NetMix: A network-structured mixture model for reduced-bias estimation of altered subnetworks

Matthew A. Reyna * [1,2], Uthsav Chitra * [1], Rebecca Elyanow [1,3], Benjamin J. Raphael [1]

[1]Department of Computer Science, Princeton University, Princeton, NJ 08540
[2]Department of Biomedical Informatics, Emory University, Atlanta, GA 30306
[3]Department of Computer Science, Brown University, Providence, RI 02912

Abstract

A classic problem in computational biology is the identification of *altered subnetworks*: subnetworks of an interaction network that contain genes/proteins that are differentially expressed, highly mutated, or otherwise aberrant compared to other genes/proteins. Numerous methods have been developed to solve this problem under various assumptions, but the statistical properties of these methods are often unknown. For example, some widely-used methods are reported to output very large subnetworks that are difficult to interpret biologically. In this work, we formulate the identification of altered subnetworks as the problem of estimating the parameters of a class of probability distributions which we call the Altered Subset Distribution (ASD). We derive a connection between a popular method, jActiveModules, and the maximum likelihood estimator (MLE) of the ASD. We show that the MLE is *statistically biased*, explaining the large subnetworks output by jActiveModules. We introduce NetMix, an algorithm that uses Gaussian mixture models to obtain less biased estimates of the parameters of the ASD. We demonstrate that NetMix outperforms existing methods in identifying altered subnetworks on both simulated and real data, including the identification of differentially expressed genes from both microarray and RNA-seq experiments and the identification of cancer driver genes in somatic mutation data.
Availability: NetMix is available online at https://github.com/raphael-group/netmix.
Contact: braphael@princeton.edu

1 Introduction

A standard paradigm in computational biology is to use interaction networks as prior knowledge in the analysis of high-throughput 'omics data, with applications in protein function prediction [79, 73, 65, 25, 18], gene expression [32, 91, 16, 48, 27], germline variants [55, 12, 56, 43, 45], somatic variants in cancer [66, 87, 57, 84, 64, 42], and other data [39, 10, 20, 89, 35, 77, 13, 60]. One classic approach is to identify *active*, or *altered*, subnetworks of an interaction network that contain outlier measurements. The altered subnetwork problem takes as input: (1) an interaction network whose nodes are biological entities (e.g., genes or proteins) and whose edges represent biological interactions (e.g., physical or genetic interactions, co-expression, etc.); and (2) a measurement or score for each node. The goal is to find high-scoring subnetworks that correspond to functionally related or correlated alterations. This problem was introduced in [48] for gene expression analysis, where gene scores were derived from p-values of differential expression. [48] developed the jActiveModules algorithm to solve this problem and identify altered subnetworks of differentially expressed genes. Subsequently, [27] introduced heinz as "the first approach that really tackles and solves the original problem raised by [48] to optimality." jActiveModules and heinz have become

* These authors contributed equally.

widely-used tools with diverse applications; a few recent examples include mass-spectrometry proteomics [51, 58], damaging *de novo* mutations in schizophrenia and other neurological disorders [36, 17], and single-cell RNA-seq [37, 85, 52].

In the past two decades, many algorithms have been developed to identify altered subnetworks in biological data (reviewed in [26, 20, 63, 64]). Each publication describing a new algorithm demonstrates the performance of their algorithm on specific biological datasets, and many of these publications also benchmark their algorithm against existing algorithms on real and/or simulated data. However, few of these publications prove theoretical guarantees for their algorithm's performance on a well-defined generative model of the data. Thus, the true performance of these algorithms is often unknown. Indeed, recent benchmarking studies (e.g., [40, 9]) of several widely used network algorithms – including jActiveModules and heinz – show considerate disagreement between subnetworks identified by different methods on the same biological datasets. Moreover, these benchmarking studies (and many others) do not compare network algorithms against single-gene tests that do not use the network; thus, the tacit assumption that interaction networks always improve gene prioritization is often not tested.

Separately, many publications in the statistics and machine learning literature investigate the problem of *detecting* whether or not a network contains an anomalous subnetwork, or a *network anomaly*, e.g., [6, 4, 1, 3, 83, 82, 81, 80, 5]. These papers describe specific generative models of network anomalies and use a rigorous hypothesis-testing framework to prove asymptotic results regarding the conditions under which it is possible to detect a network anomaly. Importantly, these papers also provide theoretical guarantees about conditions under which a network contributes to anomaly detection. However, the network anomaly literature does not specifically address the altered subnetwork problem studied in computational biology, with three key differences. First, the *detection* problem of deciding whether or not an altered subnetwork exists is not the same as the *estimation* problem of identifying the nodes in an altered subnetwork. Second, biological networks have a finite size, and it is unclear what guarantees the asymptotic results provide for finite-size networks. Finally, the topological constraints on network anomalies are different from those considered in computational biology.

In this paper, we aim to bridge the gap between the theoretical guarantees in the network anomaly literature and the practical problem of identifying altered subnetworks in biological data. We provide a rigorous formulation of the *Altered Subnetwork Problem*, the problem that jActiveModules [48], heinz [27], and other methods aim to solve. Our formulation of the Altered Subnetwork Problem is inspired by the generative model used in the network anomaly literature, but requires that the altered subnetwork is a connected subnetwork, a constraint motivated by the topology of signaling pathways [11, 50] and by the seminal works of [48] and [27].

We show that the Altered Subnetwork Problem is equivalent to estimating the parameters of a distribution which we define as the *Altered Subset Distribution (ASD)*. We prove that the jActiveModules problem [48] is equivalent to finding a maximum likelihood estimator (MLE) of the parameters of the ASD for connected subgraphs. At the same time, we demonstrate that if (1) the size of the altered subnetwork is moderately small and (2) the scores of nodes inside and outside of the altered subnetwork are not well-separated, then the MLE is a *biased* estimator of the size of the altered subnetwork. This statistical bias provides a rigorous explanation for the large subnetworks produced by jActiveModules [48]. We also show that the size of the altered subnetworks identified by heinz [27] are biased for most choices of its user-defined parameter.

We introduce a new algorithm, NetMix, that combines a Gaussian mixture model and a combinatorial optimization algorithm to identify altered subnetworks. We show that NetMix is a reduced-bias estimator of the size of the altered subnetwork. We demonstrate that NetMix outperforms other methods for identifying altered subnetworks on simulated data, gene expression data, and somatic mutation data.

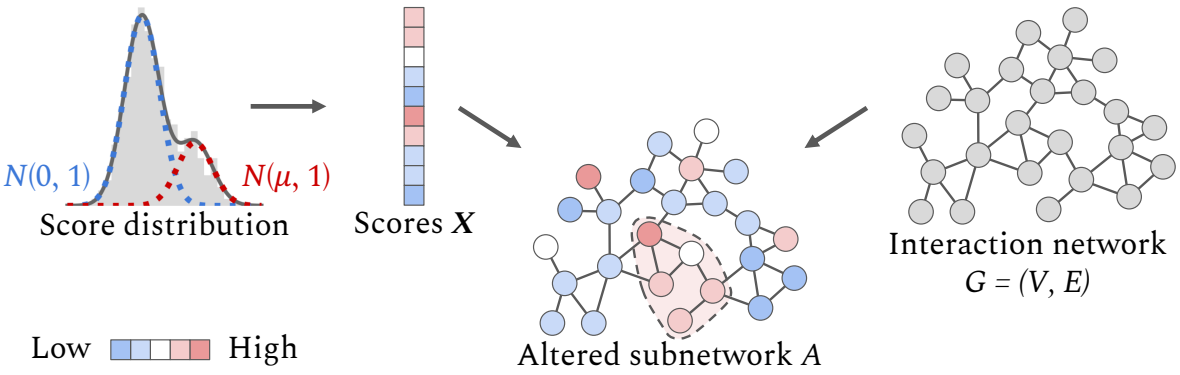

Figure 1: Altered Subnetwork Problem. Measurements, or scores, $\mathbf{X}$ from a high-throughput experiment are drawn from one of two distributions: genes/proteins in an altered subnetwork A of an interaction network $G = (V, E)$ have scores drawn from an altered distribution $N(\mu, 1)$ with $\mu > 0$, while genes/proteins not in A have scores drawn from a background distribution $N(0, 1)$. The difficulty in identifying A depends on the separation μ between the distributions and the size $|A|$ of the altered subnetwork.

