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

Matt Reyna*, Uthsav Chitra*, Rebecca Elyanow, Ben Raphael

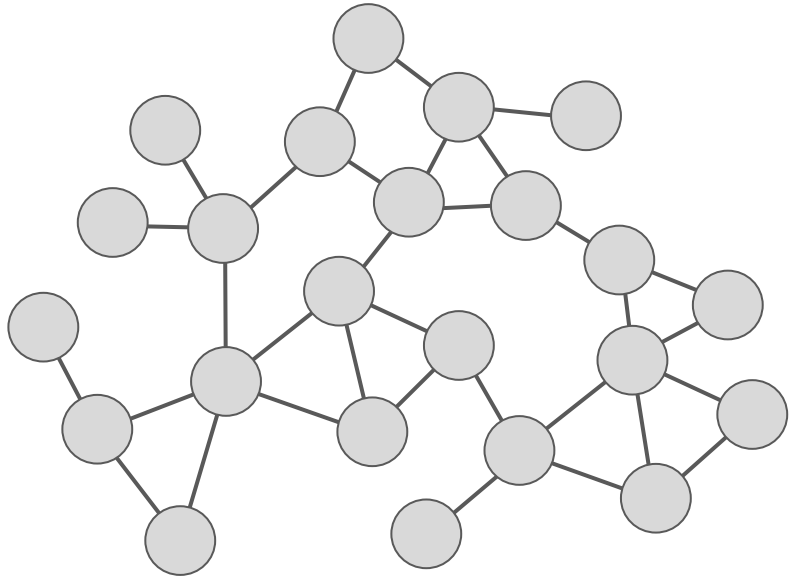
RECOMB 2020

Uthsav Chitra

PhD Student at Princeton University

Interaction Networks

A standard approach for analyzing high-throughput data is to incorporate the data with a biological **interaction network.**



Vertices: genes or proteins

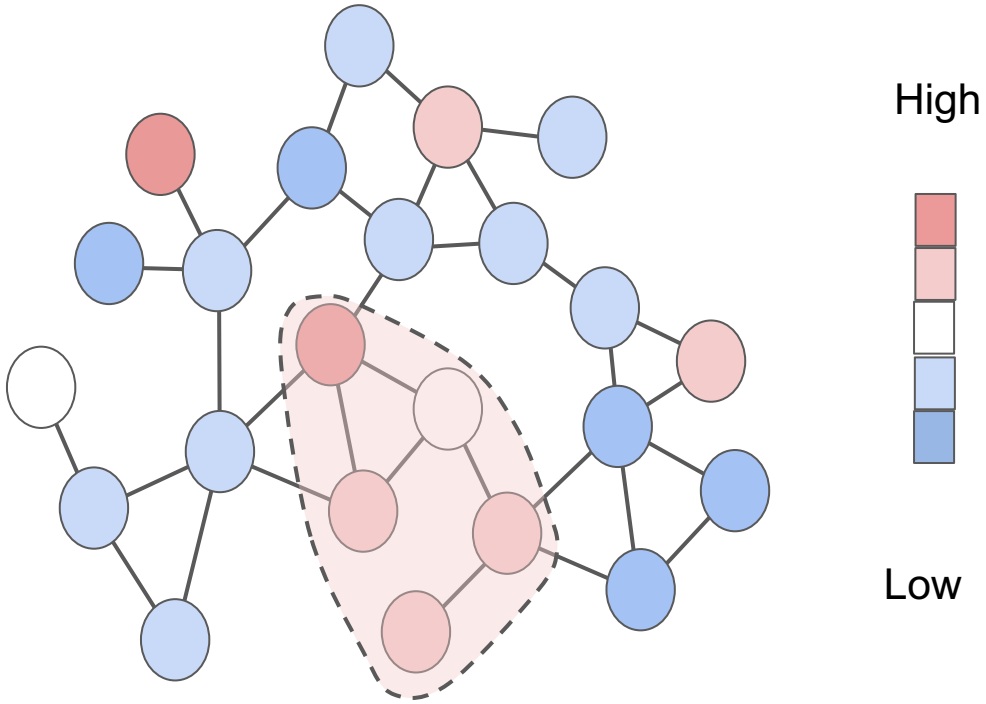
Edges: Interactions between genes/proteins

Altered Subnetwork Problem (ASP)

The underlying problem is to identify **altered subnetworks**

Given:
1) Network G = (V,E)
2) Vertex scores X_v (e.g. p-values or z-scores)

Goal: Identify high-scoring subnetworks of G



Applications of the Altered Subnetwork Problem

Differential gene expression

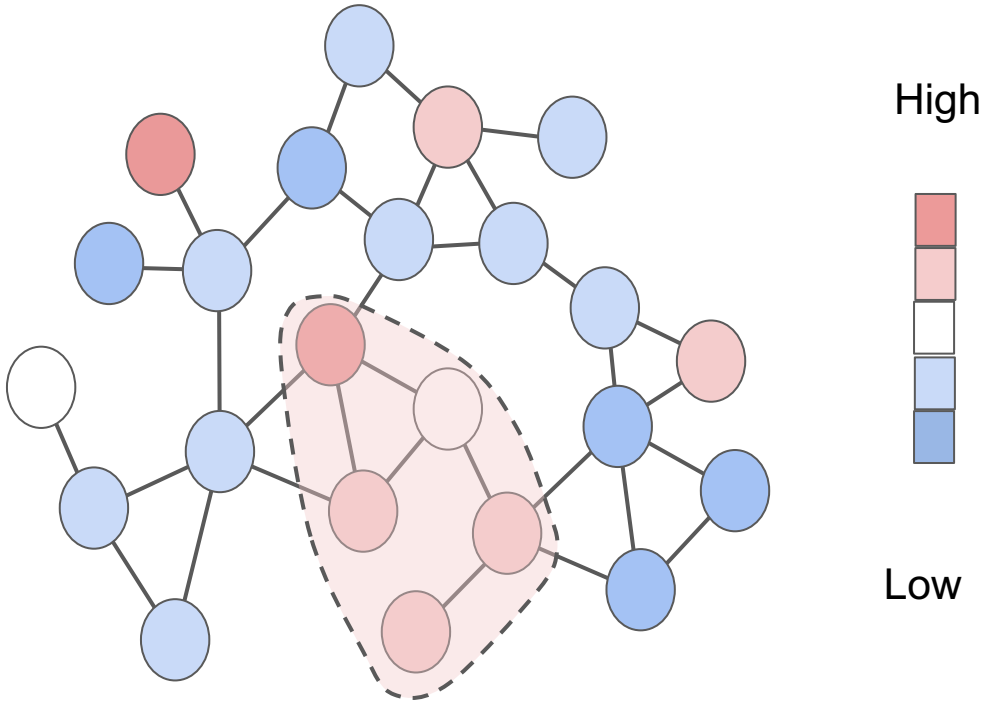
- Ideker *et al* 2002, Luscombe *et al* 2004, Dittrich *et al* 2008, …

Germline mutations from GWAS

- Lee *et al* 2011, Califano *et al* 2012, …

Somatic mutations in cancer

- Leiserson *et al* 2015, Cho *et al* 2016, Horn *et al* 2017, Reyna *et al* 2017, …



Altered Subnetwork Problem is a classic problem

BIOINFORMATICS

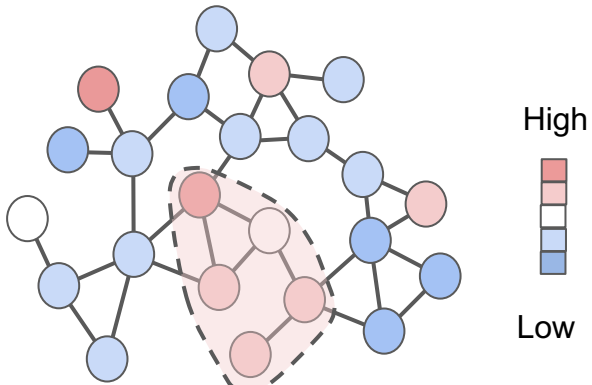
Vol. 18 Suppl. 1 2002
Pages S233–S240

Discovering regulatory and signalling circuits in molecular interaction networks

Trey Ideker [1,], Owen Ozier [1], Benno Schwikowski [2] and Andrew F. Siegel [2, 3]*

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA, [2]Institute for Systems Biology, Seattle, WA 98103, USA and [3]Departments of Management Science, Finance, Statistics, and Genome Sciences, University of Washington, Seattle, WA 98195, USA

Received on January 24, 2002; revised and accepted on April 1, 2002



Altered Subnetwork Problem:

Given:

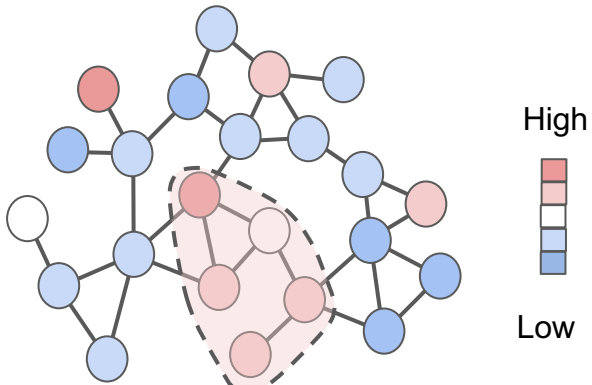
1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

Goal: Identify high-scoring subnetworks H of G

Algorithms for solving the ASP: jActiveModules

jActiveModules algorithm (Ideker et al, 2002): identifies altered subnetworks by maximizing a certain function over all connected subgraphs S

$$\arg\max_{S} \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v$$



Altered Subnetwork Problem:

