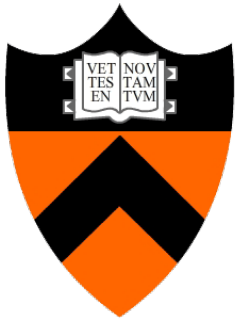
NetMix2: Unifying network propagation and altered subnetworks

Uthsav Chitra* , Tae Yoon (Tyler) Park*, Ben Raphael

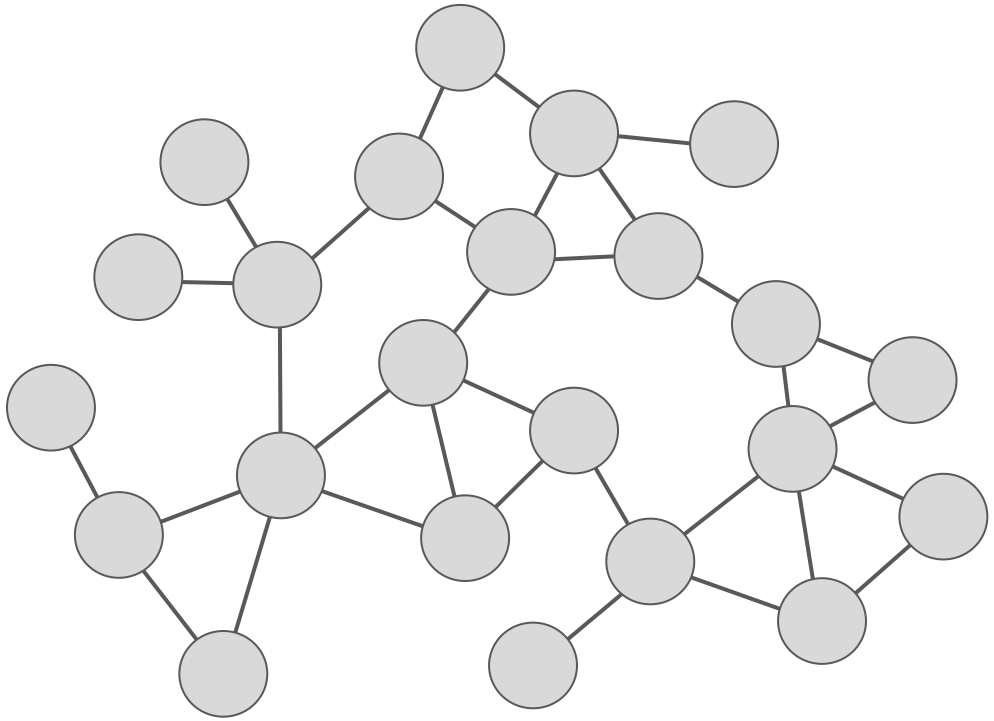
RECOMB 2022



PRINCETON UNIVERSITY

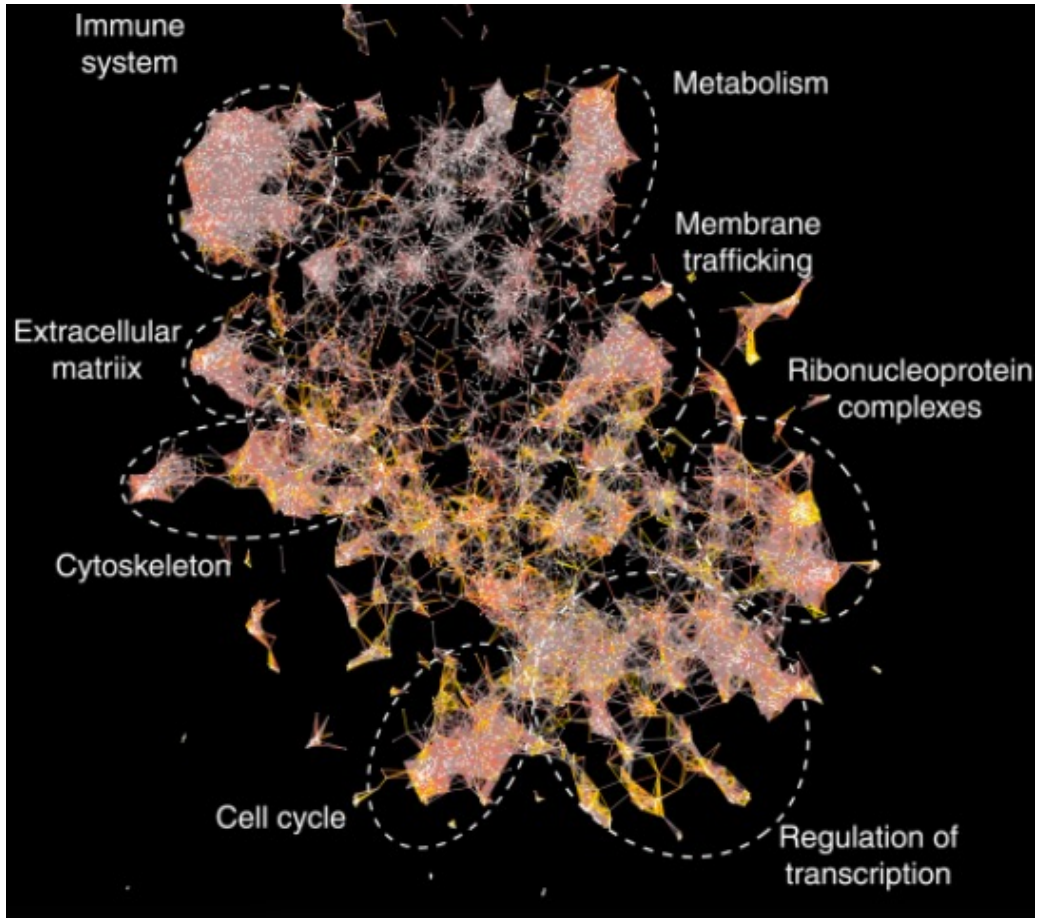
Interaction Networks

Biological interaction networks are often used as prior information when analyzing high throughput ‘omics data



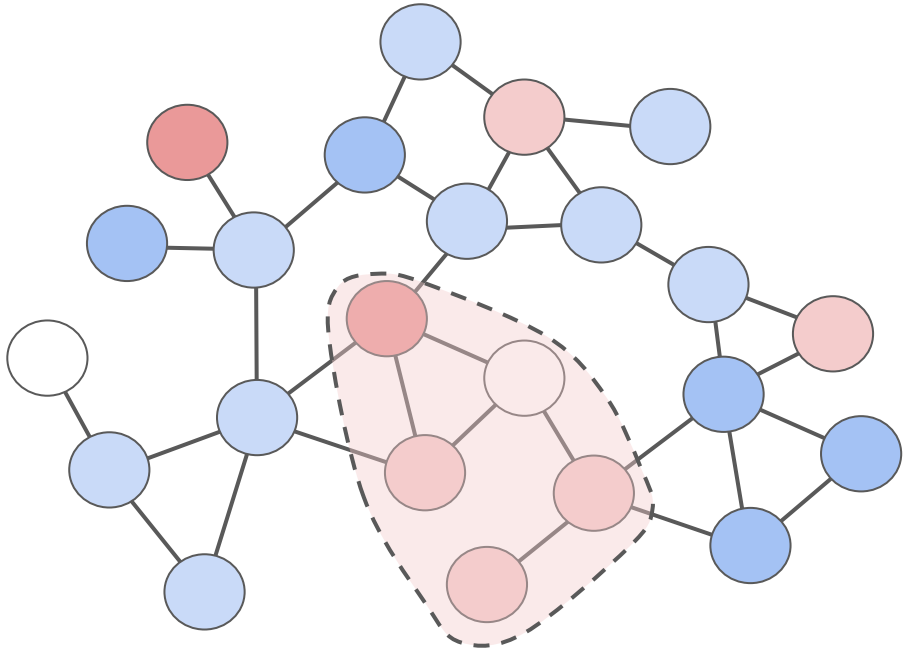
Vertices: genes or proteins

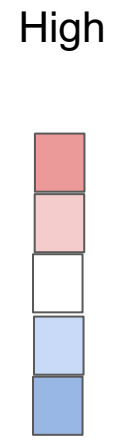
Edges: Interactions between genes/proteins



Proteins with similar functions are connected in an interaction network

Altered Subnetwork Problem (also called network modules, active subnetworks)





Given:

1) Interaction network G = (V,E)
2) Vertex scores X_v

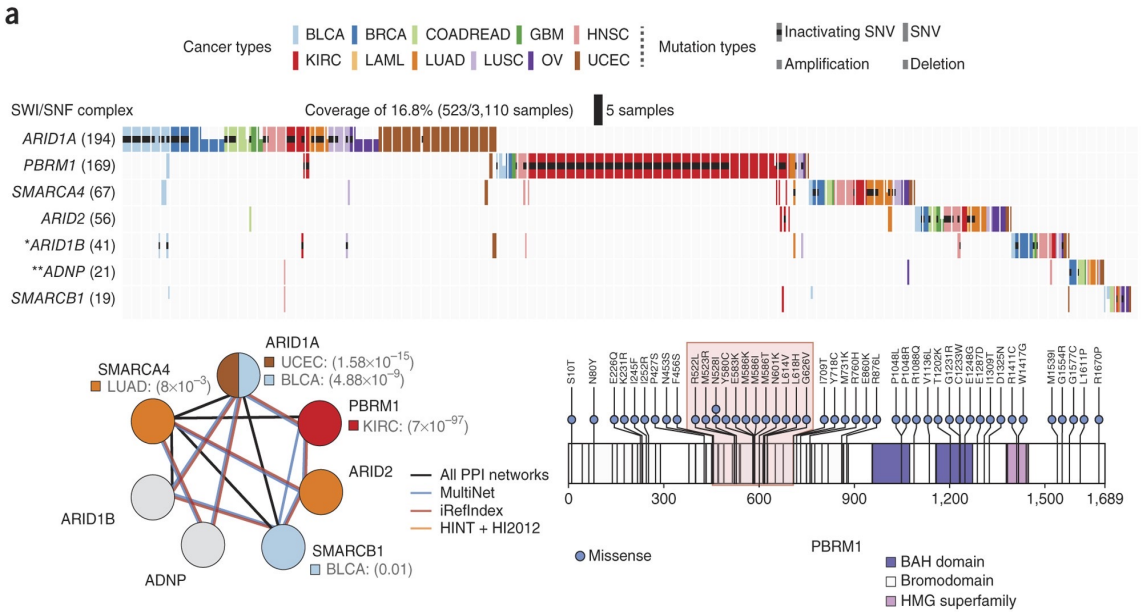
Goal: Identify **high-scoring subnetworks** of G ("altered subnetworks")

Note: Different from DREAM competition (Lenore Cowen keynote), which does not use vertex scores

Altered subnetworks reveal important pathways

Altered subnetworks = functionally related genes/proteins (eg disease genes)

Somatic mutations in cancer



Leiserson, Vandin et al (Nature Genetics 2015)

Vertex scores X_v - somatic mutation frequencies

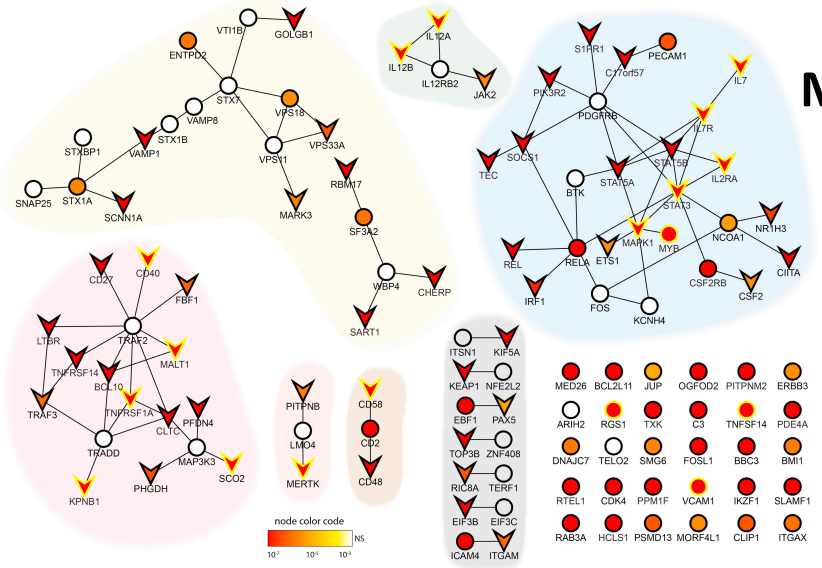
Altered Subnetwork Problem:

Given:

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

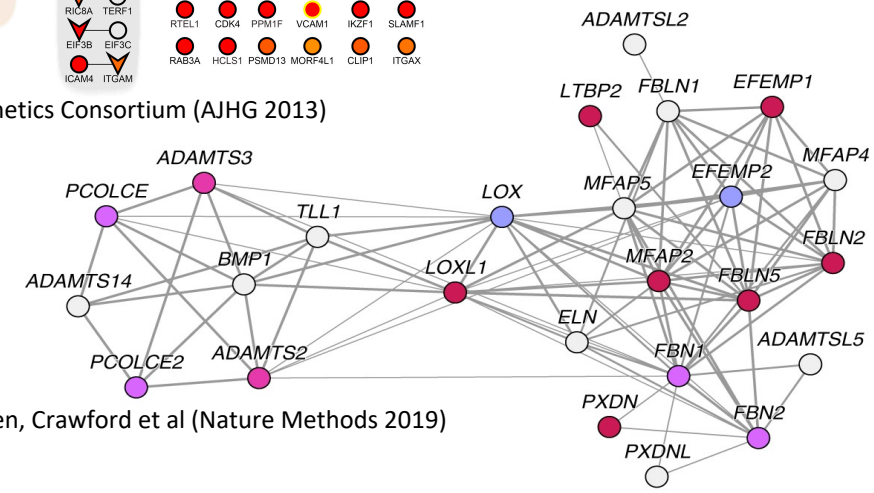
Goal: Identify high-scoring subnetworks G

Genome-wide association studies (GWAS)



Multiple Sclerosis

International Multiple Sclerosi Genetics Consortium (AJHG 2013)



Height

Choobdar, Ahsen, Crawford et al (Nature Methods 2019)

Vertex scores X_v – GWAS gene-level p-values

Many algorithms developed over past 20 years for identifying altered subnetworks

Table 1 | **Some recent bioinformatics tools for module extraction through network integration**

Tool	URL	Refs
Active-module detection through network projection of omics data		
jActiveModules	http://apps.cytoscape.org/apps/jactivemodules	48
MATISSE	http://acgt.cs.tau.ac.il/matisse	165
PinnacleZ	http://apps.cytoscape.org/apps/pinnaclez	62
GXNA	http://stat.stanford.edu/~serban/gxna	52
BioNet	http://bionet.bioapps.biozentrum.uni-wuerzburg.de	166
COSINE	http://cran.r-project.org/web/packages/COSINE/index.html	104
SANDY	http://sandy.topnet.gersteinlab.org	81
HotNet	http://ccmbweb.ccv.brown.edu/hotnet	67
PARADIGM	http://sbenz.github.com/Paradigm	70
MEMo	http://cbio.mskcc.org/memo	73
Multi-Dendrix	http://compbio.cs.brown.edu/software	37
RegMOD	http://www.biomedcentral.com/1471-2105/11/26/additional	45
NetWalk and FunWalk	http://netwalkersuite.org	76
ResponseNet	http://bioinfo.bgu.ac.il/respnet	75
ClustEx	http://www.mybiosoftware.com/pathway-analysis/5495	42
SAMBA	http://acgt.cs.tau.ac.il/samba	82
cMonkey	http://bonneaulab.bio.nyu.edu/biclustering.html	69
COBRAv2.0	http://opencobra.sourceforge.net/openCOBRA/Welcome.html	85
TieDIE	https://sysbiowiki.soe.ucsc.edu/tiedie	167
Network comparisons across species to identify conserved modules		
PathBLAST	http://www.pathblast.org	114
NetworkBLAST	http://www.cs.tau.ac.il/~bnet/networkblast.htm	168
NetworkBLAST-M	http://www.cs.tau.ac.il/~bnet/License-nbm.htm	116
IsoRankN	http://groups.csail.mit.edu/cb/mna	169
Graemlin	http://graemlin.stanford.edu	119
NeXus	http://csbio.cs.umn.edu/neXus/help.html	157
Multi-species cMonkey	http://bonneaulab.bio.nyu.edu/biclustering.html	158
Differential analysis of interaction networks to identify dynamic modules		
DDN	http://www.cbil.ece.vt.edu/software.htm	170
DNA	http://www.somnathdatta.org/Supp/DNA	171
Integration of diverse types of interaction networks to identify composite modules		
PanGIA	http://prosecco.ucsd.edu/PanGIA	147

Mitra *et al*, Nature Reviews Genetics (2013)

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)

Early algorithms model altered subnetwork as a connected subgraph

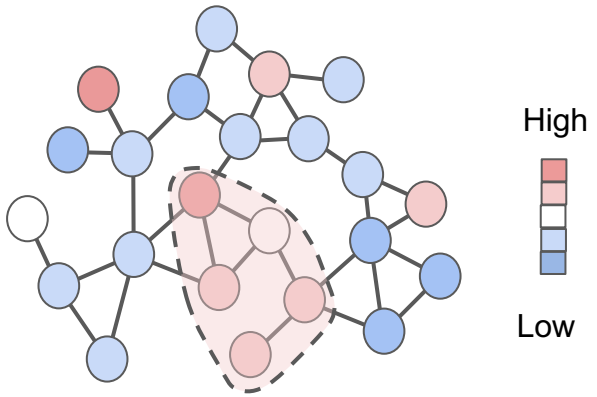
For example, seminal algorithms jActiveModules and heinz solve optimization problems over **connected subgraphs S**

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

jActiveModules/Cytoscape (Ideker et al, 2002)

$$\max_{S \subseteq V} \sum_{v \in S} w_v$$

heinz/BioNet (Dittrich, Klau et al, 2008)



Altered Subnetwork Problem:

Given:

