Quantifying and Reducing Bias in Maximum Likelihood Estimation of Structured Anomalies

Uthsav Chitra [1] **Kimberly Ding** [1] **Jasper C.H. Lee** [2] **Benjamin J. Raphael** [1]

Abstract

Anomaly estimation, or the problem of finding a subset of a dataset that differs from the rest of the dataset, is a classic problem in machine learning and data mining. In both theoretical work and in applications, the anomaly is assumed to have a specific structure defined by membership in an *anomaly family*. For example, in temporal data the anomaly family may be time intervals, while in network data the anomaly family may be connected subgraphs. The most prominent approach for anomaly estimation is to compute the Maximum Likelihood Estimator (MLE) of the anomaly; however, it was recently observed that for normally distributed data, the MLE is a *biased* estimator for some anomaly families. In this work, we demonstrate that in the normal means setting, the bias of the MLE depends on the size of the anomaly family. We prove that if the number of sets in the anomaly family that contain the anomaly is sub-exponential, then the MLE is asymptotically unbiased. We also provide empirical evidence that the converse is true: if the number of such sets is exponential, then the MLE is asymptotically biased. Our analysis unifies a number of earlier results on the bias of the MLE for specific anomaly families. Next, we derive a new anomaly estimator using a mixture model, and we prove that our anomaly estimator is asymptotically unbiased regardless of the size of the anomaly family. We illustrate the advantages of our estimator versus the MLE on disease outbreak data and highway traffic data.

[1]Department of Computer Science, Princeton University, Princeton, New Jersey, USA [2]Department of Computer Science, Brown University, Providence, Rhode Island, USA. Correspondence to: Benjamin J. Raphael <braphael@princeton.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

Anomaly identification — the discovery of rare, irregular, or otherwise anomalous behavior in data — is a fundamental problem in machine learning and data mining with numerous applications (Chandola et al., 2009). In temporal/sequential data, applications of anomaly identification include change-point detection and inference (Page, 1955; Hinkley, 1970; Adams & MacKay, 2007; Zhai et al., 2016); in matrix data, applications include bi-clustering (Hartigan, 1972; Tanay et al., 2005; Kolar et al., 2011) and gene expression analysis (Ideker et al., 2002; Dittrich et al., 2008); in spatial data, applications include disease outbreak and event detection (Neill & Moore, 2004; Neill et al., 2005; Neill, 2012); and in network data, applications include large-scale network surveillance (Arias-Castro et al., 2011; Sharpnack et al., 2013b;a) and outbreak detection (Wong et al., 2003; Leskovec et al., 2007). In many applications, the anomalous behavior is assumed to have a specific structure described by membership in an *anomaly family*. For example, in temporal data the anomaly family may be time intervals; in matrix data the anomaly family may be submatrices; and in network data the anomaly family may be connected subgraphs.

Anomaly identification can be divided into two different but closely related problems: *anomaly detection* and *anomaly estimation*. Given a dataset, the goal of anomaly detection is to decide whether or not there exists an *anomaly*, or a subset of the data, that is distributed according to a different probability distribution compared to the rest of the data. The goal of anomaly estimation is to determine the data points in the anomaly. The distinction between anomaly detection and anomaly estimation is analogous to the distinction between property testing and proper learning in statistical learning theory (Goldreich et al., 1998): just as property testing is "easier" than proper learning (with difficulty measured by sample complexity), anomaly detection is easier than anomaly estimation (with difficulty measured by the separation between the distributions of the anomaly and the rest of the data). Different choices of the anomaly family give rise to different versions of the anomaly detection and estimation problems; e.g. change-point detection versus change-point inference in temporal data (Arias-Castro et al., 2005; Hinkley, 1971; Jeng et al., 2010), or submatrix detection versus submatrix estimation in matrix data (Hajek et al.,

2017; Butucea & Ingster, 2013; Ma & Wu, 2015; Brennan et al., 2019; Chen & Xu, 2016; Banks et al., 2018; Liu & Arias-Castro, 2019; Gamarnik et al., 2019).

Most of the theoretical literature on anomaly detection and estimation focuses on *structured normal means* problems (Sharpnack et al., 2013a; Krishnamurthy, 2016). In this setting, each data point is drawn from one of two normal distributions, with the data points from the anomaly drawn from the normal distribution with the higher mean; the structure of the anomaly is determined by the anomaly family. Normal means problems have a long history in statistics and machine learning as many statistical tests commonly used in scientific disciplines are asymptotically normal, e.g. see Arias-Castro et al. (2011); Donoho & Jin (2004); Cai et al. (2007); Kolar et al. (2011); Sharpnack et al. (2013a); Chen & Xu (2016); Liu & Arias-Castro (2019). In this paper we also focus on the structured normal means setting, but we emphasize that our results algorithms can be readily extended to other probability distributions from the exponential family as in earlier works (Butucea & Ingster, 2013; Liu & Arias-Castro, 2019).

The most widely used techniques for both anomaly detection and anomaly estimation problems are likelihood models: the generalized likelihood ratio (GLR) test for the detection problem, and the maximum likelihood estimator (MLE) for the estimation problem. Both the GLR test statistic and the MLE can be expressed using a *scan statistic*, or the maximization of a function across all members of the anomaly family (Kulldorff, 1997; Glaz & Naus, 2010). In fact, as we note in Theorem 1, both the GLR test statistic and the MLE involve the maximization of the *same* function.

Despite this close relationship between the GLR test and the MLE, the two quantities have different theoretical guarantees for their respective problems. The GLR test is known to be asymptotically "near-optimal" for solving the anomaly detection problem across many different anomaly families, including intervals (Arias-Castro et al., 2005), submatrices (Butucea & Ingster, 2013), subgraphs with small cut-size (Sharpnack et al., 2013a), and connected subgraphs (Qian & Saligrama, 2014). In contrast, the MLE is known to be asymptotically near-optimal for solving the anomaly estimation problem only when the anomaly family is intervals (Jeng et al., 2010) or submatrices (Liu & Arias-Castro, 2019). In fact, Reyna et al. (2020) recently observed that the MLE is a *biased* estimator of the size of the anomaly when the anomaly family is connected subgraphs of a biological network.

These varying results for anomaly estimation across different anomaly families suggest that the bias of the MLE depends on the anomaly family, and thus raise the following two questions: (1) For which anomaly families is the MLE biased? (2) Are there anomaly estimators that are less biased than the MLE?

In this work we address both of these questions. First, we show that the bias in the MLE depends on the size of the anomaly family.[1] We prove that if the number of sets in the anomaly family that contain the anomaly is sub-exponential, then the MLE is an asymptotically unbiased estimator. We also provide empirical evidence that the converse is true by examining many common anomaly families including intervals, submatrices, connected subgraphs, and subgraphs with low-cut size. Our results unify a number of previous results in the literature including the asymptotic optimality of the MLE when the anomaly family is intervals (Jeng et al., 2010) or submatrices (Liu & Arias-Castro, 2019), and the observation that the MLE is biased when the anomaly family is connected subgraphs (Reyna et al., 2020).

Next, we derive a reduced-bias estimator of the anomaly based on a Gaussian mixture model (GMM). Our estimator is motivated by previous work that models *unstructured* anomalies using GMMs (Cai et al., 2007; Donoho & Jin, 2004). We prove that our GMM-based estimator is asymptotically unbiased, regardless of the size of the anomaly family or the number of sets containing the anomaly. We empirically demonstrate the small bias of our estimator for several anomaly families including intervals, submatrices, and connected subgraphs. We illustrate the advantages of our estimator versus the MLE on both disease outbreak data and a highway traffic dataset.

2. Background: Structured Anomalies and Maximum Likelihood Estimation

2.1. Problem Formulation

Suppose one is given observations $(X_1, \ldots, X_n)$, where a subset $A \subseteq \{1, \ldots, n\}$ of these observations, the *anomaly*, are drawn from a normal distribution $N(\mu, 1)$ with elevated mean and the remaining observations are drawn from the standard normal distribution $N(0, 1)$. Using the notation $[n] = \{1, \ldots, n\}$ and $\mathcal{P}_n$ for the power set of $[n]$, or the set of all subsets of $[n]$ for a positive integer n, we define the distribution of the observations $(X_1, \ldots, X_n)$ as follows.