2 Altered Subnetworks, Altered Subsets, and Maximum Likelihood Estimation

2.1 Altered Subnetwork Problem

Let $G = (V, E)$ be a biological interaction network with a measurement, or score, X_v for each vertex $v \in V$. We assume that there is a connected subnetwork A in G, the *altered subnetwork*, whose scores are derived from a different distribution than the scores of the vertices not in A (Figure 1). The goal of the Altered Subnetwork Problem is to find A. The problem is defined formally as follows.

Altered Subnetwork Problem (ASP). *Let $G = (V, E)$ be a graph with vertex scores $\mathbf{X} = (X_v)_{v \in V}$, and let $A \subseteq V$ be a connected subgraph of G. Suppose that*

$$X_v \overset{\text{i.i.d.}}{\sim} \begin{cases} D_A, & \text{if } v \in A, \\ D_B, & \text{if } v \in V \backslash A, \end{cases} \tag{1}$$

where D_A is the altered *distribution and D_B is the* background *distribution. Given G and $\mathbf{X}$, find A.*

Note that the ASP assumes that the network G has a *single* altered subnetwork A. When the network has multiple altered subnetworks, one can recursively solve the ASP to identify more than one altered subnetwork.

The seminal algorithm for solving the ASP is jActiveModules [48]. jActiveModules takes as input a p-value p_v for each vertex v; e.g., a p-value of differential gene expression. Under the null hypothesis, the p-values p_v across genes are distributed according to the uniform distribution $U(0, 1)$. jActiveModules transforms the p-values into scores $X_v = \Phi^{-1}(1 - p_v)$, where Φ is the CDF of a standard normal distribution. Thus, jActiveModules solves the ASP with background distribution $D_B = N(0, 1)$. jActiveModules aims to find a connected subgraph $\hat{A}$ that maximizes[1] $\Gamma(S) = \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v$, i.e.,

$$\hat{A} = \underset{\text{connected } S \subseteq V}{\operatorname{argmax}} \Gamma(S) = \underset{\text{connected } S \subseteq V}{\operatorname{argmax}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v. \tag{2}$$

[1] jActiveModules actually maximizes $\Gamma_{\text{norm}}(S) = (\Gamma(S) - \mu_{|S|})/\sigma_{|S|}$, a z-score normalized version of $\Gamma(S)$, where $\mu_{|S|}$ and $\sigma_{|S|}$ are the mean and standard deviation, respectively, of $\Gamma(T)$ over all subsets $T \subseteq V$ of size $|S|$. We show in the supplement that maximizing $\Gamma_{\text{norm}}(S)$ is equivalent to maximizing the unnormalized $\Gamma(S)$ when the data is generated from normal distributions.

The presentation of jActiveModules in [48] does not specify the altered distribution D_A. However, in Section 2.2, we argue that the choice of the objective function in (2) implicitly assumes that $D_A = N(\mu, 1)$ for some parameter $\mu > 0$. Thus, we define the normally distributed ASP as follows.

Normally Distributed Altered Subnetwork Problem. *Let $G = (V, E)$ be a graph with vertex scores $\mathbf{X} = (X_v)_{v \in V}$, and let $A \subseteq V$ be a connected subgraph of G. Suppose that for some $\mu > 0$,*

$$X_v \overset{\text{i.i.d.}}{\sim} \begin{cases} N(\mu, 1), & \text{if } v \in A, \\ N(0, 1), & \text{if } v \in V \backslash A. \end{cases} \tag{3}$$

Given G and $\mathbf{X}$, find A.

The Normally Distributed ASP has a sound statistical interpretation: if the p-values p_v of the genes are derived from an asymptotically normal test statistic, as is often the case, then the transformed p-values $X_v = \Phi^{-1}(1 - p_v)$ are distributed as $N(0, 1)$ for genes satisfying the null hypothesis and $N(\mu, 1)$ for genes satisfying the alternative hypothesis [46]. Normal distributions also have been used to model transformed p-values from differential gene expression experiments [69, 61, 90].

More generally, the Normally Distributed Altered Subnetwork Problem is related to a larger class of *network anomaly* problems, which have been studied extensively in the machine learning and statistics literature [6, 4, 1, 3, 83, 82, 81, 80, 5]. To better understand the relationships between these problems and the algorithms developed to solve them, we will describe a generalization of the Altered Subnetwork Problem. We start by defining the following distribution, which generalizes the connected subnetworks in the Normally Distributed Altered Subnetwork Problem to any family of altered subsets.

Normally Distributed Altered Subset Distribution (ASD). *Let $n > 0$ be a positive integer, let $\mathcal{S}$ be a family of subsets of $\{1, \ldots, n\}$, and let $A \in \mathcal{S}$. $\mathbf{X} = (X_1, \ldots, X_n)$ is distributed according to the* Normally Distributed Altered Subset Distribution $\text{ASD}_{\mathcal{S}}(A, \mu)$ *provided*

$$X_i \overset{\text{i.i.d.}}{\sim} \begin{cases} N(\mu, 1), & \text{if } i \in A, \\ N(0, 1), & \text{if } i \notin A. \end{cases} \tag{4}$$

Here, $\mu > 0$ is the mean *of the ASD and A is the* altered *subset of the ASD.*

More generally, the Altered Subset Distribution can be defined for any background distribution D_B and altered distribution D_A. We will restrict ourselves to normal distributions in accordance with the Normally Distributed Altered Subnetwork Problem, and we will subsequently assume normal distributions in both the Altered Subset Distribution and the Altered Subnetwork Problem.

The distribution in the Altered Subnetwork Problem is the $\text{ASD}_{\mathcal{S}}(A, \mu)$, where the family $\mathcal{S}$ of subsets are connected subgraphs of the network G. In this terminology, the Altered Subnetwork Problem is the problem of estimating the parameters A and μ of the Altered Subset Distribution given data $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ and knowledge of the parameter space $\mathcal{S}$ of altered subnetworks A. Thus, we generalize the Altered Subnetwork Problem to the ASD Estimation Problem, defined as follows.

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

The ASD Estimation Problem is a general problem of estimating the parameters of a *structured* alternative distribution. Different choices of $\mathcal{S}$ for the ASD Estimation Problem yield a number of interesting problems, some of which have been previously studied.

- $\mathcal{S} = \mathcal{P}_n$, the power set of all subsets of $\{1, \ldots, n\}$. We call the distribution $\text{ASD}_{\mathcal{P}_n}(A, \mu)$ the *unstructured* ASD.

- $\mathcal{S} = \mathcal{C}_G$, the set of all connected subgraphs of a graph $G = (V, E)$. We call $\text{ASD}_{\mathcal{C}_G}(A, \mu)$ the *connected* ASD. The connected ASD Estimation Problem is equivalent to the Altered Subnetwork Problem described above.
- $\mathcal{S} = \mathcal{D}_G(\rho)$, the set of all subgraphs of a graph $G = (V, E)$ with edge density $\geqslant \rho$. [38, 88, 7] identify altered subnetworks with high edge density, and [2] identifies altered subnetworks with edge density $\rho = 1$, i.e., cliques.
- $\mathcal{S} = \mathcal{N}_G = \{\mathcal{N}(v) : v \in V\}$, the set of all first-order network neighborhoods of a graph $G = (V, E)$. [15, 44] use first-order network neighborhoods to prioritize cancer genes.
- $\mathcal{S} \subset \mathcal{P}_n$, a family of subsets. Typically, $|\mathcal{S}| \ll |\mathcal{P}_n|$ and $\mathcal{S}$ is not defined in terms of a graph. A classic example is gene set analysis; see [47] for a review.

2.2 Bias in Maximum Likelihood Estimation of the ASD

One reasonable approach for solving the ASD Estimation Problem is to compute a maximum likelihood estimator (MLE) for the parameters of the ASD. We derive the MLE below and show that it has undesirable statistical properties. All proofs are in the supplement.

