ANOMALY DETECTION ON CONNECTED SUBGRAPHS VIA CONSTANT-FACTOR APPROXIMATION ALGORITHMS FOR THE ELEVATED MEAN PROBLEM

XIFAN YU

Abstract. In this paper, we study the Elevated Mean problem and the anomaly detection problem. In the anomaly detection problem, there is a network with a certain measure of behaviors of the nodes, and the task is to distinguish whether the nodes behave normally or an anomalous cluster exists based on observations of the nodes in the network. The anomaly detection problem has a wide range of applications such as network intrusion detection and disease outbreak detection. The Elevated Mean problem arises naturally as the optimal test for a general hypothesis-testing formulation of anomaly detection, in which vertex features follow an exponential-family distribution and candidate anomalous clusters are connected induced subgraphs of the network.

The paper is structured into two parts. In the first part, we study the Elevated Mean problem from a worst-case perspective. We present two the first non-trivial polynomial-time approximation algorithms for the Elevated Mean problem and the k-Elevated Mean problem via black-box reductions to the Quota Prize-Collecting Steiner Tree problem and the Budget Prize-Collecting Steiner Tree problem, yielding constant factor approximations.

In the second part, we focus on the statistical setting of anomaly detection. Using the constant-factor approximations for the Elevated Mean problem and the k-Elevated Mean problem in the first part, we show the first polynomial-time decision rule T_{EM} that asymptotically achieves the minmax bound for the Elevated Mean scan statistics. Besides being polynomial-time solvable, the strength of this decision rule also lies in its generality of detecting anomalous clusters among the class of connected induced subgraphs, without relying on additional assumptions used by previous work. We also prove novel, more general separability results using the rule T_{EM}.

Contents

Date: May, 2021. This paper is written in partial fulfillment of the requirements of the research-oriented joint Bx/MS program of the Department of Computer Science, the University of Chicago, under the supervision of my advisor, Prof. Lorenzo Orecchia.

1. Introduction

The task of anomaly detection over network has a wide range of applications. To better motivate the research in this paper, we start by describing three canonical applications:

- **Network intrusion:** With the advent of networks, the growing scale and broadcast ability of networks render them vulnerable to external malicious attacks, which could result in loss of valuable data or compromise of confidentiality. Thus, the task of detecting network intrusion is critical in network security. Due to the structure of the network, network intrusions often take place in a cluster of connected nodes, which often exhibit anomalous behaviors compared to the normal nodes in the network.
- **Disease outbreak:** Disease outbreak detection is of great importance in contemporary society, as the transmission of infectious disease is greatly facilitated by the large volume of population migration. In order to minimize the severity and the scope of the pandemic, early and precise monitoring of disease outbreak is crucial. Common techniques for disease surveillance incorporate data from hospital visits, pharmaceutical orders, and laboratory tests together with geographical information.
- **Ecosystem disturbance:** Disturbance in ecosystem could have negative impact on the biodiversity in the region, and the effect could take place quickly and be long-lasting, particularly when such disturbances are man-made. As a result, detecting anomalous behavior in an ecosystem in a timely manner is essential to environmental preservation and resource management, and such detection often relies on data obtained from field measurement in some geographical location.

In the simplest abstract formulation of the anomaly detection problem, we are given graph and a feature value for each vertex, which may come from an underlying statistical model. The anomalous clusters are then defined as connected induced subgraphs whose feature values significantly depart from a baseline distribution. The connectedness requirement accounts for the assumption that anomalous behavior is localized by the network topology. For example, disease outbreak would usually happen in a contiguous geographical region, and network intrusion at the initial stage would affect a small cluster of connected nodes.

A paradigmatic approach to the anomaly detection problem is the method of scan statistics [3], which has been studied not only for graph models, but also for spatio-temporal models. In this setting, anomalous clusters are detected by maximizing a score function over the feature values of the nodes in the cluster. Depending on the application, the scan statistics method can either be parametric or non-parametric. Parametric scan statistics assumes that the observations of the vertices follow certain distribution, such as the Gaussian distribution or the Poisson distribution. On the other hand, non-parametric scan statistics does not assume the underlying distribution on the graph, and instead first estimates a p-value for each vertex and then looks for a subgraph with high numbers of significant p-values [11].

In this paper, we focus on the Elevated Mean scan statistics, a parametric scan statistics arising when the feature values follow exponential-family distributions. We show how approximate solutions to the Elevated Mean problem can be computed in polynomial time to give decision rules for the anomaly detection problem. Our only assumptions on the class of potential anomalous clusters are that they are connected and that they have sizes less than or equal to some number $k \in [1, |V|]$.

1.1. Related Work.

Earlier work has either focused on statistical guarantee of solving the exact Elevated Mean problem [3] without considering the computational aspect, or convex relaxation that can be efficiently solved [2, 19, 20] but fail to achieve any approximation guarantees and lead to statistical-computational trade-offs.

There are also heuristic algorithms for the anomaly detection problem based on the approach of simulated annealing using non-compactness penalty [12], iterative convergence to local optimum of additive scan statistics [23], or other machine learning approaches.

A number of other methods based on parametric scan statistics restrict the problem to special graphs or pose relatively strong assumptions on the class of potential anomalous clusters. For example, [1] scans for axis-parallel rectangles in the plane as potential anomalous clusters, [3] considers the case where the underlying graph is embedded into the Euclidean space and the potential anomalous clusters are bi-Lipschitz deformations of balls, and the case where the graph is a finite-dimensional grid, and [19] works on 2D grid and assumes that the potential anomalous clusters form subgraphs with internal conductance at least a certain value.

1.2. **Structure of the Paper.** In this paper, we follow the generalized likelihood ratio test schema [3] based on Elevated Mean scan statistics for connected subgraph detection. The general structure of the paper is as follows. We first describe the problems of interest in Section 2, and study the computational aspect of the Elevated Mean problem in Section 3, where we show a $(3+\varepsilon)$-approximation algorithm for the k-Elevated Mean problem in Theorem 3.11, and a $\sqrt{2}$-approximation algorithm for the Elevated Mean problem in Theorem 3.12. In Section 4, we formalize the anomaly detection problem as a hypothesis testing problem, and give a polynomial time decision rule using the constant-factor approximation algorithms for the Elevated Mean problem in Section 3. In Section 5, we analyze the separability of the anomaly detection problem based on the maximum degree of the underlying graph, and we show an asymptotically tight lower bound for the expected optimum scan statistics under the null hypothesis for the Gaussian model in Section 6.

2. Problem

In this section, we introduce the Elevated Mean problem, the anomaly detection problem, and the relevant variants of the Prize-Collecting Steiner Tree problem.

2.1. **Elevated Mean Problem.** The input to the Elevated Mean problem is a pair of undirected graph $G=(V,E)$ and a vertex-valued function $p: V \to \mathbb{R}^{\geq 0}$. The objective is to maximize the following scan statistics over a class C of connected induced subgraphs of G:

$$\frac{\sum_{v\in V(H)} p(v)}{\sqrt{|V(H)|}}, \tag{2.1}$$

In this paper, we consider two types of Elevated Mean problem:

- When the class C is the set of all connected induced subgraphs of G, we refer to this problem as the (unrestricted) Elevated Mean problem.
- When the class C is of the form
$$C = C_{\leq k} = \{\textit{connected induced subgraphs of } G \textit{ of size} \leq k\},$$
where $k \in \mathbb{N}$ is an additional input parameter, we refer to this problem as the k-Elevated Mean problem.

Remark 2.1. Note that the Elevated Mean problem is a special case of the k-Elevated Mean problem. When $k=n$, $C_{\leq k} = C_{\leq n}$ is the set of all connected induced subgraphs of G. In this paper, we will use n to denote the number of vertices in a graph.

Besides the optimization problems above, we also consider the Elevated Mean problem in the context of anomaly detection. The task of anomaly detection is concerned with the following hypothesis testing problem. Under the null hypothesis $\mathcal{H}_0$, the associated value of each vertex of the graph follows some i.i.d. distribution, so there is no anomalous cluster in the graph. Under the alternative hypothesis $\mathcal{H}_1$, the associated values of the vertices in a connected induced subgraph follow a different distribution from that of the rest of the vertices in the graph, and these vertices in this anomalous cluster appear to have values higher than the rest of the graph. The precise formulation of the anomaly detection problem will be presented in Section 4, where we will optimize the scan statistics (2.1) to get a decision rule for this hypothesis testing problem.

In this paper, we will first present two constant-factor approximation algorithms for the Elevated Mean problem and the k-Elevated Mean problem, with which we derive a decision rule

for the anomaly detection problem, and then show conditions under which we can separate the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ using this rule.

2.2. **Prize-Collecting Steiner Tree Problem and Its Variants.** The input to the Prize-Collecting Steiner Tree problem (PCST) is a tuple of undirected graph $G = (V, E)$, a vertex-valued function $p : V \to \mathbb{R}^{\geq 0}$, an edge-valued function $c : E \to \mathbb{R}^{\geq 0}$, and a root vertex $r \in V$. The objective is find a subtree $T' = (V', E')$ of G that contains the root r and minimizes the following quantity:

$$\sum_{e \in E'} c(e) + \sum_{v \notin V'} p(v).$$

Intuitively, think of $p(v)$ as the prize associated with vertex v and $c(e)$ as the cost associated with edge e. The objective then becomes to miss as few prizes on the vertices not covered by the subtree T' and to spend as little cost on the edges of the subtree T' as possible.

To derive the constant-factor approximation algorithms for the Elevated Mean problem and the k-Elevated Mean problem, we consider the following variants of PCST, initially studied by Johnson, Minkoff, and Phillips [15]:

- Quota Prize-Collecting Steiner Tree:
  The input to the Quota Prize-Collecting Steiner Tree problem (Quota PCST) is a tuple of undirected graph $G = (V, E)$, a vertex-valued function $p : V \to \mathbb{R}^{\geq 0}$, an edge-valued function $c : E \to \mathbb{R}^{\geq 0}$, a root vertex $r \in V$, and a quota $Q \geq 0$. The objective is to find a subtree $T' = (V', E')$ that contains the root r and minimizes the following quantity:
  $$\sum_{e \in E'} c(e),$$
  subject to $\sum_{v \in V'} p(v) \geq Q$.
- Budget Prize-Collecting Steiner Tree:
  The input to the Budget Prize-Collecting Steiner Tree problem (Budget PCST) is a tuple of undirected graph $G = (V, E)$, a vertex-valued function $p : V \to \mathbb{R}^{\geq 0}$, an edge-valued function $c : E \to \mathbb{R}^{\geq 0}$, a root vertex $r \in V$, and a budget $B \geq 0$. The objective is to find a subtree $T' = (V', E')$ that contains the root r and maximizes the following quantity:
  $$\sum_{v \in V'} p(v),$$
  subject to $\sum_{e \in E'} c(e) \leq B$.