Anomalous Subset Distribution (ASD). *Let $\mu > 0$, let $\mathcal{S} \subseteq \mathcal{P}_n$ be a family of subsets of $[n]$ and let $A \in \mathcal{S}$. We say $\mathbf{X} = (X_1, \ldots, X_n)$ is distributed according to the* Anomalous Subset Distribution $\mathrm{ASD}_{\mathcal{S}}(A, \mu)$ *provided the X_i are independently distributed as*

$$X_i \sim \begin{cases} N(\mu, 1), & \text{if } i \in A, \\ N(0, 1), & \text{otherwise.} \end{cases} \tag{1}$$

The distribution $\mathrm{ASD}_{\mathcal{S}}(A, \mu)$ has three parameters: the *anomaly family* $\mathcal{S}$, the *anomaly* A, and the *mean* μ.

[1] All proofs are given in the Supplement.

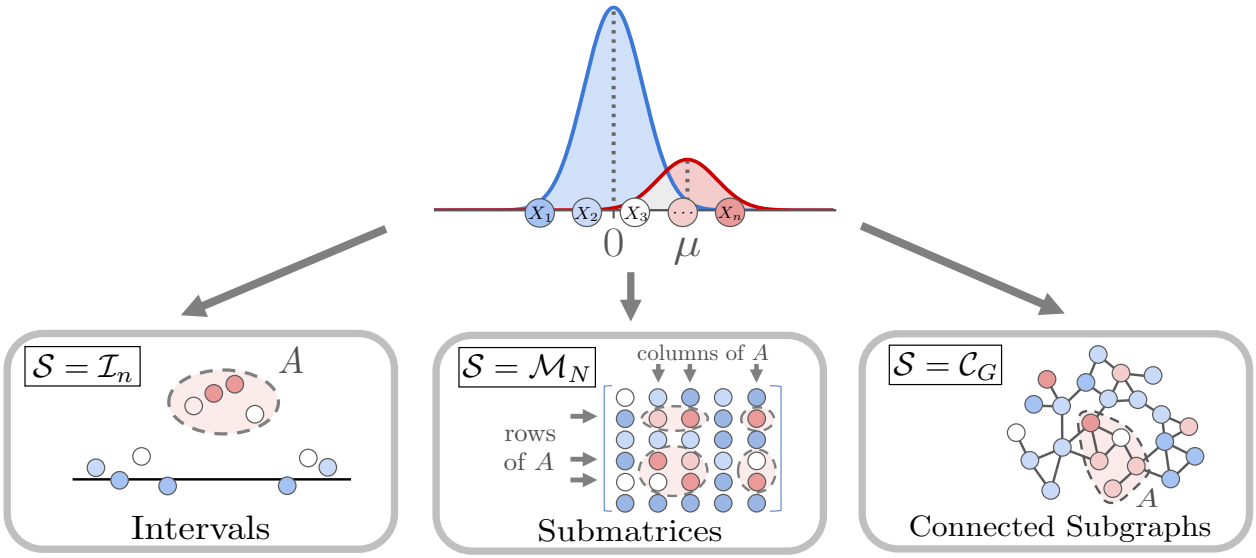

Figure 1: Observations $(X_1, \ldots, X_n)$ from the Anomalous Subset Distribution $\text{ASD}_{\mathcal{S}}(A, \mu)$ for three anomaly families $\mathcal{S}$.

The goal of anomaly estimation is to learn the anomaly A, given data $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ and anomaly family $\mathcal{S}$. We formalize this problem as the following estimation problem.

ASD Estimation Problem. *Given* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ *and* $\mathcal{S}$*, find* A*.*

A related problem is the decision problem of deciding whether or not data $\mathbf{X}$ contains an anomaly. We formalize this problem as the following hypothesis testing problem.

ASD Detection Problem. *Given* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ *and* $\mathcal{S}$*, test between the hypotheses* $H_0 : A = \varnothing$ *and* $H_1 : A \neq \varnothing$*.*

The ASD Detection and Estimation Problems are also called *structured normal means* problems (Krishnamurthy, 2016; Sharpnack et al., 2013a), where the structure comes from the choice of the anomaly family $\mathcal{S}$.

Many well-known problems in machine learning correspond to the ASD Detection and Estimation Problems for different anomaly families $\mathcal{S}$. In particular, we note the following examples.

- $\mathcal{S} = \mathcal{I}_n$, the set of all intervals $\{i, i+1, \ldots, j\} \subseteq [n]$. We call $\mathcal{I}_n$ the *interval* family, and we call $\text{ASD}_{\mathcal{I}_n}(A, \mu)$ the *interval ASD*. The interval ASD is used to model change-points, or abrupt changes, in sequential data including time-series and DNA sequences (Hinkley, 1970; 1971; Basseville & Nikiforov, 1993; Jeng et al., 2010).
- $\mathcal{S} = \mathcal{C}_G$, the set of all connected subgraphs of a graph $G = (V, E)$ with vertices $V = \{1, \ldots, n\}$. We call $\mathcal{C}_G$ the *connected* family, and we call $\text{ASD}_{\mathcal{C}_G}(A, \mu)$ the *connected ASD*. The connected ASD is used to model anomalous behavior in different types of networks including social networks, sensor networks and biological networks (Qian & Saligrama, 2014; Aksoylar et al., 2017; Ideker et al., 2002; Reyna et al., 2020). Note that the interval family $\mathcal{I}_n$ is a special case of the connected family $\mathcal{C}_{P_n}$ for the path graph P_n with n vertices.
- $\mathcal{S} = \mathcal{T}_{G,\rho}$, the set of all subgraphs H of a graph G with $|\{(i,j) \in E : i \in H, j \notin H\}| \leq \rho$. We call $\mathcal{T}_{G,\rho}$ the *graph cut family*, and we call $\text{ASD}_{\mathcal{T}_{G,\rho}}(A, \mu)$ the *graph cut ASD*. The graph cut ASD is also used to model anomalous behavior in networks (Sharpnack et al., 2013b;a; Sharpnack et al., 2016).
- $\mathcal{S} = \mathcal{E}_{G,\delta}$, the set of all subgraphs H of a graph G with edge-density at least δ. We call $\mathcal{E}_{G,\delta}$ the *edge-dense family*, and we call $\text{ASD}_{\mathcal{E}_{G,\delta}}(A, \mu)$ the *edge-dense ASD*. The edge-dense ASD is also used to model anomalous behavior in networks (Cadena et al., 2018b).
- $\mathcal{S} = \mathcal{M}_N$, the set of all submatrices of a square matrix N with n entries (each observation X_i corresponds to an entry of N). We call $\mathcal{M}_N$ the *submatrix family*, and we call $\text{ASD}_{\mathcal{M}_N}(A, \mu)$ the *submatrix ASD*. The clustering literature often uses the submatrix ASD to model biclusters in matrix data (Kolar et al., 2011; Butucea & Ingster, 2013; Brennan et al., 2019; Liu & Arias-Castro, 2019).
- $\mathcal{S} = \mathcal{B}_{P,\epsilon}$, the set of all ϵ-balls of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{R}^d$ in space. We call $\mathcal{B}_{P,\epsilon}$ the *ϵ-ball* family. The spatial scan statistic is a standard tool for solving the ASD Estimation Problem with the ϵ-ball family $\mathcal{B}_{P,\epsilon}$ (Kulldorff, 1997; Glaz & Naus, 2010).
- $\mathcal{S} = \mathcal{P}_n$, the power set of $\{1, \ldots, n\}$. We call $\mathcal{P}_n$ the *unstructured family*, and we call $\text{ASD}_{\mathcal{P}_n}(A, \mu)$ the *unstructured ASD*.

2.2. Maximum Likelihood Anomaly Estimation

A standard approach in statistics for solving a hypothesis testing problem is to use the generalized likelihood ratio (GLR) test, which the Neyman-Pearson lemma (Lehmann & Romano, 2005) shows is the most powerful test for any significance level. Likewise, a standard approach for solving an estimation problem is to compute a maximum likelihood estimator (MLE). For the ASD Detection and Estimation problems, the GLR test statistic and the MLE, respectively, have explicit formulas that involve the maximization of the same function, $\Gamma(S) = \frac{1}{\sqrt{S}} \sum_{v \in S} X_v$. We write out these formulas below; see Arias-Castro et al. (2011); Sharpnack et al. (2013a); Reyna et al. (2020) for proofs.