Theorem 1. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$. *The maximum likelihood estimators (MLEs)* $\hat{A}_{ASD}$ *and* $\hat{\mu}_{ASD}$ *of* A *and* μ, *respectively, are*

$$\hat{A}_{ASD} = \underset{S \in \mathcal{S}}{\text{argmax}}\, \Gamma(S) = \underset{S \in \mathcal{S}}{\text{argmax}}\, \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \quad \textit{and} \quad \hat{\mu}_{ASD} = \frac{1}{|\hat{A}_{ASD}|} \sum_{v \in \hat{A}_{ASD}} X_v. \tag{5}$$

The maximization of Γ over $\mathcal{S}$ in (5) is a version of the *scan statistic*, a commonly used statistic to study point processes on lines and rectangles under various distributions [53, 34]. Comparing (5) and (2), we see that jActiveModules [48] computes the scan statistic over the family $\mathcal{S} = \mathcal{C}_G$ of connected subgraphs of the graph G. Thus, although jActiveModules [48] neither specifies the anomalous distribution D_A nor provides a statistical justification for their subnetwork scoring function, Theorem 1 above shows that jActiveModules implicitly assumes that D_A is a normal distribution, and that jActiveModules aims to solve the Altered Subnetwork Problem by finding the MLE $\hat{A}_{\text{ASD}}$.

Despite this insight that jActiveModules computes the MLE, it has been observed that jActiveModules often identifies large subnetworks. [67] notes that the subnetworks identified by jActiveModules are large and "hard to interpret biologically". They attribute the tendency of jActiveModules to identify large subnetworks to the fact that a graph typically has more large subnetworks than small ones. While this observation about the relative numbers of subnetworks of different sizes is correct, we argue that this tendency of jActiveModules to identify large subnetworks is due to a more fundamental reason: the MLE $\hat{A}_{\text{ASD}}$ is a *biased* estimator of A.

First, we recall the definitions of bias and consistency for an estimator $\hat{\theta}$ of a parameter θ.

Definition 1. *Let* $\hat{\theta} = \hat{\theta}(\mathbf{X})$ *be an estimator of a parameter* θ *given observed data* $\mathbf{X} = (X_1, \ldots, X_n)$. *(a) The* bias *in the estimator* $\hat{\theta}$ *of* θ *is* $\text{Bias}_\theta(\hat{\theta}) = E[\hat{\theta}] - \theta$. *We say that* $\hat{\theta}$ *is a* biased *estimator of* θ *if* $\text{Bias}_\theta(\hat{\theta}) \neq 0$, *and is an* unbiased *estimator of* θ *otherwise. (b) We say that* $\hat{\theta}$ *is a* consistent *estimator of* θ *if* $\hat{\theta} \xrightarrow{p} \theta$, *where* $\xrightarrow{p}$ *denotes convergence in probability as* $n \to \infty$, *and is an* inconsistent *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 $\hat{\theta}$.

Let $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ be distributed according to the unstructured ASD. We observe that the estimators $|\hat{A}_{\text{ASD}}|/n$ and $\hat{\mu}_{\text{ASD}}$ are both biased and inconsistent when both $|A|/n$ and μ are moderately small (Figure 2). We summarize these observations in the following conjecture.

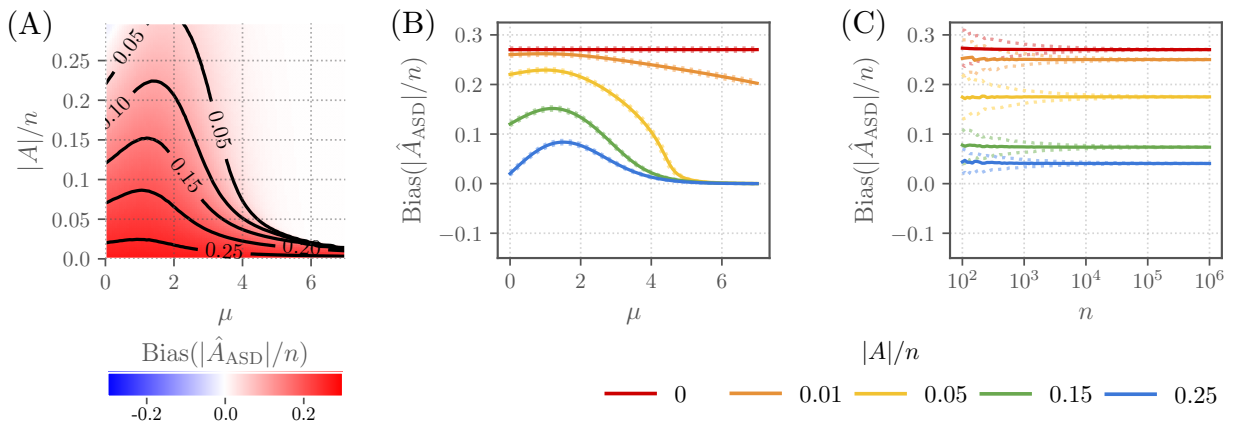

Figure 2: Scores $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ are distributed according to the unstructured ASD. (A) $\text{Bias}(|\hat{A}_{\text{ASD}}|/n)$ in the maximum likelihood estimate of $|A|/n$ as a function of the mean μ and altered subset size $|A|/n$ for $n = 10^4$. (B) $\text{Bias}(|\hat{A}_{\text{ASD}}|/n)$ for $n = 10^4$ and several values of $|A|/n$. Dotted lines indicate first and third quartiles in the estimate of the bias. (C) $\text{Bias}(|\hat{A}_{\text{ASD}}|/n)$ as a function of n for $\mu = 3$ and for several values of $|A|/n$.

Conjecture. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$. *Then there exist* $\mu_0 > 0$ *and* $\beta > 0$ *such that, if* $\mu < \mu_0$ *and* $|A|/n < \beta$, *then* $|\hat{A}_{ASD}|/n$ *and* $\hat{\mu}_{ASD}$ *are biased and inconsistent estimators of* $|A|/n$ *and* μ, *respectively.*

Note that there are many examples in the literature of biased MLEs; e.g., the MLE for the variance of a (univariate) normal distribution or the MLE for the inverse of the mean of a Poisson distribution [30]. However, examples of inconsistent MLEs are somewhat rare [29].

Although we do not have a proof of the above conjecture, we prove the following results that partially explain the bias and inconsistency of the estimators $|A_{\text{ASD}}|$ and μ_{ASD}. For the bias, we prove the following.

Theorem 2. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ *with* $A = \varnothing$. *Then* $|\hat{A}_{ASD}| = cn$ *for sufficiently large* n *and with high probability, where* $0 < c < 0.35$ *is independent of* n.

Empirically, we observe $c \approx 0.27$, i.e., $\hat{A}_{\text{ASD}}$ contains more than a quarter of the scores (Figure 2). This closely aligns with the observation in [67] that jActiveModules reports subnetworks that contain approximately 29% of all nodes in the graph. Based on Theorem 2, one may suspect that $|\hat{A}_{\text{ASD}}| \approx cn$ when μ or $|A|/n$ is sufficiently small, providing some intuition for why $|\hat{A}_{\text{ASD}}|/n$ is biased. For inconsistency, we prove that the bias is independent of n.

Theorem 3. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$, *where* $|A| = \theta(n)$. *For sufficiently large* n, *Bias*$(|\hat{A}_{ASD}|/n)$ *and Bias*$(\hat{\mu}_{ASD})$ *are independent of* n.

3 The NetMix Algorithm

Following the observation that the maximum likelihood estimators of the distribution $\text{ASD}_{\mathcal{P}_n}(A, \mu)$ are biased, we aim to find a less biased estimator by explicitly modeling the distribution of the scores $\mathbf{X}$. In this section, we derive a new algorithm, NetMix, that solves the Altered Subnetwork Problem by fitting a Gaussian mixture model (GMM) to $\mathbf{X}$.

