A Bisubmodular Approach to Event Detection and Prediction in Multivariate Social Graphs

Shuai Zhang, Haoyi Zhou, Feng Chen, and Jianxin Li

***Abstract*— A burst event on a social graph is usually framed as an anomalous and unexpected pattern that is characterized as a compact or correlated subset of affected vertices, which is a subgraph. Subgraph detection becomes a serious problem when social graphs involve multiple attributes (i.e., multivariate graph). Most existing methods are not capable of handling the feature selection and subgraph detection problems simultaneously on the multivariate graph. In this article, we propose multivariate anomalous subgraph scanning (MASS), a generic model that detects anomalous events on the multivariate social graph. First, we reformulate the traditional nonparametric statistics as a new statistical objective function that simultaneously measures the significance of a vertices subset and an attributes subset to generate an indicator of ongoing or upcoming events. Then, we reformulate the objective function as the difference between two bisubmodular functions and approximate it with a bisubmodular objective function, which can be optimized in linear time, with an analysis of its theoretical properties. We demonstrate the performance of our proposed method using two burst event detection and prediction tasks from the real world.**

***Index Terms*— Anomaly detection, bisubmodular, graph, social network.**

I. Introduction

With the development of Mobile Internet, social microblogs, such as Weibo, Twitter, Facebook, and Instagram, have played a critical role for people to get in touch and discuss daily events [1]–[4]. Furthermore, an increasing number of governments, enterprises, and individuals register social network accounts to spread and acquire particular events. Comparing with the traditional event propagation such as newspaper, notice, and message, microblogs provide a much faster way to transmit information and also an interactive tunnel cooperating with many different kinds of "super" information.

This article aims to contribute to the detection and prediction problem of domain-specific events, such as air pollution events, disease outbreak events, and crime hot-spot events.

Manuscript received November 6, 2019; revised January 3, 2020; accepted January 20, 2020. Date of publication February 25, 2020; date of current version January 29, 2021. This work was supported in part by the China NSFC Program under Grant 61872022 and Grant 61421003 and in part by the Beijing Advanced Innovation Center for Big Data and Brain Computing. *(Corresponding author: Jianxin Li.)*

Shuai Zhang, Haoyi Zhou, and Jianxin Li are with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, and also with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: lijx@buaa.edu.cn).

Feng Chen is with the School of Computer Science, The University of Texas at Dallas, Richardson, TX 75080-3021 USA.

Digital Object Identifier 10.1109/TCSS.2020.2971756

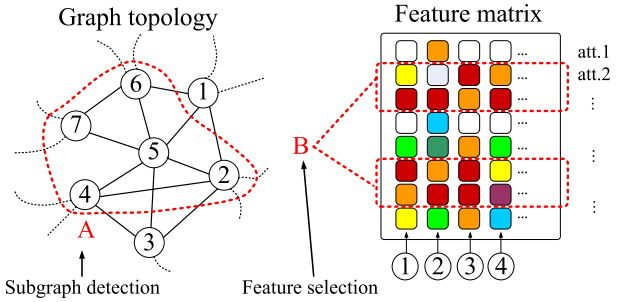

Fig. 1. Our proposed approach focuses on searching a compact subset of vertices (A) and a subset of features (B) that are jointly the most anomalous. In comparison, most existing methods assume that all the features are relevant and mainly focus on the subgraph detection process.

Social networks are naturally constructed as multivariate graphs, whose vertices are entities (e.g., users and geographic locations) and edges are relationships (e.g., follower and spatial neighborhood) and attributes as features of nodes (e.g., frequencies of domain-specific keywords). Given a multivariate social graph, events in the real world can be framed as anomalous subgraphs on the social graph. Thus, the event detection and prediction problem is to detect or forecast the most anomalous subgraph (cluster) of the social graph, and each subgraph refers to a detection of an ongoing event or a prediction of an upcoming event.

Most of the existing methods of anomalous subgraph detection search for subgraphs with the most anomalous attributes overall under the hypothesis that the set of relevant attributes is already known. Burkom [5] converted the multivariate subgraph detection problem into a univariate problem by aggregating multiple attributes of each vertex to a simple uni-variate. Kulldorff *et al.* [6] proposed a multivariate scan statistics method for disease surveillance problem, and they calculated the individual log likelihoods as scores for every attribute and summed up the scores of all attributes into a single score. Lappas *et al.* [7] proposed a brunch-and-bounding approach searching areas where the total frequency of the pre-defined terms is abnormally higher than the outer areas. Chen and Neill [1] calibrated multiple features of each node of the graph into an empirical p-value through a two-stage process. The empirical p-value of a node represents the probability of observing a new random sample with more abnormal attributes than the present feature of the node. In other words, they extract the empirical p-values as features. However, representing multiple attributes with one variable (such as p-value) has the potential to lose some useful

information of the events of interests. Shao *et al.* [2] proposed a high-efficiency evolution subgraph discovery method using tree prior based on nonparametric statistics.

Different aggregation functions of multivariate attributes are presented earlier, assuming that the predefined attributes are mostly signal attributes implicitly. However, this assumption is inappropriate for event detection and prediction in social graphs as the dynamics of event-driven attributes. Different events usually have different contexts, and the correlated attributes are unpredictable. Therefore, the detection of attributes, which are related to the ongoing or upcoming events, becomes more difficult.

Actually, a lot of keywords in a large dictionary are needed to be tracked, and usually, only a small subset of keywords are related to a specific event. An aggregation of all the attributes will be potentially dominated by the majority of noise attributes. In terms of this, we consider a different optimization approach to optimize an objective function about "interestingness" or "anomalousness" over all subsets of vertices and attributes. As shown in Fig. 1, we focus on searching a compact subset of vertices and a subset of features that are jointly the most anomalous. Moreover, this optimization task involves a critical computational challenge: the exhaustive search over all subsets of vertices and attributes is computationally infeasible and leads to an exponential scale as the number of attributes and vertices increases. To the best of our knowledge, limited work has been performed to solve this computational challenge. Neill [8] proposed a heuristic algorithm to iteratively maximize the spatial scan statistic functions over subsets of vertices and attributes in a multivariate spatial graph until convergence. This algorithm is sensitive to the initial settings and does not have known theoretical properties on the quality of the detected subsets.

This article makes the following main contributions.

1) *Proposing MASS Model:* We propose a generic model multivariate anomalous subgraph scanning (MASS) to solve the detection and prediction problem of domain-specific events on a multivariate social graph. Events are detected and predicted as multidimensional node subsets and attribute subsets, respectively. The feature significance of nodes and attributes is parameterized with a nonparametric scan statistic, which is distribution assumptions free.
2) *Designing Approximation Process for MASS:* We reformulate the problem of MASS and prove that the optimization problem is NP-hard. Then, we rewrite the original objective function as the difference of two bisubmodular functions, derive a tight bisubmodular lower bound of the original objective function, and propose a random greedy algorithm that optimizes the lower bound in linear time. Our algorithm guarantees the convergence to a local optimum within linear time under certain conditions. We believe that we are the first to detect a multivariate anomalous subgraph by optimizing an approximated bisubmodular function.
3) *Comprehensive Experiments:* The effectiveness and efficiency of MASS are validated via comprehensive experiments on the Twitter data and Weibo data. The results demonstrate that our method outperforms representative competitive methods.