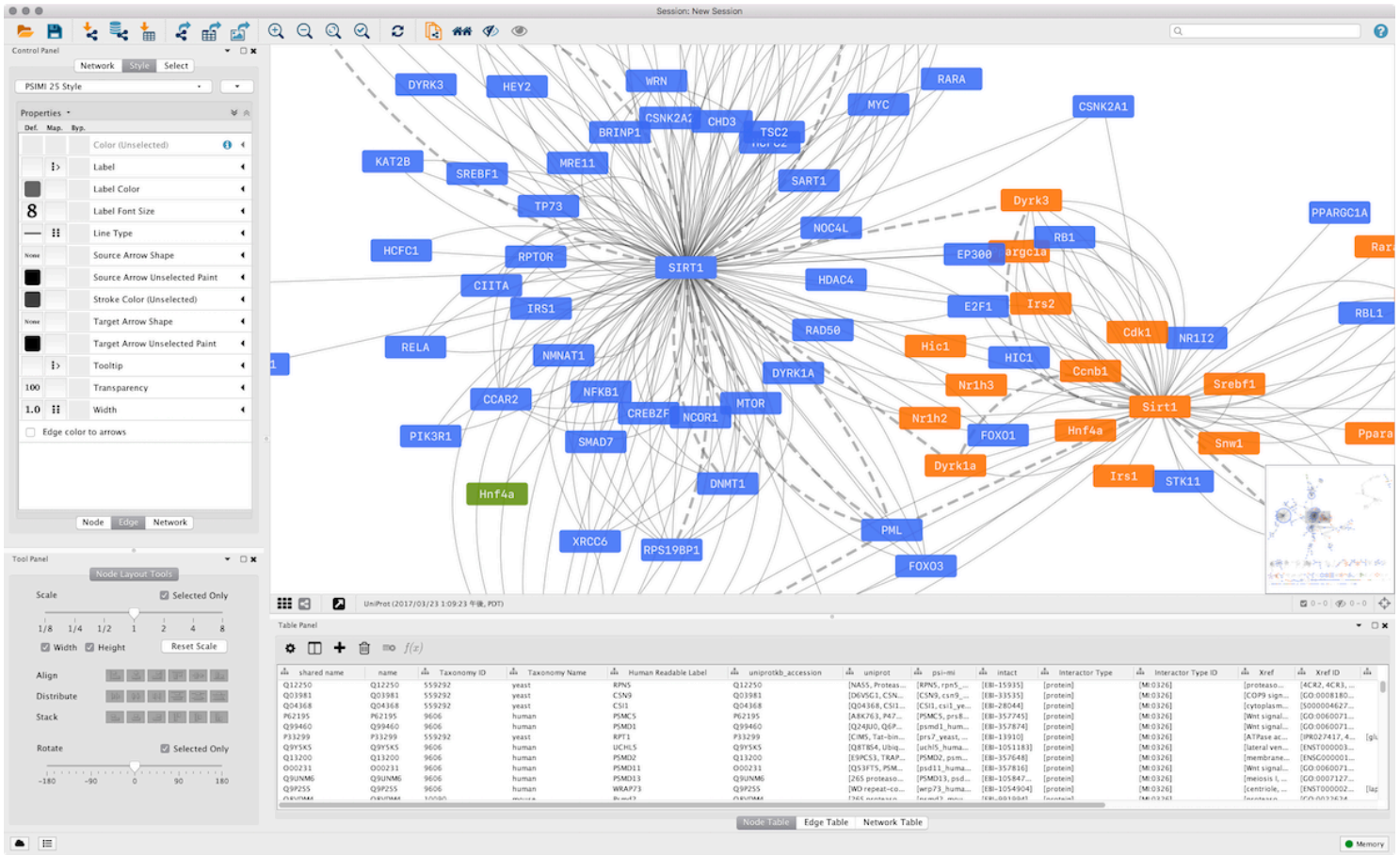
Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

Goal: Identify high-scoring subnetworks H of G

Algorithms for solving the ASP: jActiveModules

jActiveModules (Ideker et al, 2002) is implemented in the very popular **Cytoscape** platform



Algorithms for solving the ASP: heinz

Dittrich *et al* (2008) developed the **heinz** algorithm (implemented in Bionet)

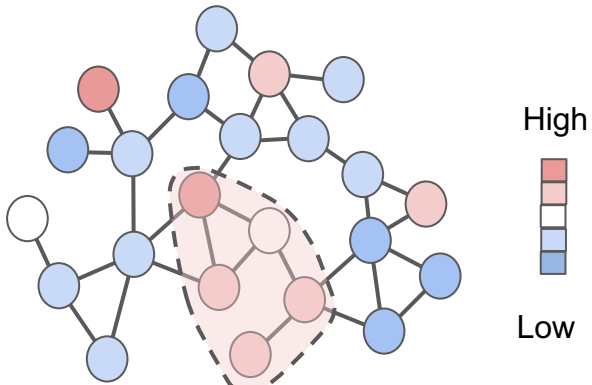
BIOINFORMATICS

Vol. 24 ISMB 2008, pages i223–i231
doi:10.1093/bioinformatics/btn161

Identifying functional modules in protein–protein interaction networks: an integrated exact approach

Marcus T. Dittrich[1,2,*,†], Gunnar W. Klau[3,4,*,†], Andreas Rosenwald[5], Thomas Dandekar[1] and Tobias Müller[1,*]

[1]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, [2]Institute of Clinical Biochemistry, University of Würzburg, Josef-Schneider-Str. 2, 97080 Würzburg, [3]Mathematics in Life Sciences Group, Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 3, 14195 Berlin, [4]DFG Research Center MATHEON, Berlin and [5]Institute of Pathology, University of Würzburg, Josef-Schneider-Str. 2, 97080 Würzburg, Germany



Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

Goal: Identify high-scoring subnetworks H of G

Algorithms for solving the ASP: heinz

Dittrich *et al* (2008) developed the **heinz** algorithm (implemented in BioNet)

> "The presented algorithm is, to our knowledge, the first approach that really tackles and solves the original problem raised by Ideker *et al.* (2002) to optimality."

Algorithms for solving the ASP: heinz

Dittrich *et al* (2008) developed the **heinz** algorithm (implemented in BioNet)

> "The presented algorithm is, to our knowledge, the first approach that really tackles and solves the original problem raised by Ideker *et al.* (2002) to optimality."

However heinz maximizes a **different** function over connected subgraphs S:

$$\arg\max_{S} \sum_{v \in S} [w_v - \tau]$$

- Weights w_v determined by fit to a Beta-Uniform Mixture distribution
- τ is user-defined parameter corresponding to False Discovery Rate (FDR)

Many subsequent algorithms have been developed for solving the ASP

Table 1 | **Software tools based on network propagation**

Tool	Goal	Type	Platform	Web site
Function prediction				
DSD[48] and capDSD[34]	Function prediction	Single network	Web server and software for download	http://dsd.cs.tufts.edu/server/ and http://dsd.cs.tufts.edu/capdsd
GeneMANIA [103]	Function prediction	Single network	Cytoscape plugin	http://apps.cytoscape.org/apps/genemania
Mashup[56]	Function prediction	Integrative	Software for download	http://mashup.csail.mit.edu/
RIDDLE[70]	Function prediction	Single network	Web server	http://www.functionalnet.org/RIDDLE/
Disease characterization				
CATAPULT[82]	Gene prioritization	Integrative	Web server and software for download	http://marcottelab.org/index.php/Catapult
Cytoscape 'diffuse' service[104]	General propagation	1D and 2D	Software for download	• http://cytoscape.org • Native in version 3.5 and greater
DADA[80]	Gene prioritization	1D	Software for download	http://compbio.case.edu/dada/
Exome Walker[72]	Gene prioritization	1D	Web server	http://compbio.charite.de/ExomeWalker
GUILD[105]	Gene prioritization	1D	Software for download	http://sbi.imim.es/web/index.php/research/software/guildsoftware
HotNet2 (REF. 30)	Module detection	2D	Software for download	http://compbio.cs.brown.edu/projects/hotnet2/
NBS[89]	Patient stratification	Integrative	Software for download	http://chianti.ucsd.edu/~mhofree/NBS/
NetQTL[79]	Gene prioritization and module detection	1D	Software for download	https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#netqtl
PRINCIPLE[106]	Gene prioritization and module detection	1D	Cytoscape plugin	http://www.cs.tau.ac.il/~bnet/software/PrincePlugin/
SNF[90]	Patient stratification	Integrative	Software for download	http://compbio.cs.toronto.edu/SNF/SNF/Software.html
TieDIE[91]	Module detection	Integrative	Software for download	https://sysbiowiki.soe.ucsc.edu/tiedie
ToppGene[107]	Gene prioritization	1D	Web server	https://toppgene.cchmc.org/

Cowen *et al*, Nature Reviews Genetics (2017)

Algorithms tend to output very large subnetworks

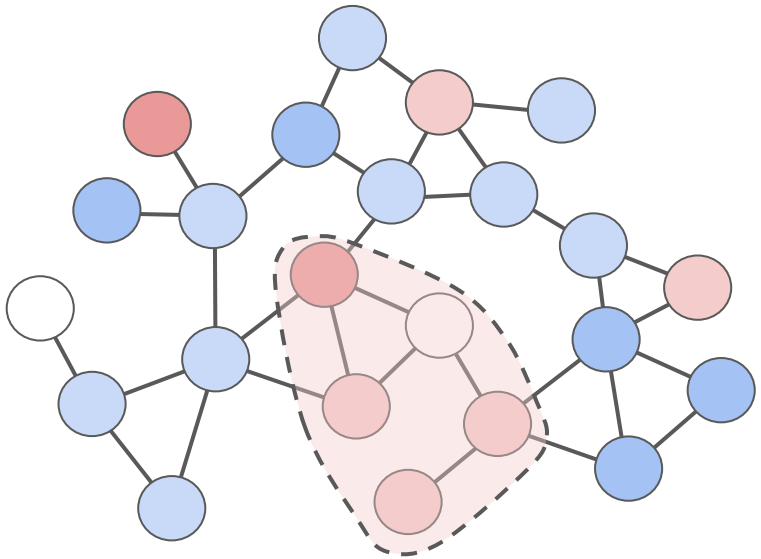
> "Many algorithms are based on the score defined by jActiveModules [8], including PANOGA [9], dmGWAS [10], EW-dmGWAS [11], PINBPA [12], GXNA [13], and PinnacleZ [14]. Others, such as BioNet [15, 16] and Sig-Mod [17] are based on a score adapted to integer linear programming. These methods are also widely applied in the current literature [18, 19, 20, 21, 22, 14, 23, 24, 25, 26], even though the above approaches have been reported to consistently result in subnetworks that are large, and therefore difficult to interpret biologically [13, 27, 28]."

