Tree Decomposition Based Anomalous Connected Subgraph Scanning for Detecting and Forecasting Events in Attributed Social Media Networks

Minglai Shao[a,b], Peiyuan Sun[a,b], Jianxin Li[a,b,*], Qiben Yan[c], Zhirui Feng[a,b]

[a] *Bejing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China*
[b] *State Key Laboratory of Software Development Environment, Beihang University, Beijing, China*
[c] *Computer Science and Engineering, Michigan State University, East Lansing, MI, USA*

Abstract

Event detection and forecasting in social media networks, such as disease outbreak and air pollution event detection, have been formulated as an anomalous connected subgraph detection problem. However, the huge search space and the sparsity of anomaly events make it difficult to solve this problem effectively and efficiently. This paper presents a general framework, namely anomalous connected subgraph scanning (GraphScan) which optimizes a large class of sophisticated nonlinear nonparametric scan statistic functions, to solve this problem in attributed social media networks. We first transform the sophisticated nonlinear nonparametric scan statistics functions into the Price-Collecting Steiner Tree (PCST) problem with provable guarantees for evaluating the significance of connected subgraphs to indicate the ongoing or forthcoming events. Then, we use tree decomposition technique to divide the whole graph into a set of smaller subgraph bags, and arrange them into a tree structure, through which we can reduce the search space dramatically. Finally, we propose an efficient approximation algorithm to solve the problem of anomalous subgraph detection using the tree of bags. With two real-world datasets from different domains, we conduct extensive experimental evaluations to demonstrate the effectiveness and efficiency of the proposed

[*] Corresponding author

Email addresses: shaoml@act.buaa.edu.cn (Minglai Shao), sunpy@act.buaa.edu.cn (Peiyuan Sun), lijx@act.buaa.edu.cn (Jianxin Li), qyan@msu.edu (Qiben Yan), fengzr@act.buaa.edu.cn (Zhirui Feng)

approach.

Keywords: anomalous subgraph detection, approximation algorithm, social media networks, nonparametric statistics, tree decomposition, event detection and forecasting.

1. Introduction

Recently, the social media, such as Twitter and Weibo, has provided an effective means for people to discern and share the events happening around them everyday [1, 2, 3, 4], and has cultivated new research problems in event detection and forecasting [5, 6, 7, 8, 9]. Due to its growing popularity, the social media can provide multiple angles in describing the events [10, 4] to inform people more comprehensively and instantaneously. For example, when the devastating earthquake took place in Sichuan province of China in 2012, Weibo firstly announced the first hand information and organized many volunteer groups across the country in aiding the survivors. Moreover, numerous recent researches have explored and demonstrated the power of social media for event forecasting, such as crime event [11], civil unrest event [12], and disease outbreak forecasting [13, 14].

This paper focuses on the problem of domain-specific event detection and forecasting in social media. In general, the social media networks are composed of nodes such as users, edges such as friend relationships, attributes such as the keywords of the text [15, 16]. In the social media networks, the events can be represented as the anomalous subgraphs, i.e., the connected subsets of nodes presenting high occurrence of keywords related to domain specific events. The event detection and forecasting problem can be generalized as the problem of finding the most anomalous subgraph in social media networks as shown in Figure 1. The identified anomalous subgraphs can be employed to discover the ongoing or forthcoming events.

However, the problem of anomalous connected subgraph detection is NP-hard [17, 18, 19]. The high complexity of social media datasets makes it challenging to detect the anomalous subgraphs effectively and efficiently [20, 21, 22, 23]: first, the **search space is huge** due to the immense amount of social media data; second, the **anomaly**

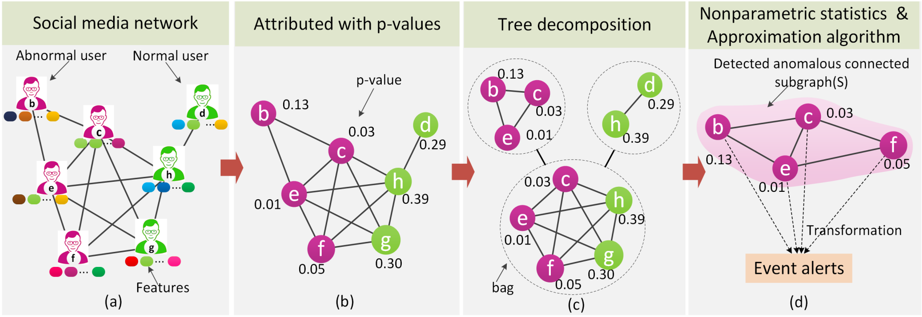


Figure 1: The proposed work focuses on the search of anomalous connected subgraph (S) in the attributed social media networks for event detection and forecasting, where each node (user) is attributed with a p-value based on a set of observed features. Specifically, (a) a social media network, (b) attributed the social media network based on p-values, (c) a tree of bags obtained from tree decomposition approach, (c) the transformation from the anomalous connected subgraph detected based on nonparametric scan statistics and an approximation approach with the tree of bags to event detection and forecasting results.

events are sparse due to the imbalanced nature of the social media data. In general, the abnormal nodes related to a domain specific event are rather sparse. These challenges make the anomalous connected subgraph detection problem computationally difficult. Most of recent researches about the connected subgraph detection problem focus on the parametric approaches and generalize this problem as a hypotheses testing problem [24, 25]. However, the proposed models may be inappropriate, in particular for the **distribution-free** data, the solutions will get even worse [26]. Moreover, most of the existing approaches only focus on the search of the significant subgraphs or constrained patterns in the entire network [27, 28, 29], and there is very **limited work capable of downscaling the search space dramatically**. Specifically, [7] proposes to detect the early emerging events in the entire social networks with the location sensitivity by considering the relevance between the event locations and user locations; [30] proposes to learn the graph structure from the entire network data by comparing the most abnormal subsets of nodes discovered with and without the constraints and the normalized log-likelihood ratio is employed to evaluate the property of a graph structure; [31] first presents to aggregate the attributes for every node in the network, and then transforms the problem into a subgraph detection problem with a large search space.

To address the technical challenges mentioned above so as to detect the anomalous connected subgraphs efficiently in social media networks, we propose a general framework, GraphScan, based on nonparametric statistics. Rather than assuming a particular distribution, such as the Poisson statistic [32], nonparametric statistics do not assume any specific distribution for normal and abnormal nodes. Instead, they first compute a p-value for every node by comparing the current node attribute observations with its historical attribute observations [17], and then maximize a score function $\mathcal{F}(G)$ of p-values of nodes in graph G. Recent studies show that the nonparametric statistics, such as Berk-Jones statistic and Tippet's statistic, can be well applied to the task of finding anomalous subgraphs [33, 34]. In this work, we optimize a large class of the nonparametric scan statistics for detecting and forecasting events. Specifically, we first transform the sophisticated nonlinear nonparametric scan statistic objective functions into the Prize-Collecting Steiner Tree (PCST) problem. Then a tree decomposition approach is introduced to divide the entire network into a set of smaller subgraph groups, namely bags, and arrange them into a tree structure. As a result, we can decrease the scale of problem dramatically and the intimate connection nodes can be arranged into the same bag via the tree decomposition. Moreover, an approximation algorithm is proposed to optimize the transformed graph scan statistics and find the most anomalous connected subgraph based on the tree of bags obtained from the tree decomposition prior.

The main contributions of this paper are summarized as follows:

- **Transforming the sophisticated nonlinear nonparametric scan stastics into the PCST problem with provable guarantees for event detection and forecasting.** We propose a generic framework, namely anomalous connected subgraph scanning (GraphScan), which transforms a large class of sophisticated nonlinear nonparametric scan statistic functions into the PCST problem for detecting and forecasting events in attributed social media networks. The events are comprised of subsets of nodes and their integrated information strengths are characterized as the scan statistics. Moreover, the proposed approach can be applied to networks with both node and edge weights.

- **Proposing an efficient approximation algorithm for anomalous connected subgraph scanning.** We first employ the tree decomposition to divide the graph into a set of smaller subgraph groups, namely bags, and arrange them into a tree structure at the same time, through which we can downscale the problem space dramatically. Then an efficient approximation algorithm is proposed based on the tree of bags for solving the transformed PCST problem to obtain the anomalous connected subgraph to indicate the ongoing or forthcoming events.

- **Conducting comprehensive experiments to validate the performance of the proposed approach.** We conduct extensive experiments to evaluate the proposed GraphScan on haze dataset and flu outbreak dataset and compare the experimental results with those obtained from the baseline approaches. The results demonstrate that GraphScan outperforms existing representative baseline approaches in both effectiveness and efficiency.

The rest of the paper is organized as follows. The related work is presented in Section 2. Section 3 introduces some definitions, such as attributed network and tree decomposition. Section 4 first presents the proposed GraphScan approach, and then performs approximation approaches for tracking the most anomalous subgraph based on the tree decomposition. Comprehensive experiments are provided in Section 5. Finally, the conclusion and future work are presented.