The structure of this article is organized as follows. Section II presents the preliminaries on multivariate graph, p-value, and nonparametric statistics. Section III presents the proposed MASS model for MASS. Section IV presents a linear time approximation algorithm to the MASS problem, with its theoretical analysis. Section V presents the experiments on the Weibo and Twitter data sets, and Section VI describes the conclusion and future work.

II. Preliminaries

In this section, three key relevant definitions are presented, including multivariate graph, nonparametric scan statistics, and bisubmodular.

A. Multivariate Graph

Definition 1 (Multivariate Graph $\mathcal{G}$): A multivariate graph $\mathcal{G}$ is defined as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f)$, where $\mathcal{V}$ represents the ground set of vertices, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the edge set (relations), and $f : \mathcal{V} \to \mathcal{R}^D$ represents a map function that maps each vertex v to a D-dimensional feature vector $f(v)$, in which D denotes the total feature number.

In this article, we consider a snapshot graph $\mathcal{G}$ of Weibo and Twitter in the present day as two case studies. In a multivariate graph $\mathcal{G}$, vertex v represents a user of Weibo or Twitter. The p-value of the dth feature of vertex v is denoted as $f_d(v)$, and it represents the frequency of keyword d in user v's tweets in the current day. For each feature d of vertex v, we estimate the importance of this feature by the statistical p-value based on its empirical distribution, denoted as $p_d(v)$. The p-value $p_d(v)$ is calculated as [1], [9]

$$p_d(v) = \frac{1}{T}\sum_{t=1}^{T} I(f_d(v^{(t)}) \geq f_d(v)), \quad d = 1, \ldots, D.$$

Intuitively, the p-value is an anomaly evaluation within the range $[0, 1]$: the smaller the p-value, the higher anomalous degree. We are ready to present the nonparametric statistics for measuring the anomalousness of a group of p-values of the vertex subset and feature subset.

B. Nonparametric Statistics

Definition 2 (Nonparametric Statistics [1]): S is a set of p-values. The aggregation function over S is $G(S)$, which is a nonparametric scoring functions measuring the joint significance of all p-values in S. $G(S)$ is defined as

$$G(S) = \phi(\alpha, N_\alpha(S), N(S)) \tag{1}$$

where α is a predefined significance level of p-values (0.05 by default), $N_\alpha(S)$ denotes the number of p-values in S which is smaller or equal to α, and the function $\phi(\alpha, N_\alpha, N)$ satisfies two intuitive properties defined in the following.

1) ϕ increases monotonically with N_α.
2) ϕ decreases monotonically with both N and α.

In our model, ϕ can be any function that satisfies the above-mentioned properties. To be specific in this article, we use Berk–Jones (BJ) statistics [10] as an illustration since many real-world applications show the effectiveness of BJ

statistic in the anomalous subgraph detection [1], [11]–[13]. It is defined as

$$\phi_{BJ}(N_\alpha(S), N(S), \alpha) = N(S) \times \mathrm{KL}(N_\alpha(S)/N(S), \alpha) \quad (2)$$

where KL is the Kullback–Leibler divergence defined as

$$\mathrm{KL}(a, b) = a \log(a/b) + (1-a) \log((1-a)/(1-b)).$$

The BJ statistics is the KL divergence between the observed and expected proportions of p-values that are less than α. It is the ratio of log likelihood on whether the empirical p-values follow a uniform distribution or a piecewise constant distribution. Berk and Jones [10] demonstrated that the BJ statistic fulfills several optimality properties, and it has greater power than any weighted Kolmogorov statistic.

III. Multivariate Anomalous Subgraph Scanning

In this section, we generalize nonparametric statistics and propose a multivariate subgraph scan statistic functions for MASS

$$F(A, B) = \phi(\alpha, \psi(A, B, \alpha), N(A) \cdot N(B)) \quad (3)$$

where $A \subseteq \mathcal{V}$ refers to a subset of vertices, $B \subseteq \{1, 2, \ldots, D\}$ refers to a subset of attributes, $N(.)$ is the cardinality function, and $N(A)$ and $N(B)$ are the sizes of subset A and subset B, respectively. $\psi(A, B, \alpha) = \sum_{v \in A, d \in B} 1(p_d(v) \leq \alpha)$ denotes the count of p-values among all $p_d(v)$ relevant to A and B that are smaller or equal to α. Function ϕ is BJ statistic, which is defined in Definition 2. We take the multivariate subgraph scan statistic function $F_{BJ}(A, B)$ that is on the strength of the BJ statistic [see (2)] as a case study. The model we proposed will also be suitable for other multivariate subgraph scan statistic functions

$$\begin{aligned} &F_{BJ}(A, B) \\ &\quad = \phi_{BJ}(\alpha, \psi(A, B, \alpha), N(A) \cdot N(B)) \\ &\quad = N(A) \cdot N(B) \times \mathrm{KL}(\psi(A, B, \alpha)/(N(A) \cdot N(B)), \alpha). \end{aligned}$$

Based on the scan statistic function as defined earlier, we consider the following problem formulation.

Problem 1 (Multivariate Anomalous Subgraph Scanning): Given a multivariate graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f)$, find a subset of vertices $A \subseteq \mathcal{V}$ and a subset of attributes $B \subseteq \{1, 2, \ldots, D\}$ that maximize the objective function

$$F_{BJ}(A, B) - \lambda \cdot R(A) \quad (4)$$

where $F(A, B)$ is a multivariate subgraph scan statistic function [see (3)] measuring the level of anomalousness of the subsets A and B, $R(A)$ is a submodular function measuring the compactness of the subgraph induced by A, the smaller the value of D the more compact the subgraph, and λ is a tradeoff specified by the user.

The set function $R(A)$ is submodular if it satisfies the diminishing return property; for every $X, Y \subseteq \mathcal{V}$ with $X \subseteq Y$ and every $x \in \mathcal{V} \setminus x$, we have that $D(X \cup \{x\}) - D(X) \geq D(Y \cup \{x\}) - D(Y)$. A number of popular compactness functions naturally satisfy the submodular property, including the graph cut function [14], the summation of distances of all pairs of vertices in A [15], and the connectivity function that is "1 minus the number of connected components in A": $R(A) = N(\mathcal{V}) - (N(A) - c(A)) + 1$, where $c(A)$ is the number of connected components in the subgraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_A, f)$, $\mathcal{E}_A$ refers to the set of edges corresponding to A, and $N(A) - c(A)$ is the number of connected components of the subgraph $\mathcal{G} = (\mathcal{V} \setminus A, \mathcal{E}_A, f)$.

