Efficient Nonparametric Subgraph Detection Using Tree Shaped Priors

Nannan Wu* **Feng Chen†** **Jianxin Li*§** **Baojian Zhou†** **Naren Ramakrishnan‡**
*Dept. of Computer Science & Engineering, Beihang University, Beijing 100191, China
†Dept. of Informatics, University at Albany, SUNY, Albany, NY 12203
‡Dept. of Computer Science, Virginia Tech, Arlington, VA 22203
{wunannan, lijx}@act.buaa.edu.cn {fchen5, bzhou6}@albany.edu naren@cs.vt.edu

Abstract

Non-parametric graph scan (NPGS) statistics are used to detect anomalous connected subgraphs on graphs, and have a wide variety of applications, such as disease outbreak detection, road traffic congestion detection, and event detection in social media. In contrast to traditional parametric scan statistics (e.g., the Kulldorff statistic), NPGS statistics are free of distributional assumptions and can be applied to heterogeneous graph data. In this paper, we make a number of contributions to the computational study of NPGS statistics. First, we present a novel reformulation of the problem as a sequence of Budget Price-Collecting Steiner Tree (B-PCST) sub-problems. Second, we show that this reformulated problem is NP-hard for a large class of non-parametric statistic functions. Third, we further develop efficient exact and approximate algorithms for a special category of graphs in which the anomalous subgraphs can be reformulated in a fixed tree topology. Finally, using extensive experiments we demonstrate the performance of our proposed algorithms in two real-world application domains (water pollution detection in water sensor networks and spatial event detection in social media networks) and contrast against state-of-the-art connected subgraph detection methods.

1 Introduction

Anomalous subgraph detection has attracted much attention in recent years (Duczmal, Kulldorff, and Huang 2006; Takahashi et al. 2008; Sharpnack, Singh, and Rinaldo 2013; Speakman, McFowland Iii, and Neill 2015; Qian, Saligrama, and Chen 2014; Li et al. 2015). We consider a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, where each $v \in \mathbb{V}$ is associated with features values x_v that follow some statistical distribution. The general goal of anomalous subgraph detection is to optimize some objective function ($F(S)$) of the abnormality of the feature values over all connected subsets of vertices ($S \subseteq \mathbb{V}$). To motivate this scenario, consider the *cholera* outbreak problem (Patil, Taillie, and others 2003) as shown in Figure 1. Suppose we have a network of counties (vertices) and each vertex has a feature referring to the number of cases of *cholera* in that county on a given day. Suppose further that two vertices are connected by an edge if they share a boundary. We wish to identify possible *cholera* outbreaks at an early stage, which requires identifying subtle patterns (e.g., a 20% increase in the number of patients with symptoms of *cholera* in four local (connected) counties) in the noisy background data. These subtle signals may not be detectable if we examine only a small part of the affected subset (e.g., a single county) or a larger connected subset containing many unaffected vertices (e.g., the aggregate count for the entire state). As a result, traditional "bottom-up" methods (which identify and aggregate individual vertices (Chandola, Banerjee, and Kumar 2009)) and "top-down" methods (which detect anomalous global trends) often have low power for detecting events (Chen and Neill 2014; Neill 2012).

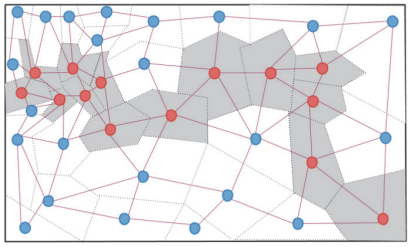

Figure 1: A potential *cholera* outbreak lead to elevated number of infected cases in counties near the river, which form an irregular shaped connected subgraph (cluster) of counties. (Redrawn from (Patil, Taillie, and others 2003).)

The underlying assumption behind anomalous pattern detection is that features of a majority of vertices are generated from the same distribution representing the (typically unknown and possibly complex) normal behavior of the system; thus, we wish to detect connected or correlated subgraphs of vertices which are unexpected given the typical data distribution (e.g. Gaussian, Poisson, or mixture of Gaussians). Existing methods can be categorized into two main groups, namely parametric and nonparametric methods. Parametric methods assume specific forms of distributions for features of normal and abnormal vertices, respectively, and formalize anomaly detection as a hypothesis testing problem. In particular, under the alternative hypothesis ($H_1(S)$), an underlying anomalous phenomenon is characterized in the following manner: features of a majority of

§Corresponding author: Jianxin Li

the vertices are generated from the same background distribution, and features of perhaps a small connected subset $S \subseteq \mathbb{V}$ of vertices are generated from a different distribution. The goal is to maximize an appropriate set function ($F(S)$), typically the likelihood ratio $F(S) = \frac{\Pr(\text{Data}|H_1(S))}{\Pr(\text{Data}|H_0)}$, over all possible connected subsets S (with H_0 being the null hypothesis). Depending on the specific forms of distributions assumed, a number of methods have been proposed, including an expectation-based Poisson statistic (Neill et al. 2005), the Kulldorff statistic (Kulldorff 1997), the elevated mean scan statistic (Qian, Saligrama, and Chen 2014), and various others.