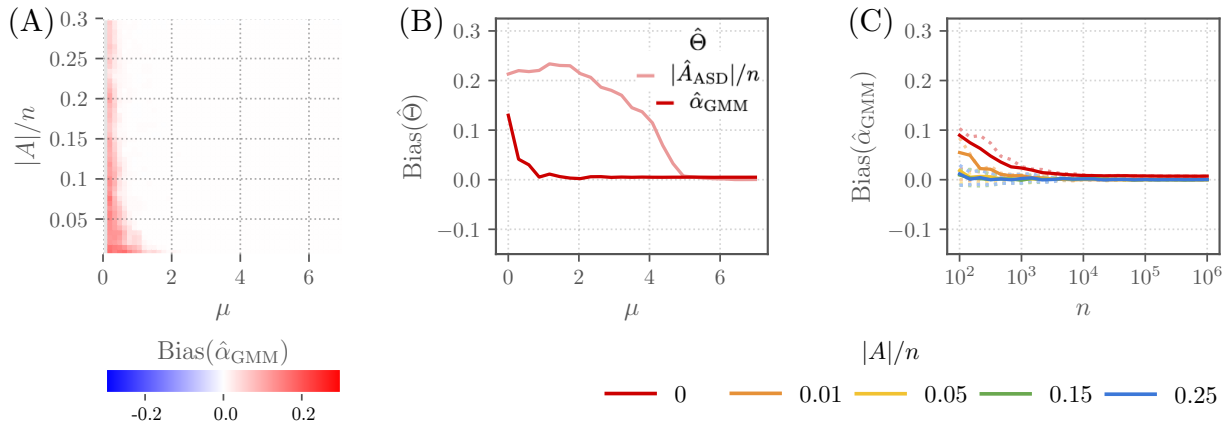

Figure 3: Scores $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ are distributed according to the unstructured ASD, and parameters $\hat{\alpha}_{\mathrm{GMM}}$ and $\hat{\mu}_{\mathrm{GMM}}$ are obtained by the EM algorithm. (A) $\mathrm{Bias}(\hat{\alpha}_{\mathrm{GMM}})$ as a function of the mean μ and altered subnetwork size $|A|/n$ for $n = 10^4$. Compare with Figure 2A. (B) $\mathrm{Bias}(\hat{\alpha}_{\mathrm{GMM}})$ and $\mathrm{Bias}(|\hat{A}_{\mathrm{ASD}}|/n)$ as functions of the mean μ for $|A|/n = 0.05$ and $n = 10^4$. (C) $\mathrm{Bias}(\hat{\alpha}_{\mathrm{GMM}})$ as a function of n for mean $\mu = 3$ and several values of $|A|/n$. Compare with Figure 2C.

3.1 Gaussian Mixture Model

We start by recalling the definition of a GMM.

Gaussian Mixture Model.. *Let $\mu > 0$ and $\alpha \in (0, 1)$. X is distributed according to the* Gaussian mixture model $\mathrm{GMM}(\alpha, \mu)$ *with parameters α and μ provided*

$$X \sim \alpha N(\mu, 1) + (1 - \alpha)N(0, 1). \tag{6}$$

An alternate interpretation of the GMM is to draw a latent variable $Z \sim \mathrm{Bernoulli}(\alpha)$ and select $X \sim N(\mu, 1)$ if $Z = 1$, and $X \sim N(0, 1)$ if $Z = 0$.

Given data $\mathbf{X} = (X_1, \ldots, X_n)$, we define $\hat{\mu}_{\mathrm{GMM}}$ and $\hat{\alpha}_{\mathrm{GMM}}$ to be the MLEs for μ and α, respectively, obtained by fitting a GMM to $\mathbf{X}$. In practice, $\hat{\mu}_{\mathrm{GMM}}$ and $\hat{\alpha}_{\mathrm{GMM}}$ are obtained by the EM algorithm, which is known to converge to the MLEs as the number of samples goes to infinity [92, 23]. Furthermore, if $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{GMM}(\mu, \alpha)$ are distributed according to the GMM with $\alpha \neq 0$, then $\hat{\mu}_{\mathrm{GMM}}$ and $\hat{\alpha}_{\mathrm{GMM}}$ are consistent (and therefore asymptotically unbiased) estimators of μ and α, respectively [14].

Analogously, by fitting a GMM to data $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ from the unstructured ASD, we observe that $\hat{\alpha}_{\mathrm{GMM}}$ is a less biased estimator of $|A|/n$ than $|\hat{A}_{\mathrm{ASD}}|/n$ (Figure 3A,B). We also observe that $\hat{\alpha}_{\mathrm{GMM}}$ is a consistent estimator of $|A|/n$ (Figure 3C). We summarize our findings in the following conjecture.

Conjecture. *Let $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ with $|A| > 0$, and let $\hat{A}_{ASD}$ be the MLE of A as defined in* (5)*. Let $\hat{\alpha}_{GMM}$ and $\hat{\mu}_{GMM}$ be the MLEs of α and μ obtained by fitting a GMM to $\mathbf{X}$. Then $\mathrm{Bias}_{|A|/n}(\hat{\alpha}_{GMM}) < \mathrm{Bias}_{|A|/n}(|\hat{A}_{ASD}|/n)$. Moreover, $\hat{\alpha}_{GMM}$ and $\hat{\mu}_{GMM}$ are consistent estimators of $|A|/n$ and μ, respectively.*

Although we do not have a proof of the above conjecture, a partial justification is the following relationship between the unstructured ASD and the GMM distribution. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be drawn from a mixture of unstructured ASDs over all possible anomalous sets A of size k, i.e., $\mathbf{X} \sim B \sum_{|A|=k} \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$, where $B = \frac{1}{\binom{n}{k}}$ is a normalizing constant. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{GMM}(\alpha, \mu)$ be independent samples from the GMM for $\mu > 0$ and $\alpha = \frac{k}{n}$ with corresponding latent variables $Z_1, \ldots, Z_n$. Then, the joint distribution of the GMM samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ conditioned on $\sum_{i=1}^n Z_i = k$ is equal to the distribution of $\mathbf{X}$:

$$\mathbf{X} \stackrel{d}{=} \left(\mathbf{Y} \,\middle|\, \sum_{i=1}^{n} Z_i = k \right). \tag{7}$$

3.2 NetMix Algorithm

We derive an algorithm, NetMix, that uses the maximum likelihood estimators (MLEs) $\hat{\mu}_{\text{GMM}}$ and $\hat{\alpha}_{\text{GMM}}$ from the GMM to solve the Altered Subnetwork Problem. Note that the GMM is *not* identical to ASD, the distribution that generated the data. Despite this difference in distributions, the above conjecture provides justification that the GMM will yield less biased estimators of A and μ than the MLEs of the ASD distribution.

Given a graph $G = (V, E)$ and scores $\mathbf{X} = (X_v)_{v \in V}$, NetMix first computes the *responsibility* $r_v = \Pr(v \in A \mid X_v)$, or the probability that $v \in A$, for each vertex $v \in V$. The responsibilities r_v are computed from the GMM MLEs $\hat{\mu}_{\text{GMM}}$ and $\hat{\alpha}_{\text{GMM}}$ (which are estimated by the EM algorithm [24]) according to the formula

$$\hat{r}_v = \frac{\hat{\alpha}_{\text{GMM}} \phi(X_v - \hat{\mu}_{\text{GMM}})}{\hat{\alpha}_{\text{GMM}} \phi(X_v - \hat{\mu}_{\text{GMM}}) + (1 - \hat{\alpha}_{\text{GMM}}) \phi(X_v)}, \tag{8}$$

where ϕ is the PDF of the standard normal distribution.

Next, NetMix aims to find a connected subgraph C of size $|C| \approx n\alpha$ that maximizes $\sum_{v \in C} r_v$. In order to find such a subgraph, NetMix assigns a weight $w(v) = \hat{r}_v - \tau$ to each vertex v, where τ is chosen so that approximately $n\hat{\alpha}_{\text{GMM}}$ nodes have non-negative weights. NetMix then computes the maximum weight connected subgraph (MWCS) $\hat{A}_{\text{NetMix}}$ in G by adapting the integer linear program in [27]. The use of τ is motivated by the observation that, if $\hat{\alpha}_{\text{GMM}} \approx \alpha$, then we expect $|\hat{A}_{\text{NetMix}}| \approx n\hat{\alpha}_{\text{GMM}} \approx n\alpha = |A|$.

We formally describe the NetMix algorithm for solving the Altered Subnetwork Problem below.