In order to reformulate the MASS problem as a new form related to bisubmodular optimization, we define the super ground set $\tilde{V} = \mathcal{V} \cup \{1, \ldots, D\}$ and allow A and B to be subsets of $\tilde{V}$. When $A \not\subset \mathcal{V}$ or $B \not\subset \{1, \ldots, D\}$, $F(A, B) = F(A \cap \mathcal{V}, B \cap \{1, \ldots, D\})$ and $R(A) = D(A \cap \mathcal{V})$. A biset function is bisubmodular if it satisfies the biset version of the diminishing return property that will be introduced next. Theorem 1 shows that the objective function of the MASS problem can be reformulated as the difference between two bisubmodular functions, and Theorem 2 presents two important theoretical properties of the MASS problem.

Theorem 1 (Bisubmodular Reformulation): The objective function of the MASS problem can be reformulated as the difference between two bisubmodular functions

$$F_{BJ}(A, B) - \lambda \cdot R(A) = F_1(A, B) - F_2(A, B) \quad (5)$$

where

$$F_1(A, B) = -N(A) \cdot N(B) \log(N(A) \cdot N(B)) - \psi(A, B, \alpha) \log \alpha$$

and

$$\begin{aligned} F_2(A, B) = &-\psi(A, B, \alpha) \cdot \log \psi(A, B, \alpha) \\ &- (N(A) \cdot N(B) - \psi(A, B, \alpha)) \log(1 - \alpha) \\ &- (N(A) \cdot N(B) - \psi(A, B, \alpha)) \\ &\cdot \log(N(A) \cdot N(B) - \psi(A, B, \alpha)) \\ &- \lambda \cdot R(A). \end{aligned}$$

Proof: To prove the equivalence in (5), it suffices to prove that each additive component in $F_1(A, B)$ and $F_2(A, B)$ is bisubmodular. The bisumodularity of a function $f : 2^{2\tilde{V}} \to \mathbb{R}$ can be proved using the following diminishing return property: $\forall (A, B) \in 2^{2\tilde{V}}, (A', B') \in 2^{2\tilde{V}}$ with $A \subseteq A'$ and $B \subseteq B'$, we have for each $v \notin A'$ and $v \notin B'$: $f(A \cup \{v\}, B) - f(A, B) > f(A' \cup \{v\}, B')$ and $f(A, B \cup \{v\}) - f(A, B) > f(A', B' \cup \{v\})$. By applying this property, we can readily prove that the functions "$-N(A) \cdot N(B)$," "$-\psi(A, B, \alpha) \log \alpha$," "$-\psi(A, B, \alpha) \log(1 - \alpha)$," "$-(N(A) \cdot N(B) - \psi(A, B, \alpha))$," and "$\lambda \cdot R(A)$" are bisubmodular, where "$N(A) \cdot N(B) - \psi(A, B, \alpha)$" refers to the number of p-values greater than α among those related to A and B. The submodularity of other additive components can be proved that using properties, such as a composition of a decreasing convex function and an increasing bisubmodular function, is still bisubmodular. ■

Theorem 2 (Theoretical Properties of the MASS Problem): The MASS problem has two main theoretical properties: 1) the MASS problem is NP-complete and 2) the MASS problem is a non-bisubmodular optimization problem.

Proof: About the NP-hardness, it suffices to show that this problem is NP-hard when λ is a finite positive value, and the compactness function $R(A)$ is the cut function of the subgraph induced by A. In this special case, this problem is equivalent

to the traditional minimum two-cut problem that is known to be NP-complete [16]. According to Theorem 1, the objective function of the MASS problem can be reformulated as the difference between two bisubmodular functions, which indicates that the objective function is neither bisubmodular nor bisupermodular. ■

IV. Approximation Algorithms

Theorem 2 indicates that the MASS problem is a non-bisubmodular optimization problem, and hence, the existing bisubmodular optimization algorithms are not directly applicable to this problem. We apply the well-known majorization–maximization framework [17]–[20] by replacing the bisubmodular function $F_2(A, B)$ using a tight bimodular upper bound function of this function. The majorization–maximization framework has a number of iterations. In a specific iteration k, suppose that the intermediate solution obtained at the previous iteration is denoted as $(A^{(k-1)}, B^{(k-1)})$. Denote the tight upper bound bimodular function of $F_2(A, B)$ as $\hat{F}_2^{(k)}(A, B)$ that satisfies the following two conditions.

1) *Tightness Condition*:

$$\hat{F}_2^{(k)}(A, B) = F_2(A^{(k-1)}, B^{(k-1)})$$

if $A = A^{(k-1)}$ and $B = B^{(k-1)}$.

2) *Upper Bound Condition*:

$$\hat{F}_2^{(k)}(A, B) \geq F_2(A, B) \quad \forall A, B \subset \tilde{\mathcal{V}}.$$

To define the upper bound function $\hat{F}_2^{(k)}(A, B)$, we first define the supergradients of $\hat{F}_2$ with respect to A and B. The supergradient of $\hat{F}_2$ with respect to A has the form

$$\hat{g}_1(j) = F_2(\tilde{V} \setminus \{j\}, B^{(k-1)}) - F_2(\tilde{V}, B^{(k-1)})$$

if $j \in A^{(k-1)}$; otherwise

$$\hat{g}_1(j) = F_2(A^{(k-1)} + \{j\}, B^{(k-1)}) - F_2(A^{(k-1)}, B^{(k-1)}).$$

The supergradient of $\hat{F}_2$, namely, $\hat{g}_2$, is similarly defined. The upper bound function of F_2 then has the form

$$\begin{aligned}\hat{F}_2^{(k)}(A, B) = {} & F_2(A^{(k-1)}, B^{(k-1)}) \\ & + \hat{g}_1(A) + \hat{g}_2(B) \\ & - \hat{g}_1(A^{(k-1)}) - \hat{g}_2(B^{(k-1)}).\end{aligned} \tag{6}$$

The objective function of the MASS problem can then be approximated by its tight lower bound bisubmodular function $F_1(A, B) - \hat{F}_2^{(k)}(A, B)$, which is identical to $F_{BJ}(A, B)$, if $A = A^{(k-1)}$ and $B = B^{(k-1)}$ and, otherwise, is lower than or equal to $F_{BJ}(A, B)$. We obtain an approximated bisubmodular maximization problem of the original MASS problem

$$\max_{A,B \subseteq \tilde{\mathcal{V}}} \quad F_1(A, B) - \hat{F}_2^{(k)}(A, B). \tag{7}$$

Expression (7) can be approximately solved using a randomized greedy algorithm that has the guaranteed approximation factor 2 [21]. The proposed overall algorithm for the MASS problem is shown in Algorithm 1. There are two main loops in this algorithm. The outer loop relates to the implementation of the majorization–maximization framework, and the inner loop relates to the implementation of the randomized greedy algorithm for approximately solving the bisubmodular maximization subproblem 7. Within the inner loop, Lines 6 and 7 calculate the marginal gains Δ_A and Δ_B of the objective function in (7) for the new vertex v with respect to A and B, respectively. Lines 8 and 9 ensure that the marginal gains Δ_A and Δ_B should be at least 0. Lines 11–16 randomly add an element v to $\tilde{A}$ (or $\tilde{B}$) with probability proportional to the resulting marginal gain Δ_A (or Δ_B) with respect to the current solution $(\tilde{A}, \tilde{B})$.