Nonparametric methods do not associate specific forms of distributions with normal and abnormal vertices. Instead, they first estimate a p-value for each vertex based on empirical calibration by comparing the current features of this vertex with its features in the historical data for the vertex (Chen and Neill 2014; McFowland, Speakman, and Neill 2013). This approach then maximizes a score function $F(S)$ of p-values in S, typically nonparametric scan statistic measuring the significance of the collection of p-values in S, over all possible connected subsets. A number of NPGS statistic functions have been proposed in recent years, including the Berk-Jones (BJ) statistic (Berk and Jones 1979), the Higher Criticism (HC) statistic (Donoho and Jin 2004), the Tippet's statistic, rank truncated statistic, and various others. Note that these nonparametric statistic functions were originally proposed to combine p-values from a set of hypothesis tests in the area of statistical meta analysis. Recent studies show that these functions can be satisfactorily used with NPGS for detecting anomalous subgraphs (McFowland, Speakman, and Neill 2013; Chen and Neill 2014; Bogdanov, Mongiovì, and Singh 2011). The main contributions of our study are summarized as follows:

- **Hardness Analysis.** We reformulate the NPGS problem as a sequence of B-PCST sub-problems, and show that this reformulated problem is NP-hard for a general category of nonparametric statistic functions. These functions satisfy two intuitive properties on the cardinality of the input subgraph S and the number of vertices in S that are significant at a confidence level α.
- **Exact and approximate algorithms for graphs with tree shaped priors.** We develop efficient algorithms to the NPGS problem that are guaranteed to find an optimal solution in worst-case time $O(N^3)$ and an approximate solution in worst-case time $O(N^2/\epsilon)$ when the connected subgraph can be reformulated in a fixed tree topology.
- **Comprehensive experiments to validate the effectiveness and efficiency of the proposed techniques.** We conduct extensive experiments on a water sensor dataset and a Weibo dataset. The results demonstrate that our proposed algorithms outperform existing representative techniques in both performance and quality.

2 Nonparametric Graph Scan Statistics

Given a graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, p)$ where $\mathbb{V} = \{v_1, ..., v_N\}$, N refers to the total number of vertices, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ refers to the set of edges, and the mapping function $p : \mathbb{V} \rightarrow [0, 1]$ defines a single empirical p-value to each vertex v, which can be calculated based on empirical calibration by comparing current features of v with its features in the historical data for v (Chen and Neill 2014; McFowland, Speakman, and Neill 2013). The general form of the Non-Parametric Graph Scan (NPGS) statistic (Chen and Neill 2014; McFowland, Speakman, and Neill 2013) is defined as:

$$F(S) = \max_{\alpha} F_\alpha(S) = \max_{\alpha} \phi(\alpha, N_\alpha(S), N(S)), \quad (1)$$

where $S \subseteq \mathbb{V}$ refers to a connected set of vertices (subgraph), $N_\alpha(S) = \sum_{v \in S} \delta(p(v) \leq \alpha)$ is the number of p-values significant at level α, $N(S) = \sum_{v \in S} 1$ is the total number of p-values in S. The function $\delta(\cdot) = 1$ if its input is True, otherwise $\delta(\cdot) = 0$. Denote $\bar{N}_\alpha(S) = \sum_{v \in S} \delta(p(v) > \alpha)$, which means that $N(S) = \bar{N}_\alpha(S) + N_\alpha(S)$. The significance level α can be optimized between 0 and some constant α_{max} (0.15 by default). We assume that the function $\phi(\alpha, N_\alpha(S), N(S))$ satisfies two intuitive properties:

(P1) ϕ is monotonically ***increasing*** w.r.t. $N_\alpha(S)$,

(P2) ϕ is monotonically ***decreasing*** w.r.t. $N(S) - N_\alpha(S)$.

These assumptions follow naturally because the score ϕ increases with the number of significant p-values and deceases with the number of insignificant p-values at the level α. The importance to consider a range of α in the function is discussed in (Chen and Neill 2014). The range of α refers to the set of all possible p-values in $\mathbb{G}$ between 0 and α.

This paper presents efficient algorithms for the large class of nonparametric scan statistics that satisfy the above two properties, such as the Berk-Jones (BJ) statistic (Berk and Jones 1979), the Higher Criticism (HC) statistic (Donoho and Jin 2004), the Kolmogorov-Smirnov statistic, the Davidov-Herman statistic, and the chi-bar squared statistic. For illustration purpose, we consider the first two functions. The BJ statistic is defined as:

$$\varphi_{BJ}(\alpha, N_\alpha(S), N(S)) = N(S) \times \text{KL}\left(\frac{N_\alpha(S)}{N(S)}, \alpha\right), \quad (2)$$

where KL is the Kullback-Liebler divergence between the observed and expected proportions of p-values less than α.

The HC statistic is defined as:

$$\varphi_{HC}(\alpha, N_\alpha(S), N(S)) = \frac{N_\alpha(S) - N(S)\alpha}{\sqrt{N(S)\alpha(1-\alpha)}}. \quad (3)$$

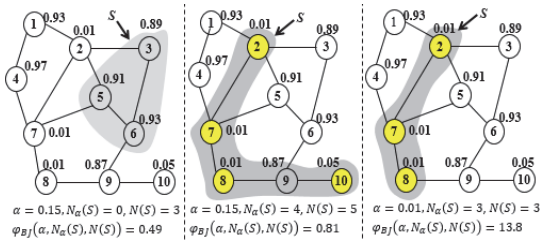

Figure 2: The BJ statistic scores of the three example subgraphs demonstrate that this score function increases with $N_\alpha(S)$ and decreases with $N(S) - N_\alpha(S)$ and α. Yellow-colored vertices refer to the vertices whose p-values are less than or equal to α.

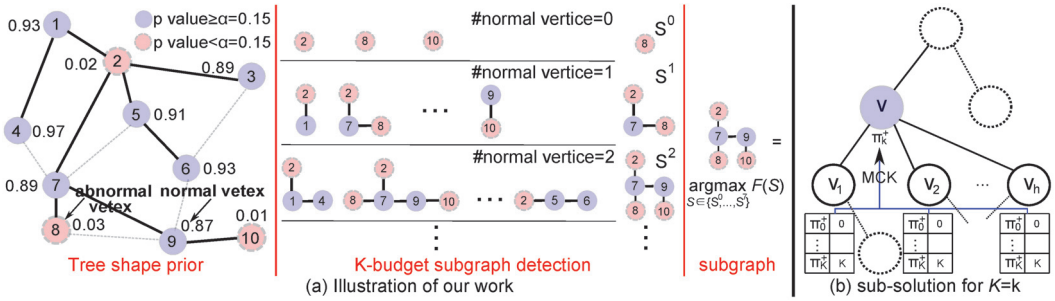

Figure 3: (a) An illustration of our work to decompose NPGS problem into a sequence of K-budget subgraph detection problems. (b) With the number of normal vertices is equal to K, including v, we aim to find a solution including more abnormal vertices. We consider assigning the value of π_K^+ in vertex V as a multiple-choice knapsack problem from π^+ of children $V_{1,\dots,h}$. MCK refers to multiple choice knapsack.