NetMix algorithm. Given a network $G = (V, E)$ and vertex scores $\mathbf{X} = (X_v)_{v \in V}$,

1. Compute $\hat{\alpha}_{\text{GMM}}$ and $\hat{\mu}_{\text{GMM}}$, the MLEs of α and μ, by fitting a GMM to $\mathbf{X}$ using expectation maximization (EM).
2. Compute the estimated responsibilities $\hat{r}_v$ for each vertex v using (8).
3. Compute τ such that $|\{v \in V : \hat{r}_v > \tau\}| = \lceil n\hat{\alpha}_{\text{GMM}} \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.
4. Find the connected subgraph $\hat{A}_{\text{NetMix}}$ defined by

$$\hat{A}_{\text{NetMix}} = \underset{\text{connected } C \subseteq V}{\operatorname{argmax}} \sum_{v \in C} (\hat{r}_v - \tau) \tag{9}$$

   using integer linear programming.

NetMix bears some similarities to heinz [27], another algorithm to identify altered subnetworks. However, there are two important differences. First, heinz does not solve the Altered Subnetwork Problem defined in the previous section. Instead, heinz models the vertex scores (assumed to be p-values) with a beta-uniform mixture (BUM) distribution. The motivation for the BUM is based on an empirical goodness-of-fit in [72]; however, later work by the same author [71] observes that the BUM tends to underestimate the number of p-values drawn from the altered distribution. Second, heinz requires that the user specify a False Discovery Rate (FDR) and shifts the p-values according to this FDR. We show below that nearly all choices of the FDR lead to a biased estimate of $|A|$. Moreover, the manually selected FDR allows users to selectively tune the value of this parameter to influence which genes are in the inferred altered subnetwork,

analogous to "p-hacking" [49, 68, 41]. Indeed, recently published analyses using heinz [17, 40, 52] use a wide range of FDR values. See the supplement for more details on the differences between heinz and NetMix. Despite these limitations, the ILP given in heinz to solve the MWCS problem is very useful for implementing NetMix and for computing the scan statistic (2) used in jActiveModules (see below).

4 Results

We compared NetMix to jActiveModules [48] and heinz [27] on simulated instances of the Altered Subnetwork Problem and on real datasets, including differential gene expression experiments from the Expression Atlas [70] and somatic mutations in cancer. jActiveModules is accessible only through Cytoscape [78, 19] and not a command-line interface, making it difficult to run on large number of a datasets. Thus, we implemented jActiveModules*, which computes the scan statistic (5) by adapting the integer linear program in heinz[2]. jActiveModules* output the global optimum of the scan statistic, while jActiveModules relies on heuristics (simulated annealing and greedy search) to find a local optimum.

4.1 Simulated Data

We compared NetMix, jActiveModules*, and heinz on simulated instances of the Altered Subnetwork Problem using the HINT+HI interaction network [57], a combination of binary and co-complex interactions in HINT [22] with high-throughput derived interactions from the HI network [76] as the graph G. For each instance, we randomly selected a connected subgraph $A \subseteq V$ with size $|A| = 0.05n$ using the random walk method of [59], and drew a sample $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{C}_G}(A, \mu)$. We ran each method on $\mathbf{X}$ and G to obtain an estimate $\hat{A}$ of the altered subnetwork A. We ran heinz with three different choices of the FDR parameter (FDR $= 0.001$, FDR $= 0.1$, and FDR $= 0.5$) to reflect the variety of FDRs used in practice.

We found that NetMix output subnetworks whose size $|\hat{A}_{\mathrm{NetMix}}|$ was very close to the true size across all values of μ in the simulations (Figure 4A). In contrast, jActiveModules* output subnetworks that were much larger than the implanted subnetwork for $\mu < 5$. This behavior is consistent with our conjectures above about the large bias in the maximum likelihood estimator $\hat{A}_{\mathrm{ASD}}$ for the unstructured ASD. Note that $\mu > 5$ corresponds to a large separation between the background and alternative distributions, and the network is not needed to separate these two distributions.

We also quantified the overlap between the true altered subnetwork A and the subnetwork $\hat{A}$ output by each method using the F-measure, finding that NetMix outperforms other methods across the full range of μ (Figure 4B). heinz requires the user to select an FDR value, and we find that the size of the output subnetwork and the F-measure varies considerably for different FDR (Figure 4A, 4B). When μ was small, a high FDR value (FDR $= 0.5$) yielded the best performance in terms of F-measure. However, when μ was large, a low FDR value (FDR $= 0.001$) gave better performance. While there are FDR values where the performance of heinz is similar to NetMix, the user *does not know what FDR value to select* for any given input, as the values of μ and the size $|A|$ of the altered subnetwork are unknown.

The bias in $|\hat{A}|/n$ observed using jActiveModules* with the interaction network (Figure 4A) was similar to the bias for the unstructured ASD (Figure 2A). Thus, we also evaluated how much benefit the network provided for each method. For small μ, we found that NetMix had a small but noticeable gain in performance when using the network; in contrast, other methods had nearly the same performance with or without the network (Figure 4C with further details in the supplement). These results emphasize the importance of evaluating network methods on simulated data *and* demonstrating that a network method outperforms a

[2]The scan statistic (2) is the maximization of a non-linear objective function, but for fixed subnetwork size $|S|$ the objective function is linear. We computed the scan statistic by modifying the ILP in heinz [27] and running this ILP over all possible subnetwork sizes.

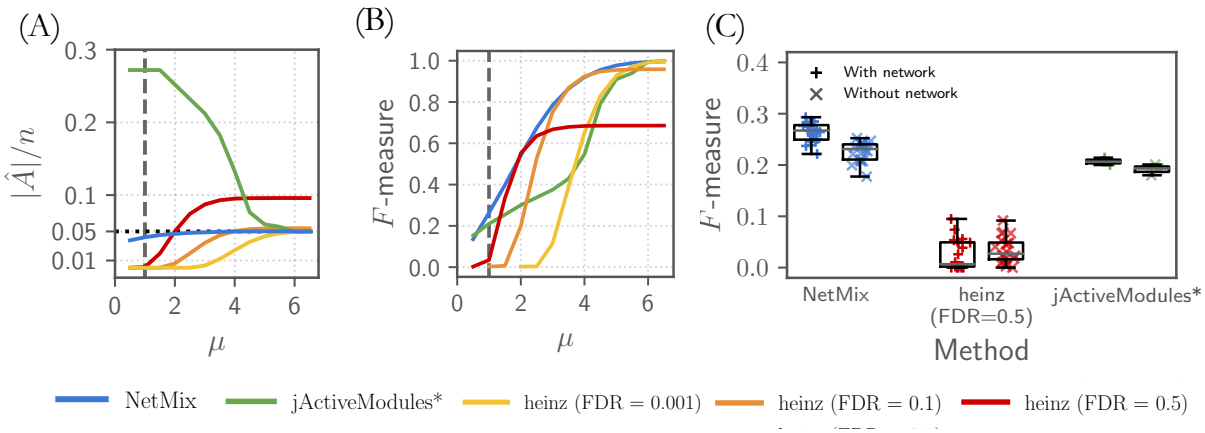

Figure 4: Comparison of altered subnetwork identification methods on simulated instances of the Altered Subnetwork Problem using the HINT+HI interaction network with $n = 15074$ nodes, and where the altered subnetwork A has size $|A| = 0.05n$. [Dashed vertical line ($\mu = 1$) represents the smallest μ such that one can detect whether G contains an altered subnetwork[3]]. (A) Size $|\hat{A}|/n$ of identified altered subnetwork $\hat{A}$ as a function of mean μ. (B) F-measure for $\hat{A}$ as a function of μ. (C) F-measure for $\hat{A}$ at $\mu = 1$, comparing performance with the network (left series for each method) and without the network (right series for each method).

single-gene test; neither of these were done in the jActiveModules [48] and heinz [27] papers, nor are they common in many other papers on biological network analysis.

4.2 Differential Gene Expression Subnetworks

We compared NetMix, jActiveModules*, and heinz on gene expression data from the Expression Atlas [70]. We analyzed 945 differential expression experiments including 292 RNA-seq experiments and 653 microarray experiments. For 84% of these experiments, the GMM used by NetMix provided a better fit to the p-value distributions than the beta-uniform mixture (BUM) [72] used by heinz (see the supplement for more details). In addition, the GMM provided a better fit in $83/85$ experiments where the null proportion (fraction of genes not differentially expressed) estimated by the GMM and BUM differed by $\geqslant 0.25$. In all 85 of these experiments, the BUM estimated a higher null proportion, consistent with the report in [71] that the BUM tends to overestimate the null proportion.