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

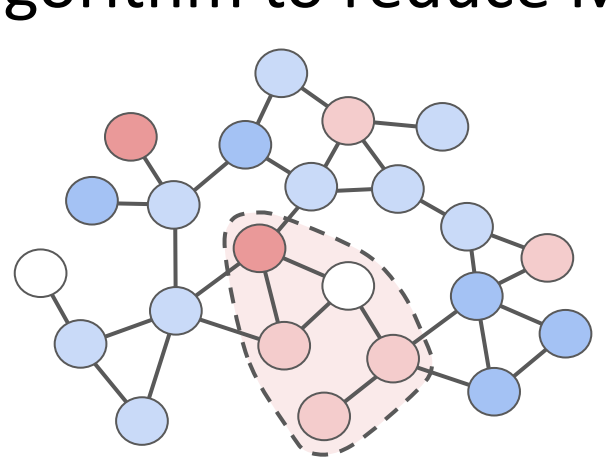
Goal: Identify high-scoring subnetworks G

Connectivity-based algorithms have **theoretical guarantees**

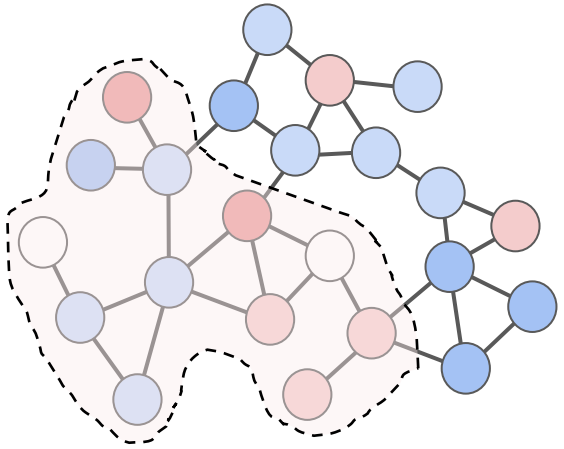
In previous work (RECOMB 2020, ICML 2021) we defined a generative model for connected altered subnetworks and:

1. Showed that existing connectivity-based methods (jActiveModules, heinz) compute *maximum likelihood estimators* (MLE), but MLE is statistically **biased** estimator of subnetwork size

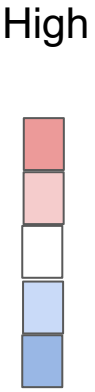
2. Derived **NetMix** algorithm to reduce MLE bias



Altered subnetwork



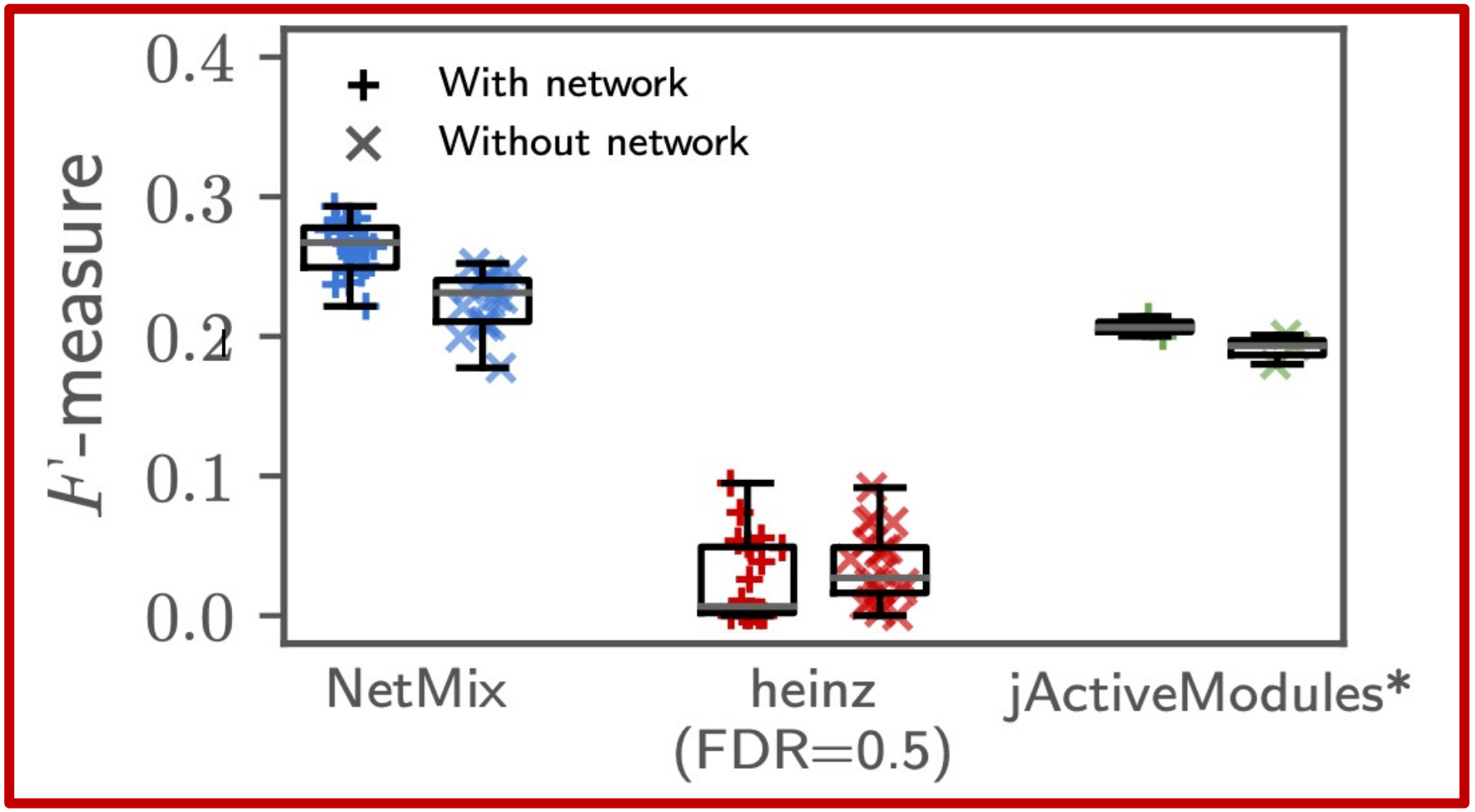
Maximum Likelihood Estimator (MLE)



Challenge: Connectivity is a weak topological constraint!

Networks have small diameter – most subnetworks are “almost connected”

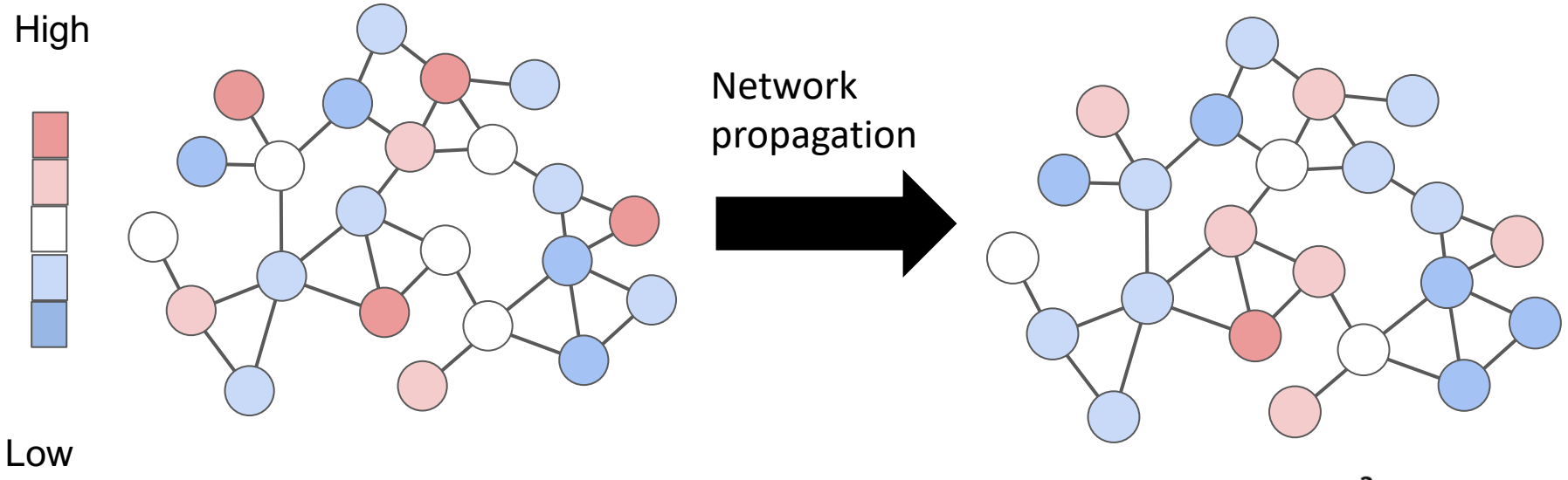
Algorithms not much better compared to not using interaction network



Simulations from our generative model where altered subnetwork is **connected subgraph**

Network propagation (network diffusion)