2. Related Work

*2.1. **Event Detection***

Recently, numerous work has been proposed for event detection in social media. Generally speaking, for event detection problems, both clustering and classification approaches are first employed to obtain the interesting information of the events. Then they are used to find the probable happening events by analyzing the temporal information or both spatial and temporal information. Specifically, in the aspect of the temporal information, [35] proposes to find the popular topic employing the measurement of "hotness" and presents an efficient approach for choosing the surrogates of these topics. [36] proposes an approach for detecting the events that happen in the twitter stream. [37] performs an efficient approach used for learning the temporal alter patterns to discover the streaming events. In the aspect of spatial and temporal information, both two kinds of the information are considered to detect the events. Such as [38] proposes to recognize the tweets that are close in both space and time, and by seeking the co-occurrence terms to check whether they represent the same event. Moreover, [5] proposes a clustering approach to select the information about the same events in history but from different space and time. An interactive approach is proposed in [39], and in this approach some queries are inserted and the tweets with the information at different space and time can be obtained. In [40], the authors perform a hierarchical clustering approach for the problem of event detection in the dataset of Twitter. In this approach, both spatial and temporal information are employed to evaluate the similarities of the tweets.

2.2. ***Event Forecasting***

For different specific event forecasting problems, there are three kinds of social event forecasting approaches, including temporal information based approaches [41, 42, 43], spatial and temporal information based approaches [44, 45, 46] and causal relationship based approaches [47, 48]. Specifically, for the first kind of the event forecasting approach, using the high dimensional or multivariate information, supervised mining approaches are generally employed to transfer event forecasting problems to the problems of classification or regression. Such as the topics are extracted from tweets and then are employed to predict the crime events based on the logistic regression [11]. For the second kind of event forecasting approaches, except the temporal information, lots of events in social media always show the spatial characteristics in different domains, such as the events of water pollution and traffic congestion. For example, [46] presents a multiple task mining approach for the problem of event forecasting by mining multiple relevant spatial information. Moreover, the approach proposed in [49] forecast the event of disease by detecting the spatial and temporal anomaly. For the third kind of event forecasting approaches, they conduct the events forecasting by using the relationships between them and other related current or historical events. Such as the approach proposed in [47] can find the conditional probabilities among the events by making use of deep models.

In summary, although there are a great deal of approaches that consider the temporal and spatial information of the social media for event detection and forecasting in social media networks, they generally only focus on the search of the significant information on the whole network and there is very limited work that is able to decrease the scale of searching dramatically, especially for event detection and forecasting based on anomalous regions detection.

3. Preliminaries

Several definitions are introduced in this section, including attributed network, tree decomposition, p-value, tree width and nonparametric statistics.

Definition 1 (Attributed Network) *An attributed network $\mathbb{G} = \{\mathbb{V}, \mathbb{E}, W\}$ is an undirected connected graph, where $\mathbb{V} = \{v_1, ..., v_N\}$ denotes the set of nodes, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ refers to the set of edges (relations), and the function $W : \mathbb{V} \rightarrow [0, 1]$ denotes a single empirical p-value for every node $v \in \mathbb{V}$, which can be calculated by employing the empirical calibration through the comparison between current features of v and its features in history.*

In this work, the anomaly of nodes in social media networks are evaluated by p-values. The two-stage empirical p-value proposed in [33] is employed for node v, denoted as $p(v)$ to evaluate degree of anomaly of node v, and its nice theoretical properties has been presented in [33]. Specifically, the $p(v)$ is measured based on a set of feature p-values. For a feature d of node v and its current feature observation, the significance of the current feature observation is measured by its statistical p-value based on the empirical distribution of this feature. Moreover, the p-value of feature d is calculated as the fraction of its historical observations in which an greater or equal observation is included on this feature.[50, 33, 17]. Intuitively, the p-value is a measurement of anomaly with the value range $[0, 1]$: *the smaller the p-value of a node is, the more abnormal this node is*.

Definition 2 (Nonparametric Scan Statistics). *Given a set of p-values S, nonparametric statistics which are also called aggregation functions of p-values denote a class of score value functions $\mathcal{F}(S)$ which evaluate the joint significance of p-values in S and have the following general formulation:*

$$\mathcal{F}(S) = \varphi(\alpha, \psi_\alpha(S), \psi(S)), \tag{1}$$

where $\psi_\alpha(S)$ denotes the number of p-values which are equal to or less than α in the subgraph S. α is an anomaly significance level of node p-values. Moreover, the function $\varphi(\alpha, \psi_\alpha, \psi)$ satisfies the properties: φ is monotonically increasing with respect to ψ_α and monotonically decreasing with respect to $\tilde{\psi}_\alpha = \psi - \psi_\alpha$.

Definition 3 (Tree Decomposition). *For the network $\mathbb{G} = (\mathbb{V}, \mathbb{E}, W)$, a tree decomposition of $\mathbb{G}$ is a pair $(\{X_i | i \in I\}, T = (I, \mathcal{H}))$, where $T = (I, \mathcal{H})$ is a tree, $I = \{1, 2, ..., |T|\}$, $|T|$ is the number of the tree nodes of T, $\{X_i | i \in I\}$ is a family of subsets nodes of $\mathbb{V}$ for each tree node $i \in I$ of T and $\mathcal{H}$ is the connection of tree nodes, such that:*

- $\bigcup_{i \in I} X_i = \mathbb{V}$;
- $\forall u, v \in \mathbb{V}$ *and* $(u, v) \in \mathbb{E}$, *there exists an* $i \in I$ *with* $u \in X_i$ *and* $v \in X_i$*;*
- *all tree nodes k on any m-n-path, $X_m \bigcap X_n \subseteq X_k$.*

In general, the subsets X_i are denoted as *bags* of nodes. The width of a tree decomposition $td = (\{X_i | i \in I\}, T = (I, \mathcal{H}))$ is defined as $tw(td) = \max_{i \in I} |X_i| - 1$, where $|X_i|$ is the nodes number of bag X_i. For graph $\mathbb{G}$, let TD refer to all tree decompositions, then the tree width of the graph $\mathbb{G}$ is denoted as $TW(\mathbb{G}) = min_{td \in TD}(tw(td))$ [51]. Several equivalent representations of tree decomposition are presented in [52, 53, 54, 55]. In this work, we use a representative tree decomposition approach, namely Gavril based on chordal graph [56]. Moreover, an illustration of the tree decomposition is shown in Figure 1 in which (b) is an original graph and (c) is the corresponding results of tree decomposition, including three bags which are built as a tree of bags.

4. Anomalous Connected Subgraph Scanning

In this section, we first introduce the anomalous connected subgraph detection problem based on nonparametric scan statistics, and then we propose to transform it into the PCST problem. Finally, an approximation algorithm is proposed to obtain the optimal approximation solution for the problem of most anomalous subgraph detection.

4.1. Nonparametric Graph Scan Statistic Problem

In order to discover the most anomalous connected subgraph in the attributed social media network $\mathbb{G}(\mathbb{V}, \mathbb{E}, W)$ so as to realize the event detection and forecasting, the generic formulation of the nonparametric graph scan statistic is shown as below.

$$\mathcal{F}(S) = \max_{\alpha \leq \alpha_{max}} \varphi(\alpha, \psi_\alpha(S), \psi(S)), \tag{2}$$

where $S \subseteq \mathbb{V}$ denotes a set of connected nodes, namely the subgraph in this work, $\psi(S)$ denotes the number of the nodes in S, $\psi_\alpha(S) = \sum_{v \in S}(\tau(p(v) \leq \alpha))$ is the number of node p-values and their anomaly significant level is α. $\tau(\cdot) = 1$ if its input is true, if not, $\tau(\cdot) = 0$. We optimize the significance level α between 0 and a constant $\alpha_{max} < 1$ ($\alpha_{max} = 0.15$ by default). Moreover, it is necessary to consider a range of α, instead of a single threshold for the anomaly significance of nodes. For a fixed α, such as α is equivalent to 0.1, the statistic may disable the ability of discovering a small amount of abnormal p-values which are smaller than 0.1 or a great deal of abnormal p-values which are slightly greater than 0.1. In practice, the value of the selected α_{max} is slightly higher than the typical significance levels which can be predefined [33].