As many experiments had a small null proportion (i.e., most genes in the experiment were differentially expressed), we restricted our analysis to the 157 experiments from the Expression Atlas with a null proportion $\geqslant 0.8$ as estimated by the GMM. We ran NetMix, jActiveModules*, and heinz on these 157 experiments with the HINT+HI network. For heinz, we used three FDR values: FDR $= 0.1$, FDR $= 0.001$, and the FDR value such that $|\hat{A}_{\text{NetMix}}|$ genes have a positive weight in the heinz scoring. These choices demonstrate how users might "p-hack" the FDR value to achieve desirable results. We also compared to a method that ignores network topology, selecting the $|\hat{A}_{\text{NetMix}}|$ genes with the lowest p-values; we call this method "top p-values". See the supplement for specific details on these methods.

Both NetMix and heinz identified subnetworks that were significantly smaller than jActiveModules* (Figure 5A), which is consistent with previous observations [67] that jActiveModules estimates overly large subnetworks. At the same time, NetMix identified subnetworks with significant overlap (FDR $\leqslant 0.01$) with more biological process GO terms than heinz ($p = 3.3 \cdot 10^{-12}$, t-test; Figure 5B) or top p-values

[3]Formally, μ is the smallest mean such that the hypotheses $H_0 : X \sim \text{ASD}_{\mathcal{C}_G}(\varnothing, 0)$ and $H_1 : X \sim \text{ASD}_{\mathcal{C}_G}(A, \mu)$ are asymptotically distinguishable. See [83] for details.

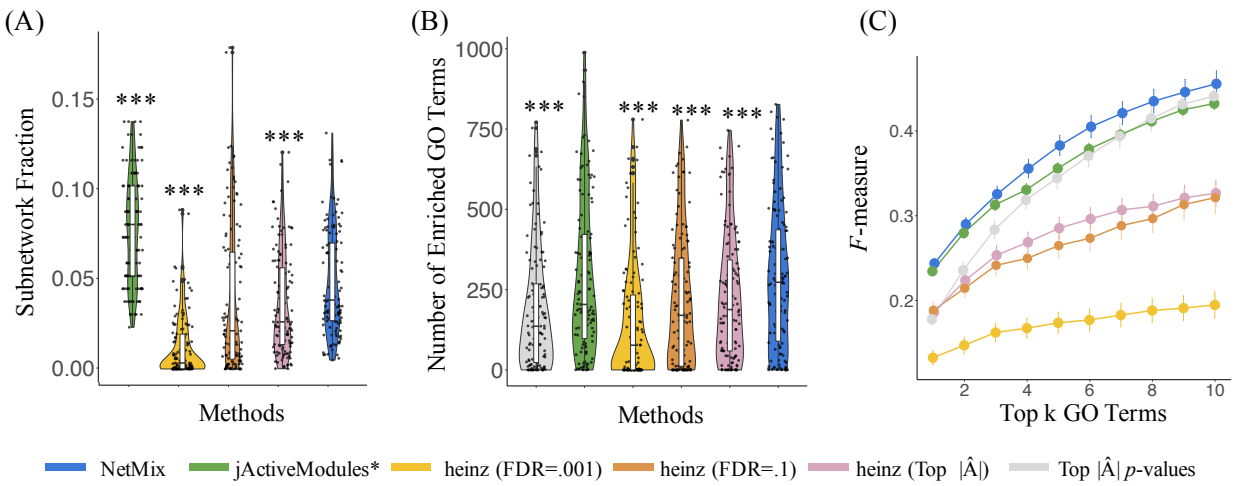

Figure 5: (A) Fraction of genes in the HINT+HI interaction network that are in the subnetwork identified by each method. $* : p \leqslant 0.01, ** : p \leqslant 0.001, *** : p \leqslant 10^{-4}$ indicate significant p-values in paired t-test between NetMix and other methods. (B) Number of enriched GO biological process terms for altered subsets identified by each method. (C) F-measure of the k most enriched GO terms.

($p < 2.2 \cdot 10^{-16}$, t-test; Figure 5B). We also found that subnetworks identified by NetMix had greater overlap (as quantified by F-measure) with genes in the top k GO terms (Figure 5C). These results suggest that NetMix identifies subnetworks that are more relevant to differential expression experiments than other methods.

We examined the experiment E-GEOD-11199 in more detail. This experiment compared *Mtb*-stimulated and unstimulated macrophages [86]. NetMix identified a subnetwork containing 706 genes, half the size of the jActiveModules* subnetwork containing 1450 genes. Both of these subnetworks contained 37 of the 42 genes whose differential expression was experimentally validated by RT-PCR [86]. Although the NetMix subnetwork was less than half the size of the jActiveModules* subnetwork, the NetMix subnetwork overlapped more GO terms (445 vs. 179). In contrast, heinz (using FDR $= 0.27$) identified a subnetwork of 382 genes containing only 25 RT-PCR validated genes. Finally, the 692 genes with the smallest p-values include only 7 validated genes. These results show that the NetMix subnetwork contains many biologically relevant genes, including most of the RT-PCR validated genes, without being overly large.

4.3 Somatic Mutations In Cancer

We compared the performance of NetMix, jActiveModules* [4, 1], jActiveModules [48], heinz [27] and Hierarchical HotNet [75] in identifying cancer driver genes, using the MutSig2CV driver p-values [54] from the TCGA PanCanAtlas project [8]. We ran all methods on the HINT+HI interaction network described above, as well as the iRefIndex 15.0 [74] and ReactomeFI 2016 [21, 28] interaction networks. See the supplement for more details on the datasets.

We evaluated the quality of the subnetwork $\hat{A}$ reported by each method by computing the overlap with the list of cancer genes in the COSMIC Cancer Gene Census (CGC) [33, 31]. We found that NetMix outperforms all other methods in F-measure across all interaction networks. For example, using the HINT+HI network, NetMix achieved an F-measure of 0.277, compared to F-measures of 0.191 for jActiveModules*, 0.216 for heinz (FDR $= 0.001$), 0.264 for heinz (FDR $= 0.1$), and 0.214 for Hierarchical HotNet[4]. Both the NetMix and Hierarchical HotNet results were statistically significant ($p < 0.01$) on all 3

[4]The jActiveModules greedy search algorithm failed to complete within 100 hours, while the jActiveModules simulated

interaction networks according to permutation tests from [75]. The modest F-measures for all methods are not surprising; the genes in CGC have diverse alterations across cancer types and thus high recall is not expected by this restricted analysis of single-nucleotide mutations in a subset of cancer types. Nevertheless, the higher performance of NetMix on this task across all networks is encouraging. Further details of these comparisons are in the supplement.

5 Discussion

In this paper, we revisit the classic problem of identifying altered subnetworks in high-throughput biological data. We formalize the Altered Subnetwork Problem as the estimation of the parameters of the Altered Subset Distribution (ASD). We show that the seminal algorithm for this problem, jActiveModules [48], is equivalent to a maximum likelihood estimator (MLE) of the ASD. At the same time, we show that the MLE is a biased estimator of the altered subnetwork, with especially large positive bias for small altered subnetworks. This bias explains previous reports that jActiveModules tends to output large subnetworks [67].

We leverage these observations to design NetMix, a new algorithm for the Altered Subnetwork Problem. We show that NetMix outperforms existing methods on simulated and real data. NetMix fits a Gaussian mixture model (GMM) to observed node scores and then finds a maximum weighted connected subgraph using node weights derived from the GMM. heinz [27], another widely used method for altered subnetwork identification, also derives node weights from a mixture model (a beta-uniform mixture of p-values) and finds a maximum weighted connected subgraph. However, heinz does not solve the Altered Subnetwork Problem in a strict sense; rather, heinz requires users to choose a parameter (an FDR estimate for the mixture fit) that implicitly constrains the size of the identified subnetwork. This user-defined parameter encourages p-hacking [49, 68, 41], and we find that nearly all values of this parameter lead to biased estimates of the size of the altered subnetwork.