Algorithm 1 MASS Algorithm

Input: The multivariate graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f)$
Result: The subsets of vertices and features: A and B.

1 $k \leftarrow 0$, $A^{(k)} \leftarrow \emptyset$, $B^{(k)} \leftarrow \emptyset$;
2 **repeat**
3 Calculate $\hat{F}_2^{(k)}(A, B)$ via Equation (6) ;
4 $\tilde{A} \leftarrow \emptyset$, $\tilde{B} \leftarrow \emptyset$;
5 **foreach** $v \in \mathcal{V}$ **do**
6 $\Delta_A \leftarrow F_1(\tilde{A} \cup \{v\}, \tilde{B}) - \hat{F}_2^{(k)}(\tilde{A} \cup \{v\}, \tilde{B}) - (F_1(\tilde{A}, \tilde{B}) - \hat{F}_2^{(k)}(\tilde{A}, \tilde{B}))$
7 $\Delta_B \leftarrow F_1(\tilde{A}, \tilde{B} \cup \{v\}) - \hat{F}_2^{(k)}(\tilde{A}, \tilde{B} \cup \{v\}) - (F_1(\tilde{A}, \tilde{B}) - \hat{F}_2^{(k)}(\tilde{A}, \tilde{B}))$
8 $\Delta_A \leftarrow \max(0, \Delta_A)$
9 $\Delta_B \leftarrow \max(0, \Delta_B)$
10 **if** $\Delta_A + \Delta_B \neq 0$ **then**
11 Let $i \in \{1, 2\}$ be chosen randomly with $\Pr[i = 1] \leftarrow \frac{\Delta_A}{\Delta_A + \Delta_B}$ and $\Pr[i = 2] \leftarrow 1 - Pr[i = 1]$
12 **if** $i = 1$ **then**
13 $\tilde{A} \leftarrow \tilde{A} \cup \{v\}$;
14 **else**
15 $\tilde{B} \leftarrow \tilde{B} \cup \{v\}$;
16 **end**
17 **end**
18 **end**
19 $A^{(k+1)} \leftarrow \tilde{A}$, $B^{(k+1)} \leftarrow \tilde{B}$;
20 $k \leftarrow k + 1$;
21 **until** $A^{(k)} \leftarrow A^{(k-1)}$ *and* $B^{(k)} \leftarrow B^{(k-1)}$;
22 **return** $A^{(k)}$ and $B^{(k)}$;

Theorem 3: The MASS algorithm is guaranteed to converge to a local maximum solution of Problem 1 if the intermediate solution $(\tilde{A}, \tilde{B})$ returned by the inner loop (Lines 5–18) at the kth iteration is a local optimum to (7). The MASS algorithm has the time complexity $O(\ell \cdot n)$, where ℓ refers to the number of iterations of the outer loop.

Proof: As $(A^{(k+1)}, B^{(k+1)})$ is an optimum to (7) at each kth iteration, we have the following inequalities:

$$\begin{aligned}& F_1(A^{(k+1)}, B^{(k+1)}) - F_2(A^{(k+1)}, B^{(k+1)}) \\ & \quad \geq F_1(A^{(k+1)}, B^{(k+1)}) - \hat{F}_2^{(k)}(A^{(k+1)}, B^{(k+1)}) \\ & \quad \geq F_1(A^{(k)}, B^{(k)}) - \hat{F}_2^{(k)}(A^{(k)}, B^{(k)}) \\ & \quad \geq F_1(A^{(k)}, B^{(k)}) - F_2(A^{(k)}, B^{(k)})\end{aligned}$$

where the fist inequality follows from the upper bound condition: $F_2(A^{(k+1)}, B^{(k+1)}) \leq \hat{F}_2^{(k)}(A^{(k+1)}, B^{(k+1)})$, the second

inequality follows from the optimality of $(A^{(k+1)}, B^{(k+1)})$, and the third inequality follows from the tightness condition of $\hat{F}^{(k)}$. According to the definition of $\hat{F}^{(k)}$ based on supergradients, it can be readily proved that the function $F_1(A^{(k+1)}, B^{(k+1)}) - F_2(A^{(k+1)}, B^{(k+1)})$ will not increase if we add an arbitrary element to (or remove an arbitrary element from) $A^{(k+1)}$ or $B^{(k+1)}$. The convergence to a local optimal solution of the MASS problem then follows. In every iteration of the outer loop of MASS, it calculates $\tilde{F}_2^{(k)}$ and the inner loop. The time complexity of $\tilde{F}_2^{(k)}$ and inner loop are both $O(n)$, and thus, the time complexity of each run of the outer loop is also $O(n)$. Since the outer loop runs a constant number of times, the time complexity of the entire MASS algorithm is $O(n)$. ■

V. Experiments

In this section, we evaluate the effectiveness and efficiency of the proposed MASS algorithm using two real-world data sets. We considered the detection and prediction of haze and flu outbreak events as two case studies. MASS shows the advantages in anomalous subgraph detection and feature selection compared with other proposed techniques.

A. Experiment Design

1) Data Sets: In this article, we consider two case study scenarios in the real world to evaluate our burst event detection and prediction method MASS.

a) Haze outbreak data set: We collected about 1.4 billion tweets from Weibo in a nine-month period starting from April 11, 2014, which is 10% of total number during that time. The data are cleaned by removing tweets with less than two terms related to haze outbreaks. The dictionary of haze-related terms contains 68 words defined by domain experts. After data cleaning, 0.35 million tweets posted by 49 655 users remain in the data set. Each user is treated as a vertex in the graph, while the edges are constructed by the comention in tweets and following relations between users. The user–user network contains 149 408 edges, and users are geocoded by province from their profiles. The 68 haze-related keywords are the attributes of a vertex. For each day d, we construct a snapshot graph and got 276 snapshot graphs across the nine-month period. In each user u in a snapshot graph, a corresponding empirical p-value is calculated for each keyword using the same approach from [1]. As for the ground truth of haze outbreak, we collected 9384 haze outbreak records (level $\geq$ 3) from official websites (MEP) as Gold Standard Reports (GSRs). An example GSR is like (Province = "Hebei," DAY = "10-06-2014").

b) Flu outbreak data set: Ten percent of all Twitter tweets in the United States across 226 weeks (January 1, 2011 to May 1, 2015) are collected as raw data. A flu outbreak dictionary of 72 keywords is defined by domain experts. After filtering by containing at least two keywords from the dictionary, 0.15 million tweets posted by 39 565 users remain. Similar to the Haze Outbreak data set, a user–user network of 49 204 edges is constructed by connecting users with comentions and following relations. Each user is geocoded by the location profile in terms of state. Over the 226 weeks, there are 226 snapshot graphs. In every snapshot graph, the corresponding empirical p-value is calculated for each week d and user u. The reason why we used weeks here rather than days such as Haze Outbreak data set is that they have different time granularities in the ground-truth data. We collected 2260 flu outbreak records (ILI $\geq$ 2000) from the official website (http://www.cdc.gov/flu/weekly/.) that is maintained by the Centers for Disease Control and Prevention (CDC) as GSRs. CDC publishes weekly influenza-like illness (ILI) activity level for each state based on the proportion of outpatient visits to healthcare providers. An example of a flu outbreak event is: (STATE = "California," COUNTRY = "U.S.," WEEK = "01-20-2013 to 01-26-2013").