Given a selected nonparametric scan statistic function $\varphi(\alpha, N_\alpha(S), N(S))$, the detection of the most anomalous connected subgraph from $\mathbb{V}$ can be formalized as the following optimization problem:

$$\max_{S\subseteq\mathbb{V}:S \text{ is connected}} \max_{\alpha\leq\alpha_{max}} \phi(\alpha, N_\alpha(S), N(S)), \quad (4)$$

which is equivalent to the problem:

$$\max_{\alpha\in\mathbb{U}(\mathbb{V},\alpha_{max})} \max_{S\subseteq\mathbb{V}:S \text{ is connected}} \phi(\alpha, N_\alpha(S), N(S)), \quad (5)$$

where $\mathbb{U}(\mathbb{V}, \alpha_{max})$ refers to the union of $\{\alpha_{max}\}$ and the set of distinct p-values less than α_{max} in $\mathbb{V}$.

3 Methodology

This section presents a new reformulation of the NPGS problem and develops efficient algorithms for a special category of graph data where the connectivity constraint of the subgraph can be reformulated in a fixed tree topology.

3.1 Problem Reformulation

The hardness analysis of the NPGS problem is difficult as it involves a nonlinear objective function, and can not be reduced from known NP-hard problems that often involve linear objective functions. The following theorem shows that the NPGS problem can be reformulated a sequence of B-PCST sub-problems, and the hardness of the reformulated problem can be readily analyzed.

Theorem 1 (NPGS Reformulation). *The NPGS problem (5) is equivalent to the following problem:*

$$(\hat{\alpha}, \hat{S}) = \max_{\alpha\in\mathbb{U}(\mathbb{V},\alpha_{max})} \max_{S_\alpha\in\{S_\alpha^0,\dots,S_\alpha^N\}} \phi(\alpha, N_\alpha(S), N(S)) \quad (6)$$

where each $S_\alpha^K = \mathbb{V}_{\mathcal{T}_\alpha^K}$, and $\mathcal{T}_\alpha^K$ is the optimal subtree of the budget node-weighted Prize-Collecting Steiner Tree problem (B-PCST):

$$\mathcal{T}_\alpha^K = \max_{\mathcal{T}\in\mathbb{T}(\mathbb{G})} \sum_{v\in\mathbb{V}_\mathcal{T}} \pi_\alpha(v), \quad s.t. \quad \sum_{v\in\mathbb{V}_\mathcal{T}} c_\alpha(v) \leq K, \quad (7)$$

where $\mathbb{T}(\mathbb{G}) \equiv \{\mathcal{T} = (\mathbb{V}_T, \mathbb{E}_T)\}$ denotes the set of sub-trees of $\mathbb{G}$, $\pi_\alpha(v) = 1$ and $c_\alpha(v) = 0$, if $p(v) \leq \alpha$; otherwise, $\pi_\alpha(v) = 0$ and $c_\alpha(v) = 1$.

Proof. This theorem can be proved by contradiction. Suppose $(\hat{\alpha}, \hat{S})$ is not an optimal solution to the NPGS problem. It follows that there exists a different solution (α^*, S^*), such that $\phi(\alpha^*, N_{\alpha^*}(S^*), N(S^*)) > \phi(\hat{\alpha}, N_{\hat{\alpha}}(\hat{S}), N(\hat{S}))$. Let $\hat{K} := N(\hat{S}) - N_{\hat{\alpha}}(\hat{S})$ and $K^* := N(S^*) - N_{\alpha^*}(S^*)$. We first observe that $N_{\alpha^*}(\mathbb{V}_{\mathcal{T}_{\alpha^*}^{K^*}}) = N_{\alpha^*}(S^*)$; Otherwise, $\mathbb{V}_{\mathcal{T}_{\alpha^*}^{K^*}}$ will be the optimal subset, instead of S^* due to (P1) and (P2). Similarly, it can be shown that $N_{\hat{\alpha}}(\mathbb{V}_{\mathcal{T}_{\hat{\alpha}}^{\hat{K}}}) = N_{\hat{\alpha}}(\hat{S})$. As $(\hat{\alpha}, \hat{S})$ is the optimal solution to the reformulated problem (6), the inequality must be true: $\phi(\alpha^*, N_{\alpha^*}(S^*), N(S^*)) \leq \phi(\hat{\alpha}, N_{\hat{\alpha}}(\hat{S}), N(\hat{S}))$, a contradiction. Therefore, the initial assumption – $(\hat{\alpha}, \hat{S})$ is not an optimal solution to the NPGS problem – must be false. □

Theorem 2 (Hardness). *The reformulated problem (6) is NP-hard for the large class of nonparametric scan statistics that satisfy (P1) and (P2).*

Proof. Each subproblem (7) is a binary-case B-PCST problem that is known to be NP-hard (Johnson, Minkoff, and Phillips 2000). It can then be readily proved that the reformulated problem (6) is NP-hard. □

3.2 Approximations with tree shaped priors

Although the B-PCST sub-problem (7) can be approximated with the factor of $O(\log N)$ in polynomial time (Bateni, Hajiaghayi, and Liaghat 2013), both the Big-O approximation factor and the polynomial time complexity of this approximation are not satisfactory for large graph analysis. To design more efficient solutions for the subproblem (7), we propose to reformulate the connectivity constraint of the subgraph S on a fixed topology. Particularly, we approximate the graph $\mathbb{G}$ as a tree $\mathcal{T}_r$ originating at a given root vertex $r \in \mathbb{V}$, and the search of the best connected subgraph S for the NPGS problem is approximated as the search of the best sub-tree in $\mathcal{T}_r$. There are several heuristics to find the tree for the input graph: (1) breadth-first tree; (2) random spanning tree; (3) steiner tree; and (4) geodesic shortest path tree. The first three tree heuristics have been successfully applied to discrepancy maximization on general graphs (Gionis, Mathioudakis, and Ukkonen 2015). The fourth tree heuristic has

been successfully applied to image segmentation and sensor networks (Stühmer, Schröder, and Cremers 2013).