Proposition 1. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ *be distributed according to the ASD. The Generalized Likelihood Ratio (GLR) test statistic* $\widehat{t}_{\mathcal{S}}$ *for the ASD Detection Problem is*

$$\widehat{t}_{\mathcal{S}} = \max_{S \in \mathcal{S}} \Gamma(S) = \max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v. \tag{2}$$

The Maximum Likelihood Estimator (MLE) $\widehat{A}_{\mathcal{S}}$ *of the anomaly* A *is*

$$\widehat{A}_{\mathcal{S}} = \operatorname*{argmax}_{S \in \mathcal{S}} \Gamma(S) = \operatorname*{argmax}_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v. \tag{3}$$

A key question in the statistics literature is: for what anomaly families $\mathcal{S}$ and means μ (i.e. parameters of the ASD) do the GLR test and the MLE solve the ASD Detection and Estimation problems, respectively?

For many anomaly families $\mathcal{S}$, it has been shown that the GLR test is asymptotically "near-optimal". This means that there exists a value $\mu_{\text{detect}} > 0$ such that the following is true: if $\mu \geq \mu_{\text{detect}}$ then the GLR test asymptotically solves the ASD Detection Problem with the probability of a type 1 or type 2 error going to 0 as $n \to \infty$, while if μ is not much smaller than μ_{detect} then there does not exist any test with such guarantees on its type 1 or type 2 error probabilities. Anomaly families $\mathcal{S}$ for which the GLR test is known to be asymptotically near-optimal include the interval family $\mathcal{S} = \mathcal{I}_n$ (Arias-Castro et al., 2005), the submatrix family $\mathcal{S} = \mathcal{M}_N$ (Butucea & Ingster, 2013), the graph cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$ (Sharpnack et al., 2013b;a), and the connected family $\mathcal{S} = \mathcal{C}_G$ (Qian & Saligrama, 2014; Qian et al., 2014).

For a few anomaly families $\mathcal{S}$, the MLE $\widehat{A}_{\mathcal{S}}$ has also been shown to optimally solve the ASD Estimation Problem. For the interval family $\mathcal{S} = \mathcal{I}_n$, Jeng et al. (2010) showed that if $\mu \geq \mu_{\text{detect}}$, then $\lim_{n \to \infty} P(\widehat{A}_{\mathcal{I}_n} \neq A) = 0$. Liu & Arias-Castro (2019) proved an analogous result for the submatrix family $\mathcal{S} = \mathcal{M}_N$ using a regularized MLE. Note that these results require $\mu \geq \mu_{\text{detect}}$, as it is not possible to estimate the anomaly without first detecting the anomaly's presence.

The MLE $\widehat{A}_{\mathcal{S}}$ is also used in the bioinformatics literature to solve the ASD Estimation Problem for the connected family $\mathcal{S} = \mathcal{C}_G$, where G is a biological network (Ideker et al., 2002; Dittrich et al., 2008). However, the MLE $\widehat{A}_{\mathcal{C}_G}$ for the connected family $\mathcal{C}_G$ does not have any theoretical guarantees, unlike the previously mentioned results for the interval and submatrix families. In fact, Nikolayeva et al. (2018) and Reyna et al. (2020) empirically observed that the size $|\widehat{A}_{\mathcal{C}_G}|$ of the MLE $\widehat{A}_{\mathcal{C}_G}$ is a biased estimate of the size $|A|$ of the anomaly, in the sense of the following definition.

Definition 1. *Given data* $\mathbf{X} = (X_1, \ldots, X_n)$*, let* $\hat{\theta} = \hat{\theta}(\mathbf{X})$ *be an estimator of a parameter* θ *of the distribution of* $\mathbf{X}$*. The quantity* $\text{Bias}_\theta(\hat{\theta}) = E[\hat{\theta}] - \theta$ *is the* bias *of the estimator* $\hat{\theta}$*. We say that* $\hat{\theta}$ *is a* biased *estimator of* θ *if* $\text{Bias}_\theta(\hat{\theta}) \neq 0$*, and that* $\hat{\theta}$ *is an* unbiased *estimator of* θ *otherwise. When it is clear from context, we omit the subscript* θ *and write* $\text{Bias}(\hat{\theta})$ *for the bias of estimator* $\hat{\theta}$*.*

Reyna et al. (2020) also empirically observed a similar bias for the MLE $\widehat{A}_{\mathcal{P}_n}$ for the unstructured family $\mathcal{S} = \mathcal{P}_n$.

We note that while many papers in statistics study the bias of MLEs for different distributions (e.g. Firth (1993); Mardia et al. (1999); Giles et al. (2013)), to our knowledge, the bias of the MLE $\widehat{A}_{\mathcal{S}}$ for the ASD has previously only been studied in Reyna et al. (2020).

3. Relating MLE Bias to Size of the Anomaly Family

The observations in the previous section lead to the following question: for which anomaly families $\mathcal{S}$ is the size $|\widehat{A}_{\mathcal{S}}|$ of the MLE $\widehat{A}_{\mathcal{S}}$ a biased estimate of the size $|A|$ of the anomaly A? In this section, we provide theoretical and experimental evidence that the key quantity that determines the bias of the MLE $\widehat{A}_{\mathcal{S}}$ is the quantity $\breve{\mathcal{S}}(A) = \{S \in \mathcal{S} : S \supseteq A\}$, or the collection of sets in the anomaly family $\mathcal{S}$ that contain the anomaly A.

First, we show that if the number $|\breve{\mathcal{S}}(A)|$ of sets containing the anomaly A is sub-exponential in n, then the size $|\widehat{A}_{\mathcal{S}}|$ of the MLE $\widehat{A}_{\mathcal{S}}$ is asymptotically unbiased.

Theorem 1. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ *where* $|\mathcal{S}| = \Omega(n)$ *and* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$*. Suppose* $|\breve{\mathcal{S}}(A)|$ *is sub-exponential in* n*. If* $\lim_{n \to \infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$*, then* $\lim_{n \to \infty} \text{Bias}(|\widehat{A}_{\mathcal{S}}|/n) = 0$*.*

A key component of our proof of Theorem 1 is the following Lemma relating the $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ of the MLE $\widehat{A}_{\mathcal{S}}$ to the number $|\breve{\mathcal{S}}(A)|$ of sets containing the anomaly A.

Lemma 1. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ *where* $|\mathcal{S}| = \Omega(n)$ *and* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$*. Assume* $\lim_{n \to \infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) =$

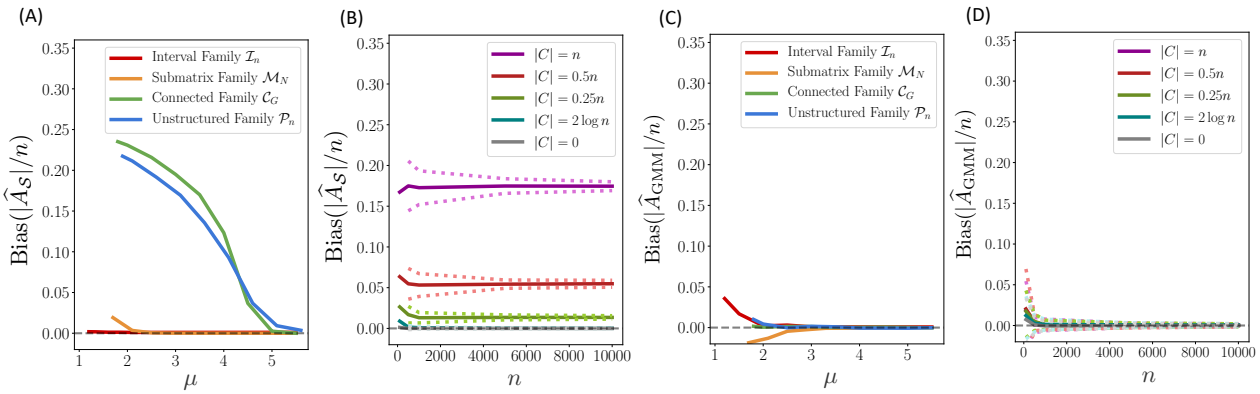

Figure 2: Bias in estimators of the size $|A|$ of the anomaly A computed from $n = 900$ samples $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ from the ASD for different choices of anomaly family $\mathcal{S}$. In each sample, the anomaly A of size $|A| = 0.05n$ is chosen uniformly at random from $\mathcal{S}$. **(A)** Bias($|\widehat{A}_{\mathcal{S}}|/n$) of the MLE vs μ for means $\mu \geq \mu_{\text{detect}}$. For small μ, the MLE shows positive bias for the connected family $\mathcal{C}_G$ and unstructured family $\mathcal{P}_n$, consistent with Conjecture 1 and Theorem 2. **(B)** Bias($|\widehat{A}_{\mathcal{S}}|/n$) vs n for $\mu = 3$ for the connected family $\mathcal{S} = \mathcal{C}_G$, where $G = (V, E)$ is a graph whose vertices $V = P \cup C$ are partitioned into a path graph P and a clique C. Dotted lines indicate first and third quartiles in the estimate of the bias. The positive bias for $|C| = \Theta(n)$ does not decrease as n increases, consistent with Conjecture 1. **(C)** Bias($|\widehat{A}_{\text{GMM}}|/n$) of GMM-based estimator vs μ for means $\mu \geq \mu_{\text{detect}}$. In contrast to (A), the bias is zero for all anomaly families and sufficiently large mean μ, consistent with Corollary 1. **(D)** Bias($|\widehat{A}_{\text{GMM}}|/n$) vs n for $\mu = 3$ and the same connected anomaly family as in (B). The GMM-based estimator appears to be unbiased for sufficiently large n, consistent with Corollary 1.