In other words, the objective of the Quota PCST is to spend the minimum amount of cost on edges of the subtree T' that collects at least a given quota Q on the vertices, while the objective of the Budget PCST is to collect the maximum amount of prizes on vertices of the subtree T' that costs at most a given budget B on the edges.

Remark 2.2. The Quota PCST and the Budget PCST problems above are the rooted versions. The approximation algorithms for the rooted versions of Quota PCST and Budget PCST also yield approximation algorithms with the same approximation guarantees for the corresponding unrooted versions, by trying all vertices as the root and picking the best solution.

By losing a factor of n in the running time of the approximation algorithms, we will use the rooted versions and the unrooted versions of the Quota PCST and the Budget PCST interchangeably from now on.

Earlier work has studied both the Quota PCST and the Budget PCST, both of which have constant-factor approximation algorithms [15, 17, 18]. We will show how these approximation algorithms yield constant-factor approximation algorithms for the Elevated Mean problem and the k-Elevated Mean problem.

3. Algorithm

In this section, we describe two constant-factor approximation algorithms for the Elevated Mean problem and the k-Elevated Mean problem respectively. Recall the definition of bi-criteria approximation algorithms:

Definition 3.1. Algorithm $\mathcal{A}$ is an (α, β) bi-criteria approximation algorithm for the k-Elevated Mean problem if for every instance (G, p, k), $\mathcal{A}$ returns a connected induced subgraph $G[V']$ that satisfies

$$\frac{1}{\alpha} \cdot OPT \leq \frac{\sum_{v \in V'} p(v)}{\sqrt{|V'|}}$$

and that $G[V'] \in C_{\leq \beta \cdot k}$, i.e., $|V'| \leq \beta \cdot k$, where OPT is the value of the optimum solution in $C_{\leq k}$.

Definition 3.2. Algorithm $\mathcal{A}$ is an (α, β) bi-criteria approximation algorithm for the Quota PCST if for every instance (G, p, c, Q), $\mathcal{A}$ returns a subtree $T' = (V', E')$ of G that satisfies

$$\sum_{e \in E'} c(e) \leq \alpha \cdot OPT,$$

and achieves quota

$$\frac{Q}{\beta} \leq \sum_{v \in V'} p(v),$$

where OPT is the value of the optimum solution that achieves quota at least Q.

Definition 3.3. Algorithm $\mathcal{A}$ is an (α, β) bi-criteria approximation algorithm for the Budget PCST if for every instance (G, p, c, B), $\mathcal{A}$ returns a subtree $T' = (V', E')$ of G that satisfies

$$\frac{OPT}{\alpha} \leq \sum_{v \in V'} p(v),$$

and spends cost

$$\sum_{e \in E'} c(e) \leq \beta \cdot B,$$

where OPT is the value of the optimum solution that spends cost at most B.

3.1. Reduction to the Budget PCST.

In what follows, we will focus the attention on the k-Elevated Mean problem. Theorem 3.4 shows how good bi-criteria approximation algorithms for the Budget PCST yield bi-criteria approximation algorithms for the k-Elevated Mean problem with good approximation guarantees.

Theorem 3.4. *An (α, β) bi-criteria approximation algorithm $\mathcal{A}$ for the Budget PCST yields an $\left(\alpha\sqrt{\beta - \frac{\beta-1}{k}}, \beta - \frac{\beta-1}{k}\right)$ bi-criteria approximation algorithm $\mathcal{B}$ for the k-Elevated Mean problem, using k calls of $\mathcal{A}$.*

Proof. Suppose $\mathcal{A}$ is an (α, β) bi-criteria approximation algorithm for the Budget PCST. Let $(G = (V, E), p, k)$ be an instance of the k-Elevated Mean problem. We first describe the algorithm $\mathcal{B}$ for the k-Elevated Mean problem.

Let $c : E \to \mathbb{R}^{\geq 0}$ be the cost function that assigns a cost of 1 to every edge.

For budget $B = 0, 1, ..., k-1$, apply $\mathcal{A}$ on the Budget PCST instance (G, p, c, B). Let $T_B = (V_B, E_B)$ be the subtree returned by $\mathcal{A}$ in the corresponding iteration. Among all the subtrees $\{T_B\}$, return the connected induced subgraph $G[V_B]$ corresponding to the subtree that maximizes the scan statistics

$$\frac{\sum_{v \in V_B} p(v)}{\sqrt{|V_B|}}$$

as the output of $\mathcal{B}$.

Now we show that $\mathcal{B}$ achieves an approximation ratio of $\alpha\sqrt{\beta - \frac{\beta-1}{k}}$ and returns a connected induced subgraph of size $\leq \left(\beta - \frac{\beta-1}{k}\right)\cdot k$. Let $G[V_{OPT}] \in C_{\leq k}$ be the optimum connected induced subgraph of the instance (G, p, k) of the k-Elevated Mean problem. Consider the iteration $B = B_0 = |V_{OPT}| - 1$ of the algorithm $\mathcal{B}$. Let $T_{B_0} = (V_{B_0}, E_{B_0})$ be the subtree returned by $\mathcal{A}$ in this iteration. Let $T' = (V', E')$ be the optimum subtree for the Budget PCST instance (G, p, c, B_0). Then, by Definition 3.3, the following holds:

$$\sum_{v\in V_{B_0}} p(v) \geq \frac{1}{\alpha}\cdot\sum_{v\in V'} p(v), \tag{3.1}$$

$$\sum_{e\in E_{B_0}} c(e) \leq \beta\cdot B_0. \tag{3.2}$$

Since the cost function c assigns unit cost to every edge, the budget inequality (3.2) translates to

$$|V_{B_0}| - 1 = |E_{B_0}| \leq \beta\cdot B_0 = \beta\cdot(|V_{OPT}| - 1), \tag{3.3}$$

$$|V_{B_0}| \leq \left(\beta - \frac{\beta-1}{|V_{OPT}|}\right)\cdot|V_{OPT}|. \tag{3.4}$$

Note that by examining (3.2) for other iterations of $\mathcal{B}$ and obtaining an inequality similar to (3.4), we know that the connected induced subgraph returned by $\mathcal{B}$ has size $\leq \left(\beta - \frac{\beta-1}{k}\right)\cdot k$.

Since $G[V_{OPT}] \in C_{\leq k}$ is a connected induced subgraph of G, there exists a subtree $\hat{T} = (V_{OPT}, \hat{E})$ of G that spans V_{OPT}. Note that $\hat{T}$ is a solution to the Budget PCST instance (G, p, c, B_0) that satisfies the budget constraint:

$$\sum_{e\in\hat{E}} c(e) = |\hat{E}| = |V_{OPT}| - 1 = B_0.$$

Thus, we have

$$\sum_{v\in V_{OPT}} p(v) \leq \sum_{v\in V'} p(v), \tag{3.5}$$

as $T' = (V', E')$ is the optimum subtree for the Budget PCST instance (G, p, c, B_0).

From (3.1), (3.4), and (3.5), we conclude that

$$\begin{aligned} VAL &\geq \frac{\sum_{v\in V_{B_0}} p(v)}{\sqrt{|V_{B_0}|}} \\ &\geq \frac{\frac{1}{\alpha}\cdot\sum_{v\in V'} p(v)}{\sqrt{\left(\beta - \frac{\beta-1}{|V_{OPT}|}\right)\cdot|V_{OPT}|}} \\ &\geq \frac{\frac{1}{\alpha}}{\sqrt{\left(\beta - \frac{\beta-1}{k}\right)}}\cdot\frac{\sum_{v\in V_{OPT}} p(v)}{\sqrt{|V_{OPT}|}} \\ &= \frac{1}{\alpha\sqrt{\beta - \frac{\beta-1}{k}}}\cdot OPT, \end{aligned}$$

where VAL is the value of the solution returned by $\mathcal{B}$ and OPT is the value of the optimum solution to the k-Elevated Mean problem instance. □

input
An undirected graph $G = (V, E)$
A function $p : V \to \mathbb{R}^{\geq 0}$
An integer $k \in [1, n]$
An (α, β) *bi-criteria approximation algorithm* $\mathcal{A}$ *for the Budget PCST*
returns
A connected induced subgraph $H \in C_{\leq k}$ *of* G

begin
Define a function $c(e) \mapsto 1$ *for every* $e \in E$
$B \leftarrow 0$
$S \leftarrow \left\{ \arg\max_{v \in V} p(v) \right\}$
while $B \leq k - 1$
$T_B = (V_B, E_B) \leftarrow \mathcal{A}(G, p, c, B)$
if $\frac{1}{\sqrt{|V_B|}} \sum_{v \in V_B} p(v) > \frac{1}{\sqrt{|S|}} \sum_{v \in S} p(v)$ **then**
$S \leftarrow V_B$
end if
$B \leftarrow B + 1$
end while
return $G[S]$
end

FIGURE 1. Algorithm in Theorem 3.4

It is also easy to get a trade-off between the running time and the approximation ratio of the algorithm $\mathcal{B}$ in Theorem 3.4. If we want a bi-criteria approximation algorithm for the k-Elevated Mean problem using fewer calls of the bi-criteria approximation algorithm $\mathcal{A}$ for the Budget PCST without sacrificing the approximation ratio too much, Proposition 3.5 provides an algorithm under the same framework of Theorem 3.4.

Proposition 3.5. *For every $\varepsilon > 0$, an (α, β) bi-criteria approximation algorithm $\mathcal{A}$ for the Budget PCST yields an $\left(\alpha\sqrt{\beta(1+\varepsilon) - \frac{\beta(1+\varepsilon)-1}{k}}, \beta - \frac{\beta-1}{k}\right)$ bi-criteria approximation algorithm $\mathcal{B}$ for the k-Elevated Mean problem, using $O(\log_{1+\varepsilon} k)$ calls of $\mathcal{A}$.*

Proof. Essentially, algorithm $\mathcal{B}$ follows the same idea as the algorithm in Theorem 3.4 but uses a geometric search over a set of $O(\log_{1+\varepsilon} k)$ budgets instead of the set of budgets $\{0, 1, ..., k-1\}$.

For budget $B = 0, 1, (1+\varepsilon), ..., (1+\varepsilon)^i$, apply $\mathcal{A}$ on the corresponding Budget PCST instance as described in Theorem 3.4. Let $T_B = (V_B, E_B)$ be the subtree returned by $\mathcal{A}$ in the corresponding iteration. When $B \geq k-1$, set $B = k-1$ and perform the last iteration before stopping the geometric search over B, so the number of iterations is $O(\log_{1+\varepsilon} k)$. Among all the subtrees $\{T_B\}$, return the connected induced subgraph $G[V_B]$ corresponding to the subtree that maximizes the scan statistics

$$\frac{\sum_{v \in V_B} p(v)}{\sqrt{|V_B|}}$$

as the output of $\mathcal{B}$. By examining the budget inequality (3.2) for each iteration, it is not hard to see that the connected induced subgraph returned by $\mathcal{B}$ has size $\leq \left(\beta - \frac{\beta-1}{k}\right) \cdot k$.