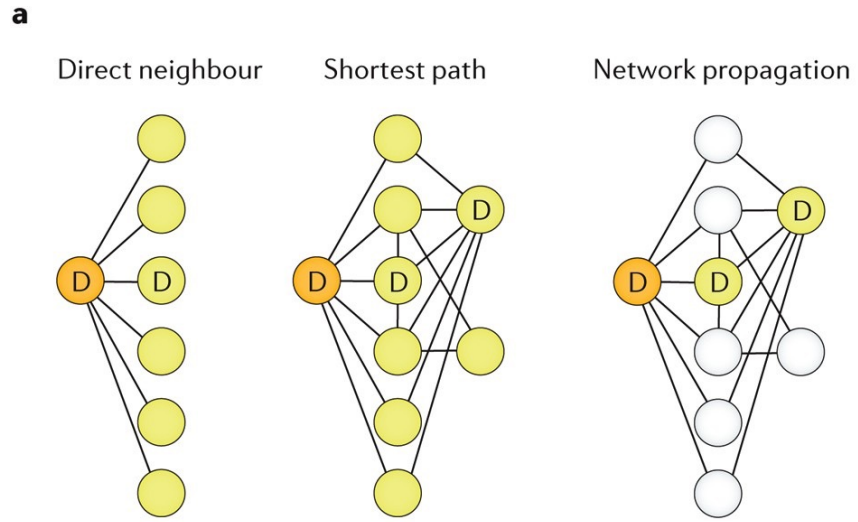
Use of random walks to “propagate”/smooth vertex scores across network



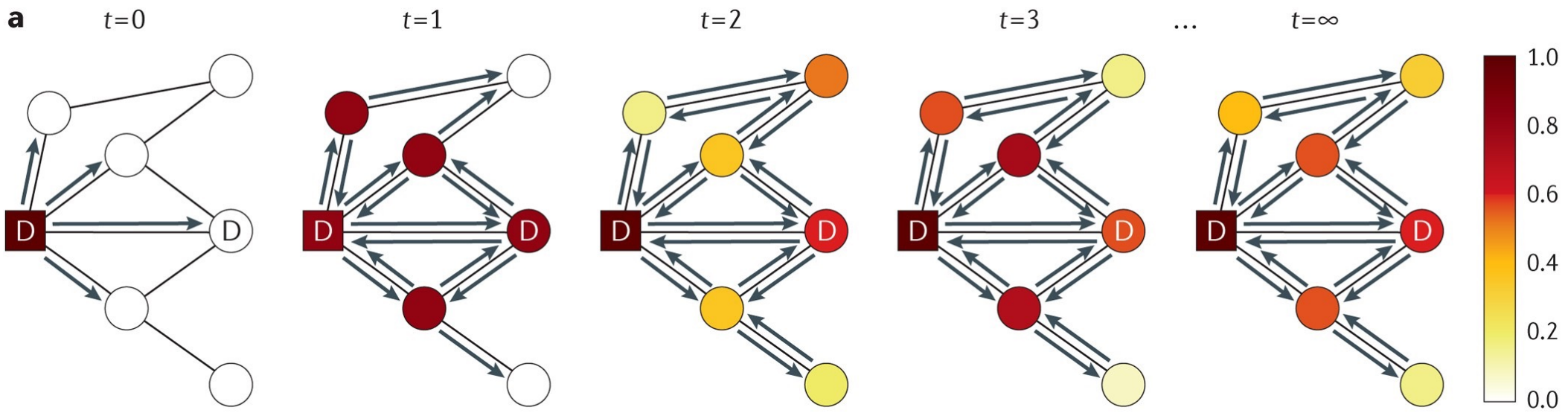
Network propagation: a universal amplifier of genetic associations

Lenore Cowen, Trey Ideker, Benjamin J. Raphael & Roded Sharan

Nature Reviews Genetics **18**, 551–562 (2017) | Cite this article

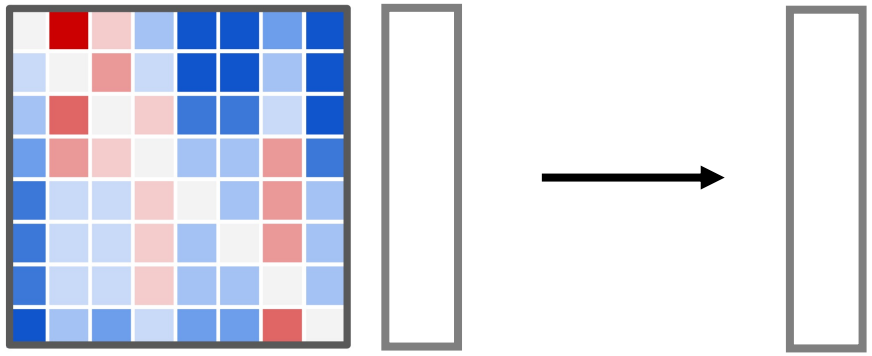


Network propagation uses global network structure



Cowen et al (Nature Reviews Genetics 2017)

Network propagation = Matrix-vector multiplication



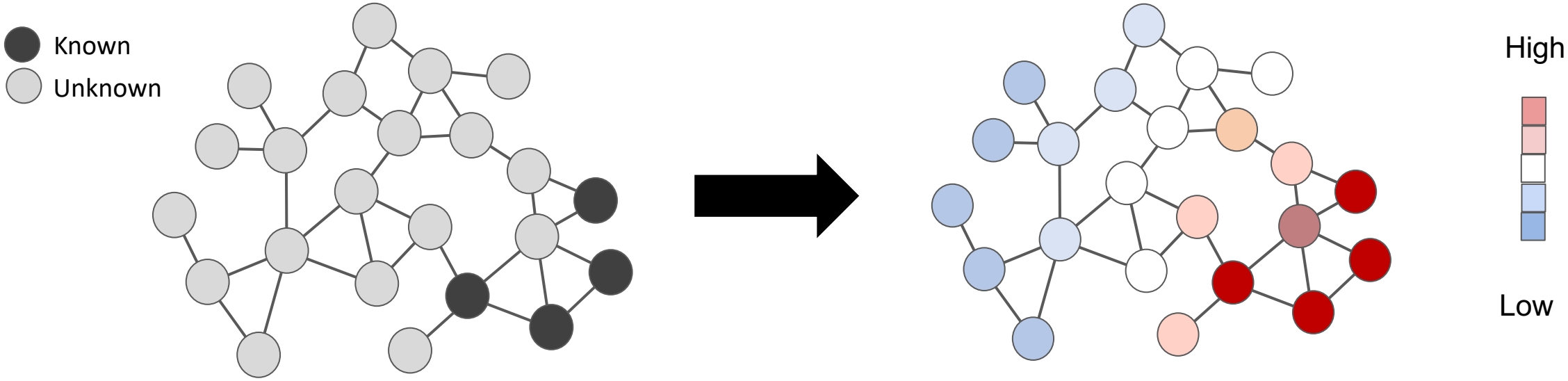
Random walk similarity matrix

Vertex scores

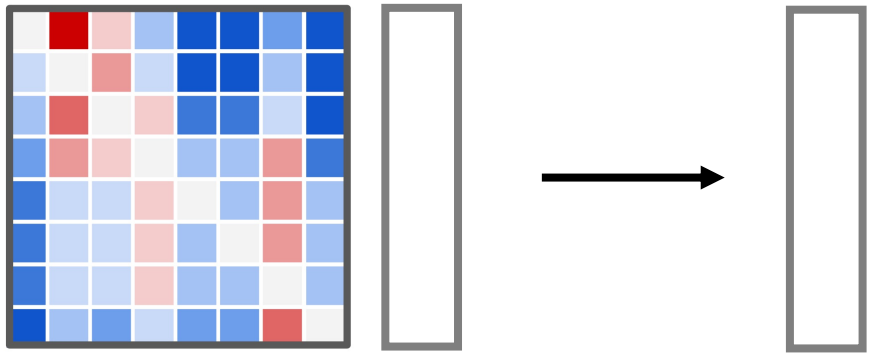
Name	Similarity matrix
Random walk	W^k
Random walk with restart	$\alpha(I-(1-\alpha)W)^{-1}$
Diffusion kernel	$e^{-\alpha W}$

Cowen et al (Nature Reviews Genetics 2017)

Network propagation is standard for ranking vertices



Rank vertices based on similarity to vertices w/ known characteristics e.g. genes associated with a specific disease (binary vertex scores X_v)