In this work, the anomalous connected subgraph scan statistic function $\mathcal{F}_{\mathrm{BJ}}(S)$ based on the Berk-Jones statistic is employed as a case study since it has shown the effective performance in some real-world applications [33, 57]. Berk-Jones statistic satisfies some optimal properties and outperforms any weighted Kolmogorov statistic [58]. The following our proposed approach is also applicative to other nonparametric scan statistic functions. Furthermore, the $F_{\mathrm{BJ}}(S)$ is shown as:

$$\begin{aligned} \mathcal{F}_{\mathrm{BJ}}(S) &= \max_{\alpha \leq \alpha_{max}} \varphi_{\mathrm{BJ}}(\alpha, \psi_\alpha(S), \psi(S)) \\ &= \max_{\alpha \leq \alpha_{max}} \psi(S) \times \mathrm{KL}(\frac{\psi_\alpha(S)}{\psi(S)}, \alpha) \end{aligned} . \tag{3}$$

where $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence (also called relative entropy) between the observed and the expected proportions with the p-values less than α. The Kullback-Leibler divergence is written as:

$$\mathrm{KL}(\vartheta, \omega) = \vartheta \log(\frac{\vartheta}{\omega}) + (1-\vartheta) \log(\frac{(1-\vartheta)}{(1-\omega)}). \tag{4}$$

Moreover, the problem of finding the most anomalous connected subgraph in attributed social media networks can be transformed into an optimization problem as shown below:

$$\max_{S \subseteq \mathbb{V}} \max_{\alpha \leq \alpha_{max}} \varphi(\alpha,\ \psi_\alpha(S), \psi(S)), \tag{5}$$

which is equivalent to:

$$\max_{\alpha \in \mathrm{U}(\mathbb{V}, \alpha_{\max})} \max_{S \subseteq \mathbb{V}} \varphi(\alpha,\ \psi_\alpha(S), \psi(S)), \tag{6}$$

where $\mathrm{U}(\mathbb{V}, \alpha_{max})$ denotes the union of $\{\alpha_{max}\}$ and different p-values which are no more than α_{max} in $\mathbb{V}$.

4.2. Problem Transformation

Because of the difficulty in solving the nonparametric graph scan statistic problem mentioned above, we propose to transform it into a PCST problem as shown in the following Theorem 1, in this work, and the transformed problem can be easily solved and analyzed.

Theorem 1 (Problem Transformation) *To obtain the most anomalous connected subgraph, the problem shown in Eq.(6) is equivalent to the problem as shown below:*

$$(\alpha^*, S^*) = \operatorname*{argmax}_{\alpha \in \mathbf{P1}} \operatorname*{argmax}_{S_\alpha^Q \in \mathbf{P2}} \varphi(\alpha, \psi_\alpha(S), \psi(S)), \tag{7}$$

where $\mathbf{P1} = \mathrm{U}(\mathbb{V}, \alpha_{max})$*,* $\mathbf{P2} = \{S_\alpha^0, ..., S_\alpha^N\}$*.* S_α^Q *refers to the solution with* $Q \in [0, N]$ *normal nodes budget of the following node-weighted PCST problem:*

$$S_\alpha^Q = \operatorname*{argmax}_S F(S) = \operatorname*{argmax}_S (\psi_\alpha(S) - \tilde{\psi}_\alpha(S)), \tag{8}$$

where $\psi_\alpha(S) = \sum_{v \in S}(\tau(p(v) \leq \alpha))$ *and* $\tilde{\psi}_\alpha(S) = \sum_{v \in S}(\pi(p(v) > \alpha))$*.* $\tau(\cdot) = 0$ *if its input is false, if not,* $\tau(\cdot) = 1$*;* $\pi(\cdot) = 0$*, if its input is false, otherwise,* $\pi(\cdot) = 1$*.*

Proof. We first denote that $(S, \alpha)'$ is the set of nodes in S and their p-values are greater than α, $(S, \alpha)''$ is the set of nodes in S and their p-values are less than or equal to α. Moreover, every $S_{\alpha}^{Q} \in \{S_{\alpha}^{0}, ..., S_{\alpha}^{N}\}$ is composed of the subset of abnormal nodes and the subset of normal nodes and satisfies the conditions: $\tilde{\psi}_{\alpha}(S) = Q$ and $S = (S, \alpha)' \bigcup (S, \alpha)''$. Assume that (α^*, S^*) is the optimal solution to the proposed problem as shown in Eq.(7)($\bigoplus$), then it can be easy inferred that $(\alpha^*, {S_{\alpha^*}^{*}}^{Q}) = (\alpha^*, S^*)$. Nevertheless, based on the properties as presented in Definition 2, there does not exist any other $(S^\star, \alpha^\star)$ with $\tilde{\psi}_{\alpha^\star}(S^\star) = Q'$, where $\alpha^\star \neq \alpha^*$ or $Q \neq Q'$, such that $\varphi(\alpha^\star, \psi_{\alpha^\star}(S^\star), \psi(S^\star)) > \varphi(\alpha^*, \psi_{\alpha^*}(S^*), \psi(S^*))$. Otherwise, this will be in contradiction to the above assumption $\bigoplus$. □

4.3. Approximations with Tree Decomposition

In order to solve the proposed event detection and forecasting problem in Eq. (8) mentioned above efficiently in social media networks, we use one of the most representative tree decomposition approaches, namely Gavril [56]. This approach can efficiently decompose the whole network $\mathbb{G}$ into a set of bags consisted of nodes and arrange them into a tree structure. This operation allows us to solve the proposed problem one bag by one bag so as to reduce the scale of the problem. However, *a following hard problem is how to guarantee the connection of the detected anomalous subgraph*.

Moreover, for the sake of obtaining the most desired solution with the connection guarantee of the detected anomalous subgraph, a dynamic programming algorithm is designed when the input network $\mathbb{G}$ is a tree of bags. Specifically, in the dynamic programming, for the bag X_k, we construct a dynamic table hash table $(s_k^p, s_k, F(s_k))$ named as D_k to record all kinds of states (e.g., overlap nodes, score function value) in the bag X_k, where $s_k = \{s_{k_1}, s_{k_2}, ..., s_{k_i}, ..., s_{k_m}\}$, and m is the number of the subset nodes in s_k, and $s_{k_i} \subseteq X_k$ refers to a subset nodes of X_k, $s_k^p = \{s_{k_1}^p, s_{k_2}^p, ..., s_{k_i}^p, ..., s_{k_m}^p\}$. $s_{k_i}^p = s_{k_i} \bigcap X_k^p$, and X_k^p refers to the parent bag of X_k, $F(s_k) = \{F(s_{k_1}), F(s_{k_2}), ..., F(s_{k_i}), ..., F(s_{k_m})\}$, and $F(s_{k_i})$ is the score function value corresponding to s_{k_i}. This table is used to store the combination information of nodes in the bag. From X_k^p, we can get the items of its child bags, namely the items in $(s_k^p, s_k, F(s_k))$, such as

$(s_{k_i}^p, s_{k_i}, F(s_{k_i}))$. Finally, a backtracking approach checks this table and gets the subset with the maximum score value. In conclusion, the proposed GraphScan framework can be summarized as Algorithm 1.

Algorithm 1 GraphScan

```
Input: Network $\mathbb{G}(\mathbb{V}, \mathbb{E}, W)$.
Output: Anomalous connected subgraph $S^*$.
Set $\alpha_{\max} = 0.15$, $\mathbf{L} = 3$.
$T = Gavril(\mathbb{G})$    ▷ $T = (X, (I, \mathcal{F}))$ is the tree decomposition.
for $l \in \{1, ..., \mathbf{L}\}$ do
    Root $T$ at tree bag $r$ and construct a post-order walk for $T$;
    for $\alpha \in \mathrm{U}(\mathbb{V}, \alpha_{max})$ do
        for $k = 1$ to $|T|$ do
            $\mathrm{ComputeDPTable}(\mathbb{G}, T, X_k, \alpha)$;    ▷ See Section 4.4
        end for
        Backtracking and seek out $S_\alpha^Q$;
    end for
    $S^l \leftarrow \underset{\alpha \in \mathrm{U}(\mathbb{V}, \alpha_{\max})}{\mathrm{argmax}} \varphi(\alpha, \psi_\alpha(S_\alpha^Q), \psi(S_\alpha^Q))$;
end for
Calculate $l^* \leftarrow \mathrm{argmax}_l \varphi(\alpha, \psi_\alpha(S^l), \psi(S^l))$;
Return $S^* \leftarrow S^{l^*}$
```

4.4. Processing Strategies of Tree Bags

After the tree decomposition, the bags can be categorized as small bags and large bags. In both types of bags, the abnormal nodes distribution can be abstracted as Figure 2. The connected abnormal nodes are independent or connected by some normal nodes. Moreover, in a bag, the nodes can be divided into overlap nodes and non-overlap nodes, where the overlap nodes mean that these nodes appear in the current bag and its parent bag simultaneously.

For different scales of bags, we employ different processing methods. We first combine the adjacent connected abnormal nodes or connected normal nodes together

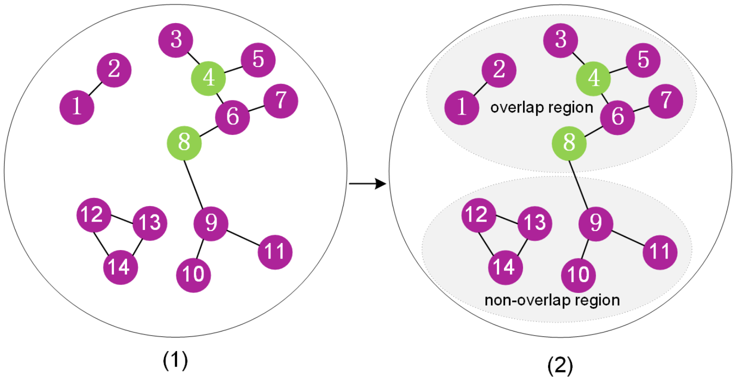


Figure 2: An illustration of abstract presentation of the abnormal nodes distribution in a bag as shown in (1), where the purple nodes, such as $1, 2, 3, 5, 6, 7$, refer to the abnormal nodes and the green nodes refer to the normal nodes. The connected abnormal nodes are independent (e.g., 1, 2) or connected by some normal nodes (e.g., $3, 5$). In a bag, the nodes can be divided into overlap nodes and non-overlap nodes as shown in (2), where the overlap nodes indicate these nodes that appear in the parent bag of this bag, and non-overlap nodes indicate other nodes.

respectively, named as the unit in this paper. As shown in Figure 3, each node refers to a unit. Moreover, in small bags, since there are a few units in each bag, we can find the results by enumerating all the possible combinations. Comparatively speaking, for a large bag X_k, we design the heuristic strategies to get the combination results, denoted as $\text{ComputeDPTable}(\mathbb{G}, T, X_k, \alpha)$ as shown below. The algorithm returns D_k, where D_k is the dynamic programming hash table of the tree bag X_k. The proposed dynamic programming is shown in Figure 3.