Breadth-First Tree (BFS-Tree): From random candidate root vertices, it generates a breadth-first tree for each root.

Random Spanning Tree (Random-ST): We get a random tree by assigning a weight (uniformly from [0, 1]) to each edge, and computing the minimum weight spanning tree.

Steiner Tree (Steiner-T): Intuitively, a tree is good if abnormal vertices are interconnected with the least number of normal vertices. For each $\alpha \in \mathbb{U}(\mathbb{V}, \alpha_{max})$, if we denote each abnormal vertex as a *terminal* vertex, and each normal vertex as a *steiner* vertex, trees can be identified by generating the steiner trees of the input graph.

Geodesic Shortest Path tree (Geodesic-SPT): For the NPGS problem, we define the optimal cost of the connecting path $\mathbf{p}$ between a fixed vertex s and other x based on the statistic: $\exp\{-\max_\alpha \phi(\alpha, N_\alpha(S_\mathbf{p}), N(S_\mathbf{p}))\}$ (Stühmer, Schröder, and Cremers 2013), where $S_\mathbf{p}$ are vertices in $\mathbf{p}$. The tree can be got through (Narvez, Siu, and Tzeng 2000).

3.3 Dynamic algorithms for the problem (7)

When the input graph $\mathbb{G}$ is a tree $\mathcal{T}(r)$ with the root vertex r, we can solve the problem (7) optimally, using dynamic programming (DP). We first introduce a few notations:

- $\mathcal{T}(v)$: a sub-tree of $\mathbb{G}$ with the root vertex v.
- π_l^{-v}: the value of the best l-budget sub-tree to the problem (7) in $T(v)$ that does not contain v.
- π_l^{+v}: the value of the best l-budget sub-tree to the problem (7) in $T(v)$ that contains v.
- π_l^v: $\pi_l^v = \max\{\pi_l^{-v}, \pi_l^{+v}\}$.
- s_l^v: a boolean value that indicates if vertex v belongs to the best l-budget sub-tree in $\mathcal{T}(v)$.
- n_l^v: a vertex pointer that indicates to which child of v to find the best l-budget sub-tree, if $s_l^v = False$.

Algorithm 1: Tree-Shaped-Priors Subgraph Detection

Input: Graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, p)$
Result: The most anomalous subgraph S^*

Set $\alpha_{max} = 0.15$ and $C = 5$;
for $c \in \{1, \cdots, C\}$ **do**
  Select seed v_0 from $\{v | v \in \mathbb{V}, p(v) \le \alpha_{max}\}$;
  Approximate the graph $\mathbb{G}$ as a tree $\mathcal{T}(v_0)$;
  for $\alpha \in \mathbb{U}(\mathbb{V}, \alpha_{max})$ **do**
    for $K = 0, \cdots, N(\mathbb{V}) - N_\alpha(\mathbb{V})$ **do**
      $S_\alpha^K \leftarrow \text{DP}(K, \mathcal{T}_{v_0}, v_0, \alpha)$ in Section 3.3;
    end
  end
  $S^c = \arg\max_{\alpha \in \mathbb{U}(\mathbb{V}, \alpha_{max}), K} \phi(\alpha, N_\alpha(S_\alpha^K), N(S_\alpha^K))$;
end
Calculate $c^* = \arg\max_c \phi(\alpha, N_\alpha(S^c), N(S^c))$;
return S^{c^*}

- $\mathcal{C}_l^v$: a set of tuples of the form (v', t). Hereby, v' is a child of v and t is an integer number that denotes the size of the sub-tree to be collected in $\mathcal{T}(v')$.
- $\mathcal{C}(v)$: the set of children of v in $\mathcal{T}(v)$.

The DP procedure is presented as follows:

Leaf vertex: We set initial values to attributes of leaf vertices. For a leaf vertex v, if $p(v) > \alpha$, set $\pi_{l\in\{0,1\}}^{-v} = 0$, $\pi_{l\in\{0,1\}}^{+v} = 0$, $\pi_{l\in\{0,1\}}^{v} = 0$ and $s_{l\in\{0,1\}}^{v} = False$, and if $p(v) \le \alpha$, set $\pi_0^{+v} = 1$, $\pi_0^v = 1$ and $s_0^v = True$.