To show that $\mathcal{B}$ achieves an approximation ratio of $\alpha\sqrt{\beta(1+\varepsilon) - \frac{\beta(1+\varepsilon)-1}{k}}$, we make one slight modification to the analysis in Theorem 3.4. Let $G[V_{OPT}] \in C_{\leq k}$ be the optimum connected induced subgraph of the instance of the k-Elevated Mean problem. Note that during the geometric search of the budget values, there is an iteration $B = B_0$ such that

$$|V_{OPT}| - 1 \leq B_0 \leq (|V_{OPT}| - 1) \cdot (1+\varepsilon). \tag{3.6}$$

Consider such iteration $B = B_0$ of the algorithm $\mathcal{B}$. Let $T_{B_0} = (V_{B_0}, E_{B_0})$ be the subtree returned by $\mathcal{A}$ in this iteration. Let $T' = (V', E')$ be the optimum subtree for the Budget PCST instance of this iteration. Then, (3.1) and (3.5) still hold, and from the budget inequality (3.2) we get

$$|V_{B_0}| - 1 = |E_{B_0}| \leq \beta \cdot B_0 \leq \beta \cdot (|V_{OPT}| - 1) \cdot (1+\varepsilon), \tag{3.7}$$

$$|V_{B_0}| \leq \left(\beta(1+\varepsilon) - \frac{\beta(1+\varepsilon)-1}{|V_{OPT}|}\right) \cdot |V_{OPT}|. \tag{3.8}$$

Combining (3.1), (3.5), and (3.7), we conclude that the algorithm $\mathcal{B}$ returns a solution of value

$$\begin{aligned}
VAL &\geq \frac{\sum_{v \in V_{B_0}} p(v)}{\sqrt{|V_{B_0}|}} \\
&\geq \frac{\frac{1}{\alpha} \cdot \sum_{v \in V'} p(v)}{\sqrt{\left(\beta(1+\varepsilon) - \frac{\beta(1+\varepsilon)-1}{|V_{OPT}|}\right) \cdot |V_{OPT}|}} \\
&\geq \frac{\frac{1}{\alpha}}{\sqrt{\left(\beta(1+\varepsilon) - \frac{\beta(1+\varepsilon)-1}{k}\right)}} \cdot \frac{\sum_{v \in V_{OPT}} p(v)}{\sqrt{|V_{OPT}|}} \\
&= \frac{1}{\alpha\sqrt{\beta(1+\varepsilon) - \frac{\beta(1+\varepsilon)-1}{k}}} \cdot OPT.
\end{aligned}$$

□

input
An undirected graph $G = (V, E)$
A function $p : V \to \mathbb{R}^{\geq 0}$
An integer $k \in [1, n]$
An (α, β) *bi-criteria approximation algorithm* $\mathcal{A}$ *for the Budget PCST*
A number $\varepsilon > 0$
returns
A connected induced subgraph $H \in C_{\leq k}$ *of* G

begin
Define a function $c(e) \mapsto 1$ *for every* $e \in E$
$B \leftarrow 1$
$S \leftarrow \left\{ \arg\max_{v \in V} p(v) \right\}$
while true
if $B \geq k - 1$ **then**
$B \leftarrow k - 1$
end if
$T_B = (V_B, E_B) \leftarrow \mathcal{A}(G, p, c, B)$
if $\frac{1}{\sqrt{|V_B|}} \sum_{v \in V_B} p(v) > \frac{1}{\sqrt{|S|}} \sum_{v \in S} p(v)$ **then**
$S \leftarrow V_B$
end if
if $B \geq k - 1$ **then**
break
end if
$B \leftarrow B \cdot (1 + \varepsilon)$
end while
return $G[S]$
end

FIGURE 2. Algorithm in Proposition 3.5

3.2. **Reduction to the Quota PCST.** Now we turn the attention to the Quota PCST. Theorem 3.6 shows how good bi-criteria approximation algorithms for the Quota PCST yield bi-criteria approximation algorithms for the k-Elevated Mean problem with good approximation guarantees.

Theorem 3.6. *For every* $\varepsilon > 0$, *an* (α, β) *bi-criteria approximation algorithm* $\mathcal{A}$ *for the Quota PCST yields a* $\left(\beta(1+\varepsilon)\sqrt{\alpha - \frac{\alpha-1}{k}}, \alpha - \frac{\alpha-1}{k}\right)$ *bi-criteria approximation algorithm* $\mathcal{B}$ *for the* k*-Elevated Mean problem, using* $O(\log_{1+\varepsilon} k)$ *calls of* $\mathcal{A}$.

Proof. Suppose $\mathcal{A}$ is an (α, β) bi-criteria approximation algorithm for the Quota PCST. Let $(G = (V, E), p, k)$ be an instance of the k-Elevated Mean problem. We first describe the algorithm $\mathcal{B}$ for the k-Elevated Mean problem.

Again, let $c : E \to \mathbb{R}^{\geq 0}$ be the function that assigns unit cost to every edge.

The algorithm $\mathcal{B}$ uses a geometric search over the quota values. Let $q = \max_{v \in V} p(v)$. For quota $Q = q, q(1+\varepsilon), ..., q(1+\varepsilon)^i$, apply $\mathcal{A}$ on the Quota PCST instance (G, p, c, Q). Let $T_Q = (V_Q, E_Q)$ be the subtree returned by $\mathcal{A}$ in the corresponding iteration. When $Q > kq$, stop the geometric search, so the number of iterations is $O(\log_{1+\varepsilon} k)$. Also, stop the geometric

search immediately when the subtree $T_Q = (V_Q, E_Q)$ returned by $\mathcal{A}$ has size

$$|V_Q| > \left(\alpha - \frac{\alpha - 1}{k}\right) \cdot k. \tag{3.9}$$

Note that this extra stopping condition of the geometric search guarantees that the connected induced subgraph returned by $\mathcal{B}$ has size $\leq \left(\alpha - \frac{\alpha-1}{k}\right) \cdot k$. Among all the subtrees $\{T_Q\}$, return the connected induced subgraph $G[V_Q]$ corresponding to the subtree that maximizes the scan statistics

$$\frac{\sum_{v \in V_Q} p(v)}{\sqrt{|V_Q|}}$$

as the output of $\mathcal{B}$.

Now we show that $\mathcal{B}$ achieves an approximation ratio of $\beta(1+\varepsilon)\sqrt{\alpha - \frac{\alpha-1}{k}}$. Let $G[V_{OPT}] \in C_{\leq k}$ be the optimum connected induced subgraph of the instance (G, p, k) of the k-Elevated Mean problem. Let $\hat{T} = (V_{OPT}, \hat{E})$ be a subtree of G that spans V_{OPT}. Note that the following holds:

$$q = \max_{v \in V} p(v) \leq \sum_{v \in V_{OPT}} p(v) \leq k \cdot \max_{v \in V} p(v) = kq. \tag{3.10}$$

Moreover, the extra stopping condition (3.9) of $\mathcal{B}$ is never met during the iterations when $Q \leq \sum_{v \in V_{OPT}} p(v)$. Because $\hat{T} = (V_{OPT}, \hat{E})$ is a solution to the Quota PCST instance (G, p, c, Q) for $Q \leq \sum_{v \in V_{OPT}} p(v)$ and has value

$$\sum_{e \in \hat{E}} c(e) = |\hat{E}| = |V_{OPT}| - 1,$$

the subtree $T_Q = (V_Q, E_Q)$ returned by the (α, β) bi-criteria approximation algorithm $\mathcal{A}$ on the corresponding Quota PCST instance, by Definition 3.2, has value

$$\sum_{e \in E_Q} c(e) \leq \alpha \cdot \sum_{e \in \hat{E}} c(e) = \alpha \cdot (|V_{OPT}| - 1),$$

so the size of T_Q is bounded by

$$\begin{aligned}
|V_Q| &= |E_Q| + 1 \\
&= \left(\sum_{e \in E_Q} c(e)\right) + 1 \\
&\leq \alpha \cdot (|V_{OPT}| - 1) + 1 \\
&= \left(\alpha - \frac{\alpha - 1}{|V_{OPT}|}\right) \cdot |V_{OPT}| \\
&\leq \left(\alpha - \frac{\alpha - 1}{k}\right) \cdot k,
\end{aligned}$$

using that c is the unit cost function on edges. This shows that the stopping condition (3.9) is not satisfied when $Q \leq \sum_{v \in V_{OPT}} p(v)$.

From (3.10) and the observation that the stopping condition (3.9) is never met when $Q \leq \sum_{v \in V_{OPT} p(v)}$, we know that there is an iteration $Q = Q_0$ during the geometric search over the quota values such that

$$\frac{1}{1+\varepsilon} \cdot \left(\sum_{v \in V_{OPT}} p(v)\right) \leq Q_0 \leq \sum_{v \in V_{OPT}} p(v). \tag{3.11}$$

Consider such iteration $Q = Q_0$ of the algorithm $\mathcal{B}$. Let $T_{Q_0} = (V_{Q_0}, E_{Q_0})$ be the subtree returned by $\mathcal{A}$ in this iteration. Let $T' = (V', E')$ be the optimum subtree for the Quota PCST

instance (G, p, c, Q_0). Then, by Definition 3.2, the following holds:

$$\sum_{e \in E_{Q_0}} c(e) \leq \alpha \cdot \sum_{e \in E'} c(e), \tag{3.12}$$

$$\sum_{v \in V_{Q_0}} p(v) \geq \frac{1}{\beta} \cdot Q_0. \tag{3.13}$$

Again, since c is the unit cost function, (3.12) translates to

$$|V_{Q_0}| - 1 = |E_{Q_0}| \leq \alpha \cdot |E'| = \alpha \cdot (|V'| - 1), \tag{3.14}$$

$$|V_{Q_0}| \leq \left(\alpha - \frac{\alpha - 1}{|V'|}\right) \cdot |V'|. \tag{3.15}$$