1. *If n is sufficiently large and* Bias($|\widehat{A}_{\mathcal{S}}|/n$) $\geq \gamma$, *then*

$$|\breve{\mathcal{S}}(A)| \geq (C_{\mu,\gamma,\alpha})^n \cdot e^{-\Theta(\sqrt{n \log n})}, \tag{4}$$

where $C_{\mu,\alpha,\gamma} = \exp\left(\frac{1}{2}\mu^2\alpha^2\left(\sqrt{1 + \frac{\gamma}{4\alpha}} - 1\right)^2\right)$.

We make two mild assumptions in Theorem 1. First, we assume the proportion $\alpha = \frac{|A|}{n}$ of anomalous observations is a positive constant independent of n. Second, we assume that the anomaly family $\mathcal{S}$ has size $|\mathcal{S}| = \Omega(n)$; this assumption is satisfied by many commonly-used anomaly families including the interval family $\mathcal{S} = \mathcal{I}_n$, the submatrix family $\mathcal{S} = \mathcal{M}_N$, the connected family $\mathcal{S} = \mathcal{C}_G$ for any graph G, and the unstructured family $\mathcal{S} = \mathcal{P}_n$.

We also require that $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$, which is a technical condition needed for the proof of Theorem 1. We conjecture that this condition can be replaced by the condition $\mu \geq \mu_{\text{detect}}$. This conjecture is based on the empirical observation that if $\mu \geq \mu_{\text{detect}}$, then the MLE $\widehat{A}_{\mathcal{S}}$ contains most of the elements in A (Figure 3), suggesting that the condition $\mu \geq \mu_{\text{detect}}$ is only a slightly weaker condition than $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$.

Theorem 1 generalizes earlier results showing that the MLEs $\widehat{A}_{\mathcal{I}_n}$ and $\widehat{A}_{\mathcal{M}_N}$ for the interval family $\mathcal{I}_n$ and submatrix family $\mathcal{M}_N$ are asymptotically unbiased (Jeng et al., 2010; Liu & Arias-Castro, 2019) as these families satisfy the conditions of Theorem 1. Moreover, Theorem 1 implies that the regularization of the MLE used in Liu & Arias-Castro (2019) is not necessary to prove asymptotic unbiasedness (see Supplement).

Informally, Theorem 1 says that if the number $|\breve{\mathcal{S}}(A)|$ of subsets that contain the anomaly A is sub-exponential in n, then the MLE $\widehat{A}_{\mathcal{S}}$ is an asymptotically unbiased estimator of the size $|A|$ of the anomaly A. Next, we prove that for the unstructured family $\mathcal{P}_n$, where $|\breve{\mathcal{S}}(A)|$ is exponential in n, the MLE $\widehat{A}_{\mathcal{P}_n}$ is asymptotically biased for all μ. This result settles a conjecture posed by Reyna et al. (2020).

Theorem 2. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ *where* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$. *Then* $\lim_{n\to\infty}$ Bias($|\widehat{A}_{\mathcal{P}_n}|/n$) > 0.

We prove Theorem 2 by deriving an explicit formula for the asymptotic bias $\lim_{n\to\infty}$ Bias($|\widehat{A}_{\mathcal{P}_n}|/n$) of the MLE.

We conjecture that Theorems 1 and 2 describe the only two possible values for the asymptotic bias of the MLE $\widehat{A}_{\mathcal{S}}$, and furthermore that the technical conditions in Theorem 1 can be relaxed to the simpler condition $\mu \geq \mu_{\text{detect}}$. Thus we conjecture that MLE $\widehat{A}_{\mathcal{S}}$ is asymptotically biased if and only if $|\breve{\mathcal{S}}(A)|$ is exponential in n.

Conjecture 1. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ *with* $\mu \geq \mu_{\text{detect}}$. *Then* $\lim_{n\to\infty}$ Bias($|\widehat{A}_{\mathcal{S}}|/n$) > 0 *if* $|\breve{\mathcal{S}}(A)|$ *is exponential in* n, *and* $\lim_{n\to\infty}$ Bias($|\widehat{A}_{\mathcal{S}}|/n$) $= 0$ *otherwise.*

Conjecture 1 generalizes Theorems 1 and 2, and is consistent with the prior work noted in Section 2.2 on the bias of the MLE $\widehat{A}_{\mathcal{S}}$ for different anomaly families $\mathcal{S}$:

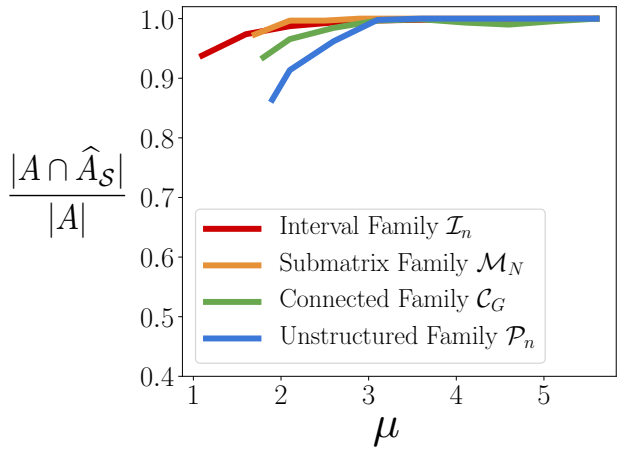

Figure 3: Normalized intersection $\frac{|A \cap \widehat{A}_{\mathcal{S}}|}{|A|}$ between anomaly A and MLE $\widehat{A}_{\mathcal{S}}$ for means $\mu \geq \mu_{\text{detect}}$ for different anomaly families $\mathcal{S}$. Data generated as in Figure 2. For $\mu \geq \mu_{\text{detect}}$ and different anomaly families $\mathcal{S}$, the normalized intersection $\frac{|A \cap \widehat{A}_{\mathcal{S}}|}{|A|} > 0.85$, i.e. the MLE $\widehat{A}_{\mathcal{S}}$ contains at least 85% of the elements in the anomaly A. This suggests that the condition $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$ in Theorem 1 is not much stronger than the condition $\mu \geq \mu_{\text{detect}}$.

- $\mathcal{S} = \mathcal{I}_n$, the interval family: $|\breve{\mathcal{S}}(A)| \leq |\mathcal{S}| = O(n^2)$ is sub-exponential, so $|\widehat{A}_{\mathcal{I}_n}|$ is asymptotically unbiased (Jeng et al., 2010).
- $\mathcal{S} = \mathcal{M}_N$, the submatrix family: $|\breve{\mathcal{S}}(A)| \leq |\mathcal{S}| = O(2^{2\sqrt{n}})$ is sub-exponential, so $|\widehat{A}_{\mathcal{M}_N}|$ is asymptotically unbiased (Liu & Arias-Castro, 2019).
- $\mathcal{S} = \mathcal{C}_G$, the connected family: When G has minimum degree 3, $|\breve{\mathcal{S}}(A)|$ is exponential (Vince, 2017), so $|\widehat{A}_{\mathcal{C}_G}|$ is asymptotically biased (Reyna et al., 2020).
- $\mathcal{S} = \mathcal{P}_n$, the unstructured family: When $|A| < 0.5n$, $|\breve{\mathcal{S}}(A)| = 2^{n\left(1-\frac{|A|}{n}\right)} = \Omega\left(2^{0.5n}\right)$ is exponential, so $|\widehat{A}_{\mathcal{P}_n}|$ is asymptotically biased.