Non-leaf vertex: Attributes n_l^v and π_l^{-v} are computed as:

$$n_l^v = \max_{v_i}\{\pi_l^{v_1}, \cdots, \pi_l^{v_h}\},\ \pi_l^{-v} = \pi_l^{n_l^v}, \tag{8}$$

where $\{v_1, \cdots, v_h\}$ refer to the h child vertices of the vertex v. As illustrated in Figure 3 (b), the computation of the attribute π_l^{+v} can be reduced to a 0-1 multiple-choice knapsack (0-1 MCK) problem that has the approximation factor $(1+\epsilon)$ and the running time $O(Kh/\epsilon)$ (Bansal and Venkaiah). The problem is to select at most one $\pi_j^{+v_i}$ from each child $i = \{1, \cdots, h\}$ such that the sum of profit is maximized without summing budget j to exceed $l - \delta(p(v) > \alpha)$. π_l^{+v} can then be calculated as:

$$\pi_l^{+v} = \max_{\mathbf{x}} \delta(p(v) \le \alpha) + \sum_{i=1}^{h}\sum_{j=0}^{K} \pi_j^{+v_i} \cdot x_{i,j} \tag{9}$$

$s.t.\ \sum_{j=0}^{K} x_{i,j} \le 1,\ i = \{1, \cdots, h\},\ \sum_{i=1}^{h}\sum_{j=0}^{K} j \cdot x_{i,j} \le l - \delta(p(v) > \alpha)$, where $\mathbf{x} \in \{0,1\}^{h\times K}$. Set $s_l^v = False$ if $\pi_l^{-v} > \pi_l^{+v}$, otherwise $s_l^v = True$. Given the result $\mathbf{x}$ from the problem (9), the set attribute $\mathcal{C}_l^v$ can be calculated as:

$$\mathcal{C}_l^v = \{(v_i, j) | x_{i,j} = 1\}. \tag{10}$$

We compute DP$(K, \mathcal{T}_{v_0}, v_0, \alpha)$ using two-stages: 1) Calculate $\{\pi_l^{-v}, \pi_l^{+v}, \pi_l^v, s_l^v, n_l^v, C_l^v\}_{l=0}^K$ for $v \in \mathbb{V}$; 2) From the root vertex v_0 down, we find the first $s_K^v = True$. Set $S_\alpha^K \leftarrow \Phi$, compute $findchild(\mathcal{T}, v, l) = \sum_{(v_i,j)\in\mathcal{C}_l^v} S_\alpha^K \cup \{v_i\} \cup findchild(\mathcal{T}, v_i, j)$ and last return $findchild(\mathcal{T}, s_K^v, K)$. The detailed information of the implementation can be found from the online appendix (Wu et al.).

Theorem 3. *1) **Exact Solution**: If each 0-1 MCK subproblem (9) is solved via dynamic programming (Pisinger 1994), Algorithm 1 is guaranteed to find the optimal solution to the tree-priors-based NPGS problem in worst-case time $O(|\mathbb{U}(\mathbb{V}, \alpha_{max})| \cdot N^3)$; 2) **Approximate Solution**: If each 0-1 MCK subproblem (9) is solved via the algorithm (Bansal and Venkaiah) that has the approximation factor $(1+\epsilon)$, then Algorithm 1 is guaranteed to find an approximate solution to the tree-priors-based NPGS problem in worst-case time $O(|\mathbb{U}(\mathbb{V}, \alpha_{max})| \cdot N^2/\epsilon)$.*

Proof. An 0-1 MCK subproblem will be solved for each vertex v_i with the time $O(K^2 h_i)$ for exact solutions via dynamic programming (Pisinger 1994) and the time $O(\frac{Kh_i}{\epsilon})$ for approximate solutions (Bansal and Venkaiah), where h_i refers to the number of child vertices of v_i. As $K \to N$ and $\sum_{v_i\in\mathbb{V}} h_i + 1 = N$, the times can be readily proved. □

Method	Noise Ratio (0%)	4%	8%	10%	30%
BFS-Tree (BJ)	0.94, 0.48 (0.64)	0.95, 0.47 (0.63)	0.93, 0.50 (0.66)	0.91, 0.47 (0.62)	0.78, 0.33 (0.47)
Random-ST (BJ)	0.94, 0.77 (0.84)	0.93, 0.75 (0.83)	0.95, 0.65 (0.77)	0.93, 0.59 (0.71)	0.79, 0.39 (0.53)
Steiner-T (BJ)	1.00, 0.99 (**1.00**)	0.98, 0.96 (**0.97**)	0.95, 0.92 (**0.94**)	0.94, 0.89 (**0.91**)	0.77, 0.52 (0.62)
Geodesic-SPT (BJ)	0.96, 0.85 (0.90)	0.92, 0.63 (0.75)	0.88, 0.65 (0.75)	0.85, 0.56 (0.68)	0.78, 0.38 (0.51)
EventTree	0.97, 1.00 (0.98)	0.89, 0.98 (0.93)	0.70, 0.98 (0.82)	0.42, 0.97 (0.59)	0.09, 0.90 (0.17)
NPHGS (BJ)	1.00, 0.92 (0.96)	0.99, 0.77 (0.84)	0.97, 0.50 (0.66)	0.97, 0.39 (0.55)	0.78, 0.06 (0.11)
LTSS (BJ)	1.00, 1.00 (**1.00**)	0.48, 0.96 (0.64)	0.34, 0.92 (0.50)	0.30, 0.90 (0.45)	0.11, 0.70 (0.20)
Graph-Laplacian	0.93, 0.87 (0.90)	0.95, 0.43 (0.60)	0.89, 0.23 (0.37)	0.68, 0.12 (0.20)	0.97, 0.50 (**0.66**)

Table 1: Comparison w.r.t. different noise levels in the water pollution dataset: Precision, Recall (F-Measure)