Recall that $\hat{T} = (V_{OPT}, \hat{E})$ is a subtree of G that spans V_{OPT}. Note that $\hat{T}$ is a solution to the Quota PCST instance (G, p, c, Q_0) that satisfies the quota constraint:

$$\sum_{v \in V_{OPT}} p(v) \geq Q_0.$$

Thus, we have

$$\sum_{e \in \hat{E}} c(e) \geq \sum_{e \in E'} c(e), \tag{3.16}$$

as $T' = (V', E')$ is the optimum subtree for the Quota PCST instance (G, p, c, Q_0). Since c is the unit cost function, (3.16) is equivalent to

$$|V_{OPT}| \geq |V'|. \tag{3.17}$$

From (3.11), (3.13), (3.15), and (3.17), we conclude that

$$\begin{aligned}
VAL &\geq \frac{\sum_{v \in V_{Q_0}} p(v)}{\sqrt{|V_{Q_0}|}} \\
&\geq \frac{\frac{1}{\beta} \cdot Q_0}{\sqrt{\left(\alpha - \frac{\alpha-1}{|V'|}\right) \cdot |V'|}} \\
&\geq \frac{\frac{1}{\beta} \cdot \frac{1}{1+\varepsilon} \cdot \left(\sum_{v \in V_{OPT}} p(v)\right)}{\sqrt{\left(\alpha - \frac{\alpha-1}{|V_{OPT}|}\right) \cdot |V_{OPT}|}} \\
&\geq \frac{\frac{1}{\beta} \cdot \frac{1}{1+\varepsilon}}{\sqrt{\alpha - \frac{\alpha-1}{k}}} \cdot \frac{\sum_{v \in V_{OPT}} p(v)}{\sqrt{|V_{OPT}|}} \\
&= \frac{1}{\beta(1+\varepsilon)\sqrt{\alpha - \frac{\alpha-1}{k}}} \cdot OPT.
\end{aligned}$$

□

Remark 3.7. So far, Theorem 3.4, Proposition 3.5, and Theorem 3.6 only show bi-criteria approximation algorithms for the k-Elevated Mean problem. If instead we do not want bi-criteria approximation algorithms, observe that Theorem 3.4 and Proposition 3.5 also produce approximation algorithms for the k-Elevated Mean problem using approximation algorithms for the Budget PCST.

In fact, the algorithms $\mathcal{B}$ in Theorem 3.4, Proposition 3.5, and Theorem 3.6 are bi-criteria approximation algorithms because the returned subgraph could potentially come from a larger class than $C_{\leq k}$. However, for the Elevated Mean problem whose objective is optimized over the class C of all connected induced subgraphs, this concern is no longer an issue. Therefore, we

input
An undirected graph $G = (V, E)$
A function $p : V \to \mathbb{R}^{\geq 0}$
An integer $k \in [1, n]$
An (α, β) *bi-criteria approximation algorithm* $\mathcal{A}$ *for the Quota PCST*
A number $\varepsilon > 0$
returns
A connected induced subgraph $H \in C_{\leq k}$ *of* G

begin
Define a function $c(e) \mapsto 1$ *for every* $e \in E$
$q \leftarrow \max_{v \in V} p(v)$
$Q \leftarrow q$
$S \leftarrow \left\{ \arg\max_{v \in V} p(v) \right\}$
while $Q \leq kq$
$T_Q = (V_Q, E_Q) \leftarrow \mathcal{A}(G, p, c, Q)$
if $|V_Q| > \left(\alpha - \frac{\alpha - 1}{k}\right) \cdot k$ **then**
break
end if
if $\frac{1}{\sqrt{|V_Q|}} \sum_{v \in V_Q} p(v) > \frac{1}{\sqrt{|S|}} \sum_{v \in S} p(v)$ **then**
$S \leftarrow V_Q$
end if
$Q \leftarrow Q \cdot (1 + \varepsilon)$
end while
return $G[S]$
end

FIGURE 3. Algorithm in Theorem 3.6

see that Theorem 3.4, Proposition 3.5, and Theorem 3.6 produce approximation algorithms $\mathcal{B}$ for the Elevated Mean problem, even if the provided algorithms $\mathcal{A}$ are bi-criteria approximation algorithms.

The observations in Remark 3.7 are summarized in Corollary 3.8 and Corollary 3.9.

Corollary 3.8. *Let* $\mathcal{A}$ *be an* α*-approximation algorithm for the Budget PCST. Then, there is*

- *an* α*-approximation algorithm for the* k*-Elevated Mean problem, using* k *calls of* $\mathcal{A}$*, given by Theorem 3.4.*
- *for every* $\varepsilon > 0$*, an* $\left(\alpha\sqrt{(1+\varepsilon) - \frac{\varepsilon}{k}}\right)$*-approximation algorithm for the* k*-Elevated Mean problem, using* $O(\log_{1+\varepsilon} k)$ *calls of* $\mathcal{A}$*, given by Proposition 3.5.*

Corollary 3.9. *Let* $\mathcal{A}_1$ *be an* (α_1, β_1) *bi-criteria approximation algorithm for the Budget PCST. Let* $\mathcal{A}_2$ *be an* (α_2, β_2) *bi-criteria approximation algorithm for the Quota PCST. Then, there is*

- *an* $\left(\alpha_1\sqrt{\beta_1 - \frac{\beta_1 - 1}{n}}\right)$*-approximation algorithm for the Elevated Mean problem, using* n *calls of* $\mathcal{A}_1$*, given by Theorem 3.4.*
- *for every* $\varepsilon > 0$*, an* $\left(\alpha_1\sqrt{\beta_1(1+\varepsilon) - \frac{\beta_1(1+\varepsilon) - 1}{n}}\right)$*-approximation algorithm for the Elevated Mean problem, using* $O(\log_{1+\varepsilon} n)$ *calls of* $\mathcal{A}_1$*, given by Proposition 3.5.*

- *for every $\varepsilon > 0$, a $\left(\beta_2(1+\varepsilon)\sqrt{\alpha_2 - \frac{\alpha_2 - 1}{n}}\right)$-approximation algorithm for the Elevated Mean problem, using $O(\log_{1+\varepsilon} n)$ calls of $\mathcal{A}_2$, given by Theorem 3.6.*

Remark 3.10. While obtaining an approximation algorithm for the Elevated Mean problem using an approximation algorithm for the Budget PCST may seem more intuitive, to the best of our knowledge, the current constant-factor approximation algorithms for the Budget PCST [15, 17] are obtained using the approximation algorithms for the Quota PCST as subroutines, and the approximation ratios of the resulting algorithms for the Budget PCST are worse than the approximation ratios of the original algorithms for the Quota PCST. Therefore, we look at algorithms for both the Budget PCST and the Quota PCST, and allow bi-criteria approximation algorithms that may provide better approximation ratios for the Elevated Mean problem.

Finally, we present two constant-factor approximation algorithms for the Elevated Mean problem and the k-Elevated Mean problem respectively.

Levin [17] showed a $(4+\varepsilon)$-approximation algorithm for the Budget PCST. Together with Corollary 3.8, we get a $(4+\varepsilon)$-approximation algorithm for the k-Elevated Mean problem

Johnson, Minkoff, and Phillips [15] noted that given an approximation algorithm for the Quota PCST with approximation ratio ≤ 2, there is a $(3+\varepsilon)$-approximation algorithm for the Budget PCST. They also showed that any approximation algorithm for the k-MST using the primal-dual schema of Goemans & Williamson [14] yield an approximation to the Quota PCST with the same approximation ratio. Garg [13] obtained a 2-approximation to the k-MST, which in turn gave a 2-approximation to the Quota PCST.

Therefore, using a $(3+\varepsilon)$-approximation algorithm for the Budget PCST, by Corollary 3.8, we have

Theorem 3.11. *There is a $(3+\varepsilon)$-approximation algorithm for the k-Elevated Mean problem for every $\varepsilon > 0$.*

Using a 2-approximation algorithm for the Quota PCST, by Corollary 3.9, we have

Theorem 3.12. *There is a $(1+\varepsilon)\sqrt{2-\frac{1}{n}}$-approximation algorithm for the Elevated Mean problem for every $\varepsilon > 0$. In particular, by choosing $\varepsilon = \frac{1}{4n}$, we get a $\sqrt{2}$-approximation algorithm for the Elevated Mean problem.*

Alternatively, we can also obtain Theorem 3.12 using a $(1+\varepsilon, 2)$ bi-criteria approximation algorithm for the Budget PCST, which can be easily obtained by applying a 2-approximation algorithm for the Quota PCST $O(\log_{1+\varepsilon} n)$ times [17]. However, with this black-box reduction from Budget PCST to Quota PCST, the resulting $\sqrt{2}$-approximation algorithm for the Elevated Mean problem has worse running time than the one that uses a 2-approximation to the Quota PCST.

4. Anomaly Detection

In this section, we formalize the anomaly detection problem as a hypothesis testing problem, and present the decision rule T_{EM} based on maximizing the Elevated Mean scan statistics (2.1).

4.1. Anomaly Detection as Hypothesis Testing.

Let $G = (V, E)$ be an undirected graph. For each vertex $v \in V$, there is an associated observation x_v, which is a random variable. We are interested in distinguishing the two hypotheses below:

- Under the null hypothesis $\mathcal{H}_0$, the observations x_v follow i.i.d. Gaussian $\mathcal{N}(0,1)$.
- Under the alternative hypothesis $\mathcal{H}_1$, there is a connected set of vertices $S \subset V$ that is the anomalous cluster, and the observations x_v independently follow

$$x_v \sim \mu \cdot \mathbf{1}_S(v) + \mathcal{N}(0,1),$$

for some $\mu > 0$.

Beside the Gaussian model (network intrusion) above, we also consider the variant of Poisson model (disease outbreak) in the anomaly detection problem:

- Under the null hypothesis $\mathcal{H}_0$, the observations x_v follow i.i.d. Poisson $Pois(1)$.
- Under the alternative hypothesis $\mathcal{H}_1$, there is a connected set of vertices $S \subset V$ that is the anomalous cluster, and the observations x_v independently follow

$$x_v \sim Pois(1 + \mu \cdot \mathbf{1}_S(v)),$$

for some $\mu > 0$.

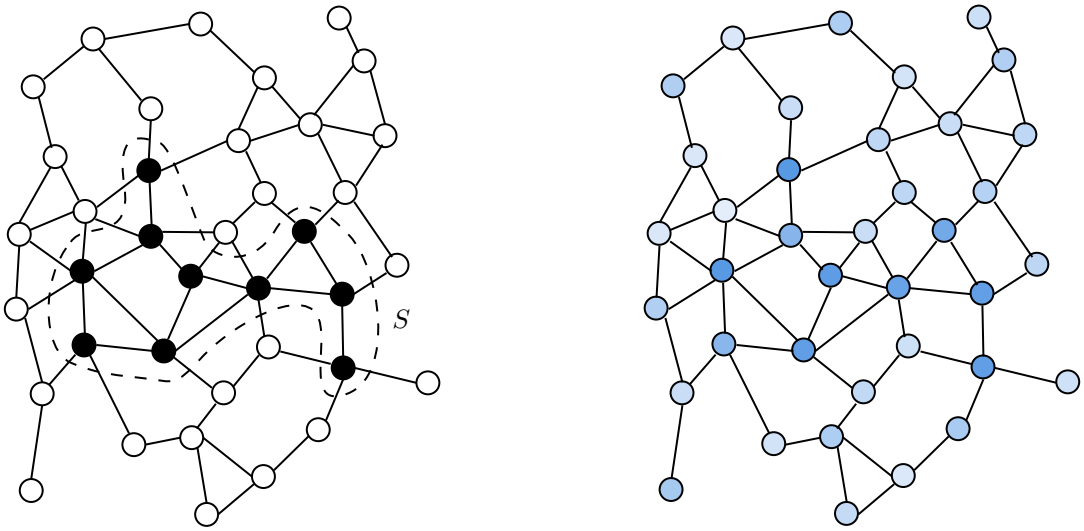

FIGURE 4. Left: the ground truth of the anomalous cluster S in the network. Right: the observations of the vertices of the network.