3.1. Experimental Evidence for Conjecture 1

We provide empirical evidence for Conjecture 1 by examining the bias of the MLE for several different anomaly families. For each anomaly family $\mathcal{S}$, we select an anomaly $A \in \mathcal{S}$ with size $|A| = 0.05n$ uniformly at random from $\mathcal{S}$. We draw a sample $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ with $n = 900$ observations, and compute the MLE $\widehat{A}_{\mathcal{S}}$. We repeat for 50 samples to estimate Bias$(|\widehat{A}_{\mathcal{S}}|/n)$. We perform this process for a range of means $\mu \geq \mu_{\text{detect}}$ (see Supplement for details on empirically calculating μ_{detect}.)

We compute Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ for the following anomaly families: $\mathcal{S} = \mathcal{I}_n$, the interval family; $\mathcal{S} = \mathcal{M}_N$, the submatrix family with matrix $N \in \mathbb{R}^{30\times 30}$; $\mathcal{S} = \mathcal{C}_G$, the connected family with an Erdős-Rényi random graph G (edge probability $= 0.01$); and $\mathcal{S} = \mathcal{P}_n$, the unstructured family. $|\breve{\mathcal{S}}(A)|$ is sub-exponential for the interval family $\mathcal{I}_n$ and the submatrix family $\mathcal{M}_N$, and is exponential for the connected family $\mathcal{C}_G$ (with high probability (Vince, 2017)) and for the unstructured family $\mathcal{P}_n$.

For the interval family $\mathcal{I}_n$ and submatrix family $\mathcal{M}_N$, where $|\breve{\mathcal{S}}(A)|$ is sub-exponential, we find that Bias$(|\widehat{A}_{\mathcal{S}}|/n) \approx 0$ for all means $\mu \geq \mu_{\text{detect}}$ (Figure 2A). In contrast, for the connected family $\mathcal{C}_G$ and unstructured family $\mathcal{P}_n$, where $|\breve{\mathcal{S}}(A)|$ is exponential, we observe that Bias$(|\widehat{A}_{\mathcal{S}}|/n) > 0$ for $\mu \in [\mu_{\text{detect}}, 5]$ (Figure 2A). (Because n is fixed, the Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ will be zero for sufficiently large μ.) These observations provide evidence in support of Conjecture 1 for these families. Moreover, although Conjecture 1 is about the Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ of the MLE $\widehat{A}_{\mathcal{S}}$, we also observe that larger Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ reduces the F-measure between the anomaly A and the MLE $\widehat{A}_{\mathcal{S}}$ (see Supplement).

Next, we examine the Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ of the MLE $\widehat{A}_{\mathcal{S}}$ in the limit $n \to \infty$, and find that the bias of the MLE $\widehat{A}_{\mathcal{S}}$ appears to converge to positive values only when $|\breve{\mathcal{S}}(A)|$ is exponential. We specifically examine the connected anomaly family $\mathcal{C}_G$ for the graph $G = (V, E)$ whose vertices $V = P \cup C$ are partitioned into two sets: a path graph P and a clique C, with $|P \cap C| = 1$. (When $|P| = |C|$, G is known as the "lollipop graph" (Zhang et al., 2009).) By varying the sizes $|P|, |C|$ of the path graph P and clique C, respectively, we can affect the value of $|\breve{\mathcal{S}}(A)|$: $|\breve{\mathcal{S}}(A)|$ is exponential if $|C| = \Theta(n)$ and is sub-exponential if $|C| = o(n)$. We observe that $\lim_{n\to\infty}$ Bias$(|\widehat{A}_{\mathcal{S}}|/n) > 0$ if $|C| = \Theta(n)$, and $\lim_{n\to\infty}$ Bias$(|\widehat{A}_{\mathcal{S}}|/n) = 0$ if $|C| = o(n)$ (Figure 2B), which aligns with Conjecture 1.

In the Supplement, we describe two more experiments that support Conjecture 1. In the first experiment, we construct an anomaly family $\mathcal{S}$ where $|\mathcal{S}|$ is exponential, while $|\breve{\mathcal{S}}(A)|$ is exponential for some anomalies A and sub-exponential for others. We observe that the MLE $\widehat{A}_{\mathcal{S}}$ is biased if and only if $|\breve{\mathcal{S}}(A)|$ is exponential, providing evidence for Conjecture 1 by showing that Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ depends on the number $|\breve{\mathcal{S}}(A)|$ of sets containing the anomaly rather than the size $|\mathcal{S}|$ of the anomaly family. In the second experiment, we empirically demonstrate that the bias of the MLE $\widehat{A}_{\mathcal{T}_{G,\rho}}$ for the graph cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$ has a strong dependence on the cut-size bound ρ. This aligns with Conjecture 1 since $|\breve{\mathcal{S}}(A)|$ is polynomial when ρ is constant in n while $|\breve{\mathcal{S}}(A)|$ is exponential when ρ is close to the number $|E|$ of edges in G (Nagamochi et al., 1994).

4. Reducing Bias using Mixture Models

In the previous section, we showed that the MLE $\widehat{A}_{\mathcal{S}}$ yields a biased estimate of the size $|A|$ of the anomaly A when the number $|\breve{\mathcal{S}}(A)|$ of sets in the anomaly family $\mathcal{S}$ that

contain A is exponential in n. In this section, we derive an anomaly estimator that is less biased than the MLE. Our anomaly estimator leverages a connection between the ASD and the Gaussian mixture model (GMM), and is motivated by previous work that uses GMMs to estimate *unstructured* anomalies (Cai et al., 2007; Donoho & Jin, 2004).

Recall the following latent variable representation of the ASD: given a sample $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ from the ASD, we define a corresponding sequence $\mathbf{Z} = (Z_1, \ldots, Z_n)$ of latent variables $Z_i = 1(i \in A)$. Estimating the anomaly A is equivalent to estimating the latent variables $\mathbf{Z}$. The bias of the MLE $\widehat{A}_{\mathcal{S}}$ corresponds to *overestimating* the sum $|A| = \sum_{i=1}^{n} Z_i$ of latent variables.

This latent variable representation of the ASD is reminiscent of the latent variable representation of a Gaussian mixture model (GMM), defined as follows.

Gaussian Mixture Model (2 components, unit variance). *Let $\mu > 0$ and $\alpha \in (0, 1)$. X is distributed according to the* Gaussian Mixture Model $\text{GMM}(\mu, \alpha)$ *provided*

$$X \sim \alpha N(\mu, 1) + (1 - \alpha) N(0, 1). \tag{5}$$

Associated with X is a latent variable Z, where $Z = 1$ if X is drawn from the $N(\mu, 1)$ distribution and $Z = 0$ if X is drawn from the $N(0, 1)$ distribution.

Note that n independent observations $X_i \overset{\text{i.i.d.}}{\sim} \text{GMM}(\mu, \alpha)$ from the GMM are not equal in distribution to a sample $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ from the ASD. In particular, in the GMM all of the data points X_i are identically distributed, while in the ASD exactly $|A|$ of the data points Y_i are drawn from the $N(\mu, 1)$ distribution. Nevertheless, we observe that the empirical distributions of the unstructured ASD and the GMM converge in Wasserstein distance as $n \to \infty$ (see Supplement). In anomaly estimation, some previous approaches model *unstructured* anomalies with a GMM (Cai et al., 2007; Donoho & Jin, 2004). However, existing work on estimating *structured* anomalies typically models the data with the ASD (Arias-Castro et al., 2011; Sharpnack et al., 2013b).

Another difference between the ASD and GMM is that one can use maximum likelihood estimation to accurately estimate the sum $\sum_{i=1}^{n} Z_i$ of latent variables Z_i from GMM observations $X_i \overset{\text{i.i.d.}}{\sim} \text{GMM}(\mu, \alpha)$ (Bishop, 2006), unlike with the ASD. Specifically, Kalai et al. (2010) showed that the following algorithm gives accurate estimates of the individual latent variables Z_i: **(1)** estimate the GMM parameters μ and α and **(2)** set $Z_i = 1$ if the estimated *responsibility* $r_i = P(Z_i = 1 \mid X_i)$, or probability of X_i being drawn from the $N(\mu, 1)$ distribution, is greater than 0.5.

In practice, the parameter estimation in step **(1)** is often done by computing the MLEs $\widehat{\mu}_{\text{GMM}}$ and $\widehat{\alpha}_{\text{GMM}}$ of the GMM parameters μ and α, respectively. For data $X_i \overset{\text{i.i.d.}}{\sim} \text{GMM}(\mu, \alpha)$ drawn from a GMM, the MLEs $\widehat{\mu}_{\text{GMM}}$ and $\widehat{\alpha}_{\text{GMM}}$ are efficiently computed via the EM algorithm (Daskalakis et al., 2017; Xu et al., 2016) and are asymptotically unbiased estimators of μ and α, respectively (Chen, 2017).