"Network module identification—a widespread theoretical bias and best practices" by Nikolayeva *et al* (Methods 2018)

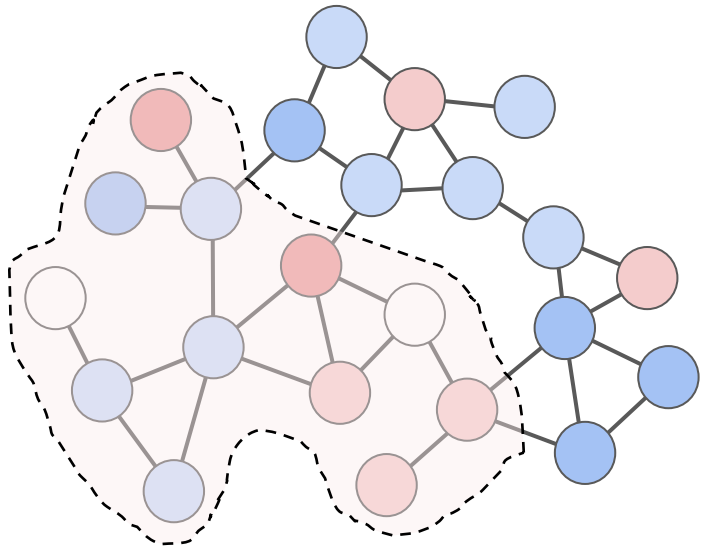
A simple simulation with an implanted subnetwork

Network has **10000** vertices, and implanted altered subnetwork has **500** vertices

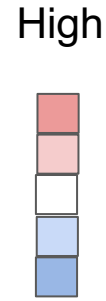
jActiveModules outputs a subnetwork with **2505** vertices (5x increase!)



True



Estimated

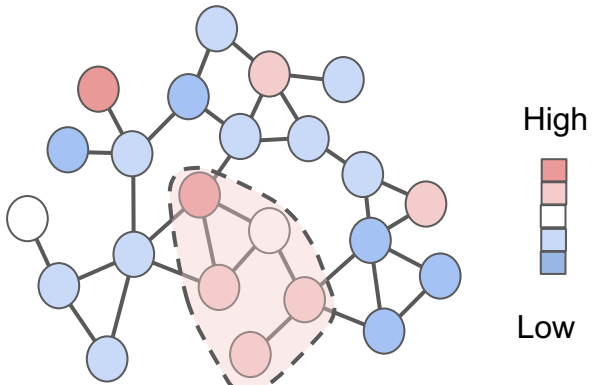


Many **heuristics** for reducing size of altered subnetworks – but their effectiveness is unclear

What does it mean to solve the Altered Subnetwork Problem to "optimality"?

Most algorithms assess their performance using real biological datasets:

- Runtime
- Recovering known biological findings
- Discovery of potentially new biological insights



Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

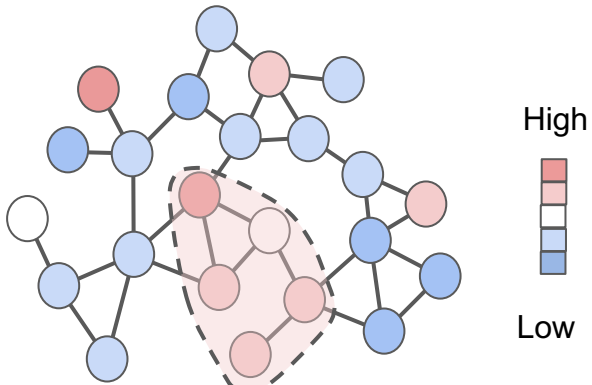
Goal: Identify high-scoring subnetworks H of G

What does it mean to solve the Altered Subnetwork Problem to "optimality"?

Most algorithms assess their performance using real biological datasets:

- Runtime
- Recovering known biological findings
- Discovery of potentially new biological insights

But most algorithms **do not** assess performance on a generative model of the data



Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

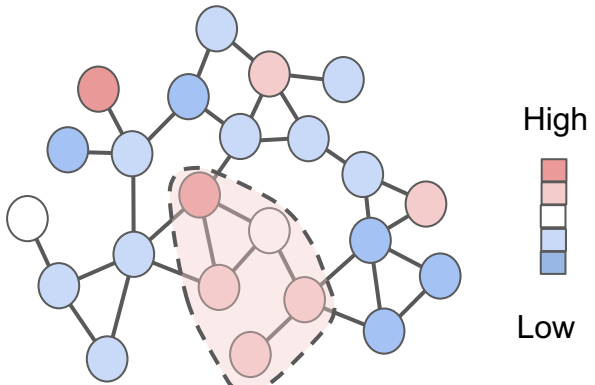
Goal: Identify high-scoring subnetworks H of G

What does it mean to solve the Altered Subnetwork Problem to "optimality"?

Most algorithms assess their performance using real biological datasets:

- Runtime
- Recovering known biological findings
- Discovery of potentially new biological insights

But most algorithms **do not** assess performance on a generative model of the data



Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

Goal: Identify high-scoring subnetworks H of G

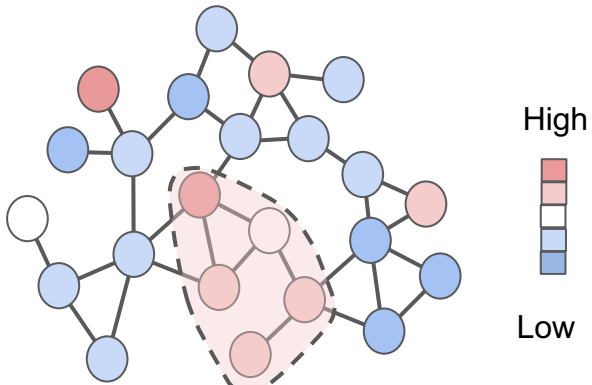
What does it mean to solve the Altered Subnetwork Problem to "optimality"?

Most algorithms assess their performance using real biological datasets:

- Runtime
- Recovering known biological findings
- Discovery of potentially new biological insights

But most algorithms **do not** assess performance on a generative model of the data

- No clear formulation of the problem being solved, so cannot be "solved to optimality"



Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

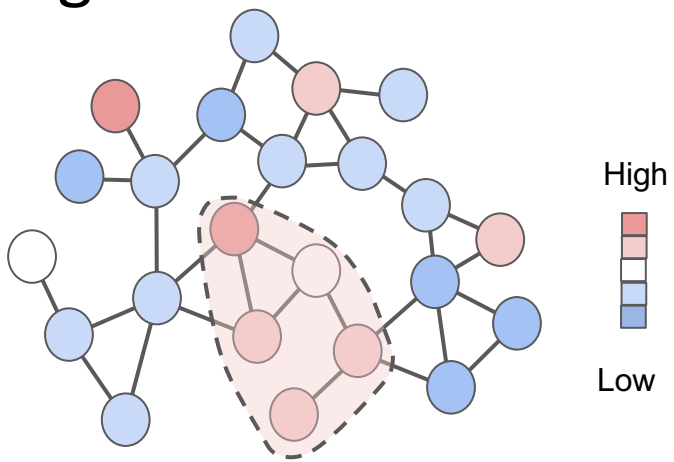
Goal: Identify high-scoring subnetworks H of G

This Work:

1. Generative model for altered subnetworks

2. Issue of identifying large subnetworks is due to **statistical bias**
   - jActiveModules = Maximum Likelihood Estimator (MLE), but MLE is biased

3. Develop NetMix algorithm, which **reduces bias** using mixture models

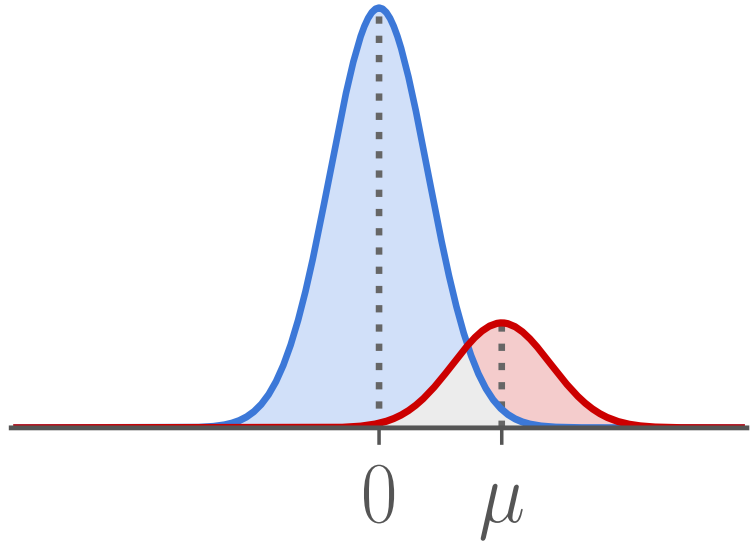


Generative model: Altered Subnetwork Distribution

- G=(V, E) is a graph
- A ⊆ V is a connected subgraph, or the altered subnetwork

Vertex scores $(X_v)_{v \in V}$ are distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$