Method	FPR (FP/Day)	TPR (Detection)	TPR (Forecast & Detect)	Lead Time (Days)	Lag Time (Days)	Run Time (Minutes)
TSPSD-Steiner HC (BJ)	0.100	**0.55** (0.49)	**0.66** (**0.66**)	**0.98** (0.97)	**3.53** (3.54)	18 (0.3) (18 (0.3))
TSPSD-Steiner HC (BJ)	0.150	**0.62** (0.61)	0.70 (**0.71**)	**0.88** (0.82)	**3.92** (4.15)	18 (0.3) (18 (0.3))
TSPSD-Steiner HC (BJ)	0.200	**0.66** (**0.66**)	**0.74** (**0.74**)	**0.87** (0.82)	**4.00** (4.15)	18 (0.3) (18 (0.3))
NPHGS HC (BJ)	0.100	0.32 (0.41)	0.47 (0.55)	0.72 (0.59)	4.35 (4.70)	3 (8)
NPHGS HC (BJ)	0.150	0.43 (0.48)	0.60 (**0.71**)	0.72 (0.70)	4.27 (4.40)	3 (8)
NPHGS HC (BJ)	0.200	0.50 (0.63)	0.70 (**0.74**)	0.71 (0.74)	4.32 (4.12)	3 (8)
EventTree	0.100	0.51	0.65	0.91	3.71	7.5
EventTree	0.150	0.57	0.68	0.70	4.40	7.5
EventTree	0.200	0.60	0.72	0.81	4.12	7.5

Table 2: Comparison between TSPSD and Other Models on the Haze outbreak dataset. The scores of HC and BJ statistics are shown in the format: $x(y)$, where x refers the score of HC, and y refers to that of BJ. For 18(0.3), 18 is the overall run time and 0.3 is the detection time.

3.4 Optimization

Algorithm 1 can be further improved in three ways. First, instead of $\bar{N}_\alpha(\mathbb{V})$ calls to the sub-procedure in Section 3.3, it suffices to call DP procedure only once with $K = \bar{N}_\alpha(\mathbb{V})$, and the returned Tree $\mathcal{T}$ with the updated attributes $\{\pi_l^{-v}, \pi_l^{+v}, \pi_l^v, s_l^v, n_l^v, C_l^v\}_{l=0}^K$ at each vertex v can be used to retrieve the solution to problem (7) for $K = 0, \cdots, \bar{N}_\alpha(\mathbb{V})$.

Second, after attributes of the vertex v are calculated: $\{\pi_l^v, \pi_l^{-v}, \pi_l^{+v}, n_l^v, \mathcal{C}_l^v\}_{l=0}^K$, we check π_l^v based on the order $l = K, \cdots, 1$. Attributes $\{\pi_l^v, \pi_l^{-v}, \pi_l^{+v}, n_l^v, \mathcal{C}_l^v\}$ related to the l-budget solution can be safely removed, if at least one of the following conditions is satisfied: 1) $\pi_l^{+v} \leq \pi_{l-1}^{+v}$; 2) $\phi(\alpha, a+\pi_l^v, a+l+\pi_l^v) \leq \phi(\alpha, a+\pi_{l-1}^v, a+\pi_{l-1}^v+l-1)$; and 3) $\phi(\alpha, a+\pi_l^v, a+\pi_l^v+l) \leq \phi(\alpha, a, a)$, where $a = N_\alpha(\mathbb{V}_\mathcal{T}) - N_\alpha(\mathbb{V}_{\mathcal{T}(v)})$.

Third, let $\mathbb{U}(\mathbb{V}, \alpha_{max}) = \{\alpha_1, \alpha_2, \cdots, \alpha_Z\}$ with ascending order based on the index. Suppose the current α is α_i. We maintain an additional attribute q in the root r that refers to an upper-bound of the number of abnormal vertices in the optimal subtree. Based on q, we can calculate an upperbound of the best subtree as follows: $\phi(\alpha_i, q, q)$. Initially $q = N_{\alpha_i}(\mathbb{V})$. As attributes of a vertex v are calculated, we apply the above optimization strategy to remove unnecessary l-budget subtrees rooted at v. Let $\mathcal{L} = \{l^1, \cdots, l^h\}$ refer to the set of l-values that have been pruned. Then q can be updated: $q = q - (\max_{l \in \{0, \cdots, \bar{N}_{\alpha_i}(\mathbb{V})\}} \{\pi_l^v\} N_\alpha(\mathbb{V}_{\mathcal{T}(v)}) - \max_l \{\pi_l^{+v}\})$. When each time q is updated, we compare the resulting upper bound $\phi(\alpha_i, q, q)$ with the best score $F_i = \max_{j \in \{1, \cdots i-1\}} \phi(\alpha_j, N_{\alpha_j}(S_{\alpha_j}), N(S_{\alpha_j}))$ calculated based on previous alpha values $\alpha_1, \cdots, \alpha_{i-1}$: If $\phi(\alpha_i, q, q) \leq F_i$, then we do not need to proceed the procedure related to α_i.

4 Experiments

4.1 Experiment Design

Datasets: 1) **Water Pollution Dataset.** The "Battle of the Water Sensor Networks" (BWSN) (Ostfeld and Salomons 2008) provides a real-world network of 12,527 nodes, and 25 nodes with chemical contaminant plumes that are distributed in four different areas. The spreads of contaminant plumes on graph were simulated using the water network simulator EPANET that was used in BWSN for a period of 8 hours. For each hour, each node has a sensor that reports 1 if it is polluted; otherwise, reports 0. We randomly selected ns percent vertices, and flipped their sensor binary values, where $ns = 0, 4, 8, 10, 30$, in order to test the robustness of methods to noises. In order to apply nonparametric graph scan methods to this data, we map the sensors whose report values are 1s to the empirical p-value 0.15 ($\leq \alpha_{max} \equiv 0.15$), and whose report values are 0s to 1.0.

2) **Event Detection Dataset.** We collected 1,433,937,815 tweets (nearly 10 percent of the whole Weibo[1] data) from Apr 11, 2014 to Jan 11, 2015. From this dataset, we selected 0.35 million tweets such that each tweet contains at least two terms from a set of 50 terms about haze outbreaks collected from domain experts, which are posted by 51,940 users. According to mentions in tweets and following relations, we construct a connected user-user network with 158,652 edges. Each user is geocoded with a province from location in profiles. For each day d and user u, we calculated the corresponding empirical p-value by the work in (Chen and Neill 2014). The methods will output a detected user subgraph, which is transformed into a province

[1]Weibo.com, the most popular online social networking services in China with more than 400 million users.