2) Data Preprocessing: Here, we describe the details of data preprocessing after the raw data collection as follows.

1) *Dictionary Definition:* Domain experts of haze and flu, respectively, defined a vocabulary of 68 keywords related to haze outbreak and a vocabulary of 72 terms related to flu outbreak.
2) *Content Filtering:* We only selected the tweets that contain more than two different keywords from the dictionary.
3) *Document Geocoding:* For each document, we chose the location under the following principles:
   (a) using locations and landmarks inside the content;
   (b) using position tag, including latitude and longitude value from users' phone;
   (c) using location information from the users' profiles.
4) *Event Record Formatting:* Every event Records (ERs) were written in the following format: ("Time (YYYY-MM-DD) / #Week," "Location (Province / State)," "Report").

3) Baselines: We considered four representative baselines, including EventTree [15], NPHGS [1], LGTA [22], and FSS [23]. We strictly followed the strategies recommended by authors in their articles to tune the related model parameters.

4) Proposed MASS: We denote the proposed approach as MASS. The tradeoff parameter λ was set to 1, and we used graph cut as the compactness function.

5) Metrics: In this article, we focus on the evaluation of event detection and prediction using different approaches. The used evaluation metrics include the following.

1) *FPR* It refers to the proportion of predicting results that correspond to no event record.
2) *True Positive Rate (TPR)* for Prediction: refers to the proportion of events that are successfully predicted.
3) *True positive rate (TPR)* for both *detection and prediction:* refers to the proportion of events that are successfully predicted or detected.
4) *Lead Time for Prediction:* refers to the time before an event is successfully predicted (longer is better);
5) *Lag time for Detection:* refers to the time after an event is successfully detected (shorter is better).

For each comparison approach, the reported alerts are structured as (date and location), where "location" is defined at the state or province level. For different GSR events, a checklist will be applied as follows.

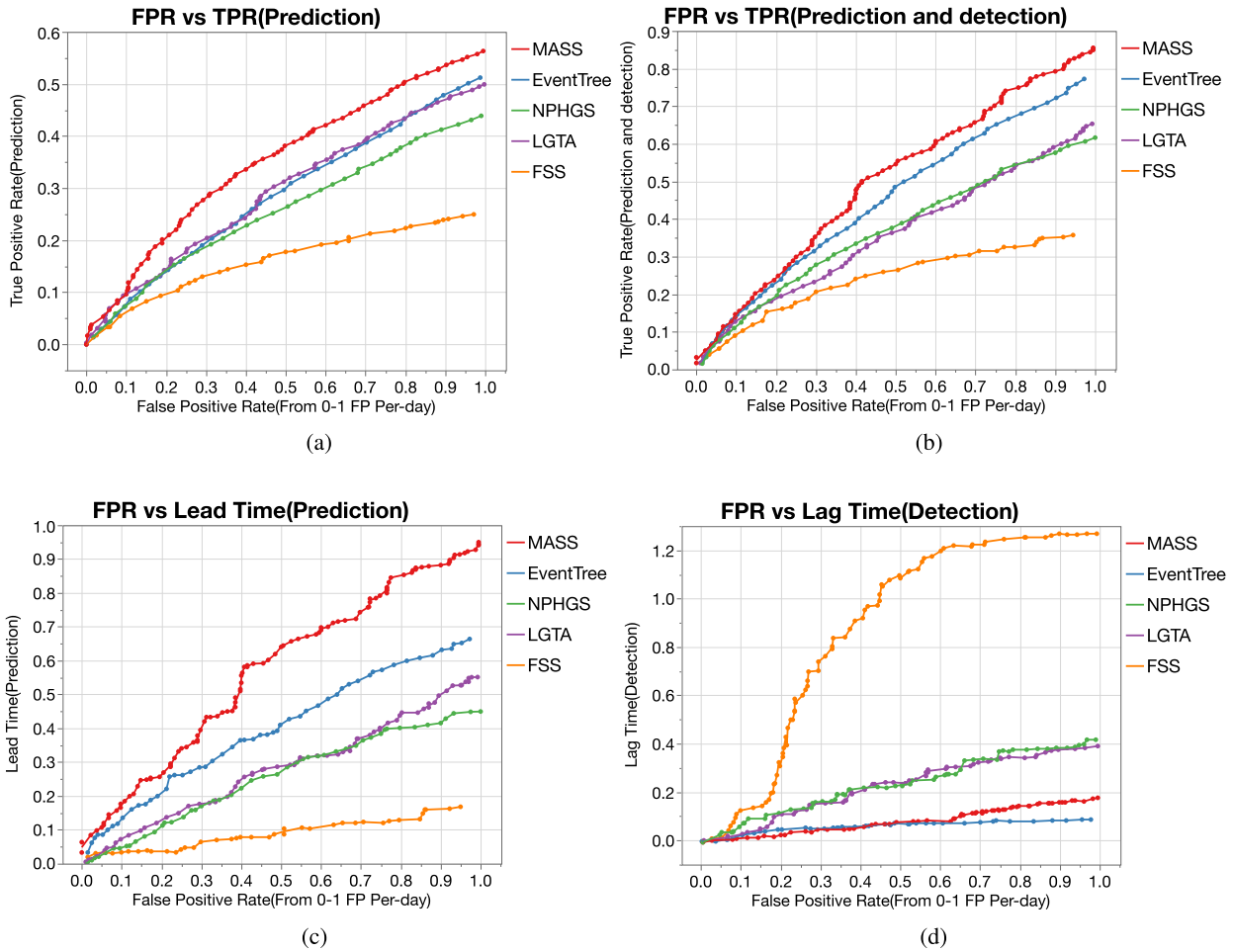

Fig. 2. FPR versus TPR between MASS and four baselines on the Haze data set. (a) FPR versus TPR (prediction). (b) FPR versus TPR (prediction and detection). (c) FPR versus lead time (prediction). (d) FPR versus lag time (detection).

1) Whether the approach raises an alert of a certain state/province within 7 days (for haze data set) 1 week (for flu data set) before that event, which is denoted "predicted."
2) Whether the approach raises an alert of a certain state/province within seven days (for haze data set) or one week (for flu data set) after that event, which is denoted "detected."
3) Whether the approach raises no alert of a certain state/province within seven days (for haze data set) nor one week (for flu data set) before or after that event, which is denoted "undetected."

For the haze data set, the time unit is "day" and for flu data set, the time unit is "week," so we talk about them separately in the above, although seven days and one week indicate for the same time span.

Transforming the Anomalous Subgraph to Event Alerts: To detect abnormal events in every time unit, both baseline and our proposed method return a discovered subgraph of users, together with an anomaly score of the subgraph. The anomaly score of our method is the maximum value of (refupperbound). The next step is to map the anomalous subgraph to outbreak alerts. The location information of users in the subgraph, such as provinces or states, represents the regions that should have an outbreak event alert.