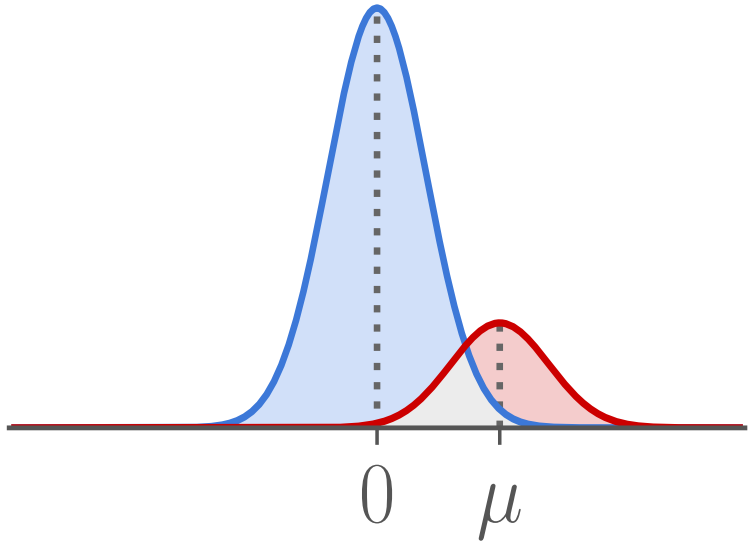
Implicitly the generative model used by jActiveModules (Ideker et al 2002)

Generative model: Altered Subnetwork Distribution

- G=(V, E) is a graph
- A ⊆ V is a connected subgraph, or the altered subnetwork

Vertex scores $(X_v)_{v \in V}$ are distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$



Statistical interpretation: Vertex scores X_v correspond to p-values p_v from an asymptotically normal test statistic:

$$X_v = \Phi^{-1}(1 - p_v)$$

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

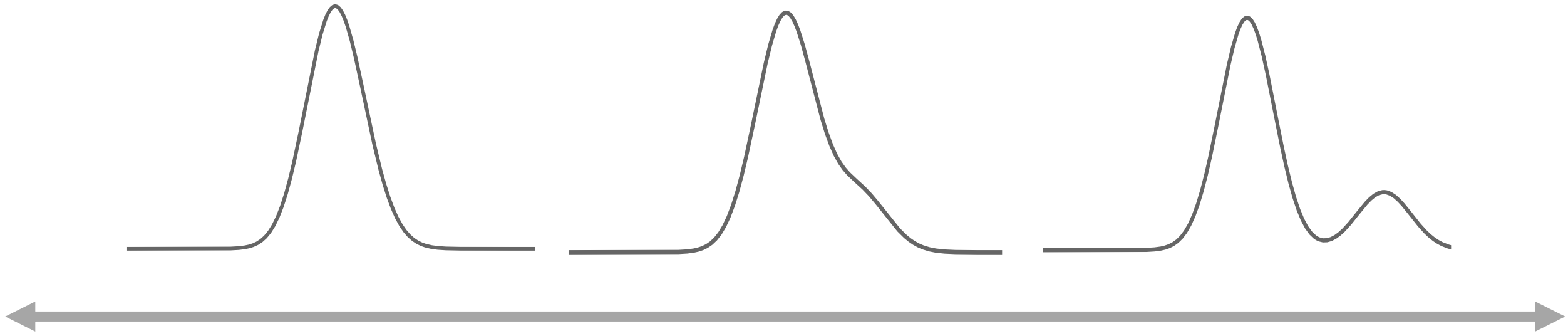
Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

Hard to solve ASP

Easy to solve ASP



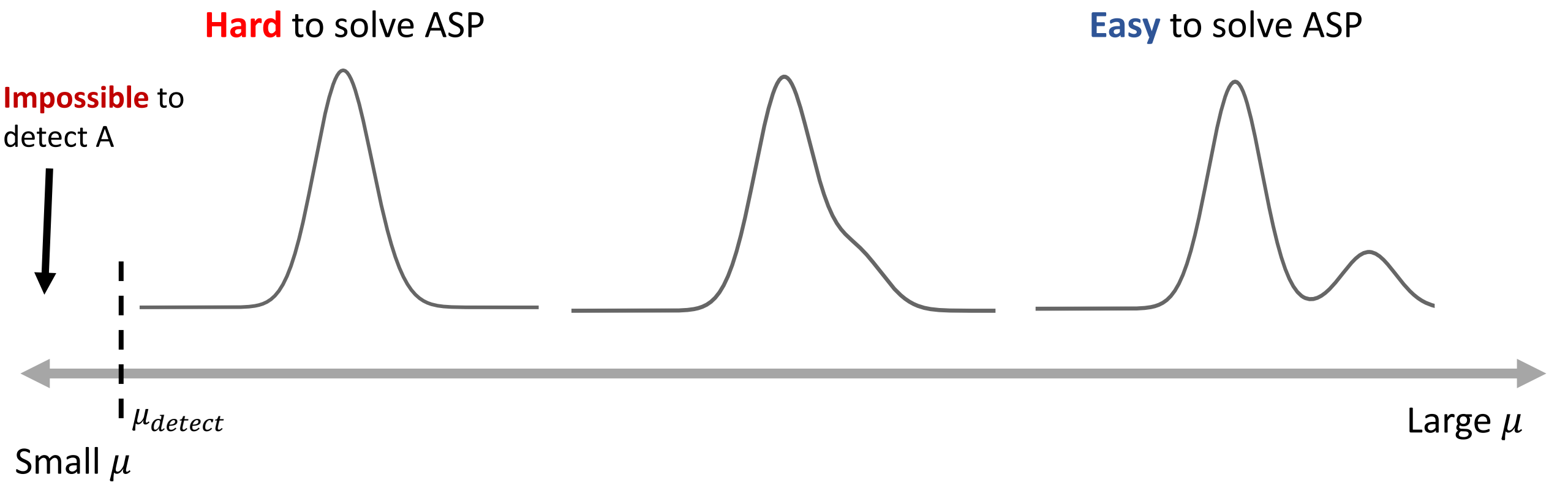
Small μ

Large μ

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

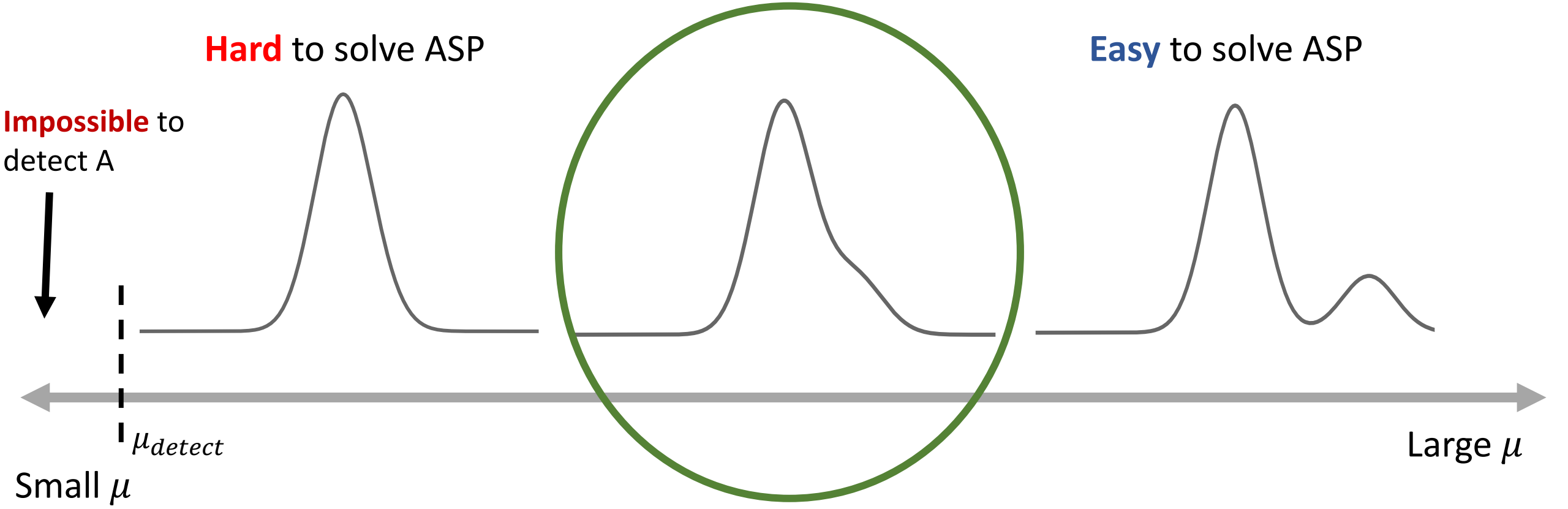
find the altered subnetwork A.



Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.



Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

Theorem: **Maximum Likelihood Estimator (MLE)** of the altered subnetwork A is:

$$\widehat{A}_{\text{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$$

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

Theorem: **Maximum Likelihood Estimator (MLE)** of the altered subnetwork A is:

$$\widehat{A}_{\text{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$$

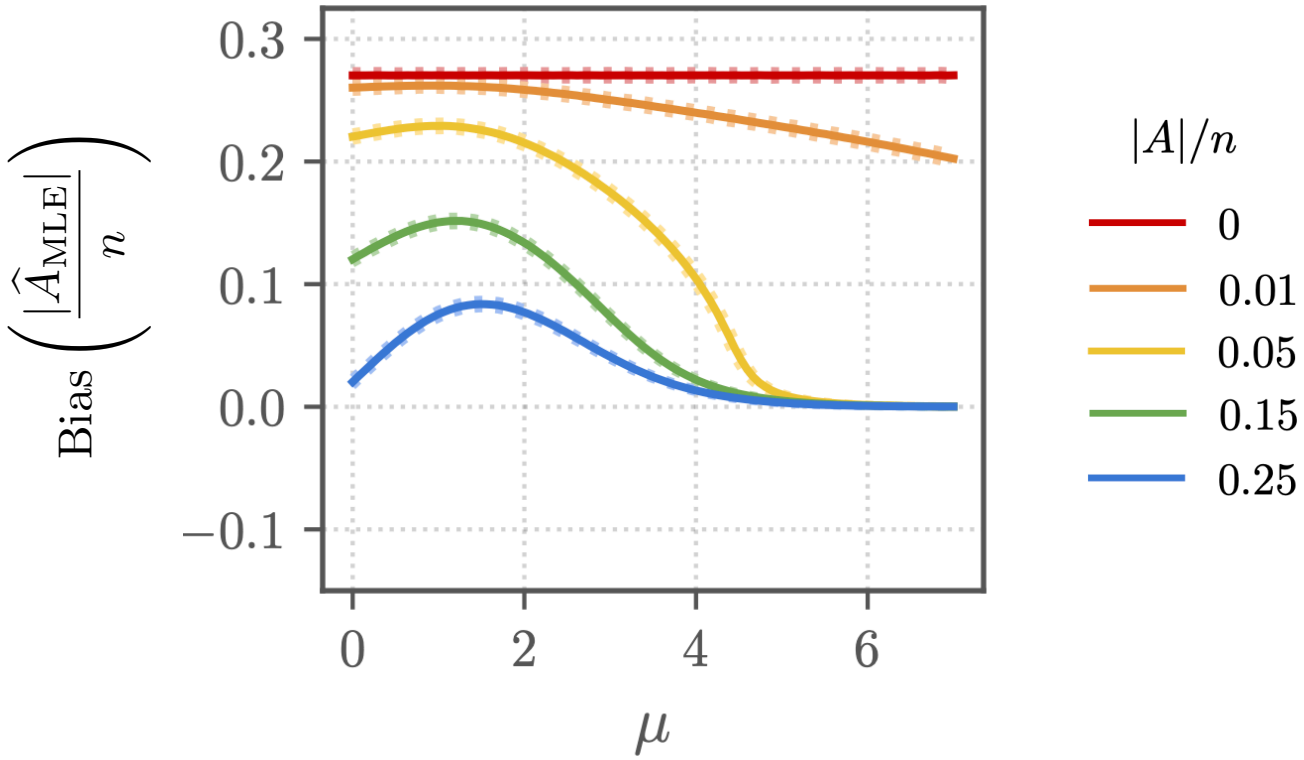
MLE = jActiveModules!

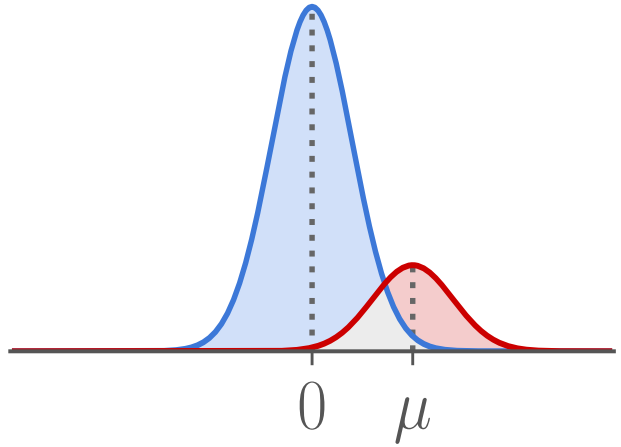
jActiveModules paper (Ideker et al, 2002) does not describe generative model nor the connection to the MLE

MLE is biased estimator

$$\text{Bias}\left(\frac{|\widehat{A}_{\text{MLE}}|}{n}\right) \triangleq E\left[\frac{|\widehat{A}_{\text{MLE}}|}{n}\right] - \frac{|A|}{n}$$

We observe that MLE has **positive bias:** MLE overestimates the size $\frac{|A|}{n}$ of the altered subnetwork on average (where n=|V|)

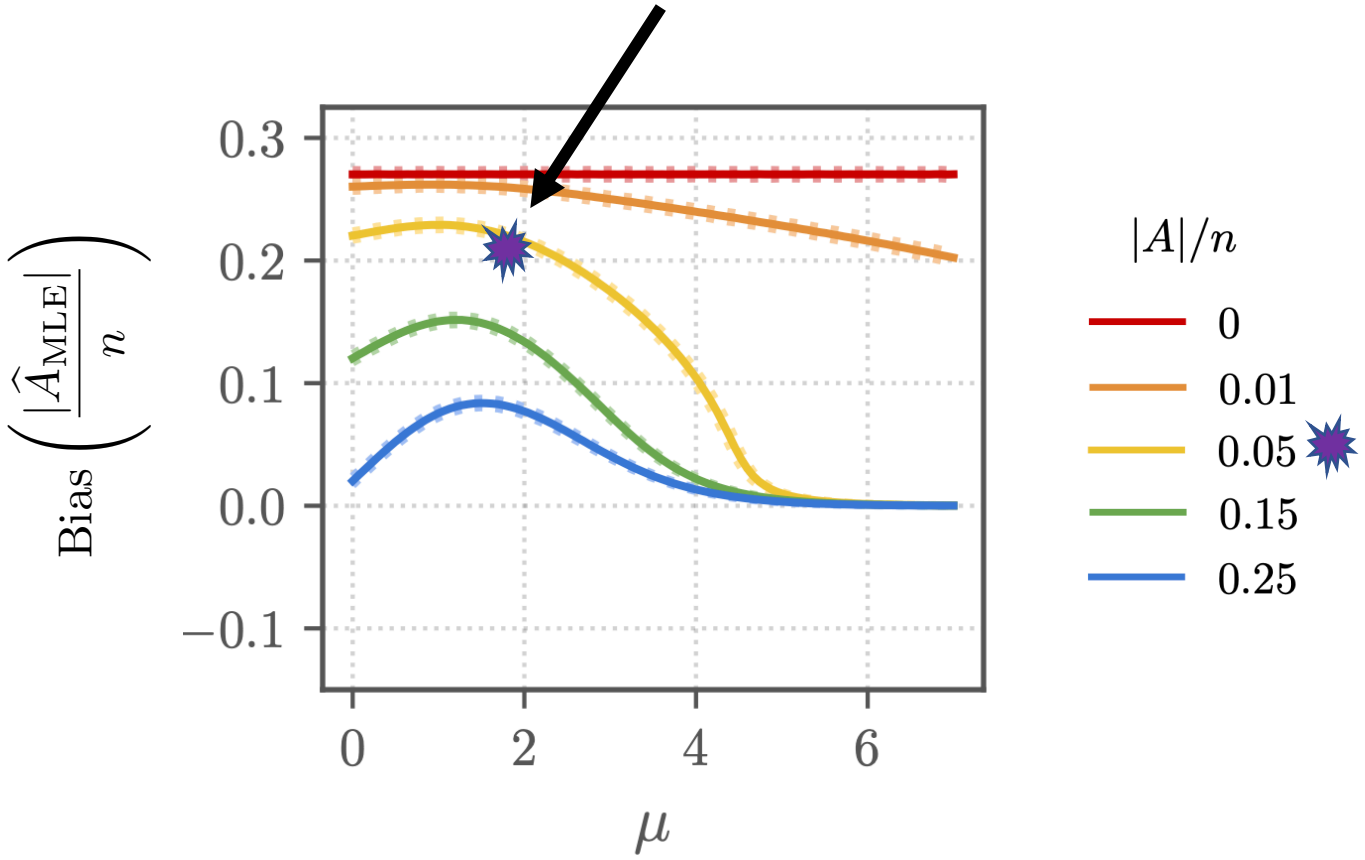


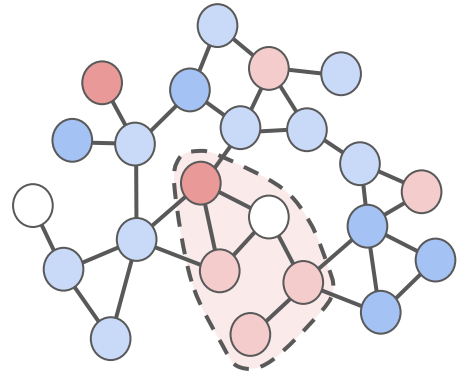


MLE is biased estimator

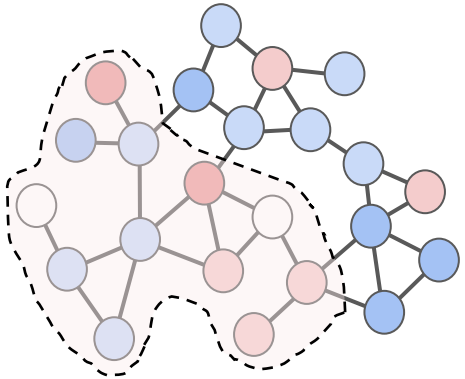
$$\text{Bias}\left(\frac{|\widehat{A}_{\text{MLE}}|}{n}\right) \triangleq E\left[\frac{|\widehat{A}_{\text{MLE}}|}{n}\right] - \frac{|A|}{n}$$

Altered subnetwork contains 5% of vertices, but MLE contains 5%+20% = 25% of vertices!