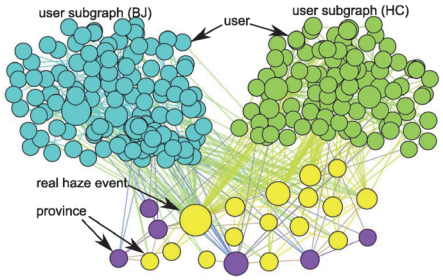

Figure 4: Haze events from Nov 27, 2014, in China. Within the 7 day window before and after that day, a yellow vertex refers to a successful forecast or detection; a blue vertex indicates an alert without a GSR record; Other color vertices consist of user subgraphs detected by BJ or HC statistics. The size of yellow and blue vertices is proportional to the count of users connected to them.

subgraph from the union of user's provinces in the subgraph. Gold Standard Reports (GSR) of 4279 official haze outbreak records (level ≥ 3) were collected from official websites (MEP), and each GSR record was formatted as ("Time(YYYYMMDD)", "Location(Province)")

Comparison Methods: We study four representative baselines: EventTree (Rozenshtein et al. 2014), NPHGS (Chen and Neill 2014), LTSS (Neill 2012), and Graph-Laplacian (Sharpnack, Singh, and Rinaldo 2013). We strictly followed strategies recommended by authors in their papers to tune the related model parameters. Specifically, for EventTree and Graph-Laplacian, we tested the set of λ values: $\{0.1, 0.2, \cdots, 1.0, 50, 100, \cdots, 1500\}$. As EventTree requires edge weights, we define the weight of an edge in the water pipeline network as the length of the pipeline segment to the edge and define the weight of an edge in the user-user network of Weibo as 1 without a better way. Two nonparametric scan statistics, HC and BJ, were evaluated. We set the parameters $\alpha_{max} \equiv 0.15$ and the number of seed nodes $C \equiv 5$ for NPHGS and our methods.

Our Methods: We denote Algorithm 1 as Tree-Shaped-Priors Subgraph Detection (TSPSD) and tested tree priors: BFS-Tree, Random-ST, Steiner-T, and Geodesic-SPT.

Performance Metrics: 1) precision, 2) recall, and 3) F-measure were employed when the true anomalous subgraphs are known. 4) FPR and 5) TPR, were used for the event detection dataset. For each GSR event, we decide whether the method: I) Had an alert in the province within 7 days before the event, which means to be "predicted"; II) Had an alert in the province within 7 days after the event, which means to be "detected"; or III) Hadno alert in the province within 7 days before or after the event, which is "undetected".

4.2 Results: Subgraph Detection

At noise levels 0%, 4%, 8% and 10% in Table 1, TSPSD with Steiner-T prior achieved the highest performance. Even if the dataset has 10% noise, TSPSD detected 89% truly contaminated vertices with the precision greater than 90%. At noise levels 30%, the values of TSPSD with Steiner-T were comparable to those of Graph-Laplacian method but slightly

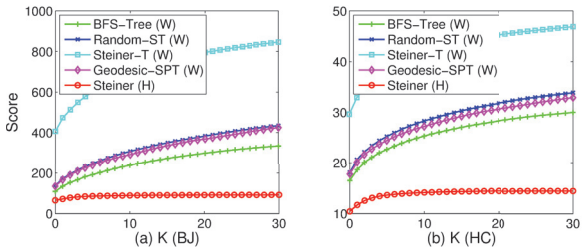

Figure 5: The average nonparametric scan statistic scores for each problem S_α^K in (7). (a) shows the BJ score for each tree prior in Water Pollution Dataset (W) and Haze Event Detection Dataset (H); (b) shows the HC score for each tree prior.

lower. From the overall performance in all different noise levels, TSPSD with Steiner-T performed more stable than baselines. The F-measures of TSPSD were higher than baselines in most cases. TSPSD at noise levels 30% has a higher F-score than baselines, EventTree, NPHGS and LTSS.

4.3 Results: Event Detection

For comparable false positive rates, TSPSD achieved the highest forecasting TPR and detection TPR than the two baseline methods in Table 2. The lead time represents how long we need to predict Haze event before it actually occurs. Our method predicting haze events is earlier than baselines, and that means the larger lead time. Haze events as natural events occur usually without exceeding a half day in China. However social events (e.g., protest events), often have trigger subevents and are driven by public sentiments, and can be potentially forecasted with a large lead time (e.g., 1 to 2 weeks). It is difficult to predict Haze events before a long time for Haze events do not have these factors. For the lag time, we use the less time to detect Haze events, and that means the less lag time. Our approach performs better than baselines. Although the run time of TSPSD was little higher than those of baseline methods, the overall time consists of tree generation and subgraph detection steps, and the first step consumes major time. For intuitively illustrating our proposed approaches effectively, we randomly select one day from the nine months to forecast or detect Haze events in Figure 4. We observe that the subgraph detected by HC statistic connected to blue vertices (*wrong alerts*) are apparently less than the subgraph detected by BJ statistic. This observation corresponds to results in Table 2. HC statistic performs better than BJ statistic. Our approach can be successfully used to predict Haze events in social media.

4.4 Sensitivity to the parameter K

The problem (7) is addressed in Section 3.3. For examining the sensitivity of selecting values of K, we plot each score $F(S_\alpha^K)$ for $K = 0, \cdots, 30$. From Figure 5 (a) and (b), we derive that $F(S_\alpha^K)$ is stable after $K = 20$. From the scores for BJ and HC nonparametric functions, we see that our approaches employing the Steiner-T prior perform best. In the Haze data set, the fewer connected users triggering Haze warnings caused to the less score for BJ and HC.