We formalize the anomaly detection problem as follows. The anomaly detection problem is a promise problem, with input being a tuple of an undirected graph $G = (V, E)$, a set of observations $x_v \in \mathbb{R}$ for the vertices, a parmeter $k \in \mathbb{N}$ and two hidden parameters $s \in \mathbb{N}$ and $\mu \in \mathbb{R}^{\geq 0}$ that are not visible to the decision rule. It is guaranteed that the observations x_v follow either the null hypothesis $\mathcal{H}_0$ or the alternative hypothesis $\mathcal{H}_1$. If the observations x_v follow $\mathcal{H}_1$, it is guaranteed that the anomalous cluster $S \subset V$ forms a connected induced subgraph $G[S] \in C_{\leq k}$, has size at least $|S| \geq s$, and has parameter μ as its elevated mean in $\mathcal{H}_1$.

The goal of the anomaly detection problem is to design a rule $\pi_G : \mathbb{R}^V \times \mathbb{N} \to \{0, 1\}$, such that given the parameter k, π_G maps the observations from $\mathcal{H}_0$ to 0 with high probability, and the observations from $\mathcal{H}_1$ to 1 with high probability. Let $\mathcal{H}_S$ denote the alternative hypothesis $\mathcal{H}_1$ in which the connected set $S \subset V$ is the anomalous cluster. We measure the risk [3] using

$$R(\pi_G) = \mathbb{P}_{\mathcal{H}_0}\left(\pi_G(\{x_v\}, k) = 1\right) + \max_{\substack{S \subset V : |S| \geq s, \\ G[S] \in C_{\leq k}}} \mathbb{P}_{\mathcal{H}_S}\left(\pi_G(\{x_v\}, k) = 0\right),$$

which combines both Type I and Type II errors.

Definition 4.1. A sequence of instances of the anomaly detection problem (G_n, k_n, s_n, μ_n) is δ-separable, if there exists a rule π_{G_n} for every instance in this sequence, such that for all sufficiently large n,

$$R(\pi_{G_n}) \leq \delta.$$

Definition 4.2. A sequence of instances of the anomaly detection problem (G_n, k_n, s_n, μ_n) is asymptotically separable, if there exists a rule π_{G_n} for every instance in this sequence, such that

$$R(\pi_{G_n}) = o_n(1).$$

In what follows, we are interested in the cases of the anomaly detection problem where $k_n, s_n = \omega(1)$.

4.2. **Decision Rule.** Now, we describe the decision rule following the generalized likelihood ratio test schema [3] based on the Elevated Mean scan statistics.

Given the input graph $G = (V, E)$, the set of observations x_v, and the parameter k, we compute a threshold θ. Then, we create a vertex-valued function $p : V \to \mathbb{R}^{\geq 0}$ by truncating the observations x_v into nonnegative values

$$p(v) = \begin{cases} x_v & \text{if } x_v \geq 0 \\ 0 & \text{if } x_v < 0 \end{cases},$$

and feed the resulting instance (G, p, k) of the k-Elevated Mean problem to the constant-factor approximation algorithm obtained in Section 3. Let VAL be the value of the solution returned by the approximation algorithm. We accept the null hypothesis $\mathcal{H}_0$ if $VAL < \theta$, and accept the alternative hypothesis $\mathcal{H}_1$ otherwise. Note that this is a polynomial time decision rule if the threshold value θ can be efficiently computed.

Moreover, this decision rule based on the Elevated Mean scan statistics

$$\frac{\sum_{v \in V(H)} p(v)}{\sqrt{|V(H)|}}$$

can be shown to correspond to the generalized likelihood ratio test [3] for the anomaly detection problem in the Gaussian model under some restrictions [3].

We will refer to this decision rule as Testing via Elevated Mean (T_{EM}).

5. Analysis of Separability

In this section, we will explore the detection power of the decision rule T_{EM}. The analysis involves two steps: to lower bound the expected optimum value of the Elevated Mean scan statistics for the alternative hypothesis $\mathcal{H}_1$, and to upper bound the expected optimum value of the Elevated Mean scan statistics for the null hypothesis $\mathcal{H}_0$. When there is a significant gap between the two expectations, we can show separability results.

5.1. **Lower Bound for the Alternative Hypothesis.** If the observations x_v follow the alternative hypothesis $\mathcal{H}_1$, we can use the ground truth of the anomalous cluster S to construct a lower bound for the expectation of the optimum scan statistics.

$$\begin{aligned} \mathbb{E}_{\mathcal{H}_1}\left[\frac{\sum_{v \in S} p(v)}{\sqrt{|S|}}\right] &\geq \mathbb{E}_{\mathcal{H}_1}\left[\frac{\sum_{v \in S} x_v}{\sqrt{|S|}}\right] \\ &= \begin{cases} \sqrt{|S|} \cdot \mu & \text{(Gaussian model)} \\ \sqrt{|S|} \cdot (1+\mu) & \text{(Poisson model)} \end{cases}. \end{aligned}$$

In fact, by central limit theorem, we know that as $|S| \to \infty$,

$$\frac{\sum_{v \in S} x_v}{\sqrt{|S|}} \xrightarrow{d} \begin{cases} \mathcal{N}\left(\sqrt{|S|} \cdot \mu, 1\right) & \text{(Gaussian model)} \\ \mathcal{N}\left(\sqrt{|S|} \cdot (1+\mu), 1+\mu\right) & \text{(Poisson model)} \end{cases},$$

and by concentration bound, with high probability we have

$$\begin{aligned} \frac{\sum_{v \in S} p(v)}{\sqrt{|S|}} &\geq \sqrt{|S|} \cdot \mu(1 - o(1)) \\ &\geq \sqrt{s} \cdot \mu(1 - o(1)). \end{aligned}$$

Since the decision rule T_{EM} uses the constant factor approximation for the k-Elevated Mean problem in Section 3, with high probability the value of the solution returned by the algorithm is lower bounded by

$$VAL \geq \frac{1}{\alpha} \cdot \sqrt{s} \cdot \mu(1 - o(1)),$$

where α is the approximation ratio.

5.2. **Upper Bound for the Null Hypothesis.** Now, we show an upper bound for the expectation of optimum scan statistics if the observations x_v follow the null hypothesis $\mathcal{H}_0$.

Let G be an undirected graph. Let C_k denote

$$C_k = \{\textit{connected induced subgraphs of } G \textit{ of size} = k\}.$$

The following is a well-known bound [7] on the number of connected induced subgraphs of size k for graphs of maximum degree $\leq d$:

$$|C_k| \leq n \cdot (e(d-1))^{k-1} \tag{5.1}$$

In what follows, we will use d to denote the maximum degree of the graph G. We will also assume the nontrivial cases when $d \geq 3$, as the analysis for smaller d is easier to go through.

Define the optimum Elevated Mean scan statistics as

$$OPT = \max_{H \in C_{\leq k}} \frac{\sum_{v \in V(H)} p(v)}{\sqrt{|V(H)|}}.$$

We want to upper bound the expectation of the optimum Elevated Mean scan statistics OPT under the null hypothesis $\mathcal{H}_0$.

5.2.1. *Gaussian model.* First, we prove a bound for the Gaussian model.

Theorem 5.1. *In the Gaussian model, under the null hypothesis $\mathcal{H}_0$, we have*

$$\mathbb{E}_{\mathcal{H}_0}[OPT] \leq \sqrt{2\log n + O(k \log d)}.$$

Proof. Consider a connected induced subgraph $H \in C_{\leq k}$. Note that the moment generating function of the scan statistics $\frac{1}{\sqrt{|V(H)|}} \cdot \sum_{v \in V(H)} p(v)$ is

$$M_H(t) = \mathbb{E}_{\mathcal{H}_0}\left[\exp\left(t \cdot \frac{1}{\sqrt{|V(H)|}} \cdot \sum_{v \in V(H)} p(v)\right)\right] \tag{5.2}$$

$$= \mathbb{E}_{\mathcal{H}_0}\left[\exp\left(\frac{t}{\sqrt{|V(H)|}} \cdot p(v)\right)\right]^{|V(H)|} \tag{5.3}$$

$$= M_p\left(\frac{t}{\sqrt{|V(H)|}}\right)^{|V(H)|}, \tag{5.4}$$

where $M_p(t)$ is the moment generating function of $p(v)$. Let $M_x(t)$ be the moment generating function of the observations x_v. Then, we can bound $M_p(t)$ using $M_x(t)$:

$$M_p(t) = \mathbb{E}_{\mathcal{H}_0}[\exp(t \cdot p(v))] \tag{5.5}$$

$$\leq \mathbb{E}_{\mathcal{H}_0}[\exp(t \cdot x_v)] + \mathbb{P}_{\mathcal{H}_0}(x_v < 0) \cdot \exp(t \cdot 0) \tag{5.6}$$

$$= M_x(t) + \frac{1}{2} \tag{5.7}$$

$$\leq \frac{3}{2} \cdot M_x(t) \tag{5.8}$$

$$= \frac{3}{2} \cdot \exp\left(\frac{t^2}{2}\right), \tag{5.9}$$

as $p(v) = \begin{cases} x_v & \text{if } x_v \geq 0 \\ 0 & \text{if } x_v < 0 \end{cases}$, and the probability that $x_v < 0$ is $\frac{1}{2}$ for i.i.d. $x_v \sim \mathcal{N}(0,1)$.

