ not converge do
    while $\theta$ not converge do
        E-step: Compute hidden variable distribution
        $p(z|d)$, Eqn.(3);
        M-step: Update topic distribution $\theta$ in
        sub-graph $S$, Eqn.(4);
    end
    Compute $f(x)$ for each location, Eqn(5);
    S ← Graph-IHT($G, s, g, t, f$); See Algorithm 1 and
    Eqn.(6-11);
end
```

5. EXPERIMENTAL EVALUATION

This section evaluates the effectiveness and efficiency of the proposed Graph-TSS based on comprehensive experiments on Twitter data. We considered the detection of civil unrest events such as protests and strikes, the detection of earthquake, and the detection of influenza outbreak as three case study scenarios, but the proposed Graph-TSS can also be directly applied to other applications, such as the detection of rare disease and local festival.

5.1 Experimental Design

5.1.1 Datasets:

1) Civil Unrest Dataset. We collect 22,728,052 tweets (nearly ten percent of all the raw Twitter data of Argentina) from April 1, 2013 to March 31, 2014. The civil unrest event labels, called Golden Standard Report (GSR), were collected and confirmed from local newspaper that are accessible from Internet. According to the geographical information of Argentina, we construct a connected city-city network with 2057 nodes and 15832 edges.

2) Earthquake Dataset. We collect 5,548,926 tweets of Chile from July 1, 2013 to June 30, 2014. The earthquake records were reported by United States Geological Survey[31]. This institution weekly publishes earthquake reports all over the world for scientific research. Similarly, we also construct a connected city-city network with 897 nodes and 4862 edges, based on the Chile geographical information.

3) Influenza disease outbreak. We collect 27,592,005 tweets during April 1, 2014 to March 31, 2015 in the United States. The influenza outbreaks event labels are reported by the Centers for Disease Control and Prevention (CDC) [2]. The CDC weekly publishes the results related to influenza-like illness (ILI) within each major region in the United States. A connected network with 3042 nodes and 21094 edges is constructed as well.

In these three datasets, we use the first six months data as training data, and use the rest data as testing data to evaluate the performance of methods. Training data is used to learn the topic-word distribution, normal area topic distribution and the criterion or threshold to raise an alert. We assume that the historical training data is sufficient to identify all the potential topics, and an emerging spatial event

Table 3: Comparison between Graph-TSS and Existing Methods on the Civil Unrest datasets

Method	Oct,2013		Nov,2013		Dec,2013		Jan,2014		Feb,2014		March,2014	
	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR
LGTA	0.073	0.23	0.075	0.226	0.066	0.192	0.066	0.115	0.065	0.13	0.068	0.20
STLocal	0.331	0.332	0.413	0.42	0.362	0.31	0.324	0.28	0.353	0.324	0.382	0.324
Graph-Laplacian	0.211	0.18	0.202	0.241	0.214	0.193	0.216	0.17	0.21	0.251	0.223	0.246
NPHGS	0.05	0.271	0.046	0.46	0.048	0.23	0.049	0.27	0.051	0.334	0.046	0.25
Event Tree	0.051	0.33	0.047	0.372	0.047	0.47	0.045	0.321	0.048	0.33	0.052	0.295
Graph-TSS	0.045	**0.561**	0.048	**0.507**	0.042	**0.516**	0.05	**0.355**	0.049	**0.476**	0.045	**0.41**

Table 4: Comparison between Graph-TSS and Existing Methods on the Chile Earthquake datasets

Method	Jan,2014		Feb,2014		March,2014		April,2014		May,2014		June,2014	
	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR
LGTA	0.118	0.228	0.116	0.202	0.126	0.195	0.134	0.202	0.099	0.216	0.10	0.192
STLocal	0.296	0.163	0.303	0.222	0.321	0.236	0.289	0.208	0.341	0.267	0.29	0.221
Graph-Laplacian	0.14	0.141	0.192	0.183	0.231	0.206	0.162	0.136	0.142	0.13	0.24	0.22
NPHGS	0.10	0.263	0.10	0.278	0.11	0.283	0.11	0.198	0.11	0.257	0.10	0.2
Event Tree	0.098	0.42	0.097	0.339	0.093	0.381	0.10	0.16	0.10	0.27	0.098	0.30
Graph-TSS	0.104	**0.49**	0.101	**0.45**	0.098	**0.44**	0.104	**0.37**	0.098	**0.42**	0.098	**0.484**

is characterized as a new topic distribution, instead of new topics.