B. Results: Event Detection and Prediction

Figs. 2 and 3 present the comparison between the proposed MASS algorithm with four competitive methods at various FPR for the task of prediction flu outbreak and haze events. The experimental results show that MASS can achieve higher detection TPR and prediction TPR than all competitive approaches. Moreover, on both prediction and detection tasks, the difference between TPR of MASS and those of other baseline approaches tends to increase as FPR increases. In particular, the difference in prediction is more than 20%, and the difference in detection is more than 10%. In addition, MASS achieves larger lead time but lower or similar Lag Time comparing to all other competitive approaches at different false positive rates. The improvement of Lead Time is more than 30%. An example illustration of detection and prediction results of MASS on the Flu data set is shown in Fig. 4.

Among the baseline methods, only FSS and LGTA were designed to conduct subgraph detection and feature selection concurrently. However, these two methods performed worse than EventTree, the competitive method that only conducted subgraph detection but performed the second best on all the metrics. Although FSS and LGTA have considered feature selection during the subgraph detection process, their strategies

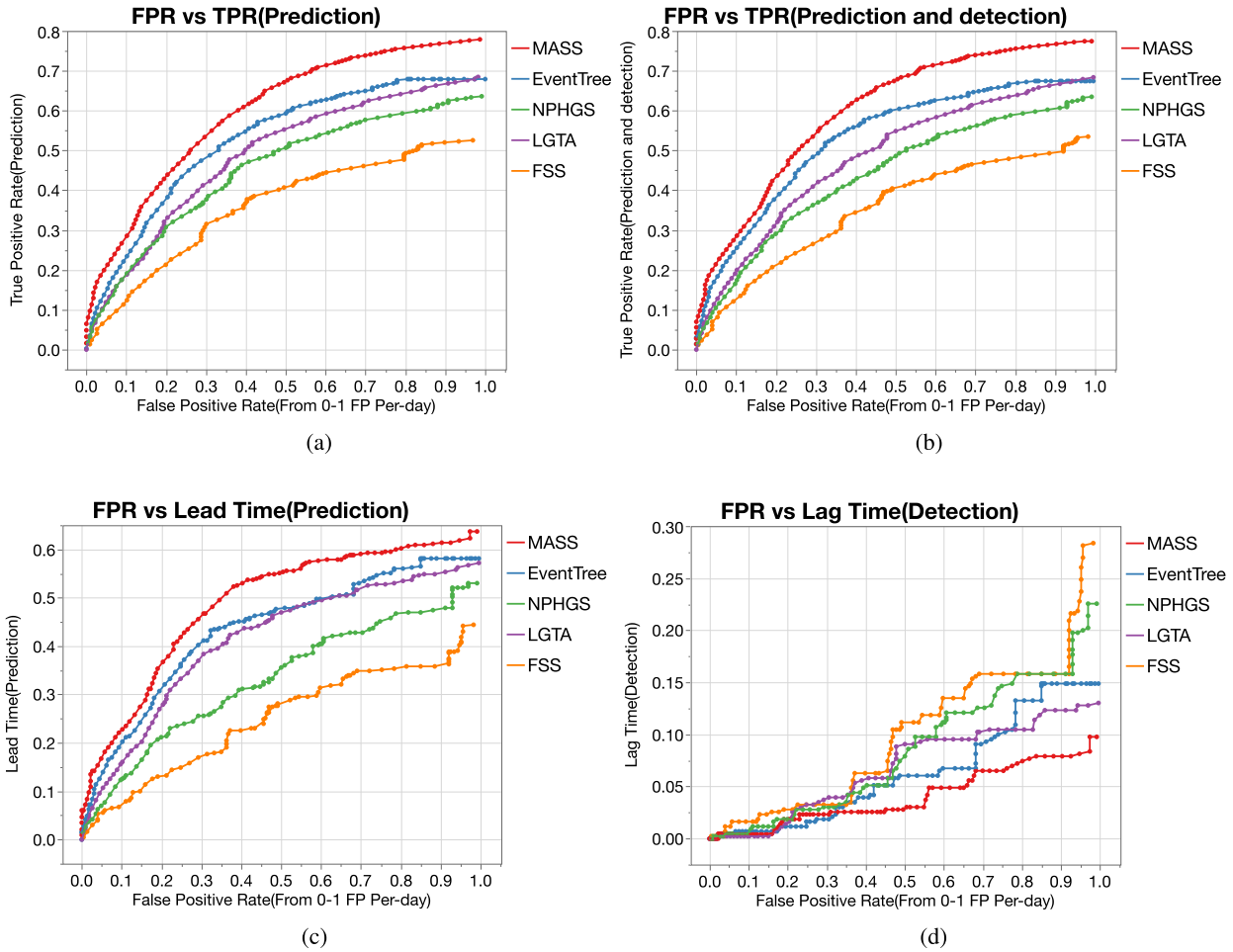

Fig. 3. FPR versus TPR between MASS and four baselines on the Flu data set. (a) FPR versus TPR (prediction). (b) FPR versus TPR (prediction and detection). (c) FPR versus lead time (prediction). (d) FPR versus lag time (detection).

TABLE I

KEYWORDS SELECTED BY MASS, LGTA, AND FSS ON THE HAZE DATA SET

	Event 1		Event 2		Event 3	
MASS	PM10	大气(atmosphere)	健康(health)	污染(pollution)	雾霾(haze)	危害(harm)
	PM2.5	感冒(cold)	危害(harm)	严重(serious)	鼻塞(nasal congestion)	哮喘(asthma)
	爆表(over-range)	环保(env-protection)	雾霾(haze)	—	呼吸(breathe)	致癌(carcinogenic)
LGTA	疾病(disease)	肺癌(lung cancer)	口罩(mask)	健康(health)	烟雾(smoke)	减排(emission reduction)
	灾情(disaster)	治理(treatment)	感冒(cold)	肺癌(lung cancer)	口罩(mask)	应急(emergency)
	污染(pollution)	阴霾(shadow)	PM2.5	烟雾(smoke)	危害(harm)	环境(environment)
	致癌(carcinogenic)	环境(environment)	预警(warning)	空气(air)	空气(air)	加湿器(humidifier)
	口罩(mask)	大气(atmosphere)	致癌(carcinogenic)	AQI	大雾(fog)	鼻炎(rhinitis)
FSS	超标(excess)	PM2.5	雾霾(haze)	健康(health)	PM2.5	环保(env-protection)
	呼吸(breathe)	鼻炎(rhinitis)	空气(air)	危害(harm)	呼吸(breathe)	健康(health)
	环境(environment)	大气(atmosphere)	肺癌(lung cancer)	环境(environment)	疾病(disease)	空气(air)
	感冒(cold)	环保(env-protection)	污染(pollution)	呼吸(breathe)	危害(harm)	雾霾(haze)
	—	—	环保(env-protection)	质量(quality)	污染(pollution)	严重(serious)

did not perform well on the quality of features and subgraph that were identified.