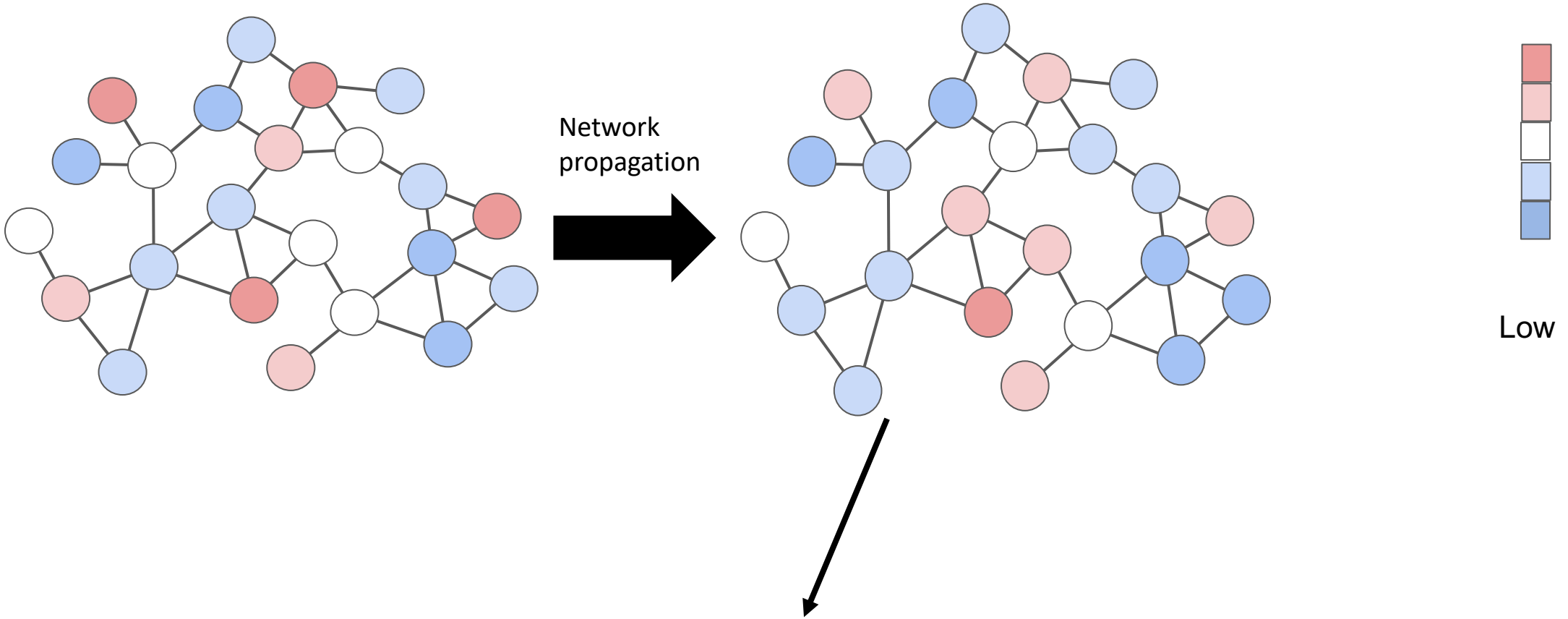
Random walk similarity matrix

Vertex scores

Google PageRank

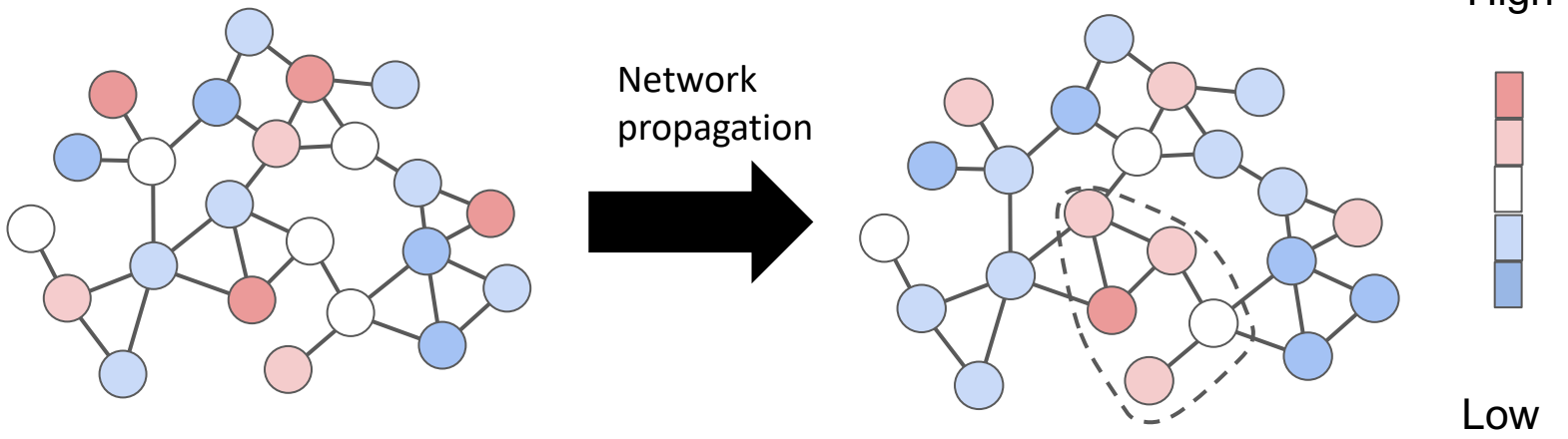
Personalized PageRank is **asymptotically optimal** for ranking in random graph models (PNAS 2017)

How to use network propagation to identify altered subnetworks?



Question: how to identify altered subnetwork from propagated gene scores?

Existing network propagation methods use ad hoc heuristics to identify altered subnetworks



HotNet2

Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes

Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding & Benjamin J Raphael

Nature Genetics **47**, 106–114 (2015) | Cite this article

39k Accesses | **500** Citations | **122** Altmetric | Metrics

PRINCE

Associating Genes and Protein Complexes with Disease via Network Propagation

Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, Roded Sharan

Published: January 15, 2010 • https://doi.org/10.1371/journal.pcbi.1000641

Ex: **PRINCE**: "We aim at inferring densely connected protein complexes that contain high scoring proteins … we start with the top 100 [propagated] scoring proteins as seeds … To each seed we iteratively add a neighboring protein with the highest score … A refinement phase takes place where proteins are removed from a putative complex to ensure that … its member proteins are densely interacting."

Issue: These algorithms lack **rigorous statistical guarantees** – hard to investigate fundamental issues like bias

Recent work shows existing approaches biased towards "high centrality" vertices

Algorithms benchmark against existing network algorithms – can hide biases shared across methods

DOMINO: a network-based active module identification algorithm with reduced rate of false calls

Hagai Levi, Ran Elkon, Ron Shamir

Author Information

Molecular Systems Biology (2021) 17: e9593 | https://doi.org/10.15252/msb.20209593

*"Our study reports on a different bias that is prevalent in AMI solutions: **their tendency to report non-specific GO terms**. …we observed that many enriched GO terms also appear on permuted datasets, suggesting that such enrichment stems from some proprieties of the network, algorithm, or the data that bias the results."*

On the limits of active module identification

Olga Lazareva, Jan Baumbach, Markus List, David B Blumenthal Author Notes

Briefings in Bioinformatics, Volume 22, Issue 5, September 2021, bbab066,
https://doi.org/10.1093/bib/bbab066
Published: 29 March 2021 **Article history**

*"Our results indicate that classical but also supposedly bias-aware [altered subnetwork algorithms] extract disease modules **based on the node degree**"*

Our work:

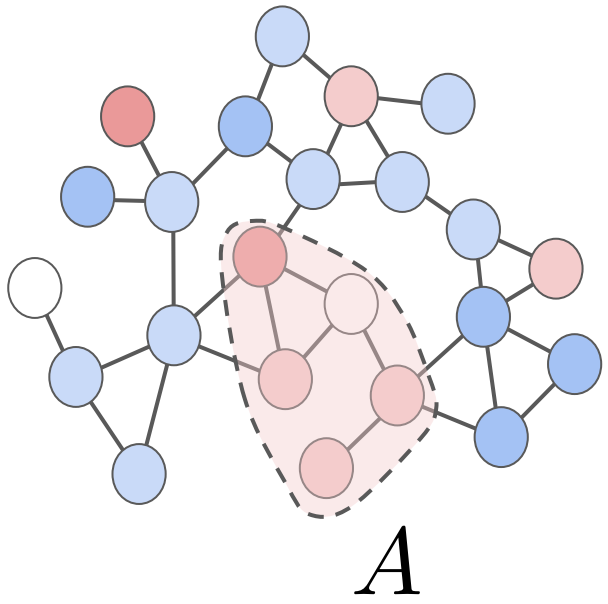
- Extend altered subnetwork generative model
  - Model different altered subnetwork topologies ("**subnetwork families**")
  - Derive propagation family – "approximates" subnetworks found by network propagation

- **NetMix2** algorithm for altered subnetwork identification with different subnetwork families
  - w/ propagation family: principled network propagation algorithm for altered subnetwork identification

- Simple baselines for evaluating network algorithms – *"scores only"* and *"network only"*

Generative model: Altered Subnetwork Distribution

- G=(V, E) is interaction network
- $\mathcal{S}$ is **subnetwork family** (set of subsets of V)
- $A \in \mathcal{S}$ is the altered subnetwork



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

$$X_v \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } v \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise} \end{cases}$$

$\mathcal{D}_{\mathrm{a}}$ = altered distribution (unknown)

$\mathcal{D}_{\mathrm{b}}$ = background distribution (typically known)