5.1.2 Data Preprocessing:

After we collected raw tweets, several preprocessing steps were conducted for our proposed method and all the comparison partners, including: **1) Tweet Geocoding:** We implemented a geocoding library for tweets based on three major rules with priorities. For each tweet, we first searched for location and landmark mentions in the tweet text, then for geotags that are available if the user enabled the geocoding function in his/her phone, and finally for location information from the users profile. The first location information identified was returned as the geographic location of this tweet; **2) Vocabulary Generation:** We first generated a vocabulary of around 1000 terms related to civil unrests, a vocabulary of around 300 terms related to earthquake, and a vocabulary of around 200 terms related to influenza from domain experts; **3) Stemming:** Python library is used for stemming and removing stop words; **4) Content Filtering:** Only the raw tweets that match more than two terms from the vocabulary were preserved, in order to remove the unrelated noise. We treat the unrelated tweets as noise and use the topic distribution in normal region to characterize the noise. This strategy is similar to the expected based scan statistics, where the unrelated counts in nodes outside the anomalous cluster are considered as noise and modeled using a specific distribution.

5.1.3 Comparison Partners:

We compare our proposed Graph-TSS with five existing representative methods, including Latent Geographical Topic Analysis (LGTA) [33], STLocal [15], Graph-Laplacian[29], NPHGS[5], EventTree[27]. We strictly followed strategies recommended by authors in their papers to tune the related model parameters. To make a fair comparison, the number of connected components g is set as 1 in Graph-TSS. The upper bound on the number of vertices s is set as 30 based on the training data.

5.1.4 Performance Metrics:

To evaluate our models quantitatively, we employ the following metrics, namely, 1) false positive rate (FPR), 2) true positive rate (TPR). For each method, the reported alerts are structured as tuples of (date, location), where "location" is defined at the city level (e.g. 2057 cities in Argentina, 897 cities in Chile, and 3042 in United States). For each Golden Standard Report Event, United States Geological Survey Event or Centers for Disease Control and Prevention Reports, we decide whether the method had an alert in the city within 7 days.

5.2 Event Detection Result

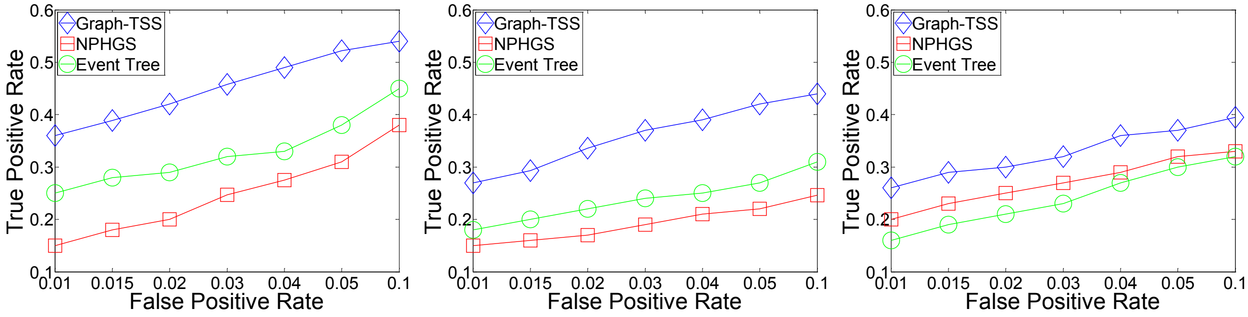
Table 3 demonstrates the comparison between the proposed Graph-TSS and other competing methods for Argentina civil unrest event detection in different months. In the civil unrest event detection, we fix the false positive rates as around 0.05 and show the best true positive rates results for NPHGS, EventTree and Graph-TSS. For the other methods, we use the optimal parameters papers and authors recommended. According to Table 3, Graph-TSS achieved much higher TPR than all competing methods, for comparable false positive rates. For each month, the TPR of Graph-TSS can outperform much more than the TPR of other methods.