The overall running times of all the methods are shown in Fig. 5 on both the haze and flu outbreak tasks. The results indicate that MASS is the fastest among all the methods on the haze task. For Flu task, MASS (23.4 min) is the second fastest and slightly slower than EventTree (20.4 min). The execution efficiency of the MASS algorithm shows the linear time complexity of our method, as shown in Theorem 3.

C. Results: Feature Selection

Tables I and II show that the features which were obtained by MASS, FSS, and LGTA approaches for six different randomly selected example GSR events.

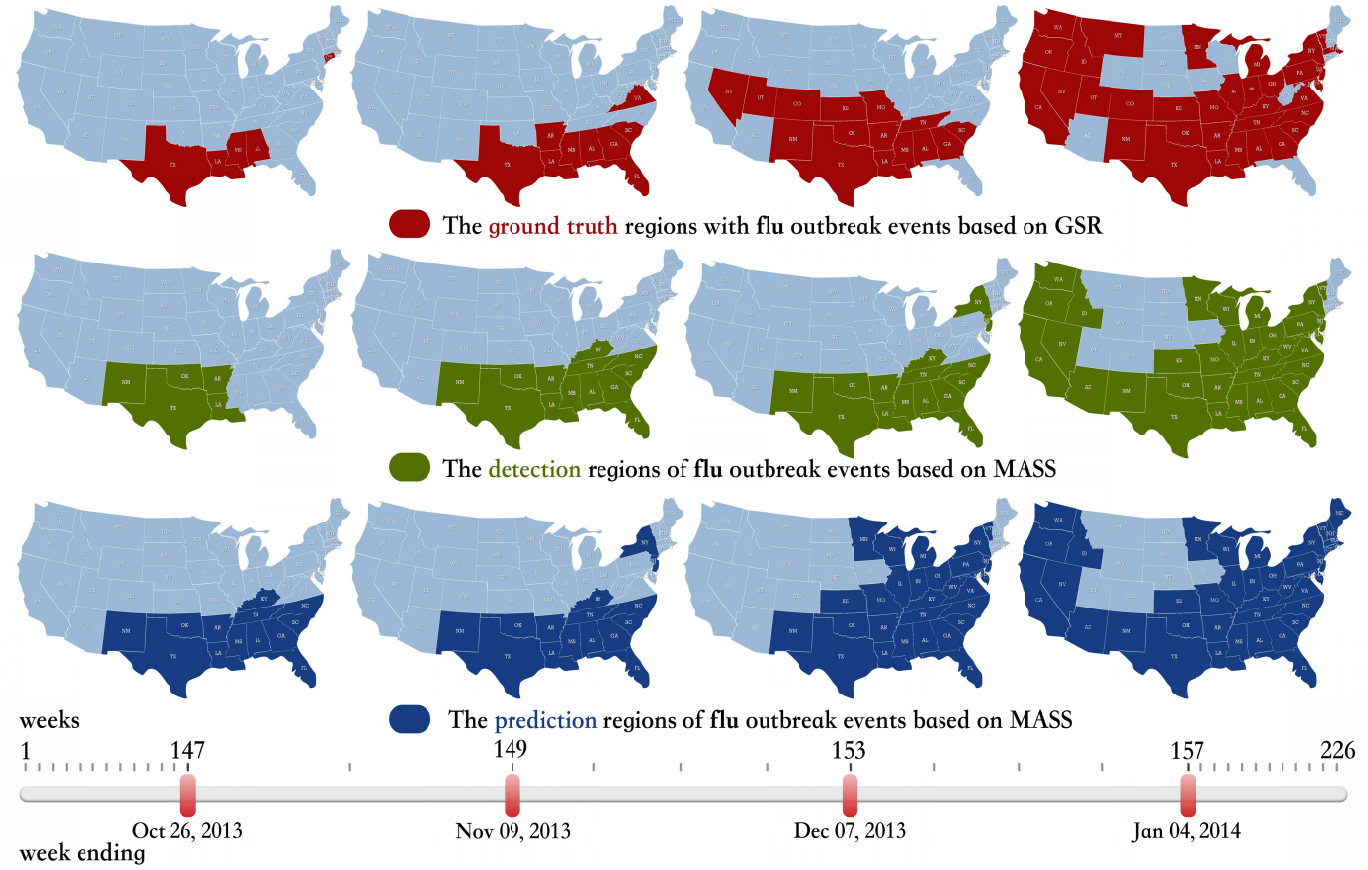

Fig. 4. Illustration of the comparison of the ground truth and MASS detection and prediction results on the flu data set from week 147 (ending October 26, 2013) to week 157 (ending January 04, 2014) in the United States. The first, second, and third rows refer to the ground truth based on GSR, the event detection alert, and the event prediction alert based on MASS. During the overall 226 weeks time span, we chose week 147, 149, 153, and 157 to illustrate the four key points of the ten-week flu outbreak events from the end of 2013 to early 2014 in the U.S.

TABLE II
KEYWORDS SELECTED BY MASS, LGTA, AND FSS ON THE FLU DATA SET

	Event 1		Event 2		Event 3	
MASS	body	flu	cold	infection	cough	sleep
	diarrhea	ache	cough	throat	headache	sneeze
	disease	stomach	sneeze	hand-washing	meds	spread
LGTA	flu	pain	flu	runny	flu	aliment
	cold	fever	cough	shoulder	cold	stomach
	virus	grippe	virus	body	cough	muscle
	cough	disease	cold	germ	runny	fever
	sore	muscle	sleep	infection	sneeze	meds
FSS	body	cold	ache	cold	ache	body
	cough	flu	cough	fever	cold	flu
	sleep	stomach	flu	head	cough	headache
	virus	disease	pain	runny	migraine	runny
	—	—	stomach	infection	sneeze	meds

In the first place, the experimental results in Tables I and II show that the number of features obtained by MASS was much less than the selected attributes based on the LGTA approach. Our proposed MASS approach is able to select different numbers of features for different events, whereas most existing methods require to predefine a fixed number of features, including the baseline LGTA approach. In the second place, the keywords obtained based on both two approaches overlap for a small number of features, which can also represent the core keywords that are relevant to the selected events. Nevertheless, the obtained keywords based on both two approaches are different for all selected events significantly. Since the MASS approach performs better than the LGTA approaches using the two data sets as discussed in Section V-B, the MASS approach can discover a small set of representative signal features which are more effective than those keywords discovered based on the LGTA approach for the task of event prediction and detection.