Motivated by the connection between the latent variable representations of the ASD and GMM, we prove an analogous result on the asymptotic unbiasedness of the MLEs $\widehat{\mu}_{\text{GMM}}$ and $\widehat{\alpha}_{\text{GMM}}$ for data drawn from the ASD. Specifically, we prove that given data $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ with sufficiently large mean μ, then the GMM MLEs $\widehat{\mu}_{\text{GMM}}$ and $\widehat{\alpha}_{\text{GMM}}$ obtained by fitting a GMM to data $\mathbf{X}$ are asymptotically unbiased estimators of μ and $|A|/n$, respectively. This result settles a conjecture of Reyna et al. (2020).

Theorem 3. *Let $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$, where $|A| = \alpha n$ for $0 < \alpha < 0.5$ and $\mu \geq C\sqrt{\log n}$ for a sufficiently large constant $C > 0$. For sufficiently large n, we have that $|\widehat{\alpha}_{GMM} - \alpha| \leq \sqrt{\frac{\log n}{n}}$ and $|\widehat{\mu}_{GMM} - \mu| \leq 3\sqrt{\frac{\log n}{n}}$ with probability at least $1 - \frac{1}{n}$.*

A sketch of our proof of Theorem 3 is as follows. Let $B = \left\{(\widehat{\alpha}, \widehat{\mu}) : |\alpha| > \sqrt{\frac{\log n}{n}} \text{ or } |\widehat{\mu} - \mu| > \sqrt{\frac{\log n}{n}}\right\}$ be the set of all "bad" estimators $(\widehat{\alpha}, \widehat{\mu})$ of the true GMM parameters (α, μ). We show that with high probability, the GMM likelihood for all $(\widehat{\alpha}, \widehat{\mu}) \in B$ is less than the GMM likelihood for (α, μ), which implies that the GMM MLE $(\widehat{\alpha}_{\text{GMM}}, \widehat{\mu}_{\text{GMM}})$ is not in B.

4.1. A GMM-based Anomaly Estimator

Motivated by Theorem 3, we use a GMM fit to derive an asymptotically unbiased anomaly estimator for any anomaly family $\mathcal{S}$. Our approach generalizes the algorithm given in (Reyna et al., 2020) for the connected family $\mathcal{S} = \mathcal{C}_G$. Our approach is inspired by both the GMM literature discussed above and by classical statistical techniques such as the False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) and the Higher Criticism (Donoho & Jin, 2004) thresholding procedures, which identify unstructured anomalies in z-score distributions by first estimating the size of the anomalies (Jin & Cai, 2007; Cai et al., 2007; Meinshausen & Rice, 2006; Benjamini, 2010; Brennan et al., 2020).

Given data $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$, we first use the EM algorithm to fit a GMM to the data $\mathbf{X}$. This fit yields estimates $\widehat{\mu}_{\text{GMM}}, \widehat{\alpha}_{\text{GMM}}$ of the GMM parameters μ, α, respectively, as well as estimates $\widehat{r}_i$ of the responsibilities $r_i = P(Z_i = 1 \mid X_i)$. Our estimator $\widehat{A}_{\text{GMM}}$ is the set $S \in \mathcal{S}$ with size $\left||S| - \widehat{\alpha}_{\text{GMM}} n\right| \leq \sqrt{\frac{\log n}{n}}$ and having the largest total responsibility:

$$\widehat{A}_{\text{GMM}} = \underset{\substack{S \in \mathcal{S} \\ \left||S| - \widehat{\alpha}_{\text{GMM}} n\right| \leq \sqrt{\frac{\log n}{n}}}}{\text{argmax}} \left(\sum_{i \in S} \widehat{r}_i \right). \tag{6}$$

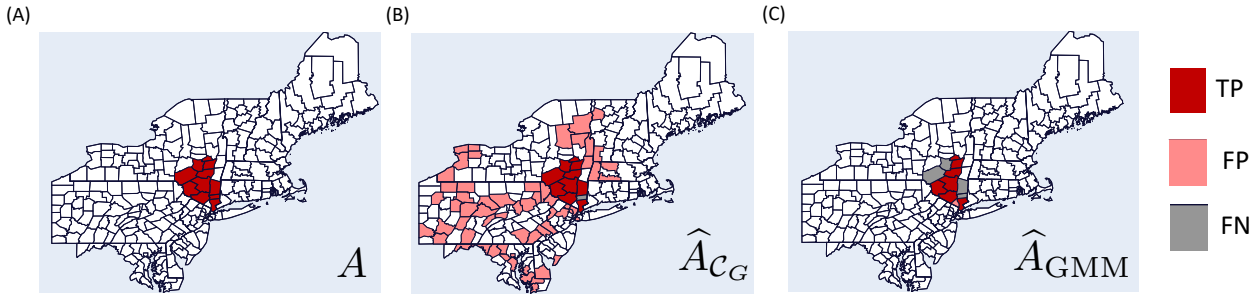

Figure 4: (A) An anomaly A containing 11 connected counties is implanted into a graph of counties in the Northeast USA (Cadena et al., 2019). (B) The MLE $\widehat{A}_{\mathcal{C}_G}$ greatly overestimates the size of the anomaly with 59 false positives (F-measure = 0.24). (C) The GMM estimator $\widehat{A}_{\text{GMM}}$ identifies 7/11 counties correctly with only 1 false positive (F-measure = 0.73).

By Theorem 3, our constraint on the size $|S|$ in (6) ensures that the size $|\widehat{A}_{\text{GMM}}|$ of the GMM-based estimator $\widehat{A}_{\text{GMM}}$ has asymptotically zero bias for sufficiently large μ. We formalize this in the following Corollary.

Corollary 1. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$, *where* $|A| = \alpha n$ *for* $0 < \alpha < 0.5$ *and* $\mu \geq C\sqrt{\log n}$ *for a sufficiently large constant* $C > 0$. *Then* $\lim_{n\to\infty} \text{Bias}(|\widehat{A}_{\mathcal{S}}|/n) = 0$.

In addition, for the unstructured family $\mathcal{S} = \mathcal{P}_n$, we show our estimator $\widehat{A}_{\text{GMM}}$ has small normalized error $\frac{|A \triangle \widehat{A}_{\text{GMM}}|}{|A|}$, as studied by Castro (2014); Castro & Tánczos (2017), where $\triangle$ is the symmetric set difference.

Corollary 2. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathit{ASD}_{\mathcal{P}_n}(A, \mu)$, *where* $|A| = \alpha n$ *for* $0 < \alpha < 0.5$ *and* $\mu \geq C\sqrt{\log n}$ *for a sufficiently large constant* $C > 0$. *Then* $\frac{|A \triangle \widehat{A}_{\mathit{GMM}}|}{|A|} \leq 2\sqrt{\frac{\log n}{n}} = o(1)$ *with probability at least* $1 - \frac{1}{n}$.

Another useful property of our estimator $\widehat{A}_{\text{GMM}}$ is that the objective $\sum_{i\in S} \widehat{r}_i$ in (6) is linear, in contrast to the non-linear objective $\sum_{i\in S} X_i/\sqrt{|S|}$ for the MLE $\widehat{A}_{\mathcal{S}}$ in Equation (3). Thus, our estimator $\widehat{A}_{\text{GMM}}$ can be efficiently computed for many anomaly families $\mathcal{S}$. For the unstructured family $\mathcal{P}_n$, $\widehat{A}_{\text{GMM}}$ can be computed in $O(n \log n)$ time by sorting the data points and returning the $\lfloor \widehat{\alpha}_{\text{GMM}} n + \sqrt{\log n/n} \rfloor$ largest ones. For the interval family $\mathcal{I}_n$, $\widehat{A}_{\text{GMM}}$ can be computed in $O(n)$ time by scanning over all intervals of size $\lfloor \widehat{\alpha}_{\text{GMM}} n + \sqrt{\log n/n} \rfloor$. For the graph cut family $\mathcal{T}_{G,\rho}$, Sharpnack et al. (2013a) shows that (6) can be efficiently solved with a convex program through the use of Lovász extensions (Bach, 2010).

More generally, when the constraint $S \in \mathcal{S}$ can be expressed with linear constraints, one can compute $\widehat{A}_{\text{GMM}}$ with an Integer Linear Program (ILP). This is true for anomaly families including the submatrix family $\mathcal{M}_N$, the graph cut family $\mathcal{T}_{G,\rho}$ (Sharpnack et al., 2013b), and the connected family $\mathcal{C}_G$ (Dittrich et al., 2008; Reyna et al., 2020). In practice, we found that directly computing (6) via ILP could sometimes be inefficient for the submatrix and connected families, and in the Supplement we derive an approximation to (6) that can be efficiently computed for these families.