Time (Min)	BFS Tree	Random ST	Steiner Tree	Geo SPT
+Opt	0.11 (0.11)	0.93 (0.1)	0.77 (0.08)	0.62 (0.11)
-Opt	3.16 (3.15)	3.89 (2.8)	2.89 (2.01)	3.79 (3.00)
Time (Min)	EventTree	NPHGS	LTSS	Graph Laplacian
Base	13.10	1.82	0.93	24.61

Table 3: Average run times of our proposed and baseline methods on the Water pollution dataset. The run time of our method consists of two parts: tree generation and subgraph detection. Such as 0.93 (0.1), 0.93 is the overall run time and 0.1 is the detection run time. Our method has two versions: 1) -Opt: Algorithm 1; 2) +Opt: Algorithm 1 + optimizations (Subsection 3.4).

Results show that most of abnormal vertices are connected from each other with a small number of normal vertices.

4.5 Runtime: Tree Shaped Priors

As shown in Table 3, run times of TSPSD were comparable to baselines but slightly higher in Random-ST and Steiner-Tree without Opt since redundant computation for $\pi_l^{-v}, \pi_l^{+v}, n_l^v, \mathcal{C}_l^v$. The speed of TSPSD with Opt was 25 times faster than TSPSD without Opt. TSPSD with optimization (Opt) performed better than all baselines. Note that results based on the HC statistic were not due to space limitations. Its performance was similar to the BJ statistic on the first dataset.

5 Conclusions

This paper has proposed efficient algorithms to anomalous subgraph detection for a large class of nonparametric scan static functions. For future work, we will extend our work to heterogeneous graphs to maximize nonparametric scan statistic over subsets of vertices, features, and edge types.

6 Acknowledgments

This work is supported by China 973 Fundamental R&D Program (No.2014CB340300) , NSFC Program (No. 61472022), China MOST project (No.2012BAH46B04), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00337. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government.

References

Bansal, M., and Venkaiah, V. Improved fully polynomial time approximation scheme for the 0-1 multiple-choice knapsack problem.

Bateni, M.; Hajiaghayi, M.; and Liaghat, V. 2013. Improved approximation algorithms for (budgeted) node-weighted steiner problems. *CoRR* abs/1304.7530.

Berk, R. H., and Jones, D. H. 1979. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* 47(1):47–59.

Bogdanov, P.; Mongiovì, M.; and Singh, A. K. 2011. Mining heavy subgraphs in time-evolving networks. In *11th ICDM, Vancouver, BC, Canada, December 11-14, 2011*, 81–90.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41(3).

Chen, F., and Neill, D. B. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *KDD '14, NY, USA - August 24 - 27, 2014*, 1166–1175.

Donoho, D., and Jin, J. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* 32(3):962–994.

Duczmal, L.; Kulldorff, M.; and Huang, L. 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. *JCGS* 15(2):428–442.

Gionis, A.; Mathioudakis, M.; and Ukkonen, A. 2015. Bump hunting in the dark: Local discrepancy maximization on graphs. In *ICDE '15, Seoul, Korea, April 13 - April 17, 2015*, 64–75.

Johnson, D. S.; Minkoff, M.; and Phillips, S. 2000. The prize collecting steiner tree problem: theory and practice. In Shmoys, D. B., ed., *SODA*, 760–769. ACM/SIAM.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics: Theory and Methods* 26:1481–1496.

Li, J.; Wen, J.; Tai, Z.; and et al. 2015. Bursty event detection from microblog: A distributed and incremental approach. *Concurrency Computat.: Pract. Exper*.

McFowland, E.; Speakman, S.; and Neill, D. B. 2013. Fast generalized subset scan for anomalous pattern detection. *JMLR* 14(1):1533–1561.

MEP, C. http://datacenter.mep.gov.cn/.

Narvez, P.; Siu, K.-Y.; and Tzeng, H.-Y. 2000. New dynamic algorithms for shortest path tree computation. *J. Netw.* 8(6):734–746.

Neill, D. B.; Moore, A. W.; Sabhnani, M.; and Daniel, K. 2005. Detection of emerging space-time clusters. In *KDD '05, Chicago, Illinois, USA, August 21-24, 2005*, 218–227.

Neill, D. B. 2012. Fast subset scan for spatial pattern detection. *JRSS* 74(2):337–360.

Ostfeld, A., U. J., and Salomons, E. 2008. The battle of water sensor networks: A design challenge for engineers and algorithms. *J. WRPM* 134(6):556–568.

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

Pisinger, D. 1994. A minimal algorithm for the multiple-choice knapsack problem. *EJOR* 83:394–410.

Qian, J.; Saligrama, V.; and Chen, Y. 2014. Connected sub-graph detection. In *AISTATS '14, Iceland, April 22-25, 2014*, 796–804.

Rozenshtein, P.; Anagnostopoulos, A.; Gionis, A.; and Tatti, N. 2014. Event detection in activity networks. In *KDD '14, New York, NY, USA - August 24 - 27, 2014*, 1176–1185.

Sharpnack, J.; Singh, A.; and Rinaldo, A. 2013. Changepoint detection over graphs with the spectral scan statistic. In *AISTATS '13, Scottsdale, AZ, USA, April 29 - May 1, 2013*, 545–553.

Speakman, S.; McFowland Iii, E.; and Neill, D. B. 2015. Scalable detection of anomalous patterns with connectivity constraints. *to appear in JCGS*.

Stühmer, J.; Schröder, P.; and Cremers, D. 2013. Tree shape priors with connectivity constraints using convex relaxation on general graphs. In *ICCV '13, Sydney, Australia, Dec 1-8*, 2336–2343.

Takahashi, K.; Kulldorff, M.; Tango, T.; and Yih, K. 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *IJHG* 7:14.

Wu, N.; Chen, F.; Li, J.; Zhou, B.; and Ramakrishnan, N. Appendix, 2005: http://www.cs.albany.edu/∼fchen/aaaiap.pdf.