We illustrate the quality of features discovered based on the MASS approach employing Event 1 for the problem

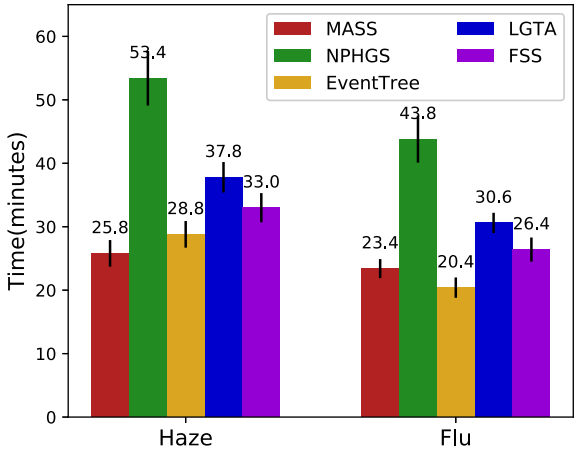

Fig. 5. Runtime of each approach using the Haze and Flu data sets.

of haze detection. Event 1 happened on December 14, 2014, and there is a corresponding news published on the same day, which reports an ongoing haze outbreak: "今天没风 (there is no wind today) 短暂回暖 (it is becoming warmer), 盘踞沈阳的雾霾 又起 (the haze around Shenyang is becoming serious again)° 沈阳市环保局空气 沈阳市环保局空气 质量发布系统显示 早上9时 除棋盘山两个点位之外 沈阳9个点位均出现 重度及以上污染 其中 张士 浑南二点位空气 质量指数爆表 (Air Quality Index is over-range),高达 500 (more than 500 degree), 首要污染物为PM2.5和PM10六级严重污染 (Primary pollutant PM2.5 and PM10 six serious pollution)…." As shown in the paragraph of the news, four of the six selected keywords were mentioned, namely 空气 (atmosphere)," "PM10," "PM2.5," and "爆表 (over-range)" where the last three keywords were not discovered by LGTA or FSS.

VI. Conclusion

In this article, we present a generic method, namely MASS, to the problem of multivariate anomalous subgraph discovery that is free of distributions as nonparametric statistics are used to measure the level of anomalousness of a multivariate anomalous subgraph. We propose a nearly linear time approximation algorithm for concurrent subgraph detection and feature selection and demonstrate that our proposed algorithm performed better than four representative state-of-the-art methods on two real-world outbreak detection and forecasting tasks. In our future work, we will try to extend the MASS approach to discover the anomalous subgraphs in a multivariate heterogeneous graph, in which the nodes or edges own different kinds of types.

References

[1] F. Chen and D. B. Neill, "Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs," in *Proc. KDD*, 2014, pp. 1166–1175.

[2] M. Shao, J. Li, F. Chen, H. Huang, S. Zhang, and X. Chen, "An efficient approach to event detection and forecasting in dynamic multivariate social media networks," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1631–1639.

[3] L. Shi, Y. Wu, L. Liu, X. Sun, and L. Jiang, "Event detection and identification of influential spreaders in social media data streams," *Big Data Mining Anal.*, vol. 1, no. 1, pp. 34–46, Mar. 2018.

[4] L.-L. Shi, L. Liu, Y. Wu, L. Jiang, J. Panneerselvam, and R. Crole, "A social sensing model for event detection and user influence discovering in social media data streams," *IEEE Trans. Comput. Soc. Syst.*, to be published.

[5] H. S. Burkom and E. Elbert, "Biosurveillance applying scan statistics with multiple, disparate data sources," *J. Urban Health*, vol. 80, no. S1, pp. i131–i132, Mar. 2003.

[6] M. Kulldorff, F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman, and R. Platt, "Multivariate scan statistics for disease surveillance," *Statist. Med.*, vol. 26, no. 8, pp. 1824–1833, Apr. 2007.

[7] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras, "On the spatiotemporal burstiness of terms," *Proc. VLDB Endowment*, vol. 5, no. 9, pp. 836–847, May 2012.

[8] D. B. Neill, "Fast Bayesian scan statistics for multivariate event detection and visualization," *Statist. Med.*, vol. 30, no. 5, pp. 455–469, Feb. 2011.

[9] M. Mongiovi, P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, and A. K. Singh, "Netspot: Spotting significant anomalous regions on dynamic networks," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 28–36.

[10] R. H. Berk and D. H. Jones, "Goodness-of-fit test statistics that dominate the Kolmogorov statistics," *Z. Wahrscheinlichkeitstheorie Verw Gebiete*, vol. 47, no. 1, pp. 47–59, 1979.

[11] F. Chen and D. B. Neill, "Non-parametric scan statistics for disease outbreak detection on twitter," *Online J. Public Health Inform.*, vol. 6, no. 1, p. e155, 2014.

[12] E. McFowland, S. Speakman, and D. B. Neill, "Fast generalized subset scan for anomalous pattern detection," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1533–1561, 2013.

[13] F. Chen and D. B. Neill, "Human rights event detection from heterogeneous social media graphs," *Big Data*, vol. 3, no. 1, pp. 34–40, Mar. 2015.

[14] J. L. Sharpnack, A. Krishnamurthy, and A. Singh, "Near-optimal anomaly detection in graphs using lovasz extended scan statistic," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1959–1967.

[15] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti, "Event detection in activity networks," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1176–1185.

[16] M. R. Garey and D. S. Johnson, *Computers and Intractability*, vol. 174. San Francisco, CA, USA: Freeman, 1979.

[17] R. K. Iyer and J. A. Bilmes, "Submodular optimization with submodular cover and submodular knapsack constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2436–2444.

[18] R. K. Iyer, S. Jegelka, and J. A. Bilmes, "Curvature and optimal algorithms for learning and minimizing submodular functions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2742–2750.

[19] R. Iyer and J. Bilmes, "Algorithms for approximate minimization of the difference between submodular functions, with applications," 2012, *arXiv:1207.0560*. [Online]. Available: https://arxiv.org/abs/1207.0560

[20] R. Iyer, S. Jegelka, and J. Bilmes, "Fast semidifferential-based submodular function optimization: Extended version," in *Proc. ICML*, 2013, pp. 1–15.

[21] J. Ward and S. Živnỳ, *Maximizing Bisubmodular and K-Submodular Functions*. Philadelphia, PA, USA: SIAM, 2014, pp. 1468–1481.

[22] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proc. WWW*, 2011, pp. 247–256

[23] D. B. Neill, E. Mcfowland, and H. Zheng, "Fast subset scan for multivariate event detection," *Statist. Med.*, vol. 32, no. 13, pp. 2185–2208, Jun. 2013.

Shuai Zhang received the B.S. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with Beihang University.

His research interests include machine learning methods and data mining on spatial temporal data.

Haoyi Zhou received the B.S. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree with Beihang University.

He is currently at Rutgers, The State University of New Jersey, New Brunswick, NJ, USA, as a Visiting Scholar. His research interests include generalization in machine learning and data mining on sequential data.

Feng Chen received the B.S. degree from Hunan University, Changsha, China, in 2001, the M.S. degree from Beihang University, Beijing, China, in 2004, and the Ph.D. degree from Virginia Tech USA, in 2012, all in computer science.

He is currently an Associate Professor of The University of Texas at Dallas, Richardson, TX, USA. His research focuses on the detection of emerging events and other relevant patterns in the mobile context and/or data mining of spatial temporal, textual, or social media data.

Jianxin Li received the Ph.D. degree from Beihang University, Beijing, China, in 2008.

He was a Visiting Scholar with the Machine Learning Department, CMU, in 2015, and a Visiting Researcher with MSRA in 2011. He is currently a Professor with the School of Computer Science and Engineering, Beihang University. His current research interests include data analysis and processing, distributed systems, and system virtualization.