We note a number of directions for future work. The first is to generalize our theoretical contributions to the identification of *multiple* altered subnetworks, a situation which is common in biological applications where multiple biological processes may be perturbed [62]. While it is straightforward to run NetMix iteratively to identify multiple subnetworks – as jActiveModules does – a rigorous assessment of the identification of multiple altered subnetworks would be of interest. Second, our results on simulated data (Section 4.1) show that altered subnetwork methods have only marginal gains over simpler methods that rank vertices without information from network interactions. We hypothesize that this is because connectivity is not a strong constraint for biological networks; indeed the biological interaction networks that we use have both small diameter and small average shortest path between nodes (see the supplement for specific statistics). Specifically, we suspect that most subsets of nodes are "close" to a connected subnetwork in such biological networks, and thus the MLE of connected altered subnetworks has similar bias as the MLE of the unstructured altered subset distribution. In contrast, for other network topologies like the line graph, connectivity is a much stronger topological constraint (see the supplement for a brief comparison of different topologies). It would be useful to investigate this hypothesis and characterize the conditions when networks provide benefit for finding altered subnetworks. In particular, other topological constraints such as dense subgraphs [38, 7], cliques [2], and subgraphs resulting from heat diffusion and network propagation processes [87, 88, 57, 20] have been used used to model altered subnetworks in biological data. Generalizing the theoretical results in this paper to these other topological constraints may be helpful for understanding the parameter regimes where these topological constraints provide signal for identification of altered subnetorks. Finally, we note that biological networks often have substantial ascertainment bias, with more interactions annotated for well-studied genes [44, 76], and these well-studied genes in turn

annealing algorithm yielded an F-measure of 0.086.

may also be more likely to have outlier measurements/scores. Thus, any network method should carefully quantify the regime where it outperforms straightforward approaches – e.g., methods based on ranking nodes by gene scores or node degree – both on well-calibrated simulations and on real data.

Acknowledgments

We thank Mohammed El-Kebir for assistance with implementing jActiveModules* by modifying the ILP in heinz. We thank David Tse for directing us to the network anomaly literature. M.A.R. was supported in part by the National Cancer Institute of the NIH (Cancer Target Discovery and Development Network grant U01CA217875). B.J.R. was supported by US National Institutes of Health (NIH) grants R01HG007069 and U24CA211000.

References

[1] Addario-Berry, L., Broutin, N., Devroye, L., Lugosi, G., et al.: On combinatorial testing problems. The Annals of Statistics **38**(5), 3063–3092 (2010)

[2] Amgalan, B., Lee, H.: Wmaxc: a weighted maximum clique method for identifying condition-specific sub-network. PloS one **9**(8), e104993 (2014)

[3] Arias-Castro, E., Candès, E.J., Durand, A.: Detection of an anomalous cluster in a network. The Annals of Statistics pp. 278–304 (2011)

[4] Arias-Castro, E., Candès, E.J., Helgason, H., Zeitouni, O.: Searching for a trail of evidence in a maze. The Annals of Statistics pp. 1726–1757 (2008)

[5] Arias-Castro, E., Castro, R.M., Tánczos, E., Wang, M.: Distribution-free detection of structured anomalies: permutation and rank-based scans. Journal of the American Statistical Association **113**(522), 789–801 (2018)

[6] Arias-Castro, E., Donoho, D.L., Huo, X.: Adaptive multiscale detection of filamentary structures in a background of uniform random points. The Annals of Statistics pp. 326–349 (2006)

[7] Ayati, M., Erten, S., Chance, M.R., Koyutürk, M.: Mobas: identification of disease-associated protein subnetworks using modularity-based scoring. EURASIP journal on bioinformatics & systems biology **2015**, 7–7 (06 2015)

[8] Bailey, M.H., Tokheim, C., Porta-Pardo, E., et al.: Comprehensive characterization of cancer driver genes and mutations. Cell **173**(2) (2018)

[9] Batra, R., Alcaraz, N., Gitzhofer, K., et al.: On the performance of de novo pathway enrichment. NPJ systems biology and applications **3**(1), 6 (2017)

[10] Berger, B., Peng, J., Singh, M.: Computational solutions for omics data. Nature reviews genetics **14**(5), 333 (2013)

[11] Bhalla, U.S., Iyengar, R.: Emergent properties of networks of biological signaling pathways. Science **283**(5400), 381–387 (1999)

[12] Califano, A., Butte, A.J., Friend, S., Ideker, T., Schadt, E.: Leveraging models of cell regulation and gwas data in integrative network-based association studies. Nature genetics **44**(8), 841–847 (2012)

[13] Chasman, D., Siahpirani, A.F., Roy, S.: Network-based approaches for analysis of complex biological systems. Current Opinion in Biotechnology **39**, 157 – 166 (2016)

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

[15] Cho, A., Shim, J.E., Kim, E., Supek, F., Lehner, B., Lee, I.: Muffinn: cancer gene discovery via network analysis of somatic mutation data. Genome Biology **17**(1), 129 (2016)

[16] Cho, D.Y., Kim, Y.A., Przytycka, T.M.: Chapter 5: Network biology approach to complex diseases. PLOS Computational Biology **8**(12), 1–11 (2012)

[17] Choi, J., Shooshtari, P., Samocha, K.E., Daly, M.J., Cotsapas, C.: Network analysis of genome-wide selective constraint reveals a gene network active in early fetal brain intolerant of mutation. PLoS genetics **12**(6), e1006121 (2016)

[18] Chua, H.N., Sung, W.K., Wong, L.: Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics **22**(13), 1623–1630 (04 2006)

[19] Cline, M.S., Smoot, M., Cerami, E., et al.: Integration of biological networks and gene expression data using cytoscape. Nature protocols **2**(10), 2366 (2007)

[20] Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics (2017)

[21] Croft, D., Mundo, A.F., et al.: The reactome pathway knowledgebase. Nucleic acids research **42**(D1), D472–D477 (2014)

[22] Das, J., Yu, H.: Hint: High-quality protein interactomes and their applications in understanding human disease. BMC systems biology **6**(1), 92 (2012)

[23] Daskalakis, C., Tzamos, C., Zampetakis, M.: Ten steps of em suffice for mixtures of two gaussians. In: Proceedings of the 2017 Conference on Learning Theory. pp. 704–710 (2017)

[24] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)

[25] Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F.: Prediction of protein function using protein–protein interaction data. Journal of Computational Biology **10**(6), 947–960 (2003)

[26] Dimitrakopoulos, C.M., Beerenwinkel, N.: Computational approaches for the identification of cancer genes and pathways. Wiley Interdisciplinary Reviews: Systems Biology and Medicine **9**(1) (2017)

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

[28] Fabregat, A., Sidiropoulos, K., et al.: The reactome pathway knowledgebase. Nucleic acids research **44**(D1), D481–D487 (2016)

[29] Ferguson, T.S.: An inconsistent maximum likelihood estimate. Journal of the American Statistical Association **77**(380), 831–834 (1982)

[30] Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**(1), 27–38 (1993)

[31] Forbes, S.A., Beare, D., Boutselakis, H., et al.: Cosmic: somatic cancer genetics at high-resolution. Nucleic acids research **45**(D1), D777–D783 (2016)

[32] de la Fuente, A.: From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. Trends in Genetics **26**(7), 326 – 333 (2010)

[33] Futreal, P.A., Coin, L., et al.: A census of human cancer genes. Nature Reviews Cancer **4**(3), 177–183 (2004)

[34] Glaz, J., Naus, J., Wallenstein, S.: Scan Statistics. Springer-Verlag New York (2001)

[35] Gligorijević, V., Pržulj, N.: Methods for biological data integration: perspectives and challenges. Journal of the Royal Society, Interface **12**(112), 20150571 (2015)

[36] Gulsuner, S., Walsh, T., Watts, A.C., et al.: Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell **154**(3), 518 – 529 (2013)

[37] Guo, M., Bao, E.L., Wagner, M., Whitsett, J.A., Xu, Y.: SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Research **45**(7), e54–e54 (12 2016)

[38] Guo, Z., Li, Y., et al.: Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. Bioinformatics **23**(16), 2121–2128 (06 2007)

[39] Halldórsson, B.V., Sharan, R.: Network-based interpretation of genomic variation data. Journal of molecular biology **425**(21), 3964–3969 (2013)

[40] He, H., Lin, D., Zhang, J., Wang, Y.p., Deng, H.w.: Comparison of statistical methods for subnetwork detection in the integration of gene expression and protein interaction network. BMC Bioinformatics **18**(1), 149 (2017)