4.2. Experiments

First, we compare the performance of our estimator $\widehat{A}_{\text{GMM}}$ to the MLE $\widehat{A}_{\mathcal{S}}$ for the anomaly families $\mathcal{S}$ from Section 3.1. We observe that $\text{Bias}(|\widehat{A}_{\text{GMM}}|/n) \approx 0$ for all means $\mu \geq \mu_{\text{detect}}$ and across many anomaly families $\mathcal{S}$ (Figure 2C). We also observe that $\lim_{n\to\infty} \text{Bias}(|\widehat{A}_{\text{GMM}}|/n) = 0$ no matter if $|\breve{\mathcal{S}}(A)|$ is exponential or sub-exponential (Figure 2D). This empirically demonstrates Theorem 3 by showing that $|\widehat{A}_{\text{GMM}}|$ is an asymptotically unbiased estimator of the anomaly size $|A|$ for sufficiently large μ regardless of the number $|\breve{\mathcal{S}}(A)|$ of sets containing the anomaly A.

Next, we simulate a disease outbreak on the Northeastern USA Benchmark (NEast) graph, a standard benchmark for estimating spatial anomalies (Cadena et al., 2018a; 2019). The NEast graph $G = (V, E)$ is a graph whose nodes are the $n = 244$ counties in the northeastern part of the USA (Kulldorff et al., 2003) with edges connecting adjacent counties. Similar to (Cadena et al., 2018a; Aksoylar et al., 2017; Qian & Saligrama, 2014), we implant a connected anomaly $A \in \mathcal{C}_G$ of size $|A| = 11$ and we draw a sample $\mathbf{X} \sim \text{ASD}_{\mathcal{C}_G}(A, 2)$. Because existing methods for estimating anomalous subgraphs typically compute the MLE $\widehat{A}_{\mathcal{C}_G}$ (Chen & Neill, 2014; Qian & Saligrama, 2014; Cadena et al., 2019), we also compare our estimator to the MLE. We find (Figure 4) that the MLE $\widehat{A}_{\mathcal{C}_G}$ greatly overestimates the size $|A|$ of the anomaly A, with many more false positives compared to the GMM estimator $\widehat{A}_{\text{GMM}}$.

We also compare our estimator $\widehat{A}_{\text{GMM}}$ and the MLE $\widehat{A}_{\mathcal{S}}$ on a real-world highway traffic dataset; similar to the NEast graph, this dataset is also often studied in the scan statistic literature (Zhou & Chen, 2016; Cadena et al., 2018a; 2019).

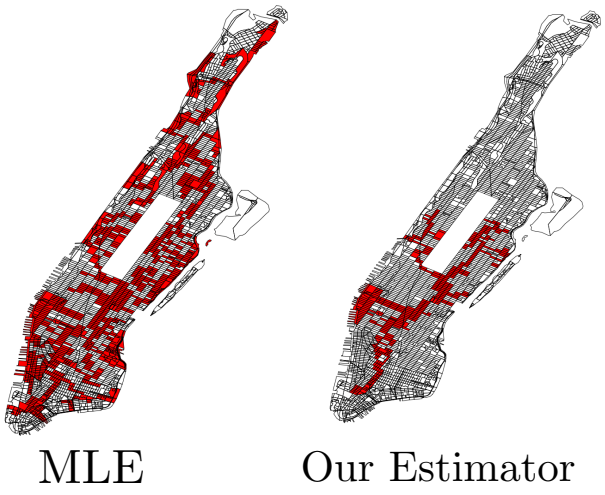

Figure 5: Comparison of the MLE, also known as the graph scan statistic (left), and our estimator (right) in estimating connected disease clusters in data of breast cancer incidence in Manhattan. Our estimator computes a smaller cluster than the MLE/graph scan statistic but with a 20% higher average incidence ratio (relative risk).

This dataset consists of a highway traffic network $G = (V, E)$ in Los Angeles County, CA with $|V| = 1868$ vertices and $|E| = 1993$ edges. The vertices V are sensors that record the speed of cars passing and the edges E connect adjacent sensors. The observations $\mathbf{X} = (X_v)_{v \in V}$ are p-values (where sensors that record higher average speeds have lower p-values) that are transformed to Gaussians using the method in Reyna et al. (2020).

For the connected family $\mathcal{C}_G$, we find that our estimator $\widehat{A}_{\text{GMM}}$ is much smaller than the MLE $\widehat{A}_{\mathcal{C}_G}$ ($|\widehat{A}_{\text{GMM}}| = 17$ versus $|\widehat{A}_{\mathcal{C}_G}| = 140$) but with higher average score (2.6 for our estimator versus 0.55 for the MLE). While there is no ground-truth anomaly in this dataset, our results show that our estimator $\widehat{A}_{\text{GMM}}$ yields a smaller anomaly but with higher average values than the MLE $\widehat{A}_{\mathcal{C}_G}$, consistent with the theoretical results in Section 3 that the MLE $\widehat{A}_{\mathcal{C}_G}$ is a biased estimator. In the Supplement, we show similar results comparing our estimator $\widehat{A}_{\text{GMM}}$ and the MLE $\widehat{A}_{\mathcal{S}}$ for the edge-dense family $\mathcal{S} = \mathcal{E}_{G,0.7}$. Since the goal of anomaly estimation in this application is to identify portions of roads with or without high traffic volume, the large and biased anomaly estimates produced by the MLE may not be useful for traffic studies.

We also compared our estimator and the MLE on a dataset of breast cancer incidence in census blocks in Manhattan (Boscoe et al., 2016) using the connected family $\mathcal{C}_G$. This dataset is typically modeled with Poisson distributions, and we accordingly adapted the MLE and our estimator to such Poisson distributions (see Supplement for details). In this setting, the MLE is also known as a graph scan statistic (Cadena et al., 2019). We find that our estimator identifies a much smaller connected cluster of breast cancer cases compared to the MLE/graph scan statistic (182 census blocks vs 382 Figure 5) but with a 20% higher cancer incidence rate, again demonstrating the bias of the MLE.

5. Conclusion

We study the problem of estimating structured anomalies. We formulate this problem as the problem of estimating a parameter of the Anomalous Subset Distribution (ASD), with the structure of the anomaly described by an anomaly family. We demonstrate that the Maximum Likelihood Estimator (MLE) of the size of this parameter is biased if and only if the number of sets in the anomaly family containing the anomaly is exponential. These results unify existing results for specific anomaly families including intervals, submatrices, and connected subgraphs. Next, we develop an asymptotically unbiased estimator using a Gaussian mixture model (GMM), and empirically demonstrate the advantages of our estimator on both simulated and real datasets.

Our work opens up a number of future directions. First, it would be highly desirable to provide a complete proof of Conjecture 1. A second direction is to generalize the ASD to more than one anomaly in a dataset by building on existing work for the interval family (Jeng et al., 2010) and the submatrix family (Chen & Xu, 2016). One potential algorithm for identifying multiple anomalies is to fit a k-component GMM to the data and sequentially compute each anomaly. A third direction is to generalize our theoretical results to other distributions, e.g. Poisson distributions, which are commonly used to model anomalies in integer-valued data (Cadena et al., 2019; Liu & Arias-Castro, 2019; Kulldorff, 1997). While our GMM-based estimator is easily adapted to other distributions, one challenge in studying bias is that the MLE does not necessarily have a simple form like it does for Gaussian distributions. These directions would strengthen the theoretical foundations for further applications of anomaly estimation.

Acknowledgments

The authors would like to thank Allan Sly for helpful discussions, and Baojian Zhou and Martin Zhu for assistance with running the Graph-GHTP code (Zhou & Chen, 2016). U.C. is supported by NSF GRFP DGE 2039656. J.C.H.L. is partially supported by NSF award IIS-1562657. B.J.R. is supported by a US National Institutes of Health (NIH) grant U24CA211000.

References

Adams, R. P. and MacKay, D. J. C. Bayesian online changepoint detection. *CoRR*, abs/0710.3742, 2007. URL http://arxiv.org/abs/0710.3742.

Aksoylar, C., Orecchia, L., and Saligrama, V. Connected subgraph detection with mirror descent on SDPs. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 51–59, 2017.