Generative model: Altered Subnetwork Distribution

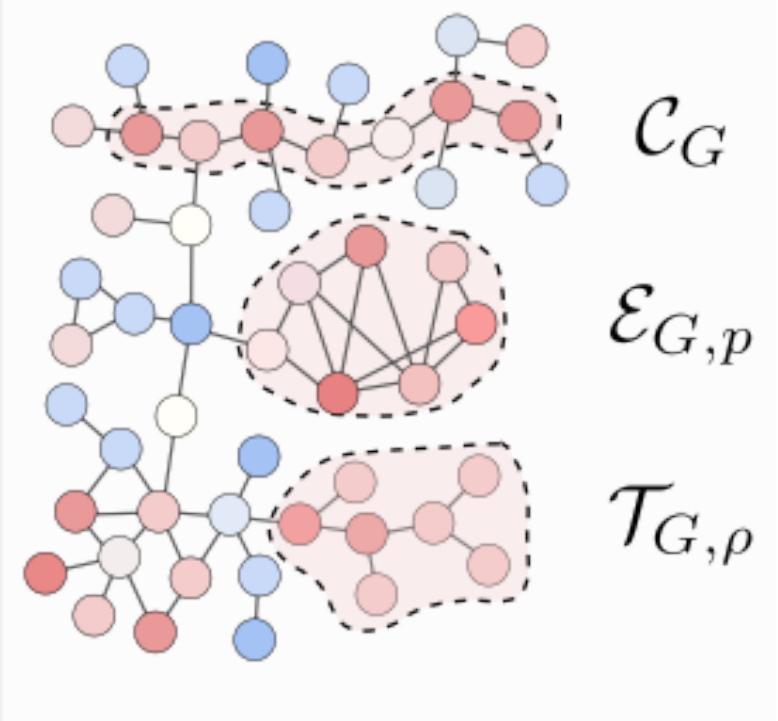
- G=(V, E) is interaction network
- $\mathcal{S}$ is **subnetwork family** (set of subsets of V)
- $A \in \mathcal{S}$ is the altered subnetwork

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

$$X_v \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } v \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise} \end{cases}$$

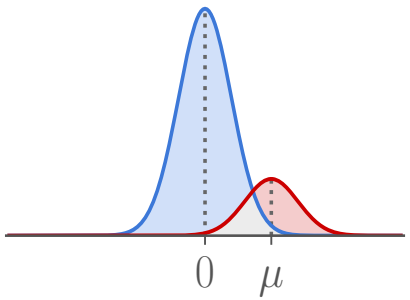
$\mathcal{D}_{\mathrm{a}}$ = altered distribution (unknown)

$\mathcal{D}_{\mathrm{b}}$ = background distribution (typically known)



Example of distributions: z-scores

$$\mathcal{D}_{\mathrm{a}} = N(\mu, 1)$$
$$\mathcal{D}_{\mathrm{b}} = N(0, 1)$$



Examples of subnetwork families:

Connected family $\mathcal{S} = \mathcal{C}_G$ = connected subgraphs S ⊆ V

Edge-dense family $\mathcal{S} = \mathcal{E}_{G,p}$ = subgraphs with density(S) > p

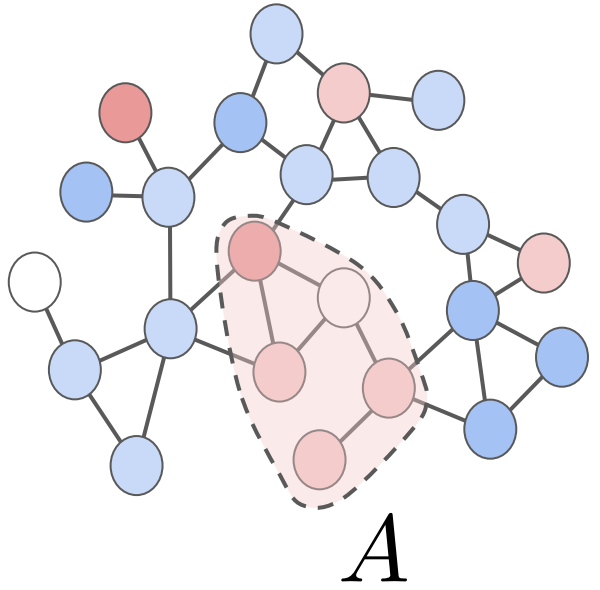
Cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$ = subgraphs with cut(S) < ρ

Generative model: Altered Subnetwork Distribution

- G=(V, E) is interaction network
- $\mathcal{S}$ is **subnetwork family** (set of subsets of V)
- $A \in \mathcal{S}$ is the altered subnetwork

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

$$X_v \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } v \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise} \end{cases}$$

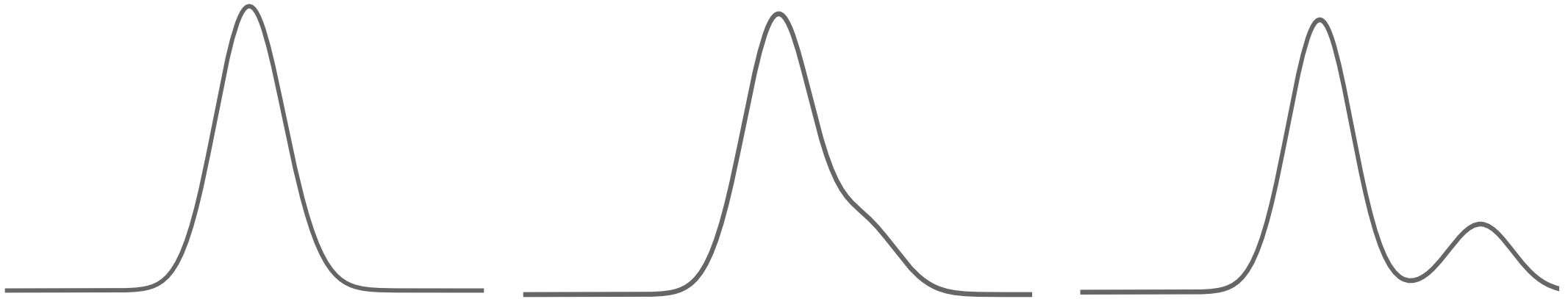


Altered Subnetwork Problem (ASP): Given graph G, subnetwork family $\mathcal{S}$ and vertex scores $(X_v)_{v \in V}$, find altered subnetwork A.

ASP = estimating parameters of distribution

Hard to solve ASP

Easy to solve ASP without network



Small distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Large distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Altered Subnetwork Distribution

Subnetwork family $\mathcal{S}$

Altered subnetwork $A \in \mathcal{S}$

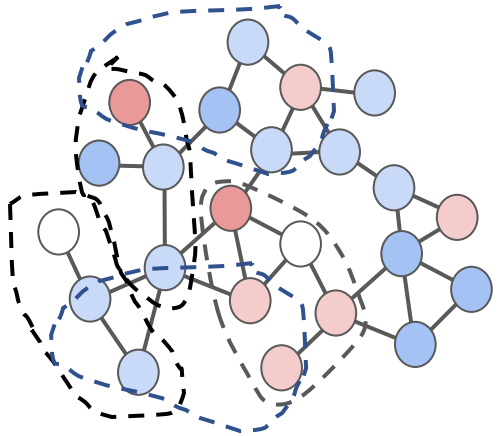
Vertex scores

$$X_v \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } v \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise} \end{cases}$$

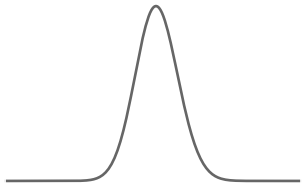
Altered subnetwork problem: Given X_v and $\mathcal{S}$ find A

$|\mathcal{S}|$ is ***large***

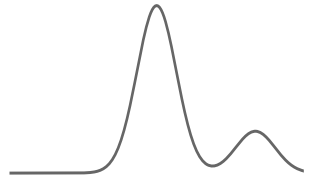
Hard to solve ASP



Hard to solve ASP



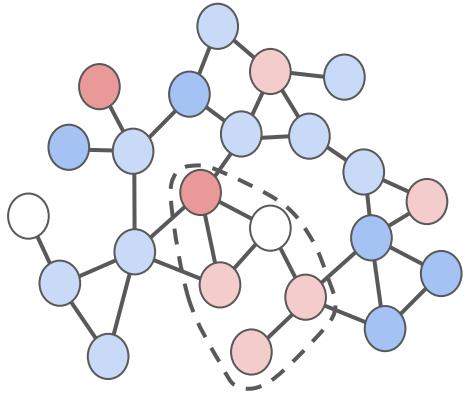
Easy to solve ASP without network



Small distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Large distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Easy to solve ASP without vertex scores



$|\mathcal{S}|$ is ***small***

Altered Subnetwork Distribution

Subnetwork family $\mathcal{S}$

Altered subnetwork $A \in \mathcal{S}$

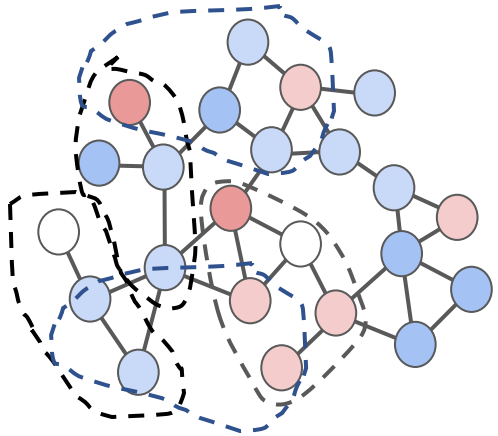
Vertex scores

$$X_v \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } v \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise} \end{cases}$$