Altered subnetwork



MLE

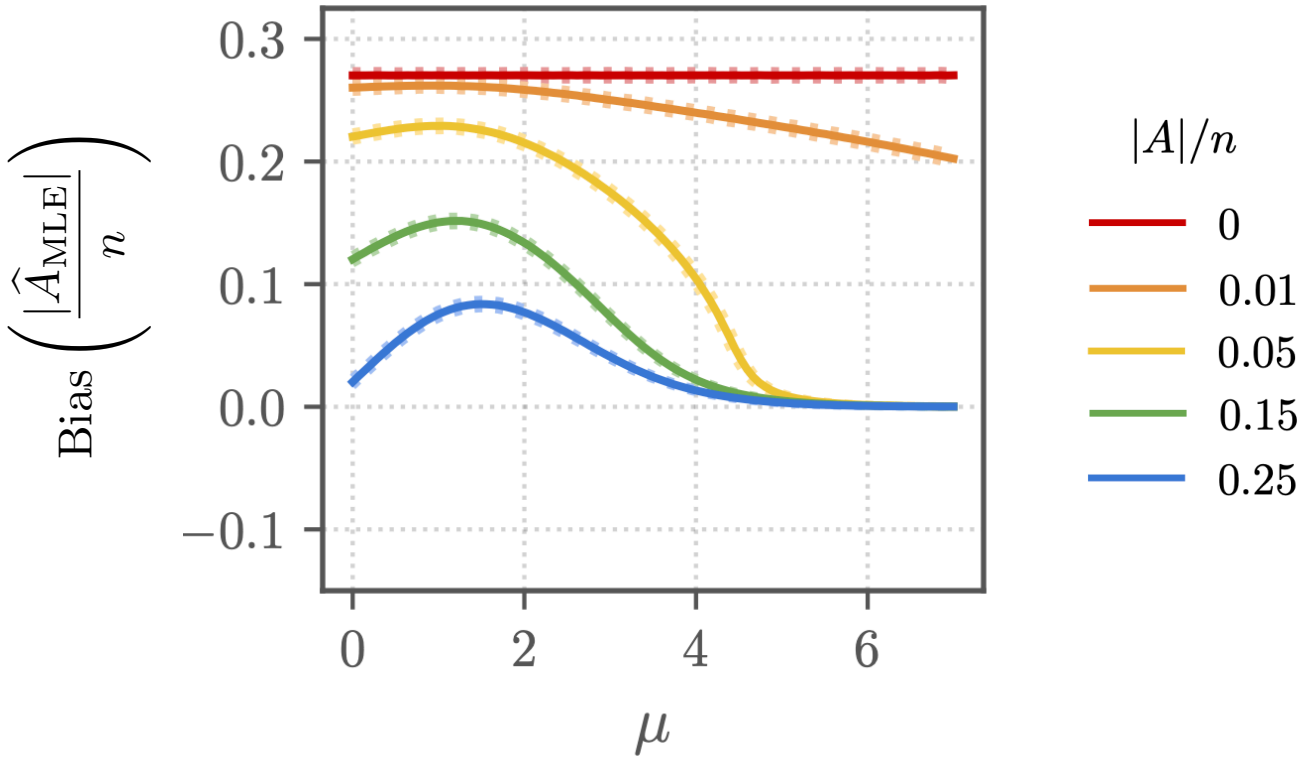
High

Low

MLE is biased estimator

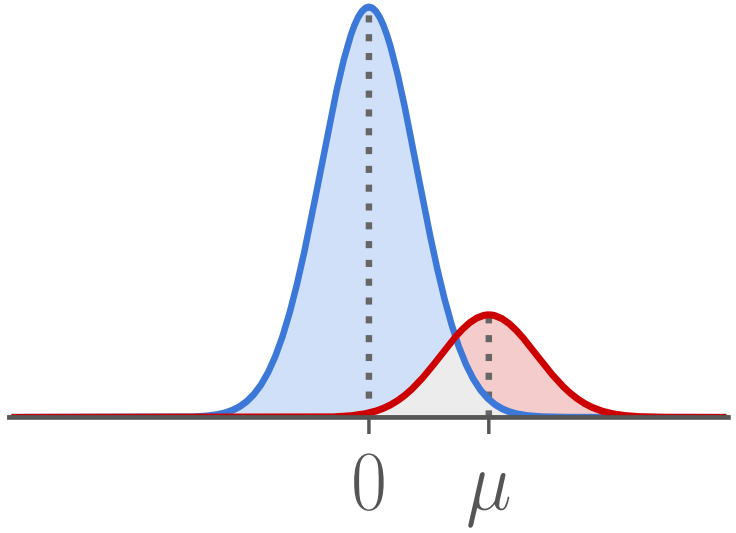
$$\text{Bias}\left(\frac{|\widehat{A}_{\text{MLE}}|}{n}\right) \triangleq E\left[\frac{|\widehat{A}_{\text{MLE}}|}{n}\right] - \frac{|A|}{n}$$

In the paper, we **prove** two results that partially show asymptotic bias of the MLE



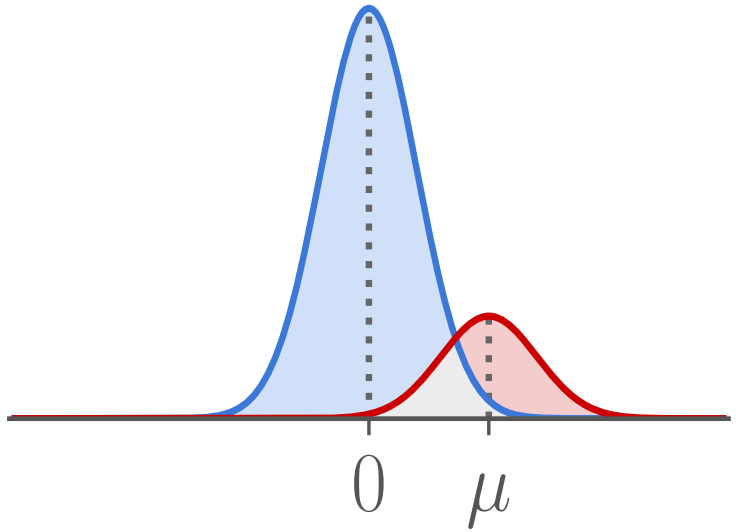
How to reduce bias?

Key idea: Model the distribution of the vertex scores before using the network



How to reduce bias?

Key idea: Model the distribution of the vertex scores before using the network



Fit vertex scores to **Gaussian Mixture Model (GMM):**

$$X_v \sim (1-\alpha)\cdot N(0,1) + \alpha \cdot N(\mu,1)$$

α = proportion of vertices in altered subnetwork

μ = mean of altered subnetwork distribution

GMM yields less biased estimate of altered subnetwork size

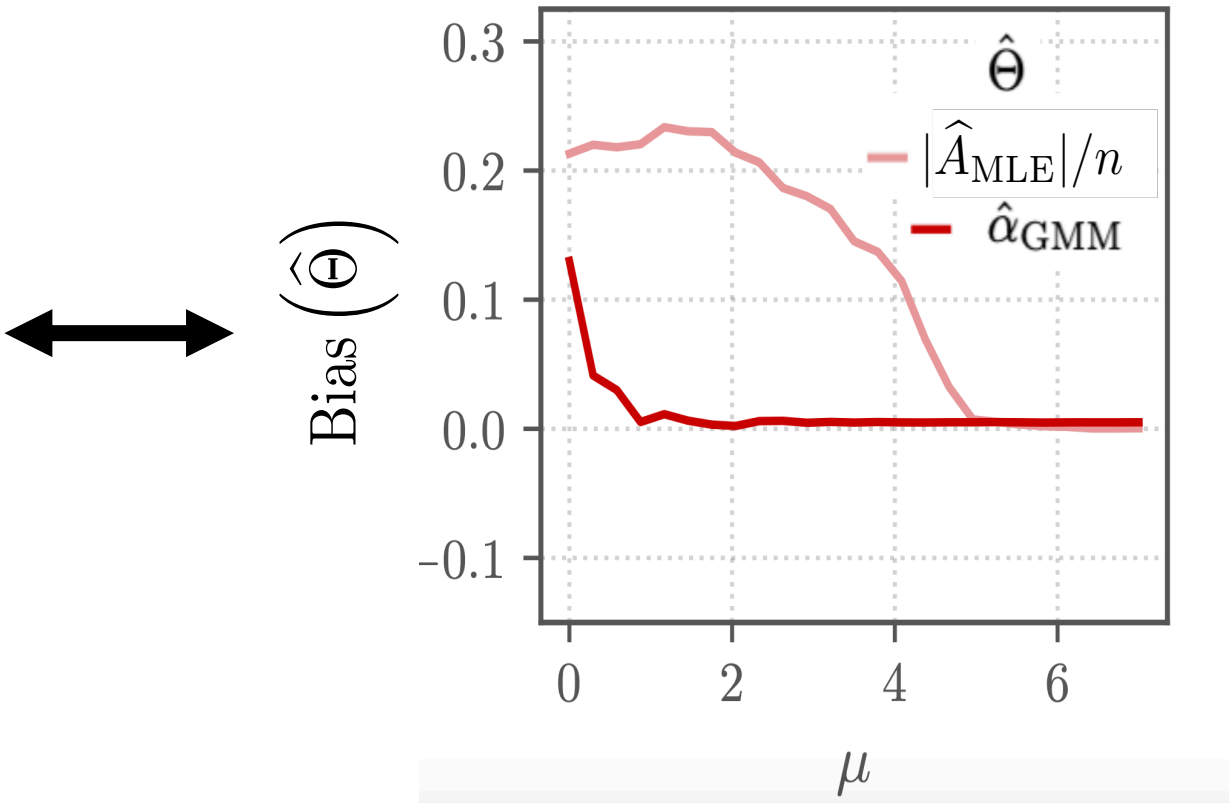
MLE: $\widehat{A}_{\mathrm{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$

vs

GMM: Fit vertex scores X_v to GMM

$$X_v \sim (1-\alpha) \cdot N(0,1) + \alpha \cdot N(\mu, 1)$$

and estimate GMM parameters $\widehat{\alpha}_{\mathrm{GMM}}, \widehat{\mu}_{\mathrm{GMM}}$