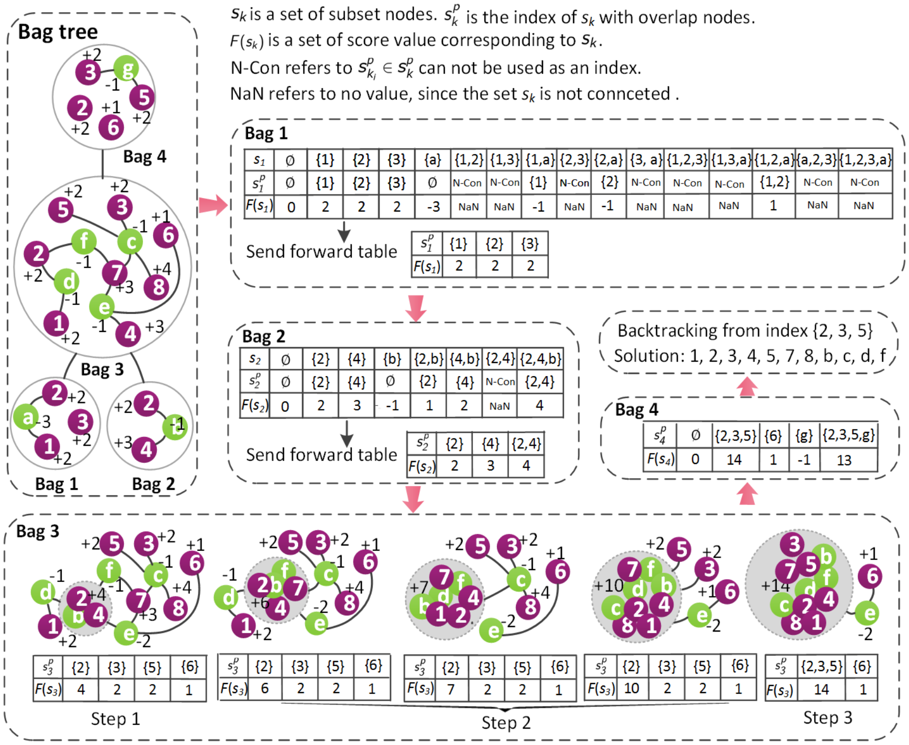
Update the connection based on the dynamic table of child bags. If the bag X_k is not a leaf bag, we first update the connection based on the dynamic table of the child bags of X_k, namely the tables $D_{c_1}, ..., D_{c_d}$ corresponding to the child bags $X_k^{c_1}, ...X_k^{c_d}$. For every child bag of X_k, such as $X_k^{c_i}$, look up every item $D_{c_{ij}}$ with $s_{c_{ij}}$ in the send forward table of D_{c_i} : $(s_{c_i}^p, s_{c_i}, F(s_{c_i}))$, $i \in \{1, 2, ..., d\}$, where D_{c_i} is the dynamic table of bag $X_k^{c_i}$. The send forward table is made up of the set of units in overlap region of $X_k^{c_i}$ and X_k and the corresponding score values as shown in Figure 3. Then we connect the unit set $s_{c_{ij}} \in X_k^{c_i}$ and units in X_k which can be connected based on the connection in $s_{c_{ij}}$ and X_k as a new unit, and update the corresponding score function value as shown in Step 1 of Figure 3.

Small bags. In small bag X_k, we consider all possible units combinations. For each combination (subset) s_k in the bag X_k, we first use the Union-Find algorithm to make sure $\mathbb{G}_{s_k}$ is a connected subgraph of $\mathbb{G}$ based on s_k. Two kinds of edges should be considered: one comes from s_k, the other connection relation comes from the child bags of s_k as mentioned above. Then we get the intersection of s_k and its parent bag s_k^p, and using it as an index to add or update the item of dynamic table D_k. For each item D_{k_i} : $(s_{k_i}^p, s_{k_i}, F(s_{k_i}))$, if $(s_{k_i}^p, s_{k_i}, F(s_{k_i})) \notin D_k$, add $(s_{k_i}^p, s_{k_i}, F(s_{k_i}))$ to D_k via $D_k = D_k \bigcup (s_{k_i}^p, s_{k_i}, F(s_{k_i}))$. Otherwise, assume $(s_{k_i}^p, s^{'}, x)$ be the current entry stored in D_k for $s_{k_i}^p$. If $F(s_{k_i}) > x$, update the value for $s_{k_i}^p$ in $D_k : s^{'} = s_{k_i}, x = F(s_{k_i})$.

Large bags. Since there are lots of nodes in the large bags, in order to decrease the time complexity, we propose a new method to avoid enumeration. **The key idea is to merge the anomalous nodes in non-overlap regions as much as possible**. For a large bag X_k, we first combine the nodes in child bags as mentioned above and initialize D_k based on the overlap region of X_k and its parent bag. Then other units in the bag X_k will be combined step by step. The details are shown as follows.

• *Combine the units in child bags and initialize* D_k. For a large bag X_k, it has a great number of child bags, so through this step, we can first combine all the connections in X_k carried by its child bags $X_k^{c_1}, ..., X_k^{c_d}$ as the operations as mentioned above. Then, we initialize all of the positive units in the overlap region of X_k and its parent bag as items individually in the dynamic table D_k as shown in Step 1 of Figure 3.

• *Connect abnormal units which is not in overlap regions of* X_k *and its parent bag and update* D_k. Let NOL refer to the set of units in non-overlap regions of X_k and their parent bag, and $nol \in$ NOL. Moreover, we let OL refer to the set of units in overlap regions of X_k and its parent bag, and $ol \in$ OL. Considering there are few abnormal units compared with normal nodes since abnormal is the minority, we first use shortest path algorithm to calculate the score function values F of combinations of every abnormal unit ol_i in overlap region OL with every unit nol_j in non-overlap region NOL. Then we choose the combination with the largest score function value



s_k is a set of subset nodes. s_k^p is the index of s_k with overlap nodes.

$F(s_k)$ is a set of score value corresponding to s_k.

N-Con refers to $s_{k_i}^p \in s_k^p$ can not be used as an index.

NaN refers to no value, since the set s_k is not connceted .

Bag 1

s_1	∅	{1}	{2}	{3}	{a}	{1,2}	{1,3}	{1,a}	{2,3}	{2,a}	{3, a}	{1,2,3}	{1,3,a}	{1,2,a}	{a,2,3}	{1,2,3,a}
s_1^p	∅	{1}	{2}	{3}	∅	N-Con	N-Con	{1}	N-Con	{2}	N-Con	N-Con	N-Con	{1,2}	N-Con	N-Con
$F(s_1)$	0	2	2	2	-3	NaN	NaN	-1	NaN	-1	NaN	NaN	NaN	1	NaN	NaN

Send forward table

s_1^p	{1}	{2}	{3}
$F(s_1)$	2	2	2

Bag 2

s_2	∅	{2}	{4}	{b}	{2,b}	{4,b}	{2,4}	{2,4,b}
s_2^p	∅	{2}	{4}	∅	{2}	{4}	N-Con	{2,4}
$F(s_2)$	0	2	3	-1	1	2	NaN	4

Send forward table

s_2^p	{2}	{4}	{2,4}
$F(s_2)$	2	3	4

Backtracking from index {2, 3, 5}
Solution: 1, 2, 3, 4, 5, 7, 8, b, c, d, f

Bag 4

s_4^p	∅	{2,3,5}	{6}	{g}	{2,3,5,g}
$F(s_4)$	0	14	1	-1	13

Bag 3

s_3^p	{2}	{3}	{5}	{6}
$F(s_3)$	4	2	2	1

Step 1

s_3^p	{2}	{3}	{5}	{6}
$F(s_3)$	6	2	2	1

s_3^p	{2}	{3}	{5}	{6}
$F(s_3)$	7	2	2	1

s_3^p	{2}	{3}	{5}	{6}
$F(s_3)$	10	2	2	1

Step 2

s_3^p	{2,3,5}	{6}
$F(s_3)$	14	1

Step 3

Figure 3: An illustration of the proposed dynamic programming for seeking out the most abnormal subgraph S. For simplicity, each node refers to a unit that merges the adjacent abnormal nodes and the normal nodes in the tree bags respectively. Each bag has a dynamic table to record the set of node, index and the corresponding score value. In small bags, we enumerate all possible combination. In large bags: step 1: combine units in child bags; step 2: connect abnormal units in overlap regions; step 3: combine the anomalous unit in overlap regions and update the dynamic table D_k for bag k.

to connect and update D_k if the value is bigger than the separate two units as shown in Step 2 of Figure 3. Finally, we iterate the above operations until there is no unit to combine. Moreover, if there are some abnormal units in non-overlap region left, we should consider the combination within them, since the optimal subset may exist among them. And in every run, we should make sure to combine the adjacent abnormal units in time.