Altered subnetwork problem: Given X_v and $\mathcal{S}$ find A

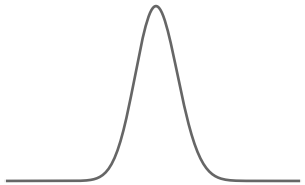
$|\mathcal{S}|$ is ***large***

Hard to solve ASP



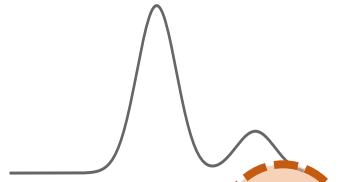
"Sweet spot" - both network (subnetwork family) and vertex scores are necessary

Hard to solve ASP



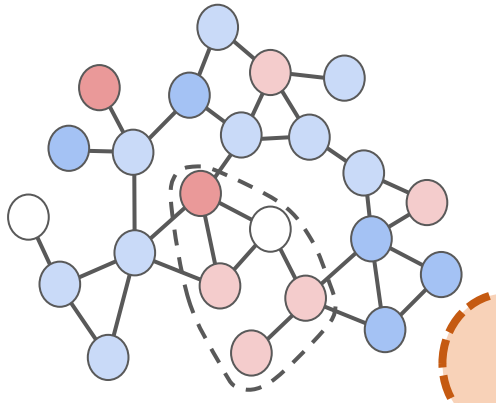
Small distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Easy to solve ASP without network



Large distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Easy to solve ASP without vertex scores



Important to compare against naïve baselines that use
(1) "scores only"
(2) "network only"

$|\mathcal{S}|$ is ***small***

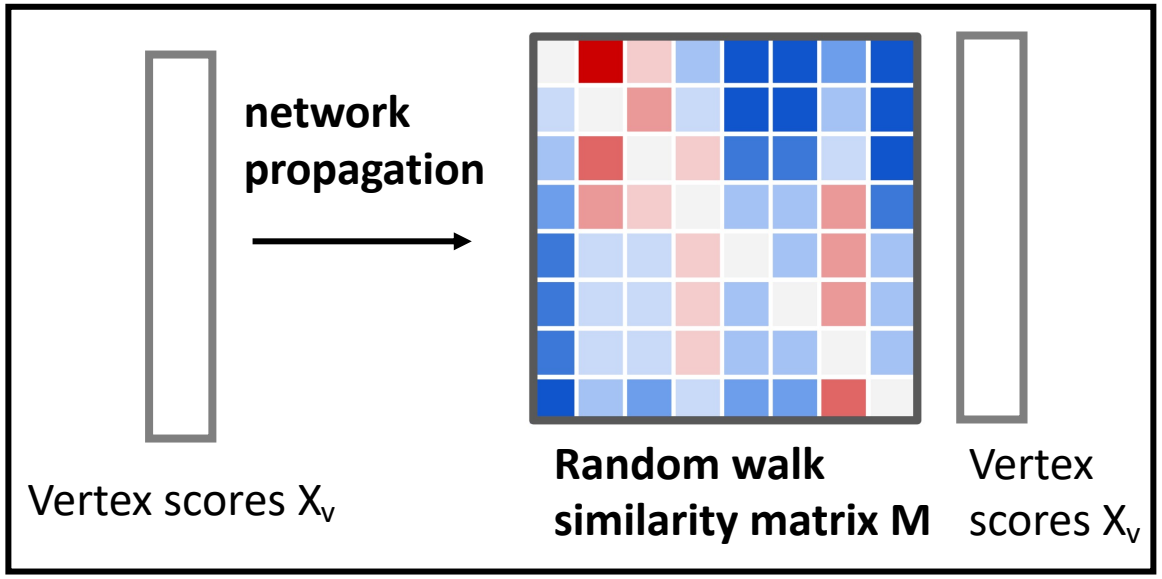
Propagation family

$\mathcal{S} = \mathcal{M}_{\delta,p}$: Subgraphs S with $M_{u,v} \geq \delta$ for p fraction of $(u, v) \in S$

Vertices are "close" via random walk

(also require $M_{v,u} \geq \delta$ if M is not symmetric, eg personalized PageRank)

In paper: theory and simulations show propagation family approximates subnetworks found by network propagation methods



Altered Subnetwork Distribution

Subnetwork family $\mathcal{S}$

Altered subnetwork $A \in \mathcal{S}$

Vertex scores

$$X_v \sim \begin{cases} \mathcal{D}_a, & \text{if } v \in A, \\ \mathcal{D}_b, & \text{otherwise} \end{cases}$$

Altered subnetwork problem: Given X_v and $\mathcal{S}$ find A

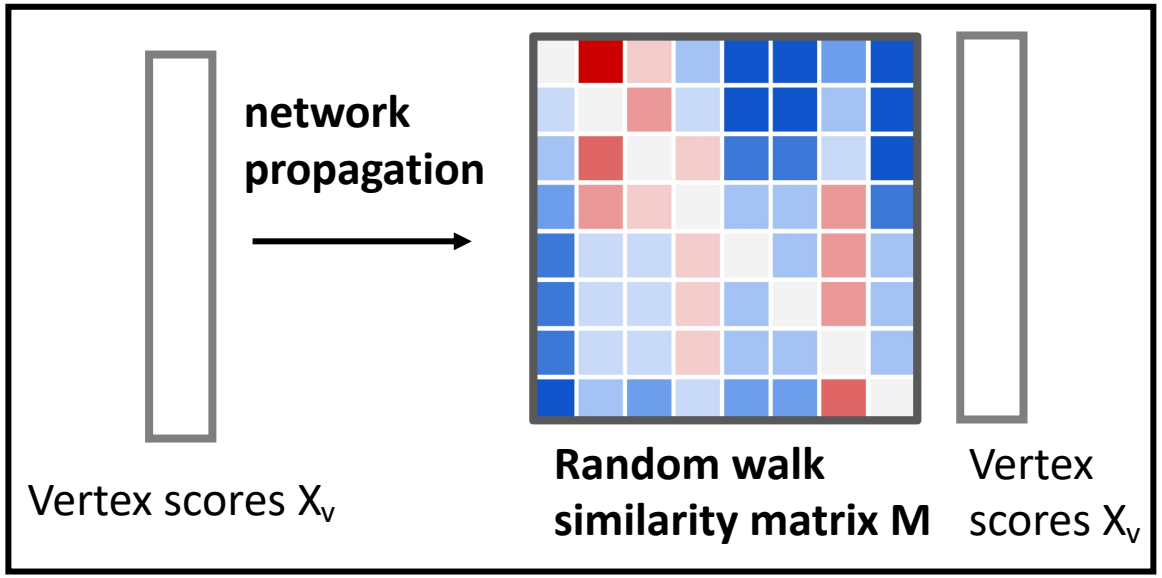
Propagation family

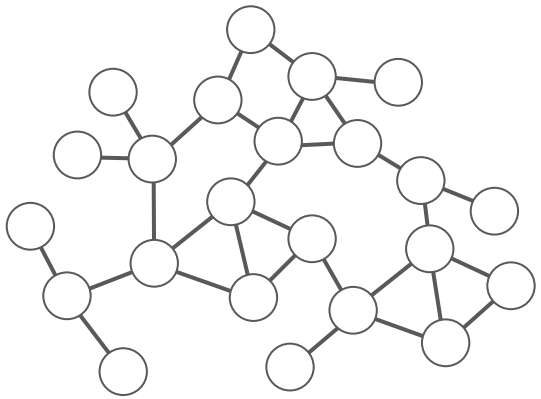
$\mathcal{S} = \mathcal{M}_{\delta,p}$: Subgraphs S with $M_{u,v} \geq \delta$ for p fraction of $(u, v) \in S$

Vertices are "close" via random walk

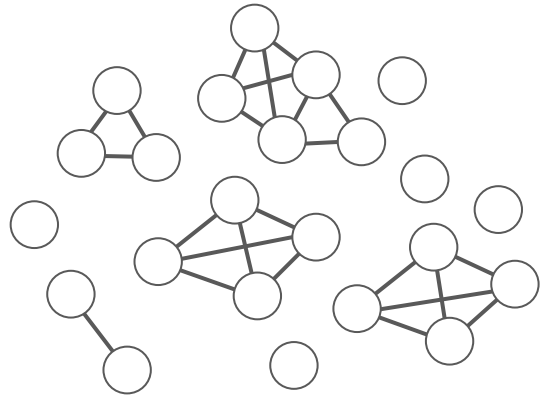
(also require $M_{v,u} \geq \delta$ if M is not symmetric, eg personalized PageRank)