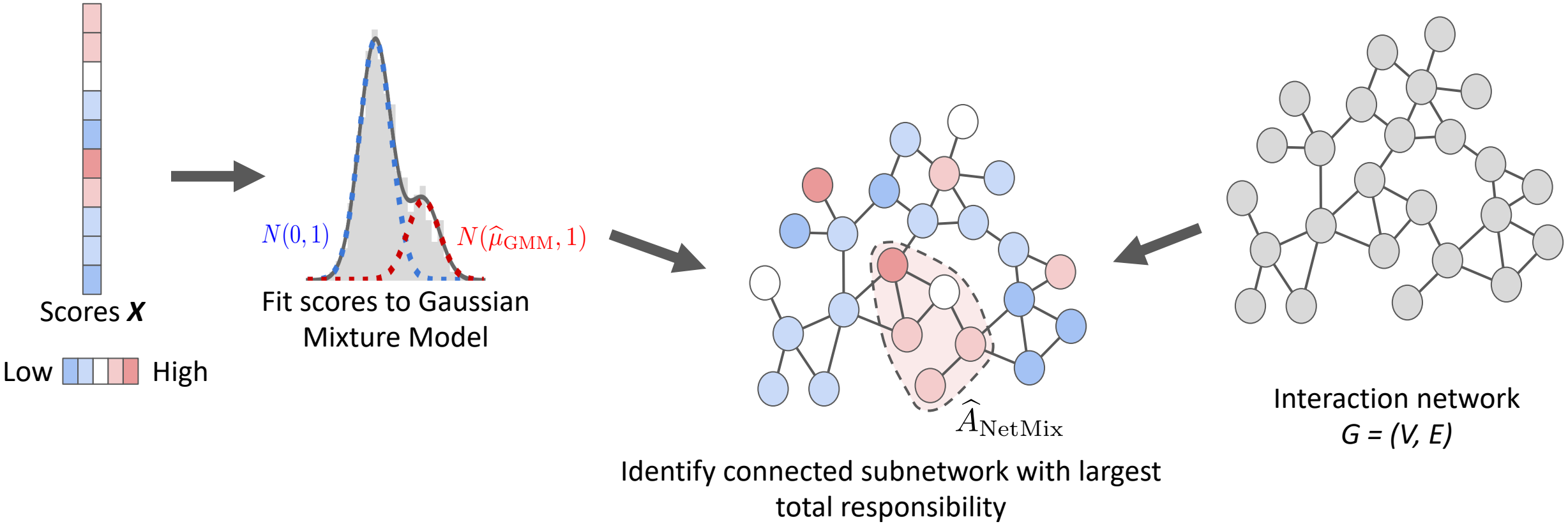
α = proportion of vertices in altered subnetwork
μ = mean of altered subnetwork distribution

Altered subnetwork A has size |A|/n = 0.05

NetMix Algorithm

Given vertex scores $(X_v)_{v \in V}$ and graph G:

1. Fit scores to GMM using EM, and compute *responsibilities* $r_v = P(v \in A \mid X_v)$
2. Find connected subnetwork $\widehat{A}_{\mathrm{NetMix}}$ with size $\approx \widehat{\alpha}_{\mathrm{GMM}} n$ and <u>largest total responsibility</u>



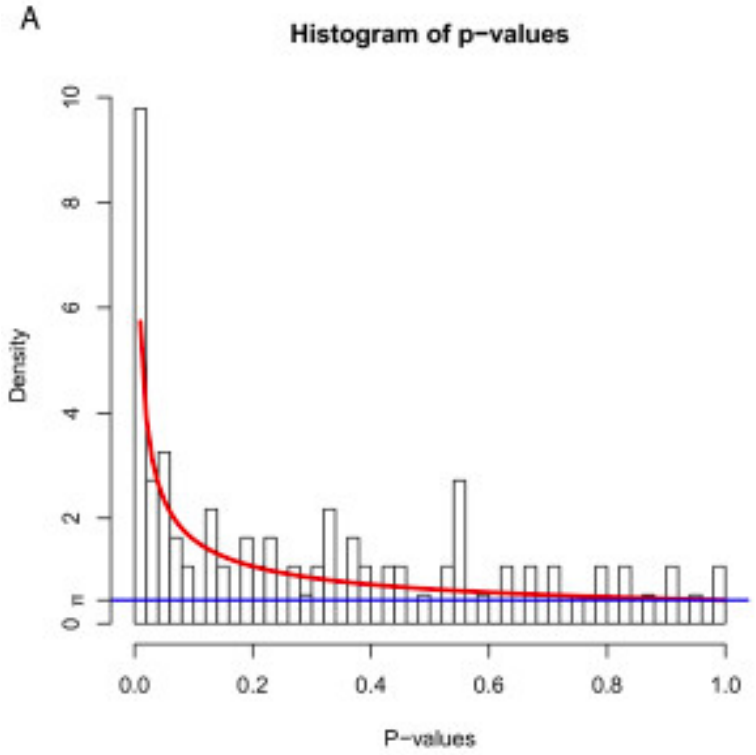
Comparison to heinz algorithm

heinz (Dittrich et al, 2008) also models distribution of vertex scores.

Two key differences:

1. **<u>Different distributions</u>:** heinz models p-values with Beta-Uniform Mixture (BUM)
   - BUM sometimes underestimates size of altered subnetwork (Pounds and Cheng 2004)

2. **<u>User-defined parameter</u>:** False Discovery Rate (FDR)
   - Most values of FDR result in **biased** estimate of size of altered subnetwork
   - Can selectively tune FDR, similar to "p-hacking"
     - In literature, FDRs range anywhere from 10^{-25} to 0.5



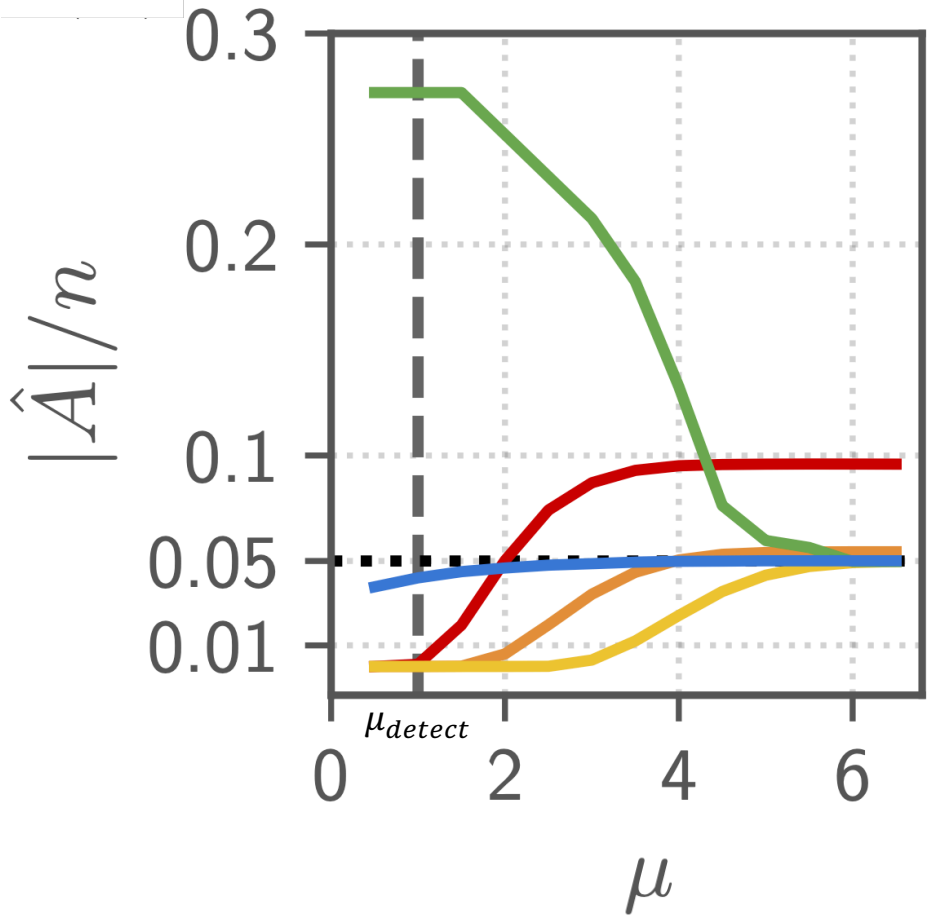
p-values fit to Beta-Uniform Mixture

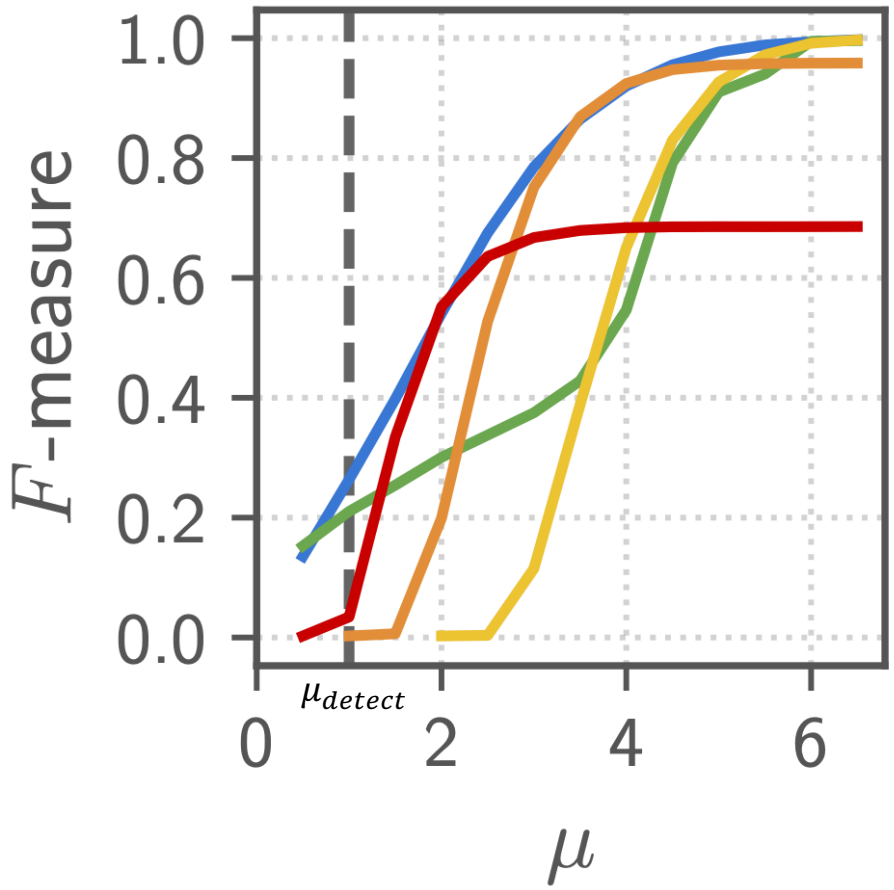
Neither is consistent with *"solving the [Altered Subnetwork Problem] to optimality"*