[41] Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D.: The extent and consequences of p-hacking in science. PLoS biology **13**(3), e1002106 (2015)

[42] Hofree, M., Shen, J.P., Carter, H., Gross, A., Ideker, T.: Network-based stratification of tumor mutations. Nature methods **10**(11), 1108–1115 (2013)

[43] Hormozdiari, F., Penn, O., Borenstein, E., Eichler, E.E.: The discovery of integrated gene networks for autism and related disorders. Genome research **25**(1), 142–154 (2015)

[44] Horn, H., Lawrence, M.S., Chouinard, C.R., Shrestha, Y., Hu, J.X., Worstell, E., Shea, E., Ilic, N., Kim, E., Kamburov, A., et al.: Netsig: network-based discovery from cancer genomes. Nature methods (2017)

[45] Huang, J.K., Carlin, D.E., Yu, M.K., Zhang, W., Kreisberg, J.F., Tamayo, P., Ideker, T.: Systematic evaluation of molecular networks for discovery of disease genes. Cell systems **6**(4), 484–495 (2018)

[46] Hung, H.M.J., O'Neill, R.T., Bauer, P., Kohne, K.: The behavior of the p-value when the alternative hypothesis is true. Biometrics **53**(1), 11–22 (1997)

[47] Hung, J.H., Yang, T.H., Hu, Z., Weng, Z., DeLisi, C.: Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in Bioinformatics **13**(3), 281–291 (Sep 2011)

[48] Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F.: Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics **18**(suppl 1), S233–S240 (2002)

[49] Ioannidis, J.P.: Why most published research findings are false. PLoS medicine **2**(8), e124 (2005)

[50] Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., Ideker, T.: Pathblast: a tool for alignment of protein interaction networks. Nucleic acids research **32**(suppl_2), W83–W88 (2004)

[51] Kim, M., Hwang, D.: Network-based protein biomarker discovery platforms. Genomics & informatics **14**(1), 2 (2016)

[52] Klimm, F., Toledo, E.M., Monfeuga, T., Zhang, F., Deane, C.M., Reinert, G.: Functional module detection through integration of single-cell rna sequencing data with protein–protein interaction networks. bioRxiv p. 698647 (2019)

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

[54] Lawrence, M.S., Stojanov, P., et al.: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature **505**(7484), 495 (2014)

[55] Lee, I., Blom, U.M., et al.: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research **21**(7), 1109–1121 (2011)

[56] Leiserson, M.D., Eldridge, J.V., Ramachandran, S., Raphael, B.J.: Network analysis of gwas data. Current opinion in genetics & development **23**(6), 602–610 (2013)

[57] Leiserson, M.D., Vandin, F., Wu, H.T., et al.: Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics **47**(2), 106–114 (2015)

[58] Liu, J.J., Sharma, K., Zangrandi, L., et al.: In vivo brain gpcr signaling elucidated by phosphoproteomics. Science **360**(6395) (2018)

[59] Lu, X., Bressan, S.: Sampling connected induced subgraphs uniformly at random. In: Scientific and Statistical Database Management. pp. 195–212. Springer (2012)

[60] Luo, Y., Zhao, X., et al.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature Communications **8**(1), 573 (2017)

[61] McLachlan, G., Bean, R., Jones, L.B.T.: A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics **22**(13), 1608–1615 (04 2006)

[62] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., Barabási, A.L.: Disease networks. uncovering disease-disease relationships through the incomplete interactome. Science (New York, N.Y.) **347**(6224), 1257601–1257601 (2015)

[63] Mitra, K., Carvunis, A.R., Ramesh, S.K., Ideker, T.: Integrative approaches for finding modular structure in biological networks. Nature Reviews Genetics **14**, 719 EP – (09 2013)

[64] the Mutation Consequences, working group of the International Cancer Genome Consortium, P.A., et al.: Pathway and network analysis of cancer genomes. Nature Methods **12**, 615 EP – (06 2015)

[65] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics **21**, i302–i310 (06 2005)

[66] Nibbe, R.K., Koyutürk, M., Chance, M.R.: An integrative-omics approach to identify functional subnetworks in human colorectal cancer. PLoS computational biology **6**(1), e1000639 (2010)

[67] Nikolayeva, I., Pla, O.G., Schwikowski, B.: Network module identification–a widespread theoretical bias and best practices. Methods **132**, 19–25 (2018)

[68] Nuzzo, R.: How scientists fool themselves–and how they can stop. Nature News **526**(7572), 182 (2015)

[69] Pan, W., Lin, J., Le, C.T.: A mixture model approach to detecting differentially expressed genes with microarray data. Functional & Integrative Genomics **3**(3), 117–124 (2003)

[70] Petryszak, R., Keays, M., Tang, Y.A., Fonseca, N.A., et al.: Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic acids research **44**(D1), D746–D752 (2015)

[71] Pounds, S., Cheng, C.: Improving false discovery rate estimation. Bioinformatics **20**(11), 1737–1745 (02 2004)

[72] Pounds, S., Morris, S.W.: Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics **19**(10), 1236–1242 (2003)

[73] Radivojac, P., Clark, W.T., et al.: A large-scale evaluation of computational protein function prediction. Nature methods **10**(3), 221 (2013)

[74] Razick, S., Magklaras, G., Donaldson, I.M.: irefindex: a consolidated protein interaction database with provenance. BMC bioinformatics **9**(1), 1 (2008)

[75] Reyna, M.A., Leiserson, M.D., Raphael, B.J.: Hierarchical hotnet: identifying hierarchies of altered subnetworks. Bioinformatics **34**(17), i972–i980 (2018)

[76] Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al.: A proteome-scale map of the human interactome network. Cell **159**(5), 1212–1226 (2014)

[77] Roy, S., Ernst, J.o.: Identification of functional elements and regulatory circuits by drosophila modencode. Science **330**(6012), 1787–1797 (2010)

[78] Shannon, P., et al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13**(11), 2498–2504 (2003)

[79] Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular systems biology **3**(1) (2007)

[80] Sharpnack, J., Rinaldo, A., Singh, A.: Detecting anomalous activity on networks with the graph fourier scan statistic. IEEE Transactions on Signal Processing **64**(2), 364–379 (2016)

[81] Sharpnack, J., Singh, A.: Near-optimal and computationally efficient detectors for weak and sparse graph-structured patterns. In: Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE. pp. 443–446. IEEE (2013)

[82] Sharpnack, J., Singh, A., Rinaldo, A.: Changepoint detection over graphs with the spectral scan statistic. In: Artificial Intelligence and Statistics. pp. 545–553 (2013)

[83] Sharpnack, J.L., Krishnamurthy, A., Singh, A.: Near-optimal anomaly detection in graphs using lovasz extended scan statistic. In: Advances in Neural Information Processing Systems. pp. 1959–1967 (2013)

[84] Shrestha, R., Hodzic, E., et al.: Hit'ndrive: patient-specific multidriver gene prioritization for precision oncology. Genome research **27**(9), 1573–1588 (2017)

[85] Soul, J., Hardingham, T.E., Boot-Handford, R.P., Schwartz, J.M.: Phenomeexpress: a refined network analysis of expression datasets by inclusion of known disease phenotypes. Scientific reports **5**, 8117 (2015)

[86] Thuong, N.T.T., Dunstan, S.J., Chau, T.T.H., et al.: Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. PLoS pathogens **4**(12), e1000229 (2008)

[87] Vandin, F., Upfal, E., Raphael, B.J.: Algorithms for detecting significantly mutated pathways in cancer. Journal of Computational Biology **18**(3), 507–522 (2011)

[88] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol **6**(1), e1000641 (2010)

[89] Wang, X., Terfve, C., Rose, J.C., Markowetz, F.: HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. Bioinformatics **27**(6), 879–880 (01 2011)

[90] Wang, Y.H., Bower, N.I., et al.: Gene expression patterns during intramuscular fat development in cattle1. Journal of Animal Science **87**(1), 119–130 (01 2009)

[91] Xia, J., Gill, E.E., Hancock, R.E.W.: Networkanalyst for statistical, visual and network-based meta-analysis of gene expression data. Nature Protocols **10**, 823 EP – (05 2015)

[92] Xu, J., Hsu, D., Maleki, A.: Global analysis of expectation maximization for mixtures of two gaussians. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 2684–2692 (2016)