• *Combine the anomalous units in overlap regions and update D_k.* Finally, since some abnormal units have been connected in the above step as shown in Step 2 of Figure 3, we then combine the anomalous units in overlap regions, such as unit3 and unit5. as shown in Step 3 of Figure 3. In this operation, we connect the overlap region anomaly units by using the same ways which are used to combine the abnormal units in non-overlap region as mentioned above. In this step, the unit3 and unit5 are combined with the unit2 since the score function value of connected units is larger than the separate value, such as the value of unit3.

Backtracking. When we complete the dynamic programming, we first find the set s which maximizes $F(s)$ in all the tables. Then we let $\mathit{opt\_sets}$ be a $|X|$ long vector of sets and construct a pre-order walk δ of T rooted by r. In the beginning, we set $\mathit{opt\_sets}[\delta[0]] = s$, then $S = s$. Then, for $k = 1$ to $|X|$, and for every child tree bag X_j of $\delta[k]$, if $s^p_{j_q} = \mathit{opt\_sets}[\delta[k]] \bigcap X_j$, we have $(s^p_{j_q}, s_{j_q}, F(s_{j_q})) \in (s^p_j, s_j, F(s_j))$ that corresponds to $s^p_{j_q}$ in table D_j. Then, $S = S \bigcup s_{j_q}$.

4.5. Theoretical Properties

Theorem 2 (Connection Transitivity). *Let $X_k^{c_1}, X_k^{c_2}, ..., X_k^{c_n}$ be the child bags of bag X_k. If $\exists\{v_1, v_2\} \cap X_k \neq \varnothing$, $\exists X_k^{c_i}$ & $\{v_1, v_2\} \subseteq X_k^{c_i}, i \in \{1, ..., n\}$, $\{v_1, v_2\} \subseteq s_{c_i} \subseteq X_k^{c_i}$, where s_{c_i} is a node subset of $X_k^{c_i}$, and there is corresponding $s^p_{c_i}$ of s_{c_i}, which is one of the forward table index of $X_k^{c_i}$, then v_1, v_2 is connected in X_k and for X_k, v_1, v_2 is the connected nodes in its child bags.*

Proof. Since $s^p_{c_i}$ is the index of s_{c_i} with overlap nodes in $X_k^{c_i}$, s_{c_i} is a connected subset which will be forward to X_k. If $v_1, v_2 \in s_{c_i}$, then v_1 and v_2 is connected. If $\{v_1, v_2\} \cap X_k \neq \varnothing$ and through the processing of combining nodes in children bags in X_k, v_1, v_2 will be merged in X_k under the connection property of s_{c_i} with the index

$s^{p}_{c_i}$. Moreover, if v_1, v_2 is connected in X_k, the score value of connected v_1, v_2 is more than the value of any one of the two, and from the index of X_k, the details of connection in child bags will be ignored, and in backtracking, v_1 and v_2 are connected anomalous nodes in child bags, even if only one of them is selected in some child bags. □

Time complexity analysis. The dynamic programming algorithm guarantees to obtain a local maximum solution and has the time complexity $O(n^{'} * \mathbb{T})$, where $n^{'} (n^{'} << n)$ refers to the number of bags, $\mathbb{T}$ is the average runtime of all bags. Most of the bags are small since the sizes of bags follow the long tail distribution [59]. For small bags, the time complexity is $O(2^m)$, where $m \leq th$ is the number of nodes in bags. th is the threshold of small bags, in this work, th is set as 10. For large bags, the time complexity is $O(m^2)$, where $m \leq tw$ and tw is the tree width. Moreover, once the tree decomposition is completed and we have a large amount of data from different time slices, with more time slices, the proposed approach can achieve shorter runtime compared with the existing baseline methods.

5. Experiments

In this section, the proposed GraphScan framework is evaluated based on two different domain datasets. We first introduce the experimental setup, and then we evaluate the performance of proposed GraphScan approach and compare it against four state-of-the-art approaches. Finally, a case study of haze event detection and forecasting is presented.

5.1. Experimental Setup

In this subsection, the datasets, the baseline approaches, performance metrics and so on are presented. The details of them are shown below.

Datasets: The experiments in this work are based on two different datasets (flu outbreak dataset and haze event dataset) as the case learning scenarios.

(1) Flu outbreak dataset. 10% of all original Twitter data from 01-01-2011 to 05-01-2015 are obtained randomly in the USA. In this dataset, we choose 0.16 million tweets in which every one has no less than two keywords about the flu outbreak event from a dictionary of 72 keywords obtained based on the experts. In terms of the relations of users and tweets, a user-user network with 39,565 users and 49,204 edges is built, where every user is characterized with a states location of USA. For every user and every day, a p-value mentioned above is computed based on the approach in [33] for every keyword. In all, correspond with the 226 weeks from 01-01-2011 to 05-01-2015, there are 226 snapshot graphs. Moreover, we obtain the Golden Standard Reports (GSR) of 2.26 thousand official records about flu outbreak (influenza-like illness (ILI) $\geq$ 2000) from the official website (`http://www.cdc.gov/flu/weekly/.`) which is controlled by the Disease Control and Prevention (CDC) Centers. The ILI level is announced by CDC weekly for every state of the USA according to the proportion of outpatient about ILI. A flu outbreak event illustration is shown as: (STATE, COUNTRY, WEEK)= ("New York", "USA", "07-14-2014 to 07-20-2014").

(2) Haze event dataset. 10% of the Weibo data from 04-11-2014 to 01-11-2015 are obtained randomly, more than 1,43 billion tweets are included. Besides, we delete the tweets which have one or zero keyword from a set of 68 keywords about haze event obtained based on the experts. Finally, 350 thousand tweets which are posted by 49,644 users are acquired. In terms of the relations of users and tweets, a user-user network which has 149,408 edges is built. For every user and every day, a p-value mentioned above is computed based on the approach in [33] for every keyword. In all, correspond with the 276 days from 04-11-2014 to 01-11-2015, there are 276 snapshot graphs. Furthermore, we obtain the GSR of 9,384 haze event records (level $\geq$ 3) from official websites. Moreover, a GSR record illustration is shown as: (Province, COUNTRY, DAY) = ("Tianjin", "China", "12-11-2014"). *The two kinds of different time slices, namely week and day, are selected corresponding to the event report interval of official in the two different datasets.*

The Proposed GraphScan Approach and Baseline Methods. The approach, namely GraphScan, is proposed in this paper based on the tree decomposition for event detection and forecasting in social media networks. In the experiments, 10-fold cross validation is employed to obtain the best relevant parameters. In detail, the threshold th is denoted as 10, the parameter α_{max} is set as 0.15. Moreover, in order to evaluate the performance of the GraphScan approach, four competition approaches are considered, including EventTree [60], Meden [22], HkS [61] and Latent Geographical Topic Analysis (LGTA)[62]. Furthermore, the related parameters of the papers are tuned strictly following the approaches mentioned by original authors in their works. Specifically, EventTree generalizes the event detection problem based on two kinds of formulations of graph theoretic. One of them obtains the correlation of the event by employing the sum of distances of all event nodes. Moreover, this can be transform to an MaxCut problem. The other one obtains the correlation by employing a minimum distance tree and results in the PCST problem which can be solved by employing the existing approximation algorithms [60]. Meden defines an efficient heuristic approach and a tight upper bound for approximating the heaviest dynamic subgraphs (most anomalous subgraphs in this work) [22]. Hsk develops an efficient approximate approach for solving the problem of heaviest k-subgraph to discover the event in a graph constructed starting from posts of users [61]. LGTA is a novel location and text joint approach which combines the geographical clustering approach and topic model together. Moreover, LGTA can find the high quality geographical anomaly and estimate the anomaly distributions in different geographical locations.

Performance Metrics. Firstly, the runtime of our proposed GraphScan approach and all baseline approaches are compared. Moreover, the related performance metrics employed for event detection and forecasting include: (1) FPR = FP/(FP+TN); (2) TPR = TP/(TP+FN) for both event detection and event forecasting; (3) Event detection lag time; (4) Event forecasting lead time, where FP, TP, TN, FN, FPR and TPR refer to false positive, true positive, true negative, false negative, false positive rate and true positive rate respectively.