Results – simulated data

G = HINT+HI interaction network with |G|=15074 nodes (Leiserson et al 2015)

Altered subnetwork A = connected subgraph of size |A|=0.05n selected uniformly at random from *G*

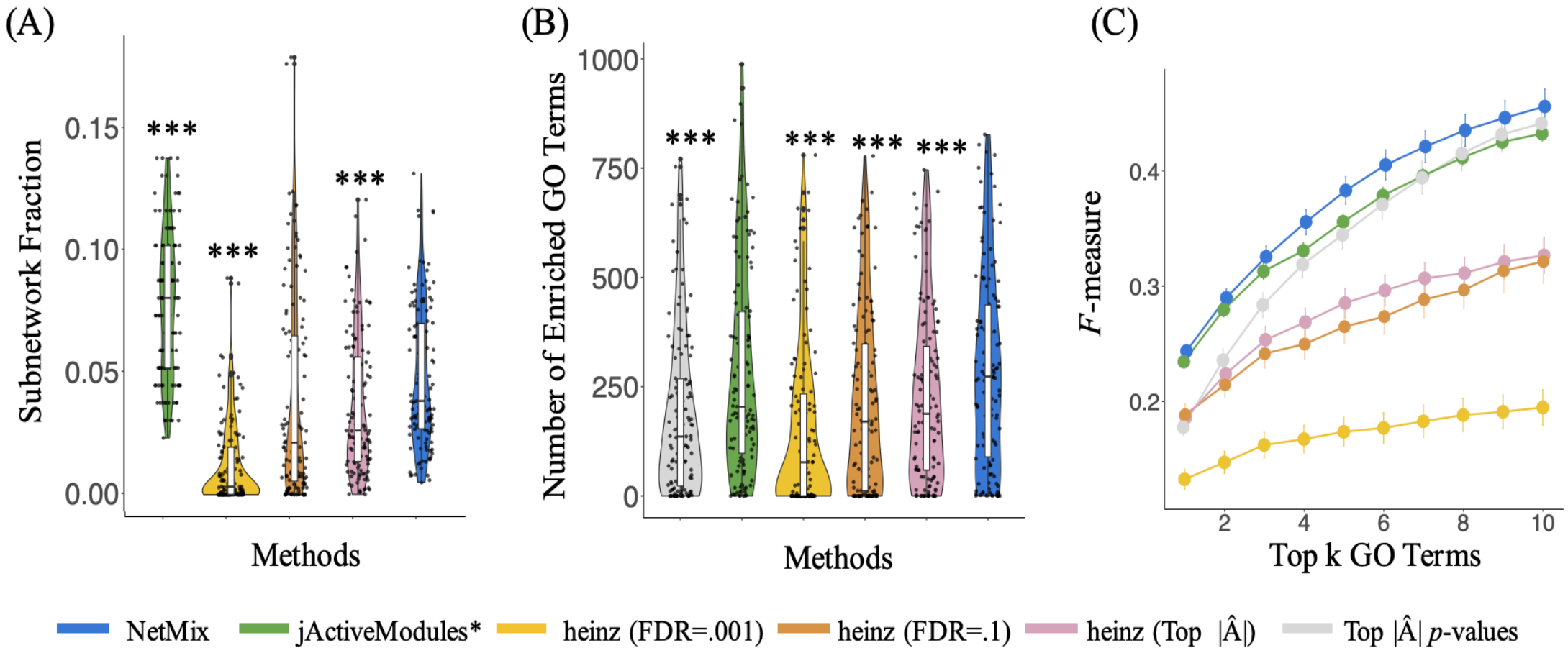




NetMix jActiveModules* heinz (FDR = 0.001) heinz (FDR = 0.1) heinz (FDR = 0.5)

Results – differential gene expression

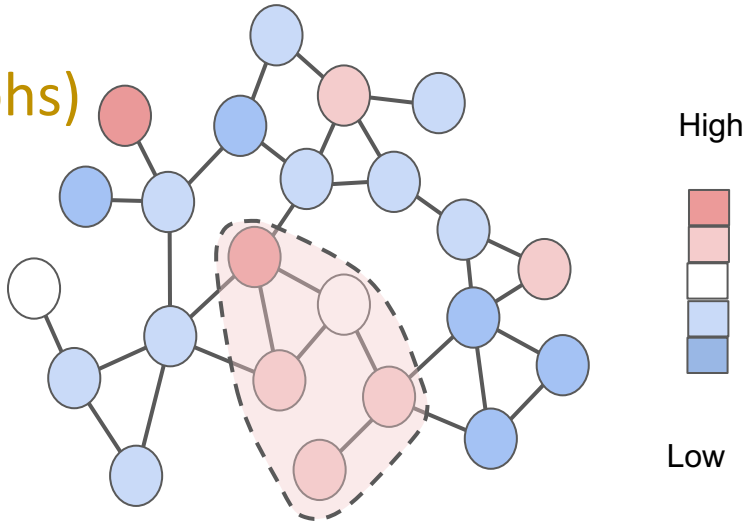
157 gene expression experiments from Expression Atlas (Petryszak et al, 2015), including both microarray and RNA-seq experiments



In the paper we also show experiments on somatic mutations in cancer

Summary + Future Directions

1. Generative model for altered subnetworks
2. jActiveModules = MLE, but MLE is statistically **biased**, explaining reports of large subnetworks.
3. Develop the **NetMix** algorithm, which uses mixture models to reduce bias.
   - Multiple altered subnetworks
   - Other topological constraints (e.g. edge-dense subgraphs)
   - Additional applications



Acknowledgements

Raphael Group:

Ben Raphael

Matt Reyna

Rebecca Elyanow

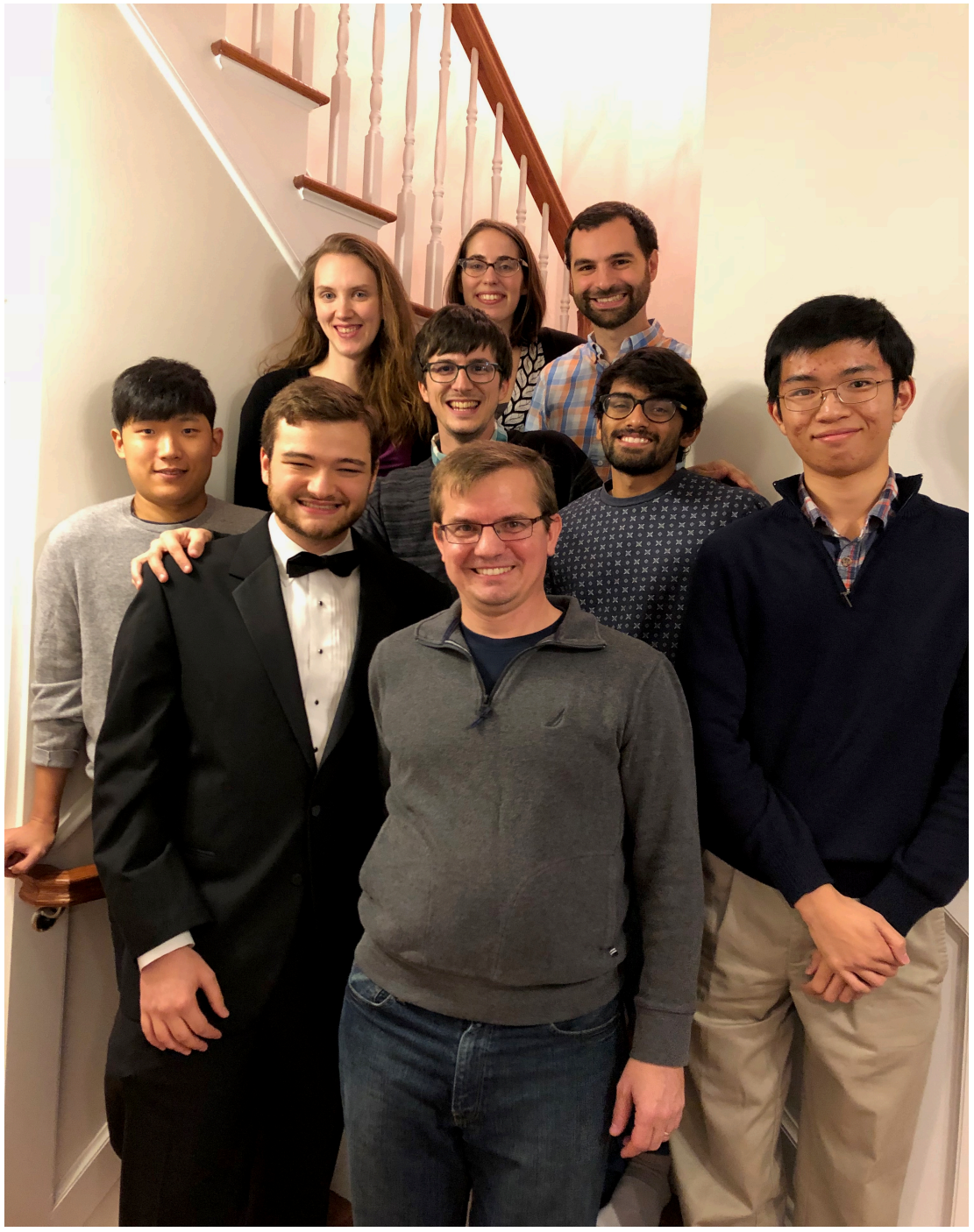
Matt Myers

Simone Zaccaria

Ron Zeira

Tyler Park

Gryte Satas



Code: https://github.com/raphael-group/netmix
Paper (bioRxiv): https://bit.ly/3ea7f3n