Therefore, plugging (5.9) into (5.4), we get

$$\begin{aligned} M_H(t) &= M_p\left(\frac{t}{\sqrt{|V(H)|}}\right)^{|V(H)|} \\ &\leq \left[\frac{3}{2}\cdot\exp\left(\frac{t^2}{2|V(H)|}\right)\right]^{|V(H)|} \\ &= \left(\frac{3}{2}\right)^{|V(H)|}\cdot\exp\left(\frac{t^2}{2}\right). \end{aligned}$$

Next, we use the folklore technique for bounding the expected maximum of a set of random variables. For any $t > 0$, we have

$$\begin{aligned} \exp\left(t\cdot\mathbb{E}_{\mathcal{H}_0}\left[OPT\right]\right) &\leq \mathbb{E}_{\mathcal{H}_0}\left[\exp\left(t\cdot OPT\right)\right] \\ &= \mathbb{E}_{\mathcal{H}_0}\left[\max_{H\in C_{\leq k}}\exp\left(t\cdot\frac{1}{\sqrt{|V(H)|}}\cdot\sum_{v\in V(H)}p(v)\right)\right] \\ &\leq \mathbb{E}_{\mathcal{H}_0}\left[\sum_{H\in C_{\leq k}}\exp\left(t\cdot\frac{1}{\sqrt{|V(H)|}}\cdot\sum_{v\in V(H)}p(v)\right)\right] \\ &\leq \sum_{H\in C_{\leq k}}\left(\frac{3}{2}\right)^{|V(H)|}\cdot\exp\left(\frac{t^2}{2}\right) \\ &\leq |C_{\leq k}|\cdot\left(\frac{3}{2}\right)^{k}\cdot\exp\left(\frac{t^2}{2}\right), \end{aligned}$$

where the first inequality is Jensen's inequality. Rewriting the inequality above and plugging in the bound (5.1), we get

$$\begin{aligned} \mathbb{E}_{\mathcal{H}_0}\left[OPT\right] &\leq \frac{\log|C_{\leq k}| + k\log\frac{3}{2} + \frac{t^2}{2}}{t} \\ &= \frac{\log\left(\sum_{i=1}^{k}|C_i|\right) + k\log\frac{3}{2}}{t} + \frac{t}{2} \\ &\leq \frac{\log\left(\sum_{i=1}^{k} n\cdot(e(d-1))^{i-1}\right) + k\log\frac{3}{2}}{t} + \frac{t}{2} \\ &\leq \frac{\log n + O(k\log d)}{t} + \frac{t}{2}. \end{aligned}$$

Setting $t = \sqrt{2}\cdot\sqrt{\log n + O(k\log d)}$, we get

$$\mathbb{E}_{\mathcal{H}_0}\left[OPT\right] \leq \sqrt{2\log n + O(k\log d)}.$$

□

5.2.2. *Poisson model.* Then, we prove a bound for the Poisson model. For the sake of the analysis, we define the following quantities:

$$Y_r = \max_{H\in C_r}\sum_{v\in V(H)}p(v), \tag{5.10}$$

$$k_i = \frac{k}{(1+\varepsilon)^i}, \tag{5.11}$$

$$Z = \max_{i\geq 0}\frac{1}{\sqrt{k_i}}Y_{k_i}, \tag{5.12}$$

where $\varepsilon > 0$ is a parameter that we will choose later. We would like to bound the expected optimum scan statistics OPT for the null hypothesis $\mathcal{H}_0$. We do so by first upper bounding the

expectation of Y_{k_i}, which in turn upper bounds the expectation of Z, and then upper bounds the expectation of OPT.

First, we bound the expected Y_r.

Lemma 5.2. *In the Poisson model, under the null hypothesis $\mathcal{H}_0$, we have*

$$\mathbb{E}_{\mathcal{H}_0}[Y_r] \leq \frac{\log n + O(r \log d)}{\log \log d}.$$

Proof. In the Poisson model, we have

$$p(v) = x_v \sim Pois(1),$$

so the moment generating function of $p(v)$ is $M(t) = \exp(e^t - 1)$. Again, using the same technique, for any $t > 0$, we have

$$\begin{aligned}
\exp\left(t \cdot \mathbb{E}_{\mathcal{H}_0}[Y_r]\right) &\leq \mathbb{E}_{\mathcal{H}_0}\left[\exp\left(t \cdot Y_r\right)\right] \\
&= \mathbb{E}_{\mathcal{H}_0}\left[\max_{H \in C_r} \exp\left(t \cdot \sum_{v \in V(H)} p(v)\right)\right] \\
&\leq \mathbb{E}_{\mathcal{H}_0}\left[\sum_{H \in C_r} \exp\left(t \cdot \sum_{v \in V(H)} p(v)\right)\right] \\
&= |C_r| \cdot \left(\exp\left(e^t - 1\right)\right)^r.
\end{aligned}$$

Rewriting the inequality above and plugging in the bound (5.1), we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}[Y_r] &\leq \frac{\log |C_r| + r(e^t - 1)}{t} \\
&\leq \frac{\log n + (r-1)\log(e(d-1)) + r(e^t - 1)}{t} \\
&\leq \frac{\log n + O(r \log d) + r(e^t - 1)}{t}.
\end{aligned}$$

Setting $t = \log \log d$, we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}[Y_r] &\leq \frac{\log n + O(r \log d) + r(\log d - 1)}{\log \log d} \\
&\leq \frac{\log n + O(r \log d)}{\log \log d}.
\end{aligned}$$

□

Next, we bound the expected value of Z using the bound on the expected values of Y_{k_i} in Lemma 5.2.

Lemma 5.3. *In the Poisson model, under the null hypothesis $\mathcal{H}_0$, the following is true for $\varepsilon = 3$ in the definition (5.11)*

$$\mathbb{E}_{\mathcal{H}_0}[Z] \leq \frac{\log n + O\left(\sqrt{k} \cdot \log d\right)}{\log \log d}.$$

Proof. First, we bound the maximum using the sum:

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}[Z] &= \mathbb{E}_{\mathcal{H}_0}\left[\max_{i \geq 0} \frac{1}{\sqrt{k_i}} Y_{k_i}\right] \\
&\leq \mathbb{E}_{\mathcal{H}_0}\left[\sum_{i \geq 0} \frac{1}{\sqrt{k_i}} Y_{k_i}\right] \\
&= \sum_{i \geq 0} \frac{1}{\sqrt{k_i}} \cdot \mathbb{E}_{\mathcal{H}_0}\left[Y_{k_i}\right].
\end{aligned}$$

By Lemma 5.2, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}[Z] &\leq \sum_{i\geq 0} \frac{1}{\sqrt{k_i}} \cdot \left(\frac{\log n + O(r \log d)}{\log \log d} \right) \\
&\leq \frac{\sum_{i\geq 0} \left(\frac{\log n}{\sqrt{k_i}} + O\left(\sqrt{k_i} \cdot \log d\right) \right)}{\log \log d} \\
&= \frac{\left(\sum_{i\geq 0} \sqrt{(1+\varepsilon)^i} \cdot \frac{\log n}{\sqrt{k}} \right) + \left(\sum_{i\geq 0} \frac{1}{\sqrt{(1+\varepsilon)^i}} \cdot O\left(\sqrt{k} \cdot \log d\right) \right)}{\log \log d} \\
&\leq \frac{1}{\log \log d} \cdot \left[\frac{\sqrt{k}-1}{\sqrt{1+\varepsilon}-1} \cdot \frac{\log n}{\sqrt{k}} + \frac{1-\frac{1}{\sqrt{k}}}{1-\frac{1}{\sqrt{1+\varepsilon}}} \cdot O\left(\sqrt{k} \cdot \log d\right) \right]
\end{aligned}$$

Setting $\varepsilon = 3$, we get

$$\mathbb{E}_{\mathcal{H}_0}[Z] \leq \frac{\log n + O\left(\sqrt{k} \cdot \log d\right)}{\log \log d}.$$

□

Finally, observe the inequality between the expectation of OPT and the expectation of Z in the following lemma:

Lemma 5.4. *Under the null hypothesis $\mathcal{H}_0$,*

$$\mathbb{E}_{\mathcal{H}_0}[OPT] \leq \sqrt{1+\varepsilon} \cdot \mathbb{E}_{\mathcal{H}_0}[Z].$$

Proof. We will actually show

$$OPT \leq \sqrt{1+\varepsilon} \cdot Z.$$

Let $G[V_{OPT}] \in C_{\leq k}$ be the optimum induced subgraph that maximizes the Elevated Mean scan statistics. By the definition (5.11) of k_i, we know there exists an index j such that

$$k_{j+1} = \frac{k}{(1+\varepsilon)^{j+1}} < |V_{OPT}| \leq \frac{k}{(1+\varepsilon)^j} = k_j, \tag{5.13}$$

so $G[V_{OPT}] \in C_{\leq k_j}$. Therefore, we can lower bound Y_{k_j} using the induced subgraph $G[V_{OPT}]$:

$$Y_{k_j} = \max_{H \in C_{k_j}} \sum_{v \in V(H)} p(v) \tag{5.14}$$

$$= \max_{H \in C_{\leq k_j}} \sum_{v \in V(H)} p(v) \tag{5.15}$$

$$\geq \sum_{v \in V_{OPT}} p(v). \tag{5.16}$$

Recall the definition of Z in (5.12):

$$Z = \max_{i \geq 0} \frac{1}{\sqrt{k_i}} Y_{k_i}.$$

Combining the inequalities (5.13) and (5.16), we get

$$\begin{aligned} OPT &= \frac{\sum_{v\in V_{OPT}} p(v)}{\sqrt{|V_{OPT}|}} \\ &\le \frac{Y_{k_j}}{\sqrt{k_{j+1}}} \\ &= \sqrt{1+\varepsilon}\cdot\frac{Y_{k_j}}{\sqrt{k_j}} \\ &\le \sqrt{1+\varepsilon}\cdot Z. \end{aligned}$$

□

Now, we are ready to state the upper bound for the expected optimum Elevated Mean scan statistics OPT under $\mathcal{H}_0$ for the Poisson model. Combining Lemma 5.3 and Lemma 5.4, we get the following theorem:

Theorem 5.5. *In the Poisson model, under the null hypothesis $\mathcal{H}_0$, we have*

$$\mathbb{E}_{\mathcal{H}_0}[OPT] \le \frac{2\log n + O\left(\sqrt{k}\cdot\log d\right)}{\log\log d}.$$

Remark 5.6. Note that this analysis of the upper bound for the null hypothesis is independent of the decision rule. The only role of the decision rule T_{EM} in this analysis is to make sure that under the alternative hypothesis $\mathcal{H}_1$, the value of the solution returned by the algorithm is off by at most a constant from the lower bound for the alternative hypothesis.

5.3. **Separability Results.** Now, we show some separability results for the anomaly detection problem following the previous analysis of the lower bound for the alternative hypothesis and the upper bound for the null hypothesis.

In what follows, we will use d_n to denote the maximum degree of the graph G_n.