Moreover, in the experiments, the event detection and forecasting results are reported as the form of (date, location), where “location” represents the state of United States in the flu outbreak dataset or “location” denotes the province of China in the Haze dataset. For every Gold Standard Reports event, test whether: (1) Each approach has a report in the state or province within seven time slices before the domain specific event, namely “predicted”; (2) Each approach has a report in the state or province within seven time slices after the domain specific event, namely “detected”; (3) Each approach has no report in the state or province within seven time slices after or before

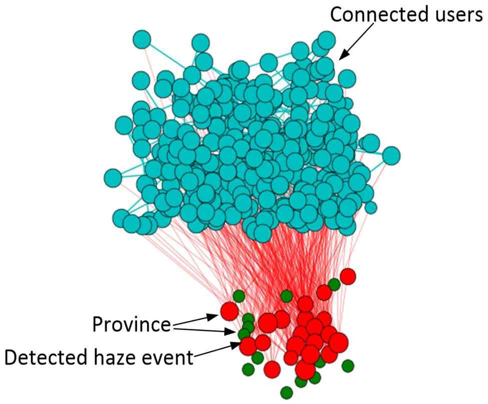


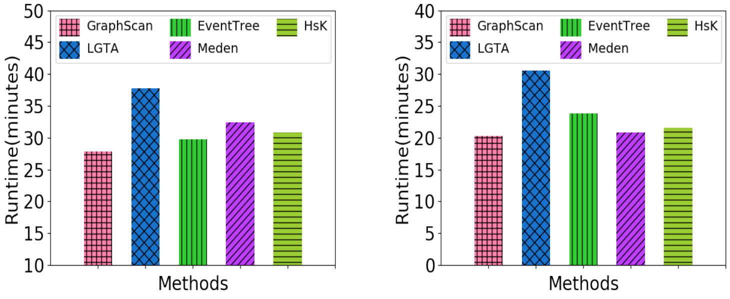
Figure 4: An illustration of the transformation from the detected anomaly subgraph to the haze event detection and forecasting reports of the provinces in China from Dec. 25, 2014. Specifically, the detected connected anomalous subgraph is comprised of the blue nodes of users and is connected by the blue lines. The relations between users and the provinces are connected by the red lines. The red nodes denote the successful detected and forecasted haze events in these provinces, the green nodes refer to the provinces that do not report haze event alerts.

the domain specific event, namely “undetected”.

5.2. Event Detection and Forecasting Results

In this subsection, we first introduce the transformation from detected anomalous subgraphs to event detection and forecasting results, and then present the comparison of the runtime and accuracy of the event detection and forecasting among all above approaches.

Transformation from Detected Anomaly Subgraphs to Reported Event Detection and Forecasting Results. Each of the above approaches (the GraphScan approach and the four baseline approaches) outputs a detected user subgraph with the maximized value of the objective function proposed in Section 4.1 for every time slice. In every subgraph, some locations are retrieved, where every location leads to a detected or forecasted event alert. As shown in Figure 4, an illustration of transformation from the detected abnormal subgraph to the reported haze event results of provinces from 12-25-2014 in China is presented. The detected subgraph is comprised of blue nodes where each node refers to a user in the social media networks, and the red nodes are the trans-



(a) Runtime based on haze event dataset.

(b) Runtime based on flu event dataset.

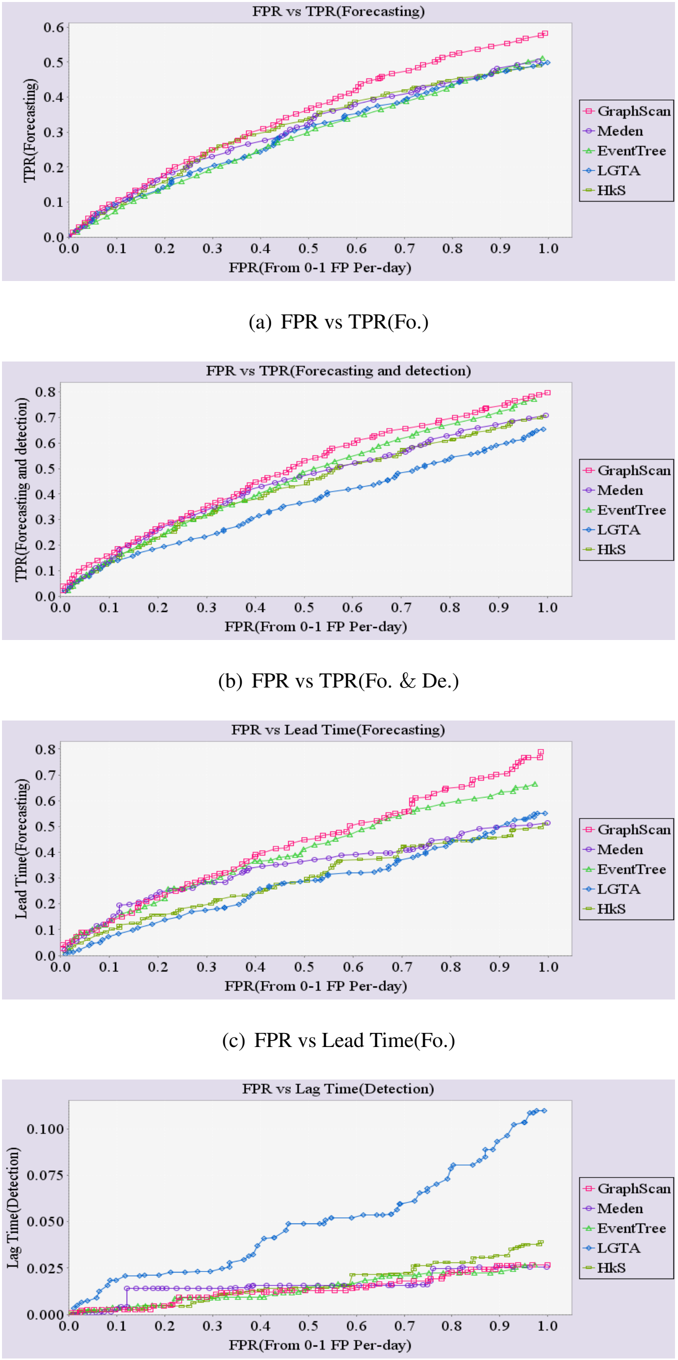
Figure 5: The runtime comparison of all competition approaches, including GraphScan, EventTree, Meden, LGTA and HsK.

formed provinces of China. Moreover, the red nodes denote the successful detected and forecasted haze events, and the green nodes denote that there are not reported results in these regions. There may be few errors since that there are few negative contents, e.g., some ongoing events may be discussed by some users in their adjacent regions.

The Runtime Comparison of Competition Approaches. The comparison of the average runtime between the GraphScan framework and all baseline approaches are presented in Figure 5. The results shows that the LGTA runtime is obvious longer than our GraphScan approach in both two datasets, and the runtime of EventTree, HsK and Meden are similar to GraphScan. GraphScan outperforms the baseline approaches. Moreover, once the tree decomposition is completed and there are lots of data from different time slices, the more time slices, the shorter run time compared with the existing baseline approaches. The main reason is that the tree decomposition downscales the search space, and as a result, the anomalous subgraphs can be obtained with less time based on the tree of bags. Moreover, as time slices increase, the proportion of time occupied by tree decomposition decreases.

The Analysis of Event Detection and Forecasting Results. We evaluate the performance of the GrapScan approach and all other approaches using two different datasets. The results are shown in Figure 6 and Figure 7 respectively.

(1) For haze event dataset, the results of haze event detection and forecasting of



(a) FPR vs TPR(Fo.)

(b) FPR vs TPR(Fo. & De.)

(c) FPR vs Lead Time(Fo.)