Arias-Castro, E., Donoho, D. L., and Xiaoming Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory*, 51 (7):2402–2425, July 2005. ISSN 1557-9654.

Arias-Castro, E., Candès, E. J., and Durand, A. Detection of an anomalous cluster in a network. *Ann. Statist.*, 39(1): 278–304, 02 2011.

Bach, F. Convex analysis and optimization with submodular functions: a tutorial. *CoRR*, abs/1010.4207, 2010. URL http://arxiv.org/abs/1010.4207.

Banks, J., Moore, C., Vershynin, R., Verzelen, N., and Xu, J. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, July 2018. ISSN 1557-9654.

Basseville, M. and Nikiforov, I. V. *Detection of Abrupt Changes - Theory and Application*. Prentice Hall, Inc., 1993.

Benjamini, Y. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):405–416, 2010.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Boscoe, F. P., Talbot, T. O., and Kulldorff, M. Public domain small-area cancer incidence data for new york state, 2005-2009. *Geospatial health*, 11(1):304–304, 04 2016.

Brennan, J., Vinayak, R. K., and Jamieson, K. Estimating the number and effect sizes of non-null hypotheses. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1123–1133. PMLR, 13–18 Jul 2020.

Brennan, M., Bresler, G., and Huleihel, W. Universality of computational lower bounds for submatrix detection. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 417–468, 2019.

Butucea, C. and Ingster, Y. I. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B): 2652–2688, 11 2013.

Cadena, J., Basak, A., Vullikanti, A. K. S., and Deng, X. Graph scan statistics with uncertainty. In *AAAI*, 2018a.

Cadena, J., Chen, F., and Vullikanti, A. Graph anomaly detection based on steiner connectivity and density. *Proceedings of the IEEE*, 106(5):829–845, 2018b.

Cadena, J., Chen, F., and Vullikanti, A. Near-optimal and practical algorithms for graph scan statistics with connectivity constraints. *ACM Trans. Knowl. Discov. Data*, 13 (2):20:1–20:33, April 2019. ISSN 1556-4681.

Cai, T. T., Jin, J., and Low, M. G. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 12 2007.

Castro, R. M. Adaptive sensing performance lower bounds for sparse signal detection and support estimation. *Bernoulli*, 20(4):2217–2246, 11 2014.

Castro, R. M. and Tánczos, E. Adaptive compressed sensing for support recovery of structured sparse sets. *IEEE Transactions on Information Theory*, 63(3):1535–1554, 2017.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), July 2009. ISSN 0360-0300.

Chen, F. and Neill, D. B. 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*, KDD '14, pp. 1166–1175, 2014.

Chen, J. Consistency of the mle under mixture models. *Statist. Sci.*, 32(1):47–63, 02 2017.

Chen, Y. and Xu, J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, 17(1):882–938, January 2016. ISSN 1532-4435.

Daskalakis, C., Tzamos, C., and Zampetakis, M. Ten steps of em suffice for mixtures of two gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 704–710, 2017.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, 2008. ISSN 1367-4803.

Donoho, D. and Jin, J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 06 2004.

Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 03 1993. ISSN 0006-3444.

Gamarnik, D., Jagannath, A., and Sen, S. The overlap gap property in principal submatrix recovery. *CoRR*, abs/1908.09959, 2019. URL http://arxiv.org/abs/1908.09959.

Giles, D. E., Feng, H., and Godwin, R. T. On the bias of the maximum likelihood estimator for the two-parameter lomax distribution. *Communications in Statistics - Theory and Methods*, 42(11):1934–1950, 2013. doi: 10.1080/03610926.2011.600506.

Glaz, J. and Naus, J. I. *Scan Statistics*, pp. 1–7. American Cancer Society, 2010. ISBN 9780471667193.

Goldreich, O., Goldwasser, S., and Ron, D. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, July 1998. ISSN 0004-5411.

Hajek, B., Wu, Y., and Xu, J. Information limits for recovering a hidden community. *IEEE Transactions on Information Theory*, 63(8):4729–4745, Aug 2017. ISSN 1557-9654.

Hartigan, J. A. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

Hinkley, D. V. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 04 1970. ISSN 0006-3444.

Hinkley, D. V. Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523, 12 1971. ISSN 0006-3444.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl_1): S233–S240, 2002. ISSN 1367-4803.

Jeng, X. J., Cai, T. T., and Li, H. Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association*, 105(491):1156–1166, 2010.

Jin, J. and Cai, T. T. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478): 495–506, 06 2007.

Kalai, A. T., Moitra, A., and Valiant, G. Efficiently learning mixtures of two gaussians. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, pp. 553–562, New York, NY, USA, 2010. ISBN 9781450300506.

Kolar, M., Balakrishnan, S., Rinaldo, A., and Singh, A. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems 24*, pp. 909–917. 2011.

Krishnamurthy, A. Minimax structured normal means inference. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 960–964, July 2016.

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

Kulldorff, M., Tango, T., and Park, P. J. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665 – 684, 2003.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 420–429, 2007.

Liu, Y. and Arias-Castro, E. A multiscale scan statistic for adaptive submatrix localization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 44–53, 2019.

Ma, Z. and Wu, Y. Computational barriers in minimax submatrix detection. *Ann. Statist.*, 43(3):1089–1116, 06 2015.

Mardia, K., Southworth, H., and Taylor, C. On bias in maximum likelihood estimators. *Journal of Statistical Planning and Inference*, 76(1):31–39, 1999. ISSN 0378-3758.

Meinshausen, N. and Rice, J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1):373–393, 02 2006.

Nagamochi, H., Nishimura, K., and Ibaraki, T. Computing all small cuts in undirected networks. In Du, D.-Z.

and Zhang, X.-S. (eds.), *Algorithms and Computation*, pp. 190–198, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg. ISBN 978-3-540-48653-4.

Neill, D. B. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

Neill, D. B. and Moore, A. W. A fast multi-resolution method for detection of significant spatial disease clusters. In *Advances in Neural Information Processing Systems 16*, pp. 651–658. 2004.

Neill, D. B., Moore, A. W., and Cooper, G. F. A bayesian spatial scan statistic. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pp. 1003–1010, Cambridge, MA, USA, 2005. MIT Press.

Nikolayeva, I., Pla, O. G., and Schwikowski, B. Network module identification—a widespread theoretical bias and best practices. *Methods*, 132:19 – 25, 2018. ISSN 1046-2023.

Page, E. S. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3-4):523–527, 12 1955. ISSN 0006-3444.

Qian, J. and Saligrama, V. Efficient minimax signal detection on graphs. In *Advances in Neural Information Processing Systems 27*, pp. 2708–2716. 2014.

Qian, J., Saligrama, V., and Chen, Y. Anomalous cluster detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3854–3858, May 2014.

Reyna, M. A., Chitra, U., Elyanow, R., and Raphael, B. J. Netmix: A network-structured mixture model for reduced-bias estimation of altered subnetworks. In *Research in Computational Molecular Biology*, pp. 169–185, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45257-5.

Sharpnack, J., Krishnamurthy, A., and Singh, A. Near-optimal anomaly detection in graphs using lovász extended scan statistic. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 1959–1967, 2013a.

Sharpnack, J., Singh, A., and Rinaldo, A. Changepoint detection over graphs with the spectral scan statistic. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 545–553, 2013b.

Sharpnack, J., Rinaldo, A., and Singh, A. Detecting anomalous activity on networks with the graph fourier scan statistic. *IEEE Transactions on Signal Processing*, 64(2): 364–379, Jan 2016. ISSN 1941-0476.

Tanay, A., Sharan, R., and Shamir, R. Biclustering algorithms: A survey. In *In Handbook of Computational Molecular Biology*, 2005.

Vince, A. Counting connected sets and connected partitions of a graph. *Australasian Journal Of Combinatorics*, 67 (2):281–293, 2017.

Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 808–815. AAAI Press, 2003. ISBN 1577351894.

Xu, J., Hsu, D., and Maleki, A. Global analysis of expectation maximization for mixtures of two gaussians. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2684–2692, 2016. ISBN 9781510838819.

Zhai, S., Cheng, Y., Lu, W., and Zhang, Z. Deep structured energy based models for anomaly detection. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1100–1109, 2016.

Zhang, Y., Liu, X., Zhang, B., and Yong, X. The lollipop graph is determined by its q-spectrum. *Discrete Mathematics*, 309(10):3364 – 3369, 2009. ISSN 0012-365X.

Zhou, B. and Chen, F. Graph-structured sparse optimization for connected subgraph detection. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 709–718, 2016.