Table 4 demonstrates the comparison between the proposed Graph-TSS and other competing methods for Chile earthquake event detection in different months, and table 5 demonstrates the comparison between the proposed Graph-TSS and other competing methods for United States influenza outbreak detection in six months. The results indicate consistent patterns as observed in Table 3. These two tables show that our proposed Graph-TSS performs the best on both Chile earthquake and United States influenza outbreak data sets in the six months.

We note that the true positive rates of all tested methods were lower than 55%, perhaps because some GSR events or other events did not produce strong signals in the noisy Twitter data, or because an alert was only considered "cor-

Table 5: Comparison between Graph-TSS and Existing Methods on the Influenza disease outbreak datasets

Method	Jan,2014		Feb,2014		March,2014		April,2014		May,2014		June,2014	
	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR
LGTA	0.128	0.217	0.113	0.223	0.136	0.197	0.137	0.199	0.102	0.204	0.111	0.212
STLocal	0.316	0.203	0.283	0.212	0.291	0.219	0.319	0.228	0.323	0.234	0.301	0.216
Graph-Laplacian	0.152	0.191	0.178	0.213	0.221	0.236	0.183	0.186	0.17	0.17	0.22	0.2
NPHGS	0.10	0.313	0.10	0.298	0.11	0.353	0.11	0.338	0.11	0.347	0.10	0.29
Event Tree	0.099	0.34	0.098	0.333	0.097	0.29	0.10	0.31	0.10	0.34	0.10	0.36
Graph-TSS	0.101	**0.43**	0.102	**0.38**	0.100	**0.41**	0.099	**0.37**	0.099	**0.36**	0.099	**0.41**



(a) Civil unrest dataset (b) Chile earthquake dataset (c) Influenza outbreak dataset

Figure 4: TPR in various false positive rates

rect” if it matched both the date and location of a labeled event.

Figure 4 shows the true positive rate results at various false positive rates on the Argentina civil unrest dataset, Chile earthquake dataset and Uunted States influenza outbreak dataset. We averaged over the results of the six months. With increasing value of FPR, the TPR increases and then plateaus. We observe that Graph-TSS consistently outperform the comparison partners for different FPR levels. Compared to the most competitive state of the art method, Graph-TSS improves the TPR by 22%(Argentina civil unrest), 42%(Chile earthquake), and 18% (United States Influenza outbreak) in 10% FPR level, respectively. The results confirm our idea that the latent semantics of textual information in social media can contribute to the spatial event detection.

Table 6: Run times results on Civil Unrest datasets

Method	LGTA	STLocal	Graph Laplacian
Time (Mins)	141.4	135.1	284.4
Method	NPHGS	Event Tree	Graph-TSS
Time (Mins)	35.8	69.8	45.5

Table 7: Run times results on Chile Earthquake datasets

Method	LGTA	STLocal	Graph Laplacian
Time (Mins)	108.4	95.3	204.2
Method	NPHGS	Event Tree	Graph-TSS
Time (Mins)	24.8	35.87	17.16

5.3 Runtime Result

The run time of our proposed method consists of two parts: latent semantics of textual information learning and subgraph detection. As shown in Table 6, Table 7 and Table 8, run times of Graph-TSS are comparable to baselines methods on Argentina civil unrest, Chile earthquake, and United States influenza outbreak data sets. It is only slightly higher than NPHGS on civil unrest and influenza outbreak data sets, because NPHGS is a greedy method, rather than an approximation algorithm. Our approximation algorithm Graph-IHT can detect the sub-graph in nearly-linear time $O(|\mathbb{E}| \log^3 |\mathbb{V}|)$.

6. CONCLUSION

This paper presents the Graph Topic Scan Statistic for spatial event detection by using textual information in social media. Because the spatial event detection under Graph-TSS is NP-hard, this paper also designs an efficient approximation algorithm to efficiently maximize the Graph-TSS over connected sub-graphs to identify the most anomalous region. This work performs experiments on real Twitter data. The empirical results demonstrate that Graph-TSS can effectively detect Argentina civil unrest events, Chile earthquake outbreak events, and United States influenza events, outperforming the competing methods. For future work, we will extend our work to heterogeneous graphs to detect event regions. In addition, we plan to extend a Bayesian framework such that rich domain knowledge can be naturally integrated.