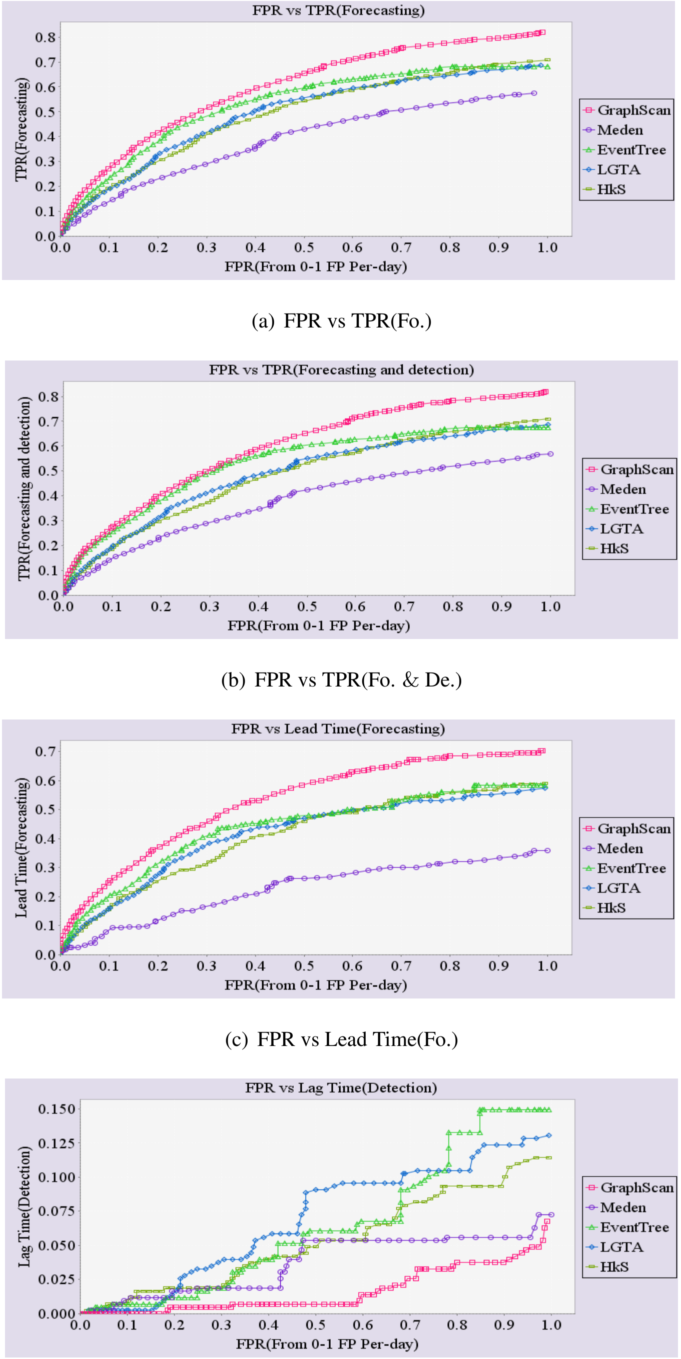
(d) FPR vs Lag Time(De.)

Figure 6: The comparison between the GraphScan and baseline approaches using haze event dataset.

all above approaches are presented in Figure 6. This figure shows that the comparison of the results at different false positive rate for haze event detection and forecasting tasks. The results show that the GraphScan approach has achieved higher true positive rate for haze event forecasting compared with the baseline methods, and higher true positive rate for both haze event forecasting and detection than all the comparison approaches. Moreover, there is a trend that the margin between the true positive rate results of GraphScan and the results obtained from all comparison approaches continuously increases for haze event detection and forecasting when the false positive rate increases. Specifically, the margin of forecasting is nearly 10% as presented in Figure 6(a). The proposed GraphScan approach achieves the longer lead time for haze event forecasting as shown in Figure 6(c). Moreover, the GraphScan approach also obtain the shorter lag time for haze event detection than comparison approaches as shown in Figure 6(d).

(2) For flu outbreak dataset, results of the flu outbreak event detection and forecasting of all above approaches are presented in Figure 7. The results also show that the proposed GraphScan approach achieves higher true positive rate for flu outbreak event forecasting than baseline approaches, and higher true positive rate for flu outbreak event forecasting and detection compared with all comparison approaches. Furthermore, the margin between the true positive rate of GraphScan and that of all comparison methods continuously increases for flu outbreak event detection and forecasting when the false positive rate increases. As presented in Figure 7(a), the margin of flu outbreak event forecasting is greater than 10%. The margin of event detection and forecasting is greater than 10% as presented in Figure 7(b). Moreover, the proposed GraphScan approach obtains longer lead time for flu outbreak event forecasting as shown in Figure 7(c). Finally, Figure 7(d) shows that GraphScan also gets the shorter lag time for flu outbreak event detection than all baseline approaches at different false positive rates.

To sum up, the proposed GraphScan approach decreases the scale of problem dramatically and outperforms the baseline approaches. Moreover, besides the two networks used in the experiments, the GraphScan approach can be used for a wide variety of other types of networks and applications, e.g., for congestion detection in road traffic network.



(a) FPR vs TPR(Fo.)

(b) FPR vs TPR(Fo. & De.)

(c) FPR vs Lead Time(Fo.)

(d) FPR vs Lag Time(De.)

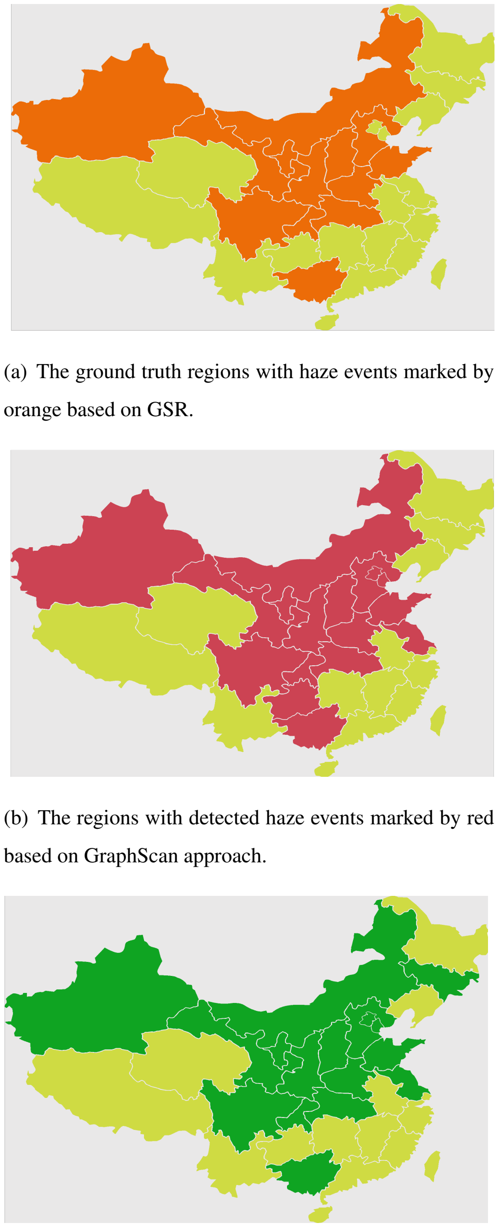
Figure 7: The comparison between GraphScan and other approaches based on the flu outbreak dataset.

5.3. Case Study: Haze Event Detection and Forecasting

In order to better present the experimental results of the proposed GraphScan framework, we show an example of haze event detection and forecasting on 2015-01-02 in China in Figure 9. In this figure, the ground truth and the haze event detection and forecasting results obtained from GraphScan approach are presented. Moreover, in the subfigures, each individual region denotes a province of China. As shown in this figure, almost all of the haze events from different provinces have been successful detected and forecasted. There are few deviations which may be caused by the existence of some negative posts, e.g., some events may be discussed by some users in their adjacent areas. On the other hand, this case vividly reflects the performance of the proposed approach for both event detection and event forecasting in attributed social media networks.

6. Conclusion and Future Work

This paper proposes an efficient approach, GraphScan, for event detection and forecasting using the anomalous connected subgraph detection in social media networks. In GraphScan, we first generalize the scan statistics and then make several contributions to find the anomalous subgraph. Specifically, we first employ tree decomposition to divide the graph into a set of smaller groups, namely bags, and at the same time we arrange them in a tree structure, through which we can decrease the scale of problem dramatically. Then an efficient approximation algorithm is proposed for solving the problem of the anomalous subgraph detection based on the tree of bags obtained from the tree decomposition approach. We evaluate the proposed approach using two different datasets. The results show that GraphScan outperforms existing approaches in both event detection and forecasting. In future, we will extend GraphScan to discover the abnormal subgraphs that evolve over time by leveraging time series heterogeneous graphs.



(a) The ground truth regions with haze events marked by orange based on GSR.

(b) The regions with detected haze events marked by red based on GraphScan approach.

(c) The regions with forecasted haze events marked by green based on GraphScan approach.

Figure 8: An illustration of the comparison of the detected and forecasted haze event results based on the proposed GraphScan approach and the haze event ground truth based on GSR on 2015-01-02 in China. In these subfigures, every individual region denotes a province of China.

References

[1] J. Li, J. Wen, Z. Tai, R. Zhang, W. Yu, Bursty event detection from microblog: a distributed and incremental approach, Concurrency and Computation Practice and Experience 28 (11) (2016) 3115–3130.

[2] D. Wang, A. Al-Rubaie, S. S. Clarke, J. Davies, Real-time traffic event detection from social media, ACM Transactions on Internet Technology (TOIT) 18 (1) (2017) 9.

[3] F. Chen, D. B. Neill, Human rights event detection from heterogeneous social media graphs, Big Data 3 (1) (2015) 34–40.

[4] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, C. Rong, Detecting spammers on social networks, Neurocomputing 159 (2015) 27–34.

[5] D. Xiaowen, M. Dimitrios, C. Francesco, F. Pasca, Multiscale event detection in social media, Data Mining and Knowledge Discovery 29 (5) (2015) 1374–1405.

[6] S. Speakman, M. F. Iii, D. B. Neill, Scalable detection of anomalous patterns with connectivity constraints, Journal of Computational and Graphical Statistics 24 (4) (2015) 1014–1033.

[7] S. Unankard, X. Li, M. A. Sharaf, Emerging event detection in social networks with location sensitivity, World Wide Web-internet and Web Information Systems 18 (5) (2015) 1393–1417.