Theorem 5.7. *There exists a constant c such that the following is true.*

In the Gaussian model, a sequence of instances (G_n, k_n, s_n, μ_n) of the anomaly detection problem is asymptotically separable if

$$\mu_n \ge c\cdot\frac{\sqrt{\log n + k_n\log d_n}}{\sqrt{s_n}},$$

and moreover, T_{EM} is a polynomial time decision rule that asymptotically separates $\mathcal{H}_0$ and $\mathcal{H}_1$ if the above inequality holds.

Proof. By Theorem 5.1, the expected optimum scan statistics for the null hypothesis $\mathcal{H}_0$ is upper bounded by

$$\mathbb{E}_{\mathcal{H}_0}[OPT_n] \le \sqrt{2\log n + O(k_n\log d_n)}.$$

Here we will show that the same asymptotic bound holds for OPT_n with high probability. Note that for each subgraph $H \in C_{\le k_n}$, the associated scan statistics is upper bounded by

$$\begin{aligned} \frac{\sum_{v\in V(H)} p(v)}{\sqrt{V(H)}} &\le \frac{\sum_{v\in V(H)} |x_v|}{\sqrt{V(H)}} \\ &= \max_{(\xi_v)\in\{\pm1\}^{V(H)}} \frac{\sum_{v\in V(H)}\xi_v\cdot x_v}{\sqrt{V(H)}}, \end{aligned}$$

as $p(v) = \begin{cases} x_v & \text{if } x_v \ge 0 \\ 0 & \text{if } x_v < 0 \end{cases}$. Since x_v follow i.i.d. $\mathcal{N}(0,1)$ under $\mathcal{H}_0$, each term $\frac{\sum_{v\in V(H)}\xi_v\cdot x_v}{\sqrt{V(H)}}$ in the maximum above also follows $\mathcal{N}(0,1)$, and there are at most 2^{k_n} terms. As a result, the

optimum scan statistics $OPT_n|_{\mathcal{H}_0}$ under the null hypothesis is upper bounded by

$$OPT_n|_{\mathcal{H}_0} = \max_{H \in C_{\leq k_n}} \frac{\sum_{v \in V(H)} p(v)}{\sqrt{|V(H)|}} \tag{5.17}$$

$$\leq \max_{H \in C_{\leq k_n}} \left(\max_{(\xi_v) \in \{\pm 1\}^{V(H)}} \frac{\sum_{v \in V(H)} \xi_v \cdot x_v}{\sqrt{V(H)}} \right), \tag{5.18}$$

which is the maximum of a collection of at most $2^{k_n} \cdot |C_{\leq k_n}|$ random variables following $\mathcal{N}(0,1)$. Let Z_n denote this maximum, i.e.,

$$Z_n = \max_{H \in C_{\leq k_n}} \left(\max_{(\xi_v) \in \{\pm 1\}^{V(H)}} \frac{\sum_{v \in V(H)} \xi_v \cdot x_v}{\sqrt{V(H)}} \right).$$

Using the same proof as in Theorem 5.1, we have

$$\mathbb{E}_{\mathcal{H}_0}[Z_n] \leq \sqrt{2 \log n + O(k_n \log d_n)}. \tag{5.19}$$

Since Z_n is the maximum of a collection of $\mathcal{N}(0,1)$ Gaussian random variables, the following concentration inequality [24] holds

$$\mathbb{P}_{\mathcal{H}_0} \left(Z_n - \mathbb{E}\left[Z_n \right] \geq t \right) \leq c_1 \cdot e^{-c_2 t}, \tag{5.20}$$

where $c_1, c_2 > 0$ are constants. Therefore, from (5.18), (5.19), and (5.20), with high probability we have

$$OPT_n|_{\mathcal{H}_0} \leq Z_n|_{\mathcal{H}_0} \tag{5.21}$$

$$\leq \sqrt{2 \log n + O(k_n \log d_n)} \cdot (1 + o(1)) \tag{5.22}$$

$$\leq O\left(\sqrt{\log n + k_n \log d_n} \right). \tag{5.23}$$

On the other hand, recall that the optimum scan statistics for the alternative hypothesis $\mathcal{H}_1$ is lower bounded by

$$OPT_n|_{\mathcal{H}_1} \geq \frac{\sum_{v \in S} p(v)}{\sqrt{|S|}} \geq \sqrt{s_n} \cdot \mu_n (1 - o(1)) \tag{5.24}$$

with high probability, where $S \subset V$ is the anomalous cluster.

Combining the inequalities (5.23) and (5.24), we conclude that by choosing an appropriate constant c, if

$$\mu_n \geq c \cdot \frac{\sqrt{\log n + k_n \log d_n}}{\sqrt{s_n}}$$

holds, then for all sufficiently large n, with high probability we have

$$OPT_n|_{\mathcal{H}_0} < \frac{0.99 \cdot c}{\alpha} \cdot \sqrt{\log n + k_n \log d_n} \leq 0.99 \cdot c \cdot \sqrt{\log n + k_n \log d_n} \leq OPT_n|_{\mathcal{H}_1}, \tag{5.25}$$

where α is the approximation ratio of the algorithm we used for the Elevated Mean problem.

Clearly, if the inequality (5.25) above holds with high probability, the decison rule T_{EM} asymptotically separates $\mathcal{H}_0$ and $\mathcal{H}_1$ by setting the threshold value to

$$\theta_n = \frac{0.99 \cdot c}{\alpha} \cdot \sqrt{\log n + k_n \log d_n}.$$

□

Theorem 5.8. *In the Poisson model, a sequence of instances (G_n, k_n, s_n, μ_n) of the anomaly detection problem is asymptotically separable if*

$$\mu_n = \omega \left(\frac{\log n + \sqrt{k_n} \cdot \log d_n}{\log \log d_n \cdot \sqrt{s_n}} \right),$$

and moreover, T_{EM} is a polynomial time decision rule that asymptotically separates $\mathcal{H}_0$ and $\mathcal{H}_1$ provided with a function $g : \mathbb{N} \to \mathbb{R}$ such that $g(n) = \omega_n(1)$ and

$$\mu_n = \Omega\left(g(n) \cdot \frac{\log n + \sqrt{k_n} \cdot \log d_n}{\log\log d_n \cdot \sqrt{s_n}}\right).$$

Proof. The proof idea is similar to the proof for the Gaussian model in Theorem 5.7, but without the concentration inequality for the maximum, we apply Markov's inequality. □

As a special case, we have the following corollary if the graphs G_n have bounded degree and if the search space $C_{\leq k_n}$ is not too far from the size of the anomalous cluster s_n.

Corollary 5.9. *Suppose $s_n = \Omega(k_n)$ and d_n is bounded above by a constant. Then, there exists a constant c' such that the following is true.*

A sequence of instances (G_n, k_n, s_n, μ_n) of the anomaly detection problem is asymptotically separable

- *in the Gaussian model if*

$$\mu_n \geq c' \cdot \left(\frac{\sqrt{\log n}}{\sqrt{k_n}} + 1\right).$$

- *in the Poisson model if*

$$\mu_n = \omega\left(\frac{\log n}{\sqrt{k_n}} + 1\right).$$

6. Discussion

We have shown separability bounds for the anomaly detection problem in the previous section using upper bound for the null hypothesis and lower bound for the alternative hypothesis. In this section, we show that the upper bound for the expected Elevated Mean scan statistics in the null hypothesis $\mathcal{H}_0$ we established in the previous section is asymptotically tight for the Gaussian model.

First, we show a simple degree-independent lower bound for the null hypothesis in the Gaussian model.

Lemma 6.1. *In the Gaussian model, under the null hypothesis $\mathcal{H}_0$, we have*

$$\mathbb{E}_{\mathcal{H}_0}[OPT] = \Omega\left(\sqrt{\log n}\right).$$

Proof. Note that we have

$$OPT \geq \max_{v \in V} p(v) \geq \max_{v \in V} x_v,$$

and x_v are i.i.d. $\mathcal{N}(0,1)$ random variables under the null hypothesis $\mathcal{H}_1$. As the expected $\max_{v \in V} x_v$ is lower bounded [16] by $\Omega\left(\sqrt{\log n}\right)$, we conclude that

$$\mathbb{E}_{\mathcal{H}_0}[OPT] = \Omega\left(\sqrt{\log n}\right).$$

□

Next, we show another lower bound for the null hypothesis in the Gaussian model based on the maximum degree. We will use a lower bound on the maximum size of set systems with restricted intersections [5] and the stability of the maximum of weakly dependent Gaussian random variables [6].

Lemma 6.2. *Let $(G = (V, E), k, s, \mu)$ be an instance of the anomaly detection problem in the Gaussian model. Let $v \in V$ be a vertex of degree d. Suppose $d \geq 2(k-1)^2$. Then, under the null hypothesis $\mathcal{H}_0$, we have*

$$\mathbb{E}_{\mathcal{H}_0}[OPT] = \Omega\left(\sqrt{k \log d}\right).$$

Proof. Consider the neighborhood $N(v)$ of the vertex v. Let h be a parameter that we will later determine. We would like to find a $(k-1)$-uniform family $\mathcal{F}$ in the neighborhood $N(v)$ of large size, such that the intersection of any two distinct sets $E, F \in \mathcal{F}$ is

$$|E \cap F| \leq h - 1,$$

and that

$$\frac{h}{k} = o(1). \tag{6.1}$$

It is a known result [5] that for a given h, if $d \geq 2(k-1)^2$ and $k \geq h$, then there exists a $(k-1)$-uniform family $\mathcal{F}$ on a ground set of d elements, such that for any two distinct sets $E, F \in \mathcal{F}$,

$$|E \cap F| \leq h - 1,$$

and that the size of $\mathcal{F}$ is

$$|\mathcal{F}| > \left(\frac{d}{2(k-1)}\right)^h \geq \left(\frac{d}{2}\right)^{h/2}. \tag{6.2}$$

In our case, there exists a $(k-1)$-uniform family $\mathcal{F}$ in the neighborhood $N(v)$ of size $> \left(\frac{d}{2}\right)^{h/2}$ such that the pairwise intersection of sets in $\mathcal{F}$ is $\leq h-1$.

Now fix a monotone function $g : \mathbb{N} \to \mathbb{R}$ such that $g(n) = \omega(1)$. Then, the inequality (6.1) is satisfied if we set $h = \frac{k}{g(n)}$.

Consider the family $\mathcal{F}_v = \{E \cup \{v\} : E \in \mathcal{F}\}$. Clearly, $\mathcal{F}_v$ is k-uniform, the sets in $\mathcal{F}_v$ are connected in the graph G, and the pairwise intersection of sets in $\mathcal{F}_v$ is $\leq h$. Therefore, we have established the existence of a k-uniform family $\mathcal{F}_v$ in $N(v) \cup \{v\}$ containing sets of connected vertices, such that for any two distinct sets $A, B \in \mathcal{F}_v$,

$$|A \cap B| \leq h,$$

and the size of $\mathcal{F}_v$ is

$$|\mathcal{F}_v| > \left(\frac{d}{2}\right)^{h/2},$$

where $h = \frac{k}{g(n)}$.