Acknowledgments

The paper is partially supported by GRF Grant 17201414 from Hong Kong Research Grant Council.

This paper is also supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to

Table 8: Run times results on Influenza disease outbreak datasets

Method	LGTA	STLocal	Graph Laplacian
Time (Mins)	183.5	175.8	394.2
Method	NPHGS	Event Tree	Graph-TSS
Time (Mins)	43.1	78.87	53.36

reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

7. REFERENCES

[1] How fast the news spreads through social media. *http://blog.sysomos.com/2011/05/02/how-fast-the-newsspreads-through-social-media/*, 2012.

[2] Centers for disease control and prevention. *http://www.cdc.gov/flu/weekly/*, 2016.

[3] A. Agovic and A. Banerjee. Gaussian process topic models. *arXiv preprint arXiv:1203.3462*, 2012.

[4] A. Ahmed, L. Hong, and A. J. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36. WWW, 2013.

[5] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1166–1175. ACM, 2014.

[6] Digital-Trend. http://www.digitaltrends.com/social-media/weibo-monitoring-bird-flu-china/. 2015.

[7] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, pages 181–192. VLDB Endowment, 2005.

[8] C. Hegde, P. Indyk, and L. Schmidt. A nearly-linear time framework for graph-structured sparsity. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 928–937, 2015.

[9] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM, 2012.

[10] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[11] C. C. Kling, J. Kunegis, S. Sizov, and S. Staab. Detecting non-gaussian geographical topics in tagged photo collections. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 603–612. ACM, 2014.

[12] J. Konemann, S. Sadeghian, and L. Sanita. An lmp o (log n)-approximation algorithm for node weighted prize collecting steiner tree. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 568–577. IEEE, 2013.

[13] M. Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[14] M. Kulldorff, F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman, and R. Platt. Multivariate scan statistics for disease surveillance. *Statistics in medicine*, 26(8):1824–1833, 2007.

[15] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras. On the spatiotemporal burstiness of terms. *Proceedings of the VLDB Endowment*, 5(9):836–847, 2012.

[16] Y. Liu, M. Ester, B. Hu, and D. W. Cheung. Spatio-temporal topic models for check-in data. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 889–894. IEEE, 2015.

[17] Y. Liu, G. Luo, and Y. Zhang. Response surface modeling by local kernel partial least squares. In *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, pages 269–276. IEEE, 2012.

[18] E. McFowland, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research*, 14(1):1533–1561, 2013.

[19] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web*, pages 533–542. ACM, 2006.

[20] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239, 2010.

[21] D. B. Neill. *Detection of spatial and spatio-temporal clusters*. PhD thesis, University of South Carolina, 2006.

[22] D. B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, 8(1):1, 2009.

[23] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 218–227. ACM, 2005.

[24] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng. On detection of emerging anomalous traffic patterns using gps data. *Data & Knowledge Engineering*, 87:357–373, 2013.

[25] G. Patil, C. Taillie, et al. Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science*, 18(4):457–465, 2003.

[26] J. Qian, V. Saligrama, and Y. Chen. Connected sub-graph detection. In *AISTATS*, volume 14, pages 22–25, 2014.

[27] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1176–1185. ACM, 2014.

[28] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[29] J. Sharpnack, A. Rinaldo, and A. Singh. Changepoint detection over graphs with the spectral scan statistic. *arXiv preprint arXiv:1206.0773*, 2012.

[30] S. Shiode. Street-level spatial scan statistic and stac for analysing street crime concentrations. *Transactions in GIS*, 15(3):365–383, 2011.

[31] USGS. http://earthquake.usgs.gov/. 2014.

[32] N. Wu, F. Chen, J. Li, B. Zhou, and N. Ramakrishnan. Efficient nonparametric subgraph detection using tree shaped priors. 2016.

[33] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, pages 247–256. ACM, 2011.

[34] B. Zhou and F. Chen. Technical report: Graph-structured sparse optimization for connected subgraph detection. In *arXiv preprint*, 2016.