[8] Z. Guan, X. Yan, L. M. Kaplan, Measuring two-event structural correlations on graphs, Proceedings of the VLDB Endowment 5 (11) (2012) 1400–1411.

[9] Z. Guan, J. Wu, Q. Zhang, A. Singh, X. Yan, Assessing and ranking structural correlations in graphs, in: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, ACM, 2011, pp. 937–948.

[10] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users:real-time event detection by social sensors, in: Proc. International World Wide Web Conference, 2010, pp. 851–860.

[11] X. Wang, M. S. Gerber, D. E. Brown, Automatic crime prediction using events extracted from twitter posts., Springer Berlin Heidelberg 12 (2012) 231–238.

[12] A. Tumasjan, T. O. Sprenger, P. G. Sandner, I. M. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, in: Fourth international AAAI conference on weblogs and social media, 2010, pp. 178–185.

[13] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, B. Liu, Predicting flu trends using twitter data, in: INFOCOM WKSHPS, IEEE, 2011, pp. 702–707.

[14] L. Zhao, J. Wang, F. Chen, C. T. Lu, N. Ramakrishnan, Spatial event forecasting in social media with geographically hierarchical regularization, Proceedings of the IEEE 105 (10) (2017) 1953–1970.

[15] S. v. d. Beukel, S. H. Goos, J. Treur, An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle, Neurocomputing 338 (1) (2019) 361–371.

[16] Q. Qu, S. Liu, F. Zhu, C. S. Jensen, Efficient online summarization of large-scale dynamic networks, IEEE Transactions on Knowledge and Data Engineering 28 (12) (2016) 3231–3245.

[17] M. Mongiovi, P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, A. K. Singh, Netspot: Spotting significant anomalous regions on dynamic networks, in: ICDM, 2013, pp. 28–36.

[18] L. Akoglu, H. Tong, D. Koutra, Graph based anomaly detection and description: a survey, Data mining and knowledge discovery 29 (3) (2015) 626–688.

[19] J. Cadena, F. Chen, A. Vullikanti, Near-optimal and practical algorithms for graph scan statistics with connectivity constraints, ACM Transactions on Knowledge Discovery from Data (TKDD) 13 (2) (2019) 20.

[20] J. Gao, C. Zhou, J. X. Yu, Toward continuous pattern detection over evolving large graph with snapshot isolation, The VLDB Journal 25 (2) (2016) 269–290.

[21] M. Mongiovi, P. Bogdanov, A. K. Singh, Mining evolving network processes, in: ICDM, IEEE, 2013, pp. 537–546.

[22] P. Bogdanov, M. Mongiovì, A. K. Singh, Mining heavy subgraphs in time-evolving networks, in: ICDM, IEEE, 2011, pp. 81–90.

[23] X. Zhou, L. Chen, Event detection over twitter social media streams, The VLDB journal 23 (3) (2014) 381–400.

[24] H. Djidjev, G. Sandine, C. Storlie, S. Vander Wiel, Graph based statistical analysis of network traffic, in: Proceedings of the Ninth Workshop on Mining and Learning with Graphs, 2011.

[25] J. Neil, C. Hash, A. Brugh, M. Fisk, C. B. Storlie, Scan statistics for the online detection of locally anomalous subgraphs, Technometrics 55 (4) (2013) 403–414.

[26] D. B. Neill, J. Lingwall, A nonparametric scan statistic for multivariate disease surveillance, Advances in Disease Surveillance 4 (2007) 106.

[27] S. Velampalli, V. M. Jonnalagedda, Frequent subgraph mining algorithms: Framework, classification, analysis, comparisons, in: Data Engineering and Intelligent Computing, Springer, 2018, pp. 327–336.

[28] F. Zhu, Z. Zhang, Q. Qu, A direct mining approach to efficient constrained graph pattern discovery, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, ACM, 2013, pp. 821–832.

[29] N. Li, Z. Guan, L. Ren, J. Wu, J. Han, X. Yan, giceberg: Towards iceberg analysis in large graphs, in: 2013 IEEE 29th International Conference on Data Engineering (ICDE), IEEE, 2013, pp. 1021–1032.

[30] S. Somanchi, D. B. Neill, Graph structure learning from unlabeled data for event detection, arXiv preprint arXiv:1701.01470.

[31] H. S. Burkom, Biosurveillance applying scan statistics with multiple, disparate data sources, Journal of Urban Health 80 (1) (2003) i57–i65.

[32] D. B. Neill, A. W. Moore, M. Sabhnani, K. Daniel, Detection of emerging space-time clusters, in: KDD, ACM, 2005, pp. 218–227.

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

[34] M. Shao, J. Li, F. Chen, X. Chen, An efficient framework for detecting evolving anomalous subgraphs in dynamic networks, in: IEEE INFOCOM 2018-IEEE Conference on Computer Communications, IEEE, 2018, pp. 2258–2266.

[35] E. Schubert, M. Weiler, H.-P. Kriegel, Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 871–880.

[36] M. Walther, M. Kaisser, Geo-spatial event detection in the twitter stream, in: European conference on information retrieval, Springer, 2013, pp. 356–367.

[37] M. Adedoyin-Olowe, M. M. Gaber, C. M. Dancausa, F. Stahl, J. B. Gomes, A rule dynamics approach to event detection in twitter with its application to sports and politics, Expert Systems with Applications 55 (2016) 351–360.

[38] K. Watanabe, M. Ochi, M. Okabe, R. Onai, Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 2541–2544.

[39] T. Baldwin, P. Cook, B. Han, A. Harwood, S. Karunasekera, M. Moshtaghi, A support platform for event detection using social intelligence, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 69–72.

[40] T. Sugitani, M. Shirakawa, T. Hara, S. Nishio, Detecting local events by analyzing spatiotemporal locality of tweets, in: International Conference on Advanced Information NETWORKING and Applications Workshops, 2013, pp. 191–196.

[41] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of computational science 2 (1) (2011) 1–8.

[42] M. J. Fard, P. Wang, S. Chawla, C. K. Reddy, A bayesian perspective on early stage event prediction in longitudinal data, IEEE Transactions on Knowledge and Data Engineering 28 (12) (2016) 3126–3139.

[43] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, N. Ramakrishnan, Combining heterogeneous data sources for civil unrest forecasting, in: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ACM, 2015, pp. 258–265.

[44] L. Zhao, J. Ye, F. Chen, C.-T. Lu, N. Ramakrishnan, Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 2085–2094.

[45] Q. Zhang, N. Perra, D. Perrotta, M. Tizzoni, D. Paolotti, A. Vespignani, Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model, in: WWW, 2017, pp. 311–319.

[46] L. Zhao, F. Chen, C.-T. Lu, N. Ramakrishnan, Spatiotemporal event forecasting in social media, in: Proceedings of the 2015 SIAM International Conference on Data Mining, 2015, pp. 963–971.

[47] L. Hu, J. Li, L. Nie, X.-L. Li, C. Shao, What happens next? future subevent prediction using contextual hierarchical lstm., in: AAAI, 2017, pp. 3450–3456.

[48] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, H. Wang, et al., Predicting social unrest events with hidden markov models using gdelt, Discrete Dynamics in Nature and Society 2017 (2017) 1–13.

[49] T. Rekatsinas, S. Ghosh, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, N. Ramakrishnan, Sourceseer: Forecasting rare disease outbreaks using multiple data sources, in: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 379–387.

[50] F. Chen, D. B. Neill, Non-parametric scan statistics for disease outbreak detection on twitter, Online journal of public health informatics 6 (1) (2014a) e155.

[51] C. Groër, B. D. Sullivan, D. Weerapurage, Inddgo: Integrated network decomposition & dynamic programming for graph optimization, ORNL/TM-2012/176.

[52] H. L. Bodlaender, A partial k-arboretum of graphs with bounded treewidth, Theoretical computer science 209 (1) (1998) 1–45.

[53] H. L. Bodlaender, Treewidth: Structure and algorithms, in: SIROCCO, Vol. 4474, Springer, 2007, pp. 11–25.

[54] K.-i. Kawarabayashi, B. Mohar, Some recent progress and applications in graph minor theory, Graphs and combinatorics 23 (1) (2007) 1–46.

[55] L. Lovász, Graph minor theory, Bulletin of the American Mathematical Society 43 (1) (2006) 75–86.

[56] F. Gavril, The intersection graphs of subtrees in trees are exactly the chordal graphs, Journal of Combinatorial Theory, Series B 16 (1) (1974) 47–56.

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

[58] R. H. Berk, D. H. Jones, Goodness-of-fit test statistics that dominate the kolmogorov statistics, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47 (1) (1979) 47–59.

[59] A. B. Adcock, B. D. Sullivan, M. W. Mahoney, Tree decompositions and social graphs, Internet Mathematics 12 (5) (2016) 315–361.

[60] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, N. Tatti, Event detection in activity networks, in: KDD, 2014, pp. 1176–1185.

[61] M. Letsios, O. D. Balalau, M. Danisch, E. Orsini, M. Sozio, Finding heaviest k-subgraphs and events in social media, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), IEEE, 2016, pp. 113–120.

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