Now consider a collection $\mathcal{G}$ of connected induced subgraphs of G given by the vertex sets in $\mathcal{F}_v$, i.e., $\mathcal{G} = \{G[A] : A \in \mathcal{F}_v\}$. The Elevated Mean scan statistics associated with each subgraph $H = G[A] \in \mathcal{G}$

$$p_H = \frac{\sum_{v \in A} p(v)}{\sqrt{|A|}}$$

is lower bounded by the random variable

$$x_H = \frac{\sum_{v \in A} x_v}{\sqrt{|A|}},$$

and the variables x_H each follow the normal distribution $\mathcal{N}(0, 1)$.

Note that the covariance of x_{H_1} and x_{H_2} for two distinct subgraphs $H_1 = G[A], H_2 = G[B] \in \mathcal{G}$ is

$$\begin{aligned} Cov(x_{H_1}, x_{H_2}) &= Cov\left(\frac{1}{\sqrt{k}}\sum_{u\in A} x_u, \frac{1}{\sqrt{k}}\sum_{v\in B} x_v\right) \\ &= \frac{1}{k}\cdot \sum_{\substack{u\in A,\\ v\in B}} Cov(x_u, x_v) \\ &= \frac{1}{k}\cdot \sum_{v\in A\cap B} Var(x_v) \\ &= \frac{|A\cap B|}{k} \\ &\le \frac{h}{k} \le \frac{1}{g(n)} = o(1). \end{aligned}$$

As a consequence, by the stability result of the maximum of weakly dependent Gaussian random variables [6], we have

$$\mathbb{E}_{\mathcal{H}_0}\left[\max_{H\in\mathcal{G}} x_H\right] = \Omega\left(\sqrt{\log|\mathcal{G}|}\right),$$

where $|\mathcal{G}| = |\mathcal{F}_v| > \left(\frac{d}{2}\right)^{h/2}$. Therefore, we have the following lower bound on the expected OPT under the null hypothesis

$$\begin{aligned} \mathbb{E}_{\mathcal{H}_0}[OPT] &= \mathbb{E}_{\mathcal{H}_0}\left[\max_{H\in C_{\le k}} \frac{\sum_{v\in V(H)} p(v)}{\sqrt{|V(H)|}}\right] \\ &\ge \mathbb{E}_{\mathcal{H}_0}\left[\max_{H\in\mathcal{G}} \frac{\sum_{v\in V(H)} p(v)}{\sqrt{|V(H)|}}\right] \\ &\ge \mathbb{E}_{\mathcal{H}_0}\left[\max_{H\in\mathcal{G}} \frac{\sum_{v\in V(H)} x_v}{\sqrt{|V(H)|}}\right] \\ &= \mathbb{E}_{\mathcal{H}_0}\left[\max_{H\in\mathcal{G}} x_H\right] \\ &= \Omega\left(\sqrt{\log|\mathcal{G}|}\right) \\ &= \Omega\left(\sqrt{\frac{h}{2}\cdot\log\left(\frac{d}{2}\right)}\right) \\ &= \Omega\left(\sqrt{\frac{k\log d}{g(n)}}\right). \end{aligned}$$

Since $g:\mathbb{N}\to\mathbb{R}$ is an arbitrary monotone function such that $g(n) = \omega(1)$, we conclude that

$$\mathbb{E}_{\mathcal{H}_0}[OPT] = \Omega\left(\sqrt{k\log d}\right).$$

□

Combining Lemma 6.1 and Lemma 6.2, we get the following lower bound that matches the upper bound in Theorem 5.1.

Theorem 6.3. *Let $(G = (V, E), k, s, \mu)$ be an instance of the anomaly detection problem in the Gaussian model. Let $v \in V$ be a vertex of degree d. Suppose $d \ge 2(k-1)^2$. Then, under the null hypothesis $\mathcal{H}_0$, we have*

$$\mathbb{E}_{\mathcal{H}_0}[OPT] = \Omega\left(\sqrt{\log n + k\log d}\right).$$

7. Conclusion

In summary, we described a black-box reduction schema from the Elevated Mean problem to the Quota Prize-Collecting Tree problem and the Budget Prize-Collecting Tree problem, and showed a $\sqrt{2}$-approximation algorithm for the Elevated Mean problem, and a $(3+\varepsilon)$-approximation algorithm for the k-Elevated Mean problem. We also provided separability bounds based on the maximum degree of the underlying graph for the anomaly detection problem, and used the constant-factor approximation algorithms shown earlier to show a polynomial time decision rule T_{EM} that achieves the separability bounds we proved.

Table 1 below shows a comparison between our separability results and some earlier work for the Gaussian model of the anomaly detection problem.

Work	*Separability Bound*	*Class of Potential Anomalous Clusters*	*Assumptions*	*Computation*
This	$\mu_n \geq c \cdot \frac{\sqrt{\log n + k_n \log d_n}}{\sqrt{s_n}}$	*connected induced subgraphs of size* $\leq k_n$	$\Delta(G_n) = d_n$	*polynomial time solvable*
[19]	$\mu_n = \omega\left(\log k_n \sqrt{\log n}\right)$	*connected induced subgraphs of 2D grid of size* $\leq k_n$	$k_n = s_n$	*solving convex objective subject to LMI constraints*
[19]	$\mu_n = \omega\left(\frac{\log k_n}{\phi_n \cdot k_n}\right)$	*connected induced subgraphs of 2D grid containing a fixed vertex a, of size* $\leq k_n$*, and of internal conductance* $\geq \phi_n$	$k_n = s_n$	*solving convex objective subject to LMI constraints*
[3]	$\mu_n \geq \frac{\sqrt{(2+o(1))\log n}}{\sqrt{k_n}}$	*connected induced subgraphs of d-dimensional grid of size* $\leq k_n$	$k_n = s_n$ $s_n = o(\log n)$ $d = O(1)$	*no efficiency guarantee*

Table 1. Comparison Between the Separability Results

In particular, our separability bound for the Gaussian model matches that of [3] up to constant factor if the underlying graph has bounded degree, and generalizes the results to all graphs. Our decision rule is also the first computationally tractable algorithm that achieves such separability bound.

We conclude the paper by pointing to some future directions.

- Simpler and faster approximation algorithms for the Elevated Mean problem are needed for more practical applications of the decision rule T_{EM}.
- Better analysis for the Poisson model of the anomaly detection problem is needed in order to get tighter separability results.
- Non-asymptotic bounds and precise constants in the bounds are worth examining.

Acknowledgments

This paper is written in partial fulfillment of the requirements of the research-oriented joint Bx/MS program of the Department of Computer Science, the University of Chicago, under the supervision of my advisor, Prof. Lorenzo Orecchia.

I would like to thank Prof. Lorenzo Orecchia for introducing me to this problem and helping me with research. I would also like to thank Chris Jones and Goutham Rajendran for useful discussions on related topics. Lastly, I would like to thank my friends, the members of Prof. Orecchia's research group, the professors who taught me various subjects, and my family for their continuous support during my study at the University of Chicago.

References

[1] Deepak Agarwal, Andrew McGregor, Jeff M Phillips, Suresh Venkatasubramanian, and Zhengyuan Zhu. Spatial scan statistics: approximations and performance study. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 24–33, 2006.

[2] Cem Aksoylar, Lorenzo Orecchia, and Venkatesh Saligrama. Connected subgraph detection with mirror descent on sdps. In *International Conference on Machine Learning*, pages 51–59. PMLR, 2017.

[3] Ery Arias-Castro, Emmanuel J Candes, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304, 2011.

[4] Sanjeev Arora and George Karakostas. A 2+ ε approximation algorithm for the k-mst problem. In *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 754–759, 2000.

[5] László Babai and Péter Frankl. *Linear algebra methods in combinatorics*.

[6] Simeon M Berman. Limit theorems for the maximum term in stationary sequences. *The Annals of Mathematical Statistics*, pages 502–516, 1964.

[7] Béla Bollobás. *The art of mathematics: Coffee time in Memphis*. Cambridge University Press, 2006.

[8] Keith M Briggs, Linlin Song, and Thomas Prellberg. A note on the distribution of the maximum of a set of poisson random variables. *arXiv preprint arXiv:0903.4373*, 2009.

[9] Jose Cadena, Feng Chen, and Anil Vullikanti. Graph anomaly detection based on steiner connectivity and density. *Proceedings of the IEEE*, 106(5):829–845, 2018.

[10] Jose Cadena, Feng Chen, and Anil Vullikanti. Near-optimal and practical algorithms for graph scan statistics with connectivity constraints. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–33, 2019.

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

[12] Luiz Duczmal, Martin Kulldorff, and Lan Huang. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2):428–442, 2006.

[13] Naveen Garg. Saving an epsilon: a 2-approximation for the k-mst problem in graphs. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 396–402, 2005.

[14] Michel X Goemans and David P Williamson. A general approximation technique for constrained forest problems. *SIAM Journal on Computing*, 24(2):296–317, 1995.

[15] David S Johnson, Maria Minkoff, and Steven Phillips. The prize collecting steiner tree problem: theory and practice. In *SODA*, volume 1, page 4. Citeseer, 2000.

[16] Gautam Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www. gautamkamath. com/writings/gaussian max. pdf*, 2015.

[17] Asaf Levin. A better approximation algorithm for the budget prize collecting tree problem. *Operations Research Letters*, 32(4):316–319, 2004.

[18] Lehilton LC Pedrosa and Hugo KK Rosado. A 2-approximation for the k-prize-collecting steiner tree problem. In *Latin American Symposium on Theoretical Informatics*, pages 76–88. Springer, 2021.

[19] Jing Qian and Venkatesh Saligrama. Efficient minimax signal detection on graphs. *arXiv preprint arXiv:1411.6203*, 2014.

[20] Jing Qian, Venkatesh Saligrama, and Yuting Chen. Connected sub-graph detection. In *Artificial Intelligence and Statistics*, pages 796–804. PMLR, 2014.

[21] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813:814, 2015.

[22] James Sharpnack, Alessandro Rinaldo, and Aarti Singh. Detecting anomalous activity on networks with the graph fourier scan statistic. *IEEE Transactions on Signal Processing*, 64(2):364–379, 2015.

[23] Skyler Speakman, Yating Zhang, and Daniel B Neill. Dynamic pattern detection with temporal consistency and connectivity constraints. In *2013 IEEE 13th International Conference on Data Mining*, pages 697–706. IEEE, 2013.

[24] Kevin Tanguy. Some superconcentration inequalities for extrema of stationary gaussian processes. *Statistics & Probability Letters*, 106:239–246, 2015.