Alternatively: edge-dense subnetworks of *"similarity threshold graph"*





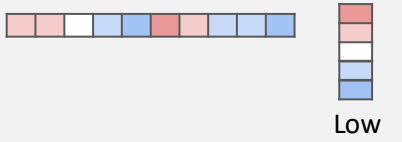
Interaction network G



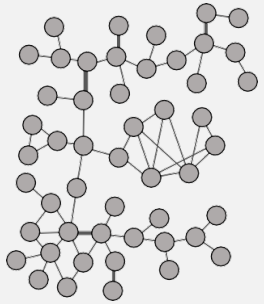
Similarity threshold graph G_δ

Input

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

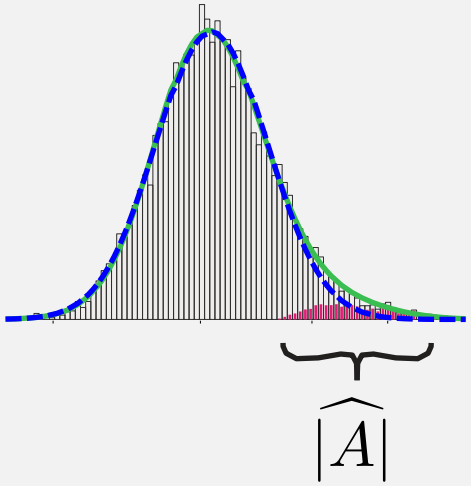


Interaction network
G=(V,E)

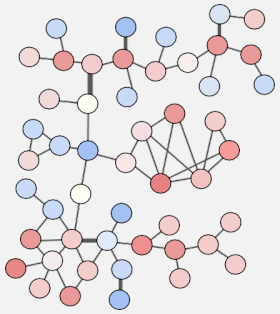


NetMix2

Step 1: Estimate size $\widehat{|A|}$ of altered subnetwork *A* using local FDR (non-parametric method)



Step 2: Compute subnetwork $S \in \mathcal{S}$ with size $|S| = \widehat{|A|}$ and largest total vertex score X_v



$$\widehat{A}_{\text{NetMix2}} = \underset{\substack{S \in \mathcal{S} \\ |S| = \widehat{|A|}}}{\operatorname{argmax}} \sum_{v \in S} X_v$$

Subnetwork family $\mathcal{S}$

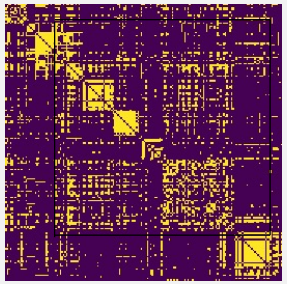
Connected family $\mathcal{C}_G$

Edge-dense family $\mathcal{E}_{G,p}$

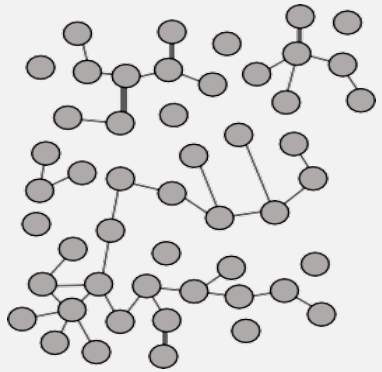
Cut family $\mathcal{T}_{G,\rho}$

Propagation family $\mathcal{M}_{\delta,p}$

$\vdots$



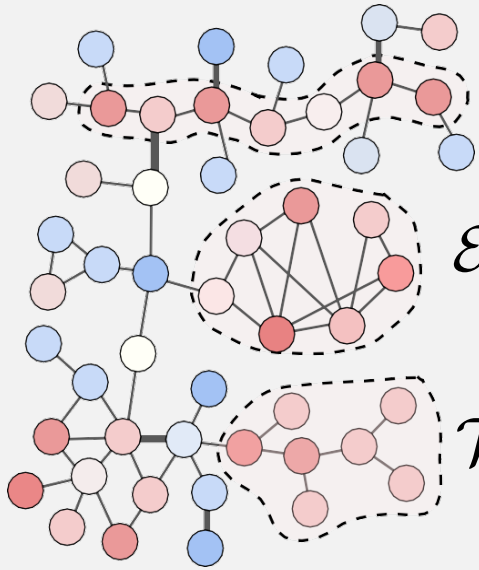
Similarity matrix
(Personalized PageRank)

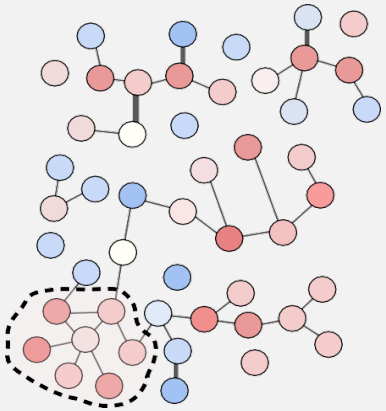


Similarity threshold graph

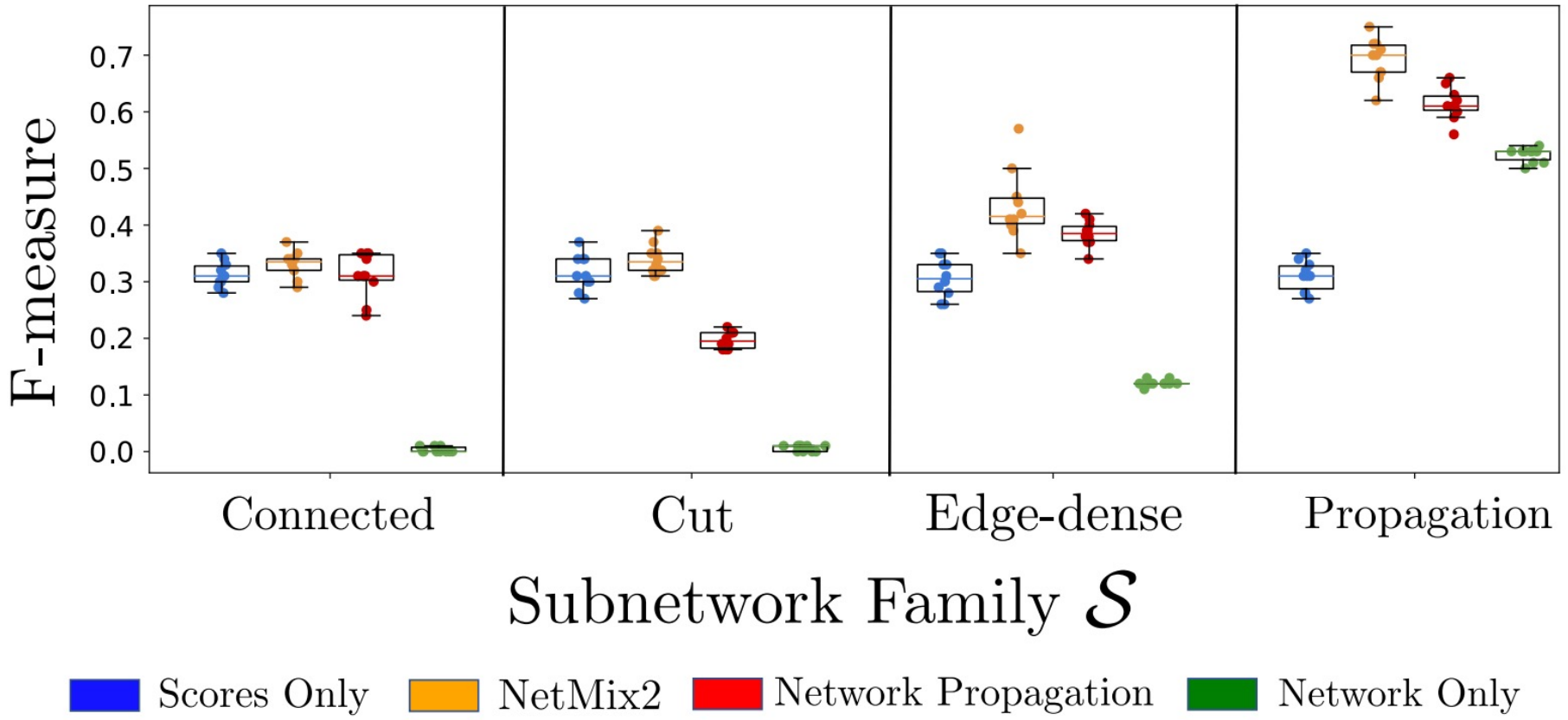
Output

Altered subnetwork $A \in \mathcal{S}$





Simulations: Propagation family corresponds to the subnetworks identified by network propagation



Scores only = {vertices w/ top-|A| scores}

Network only = {vertices w/ top-|A| vertex centrality}

Network propagation = {vertices w/ top-|A| propagated scores}

G = HINT+HI interaction network with |G|≈15000 nodes (Leiserson et al 2015)

Altered subnetwork A of size |A|=0.01n selected uniformly at random from subnetwork family $\mathcal{S}$

Results: somatic mutations in cancer

NetMix2 outperforms other methods at identifying previously reported driver mutations in cancer.

	STRING network						
		CGC		OncoKB		TCGA	
Method	**Subnetwork size**	*Number*	*F-measure*	*Number*	*F-measure*	*Number*	*F-measure*
NetMix2	280	132	**0.3**	133	**0.313**	151	**0.546**
NetMix	313*	129	0.282	130	0.295	147	0.502
Heinz (FDR=0.01)	335	139	0.297	138	0.306	156	0.513
NetSig	773	145	0.211	172	0.257	84	0.161
Hierarchical HotNet	246	73	0.172	70	0.172	74	0.285
Network Propagation	280	86	0.195	89	0.210	98	0.354
Scores-only	280	126	0.286	127	0.3	145	0.524
Network-only	280	77	0.175	83	0.196	55	0.199

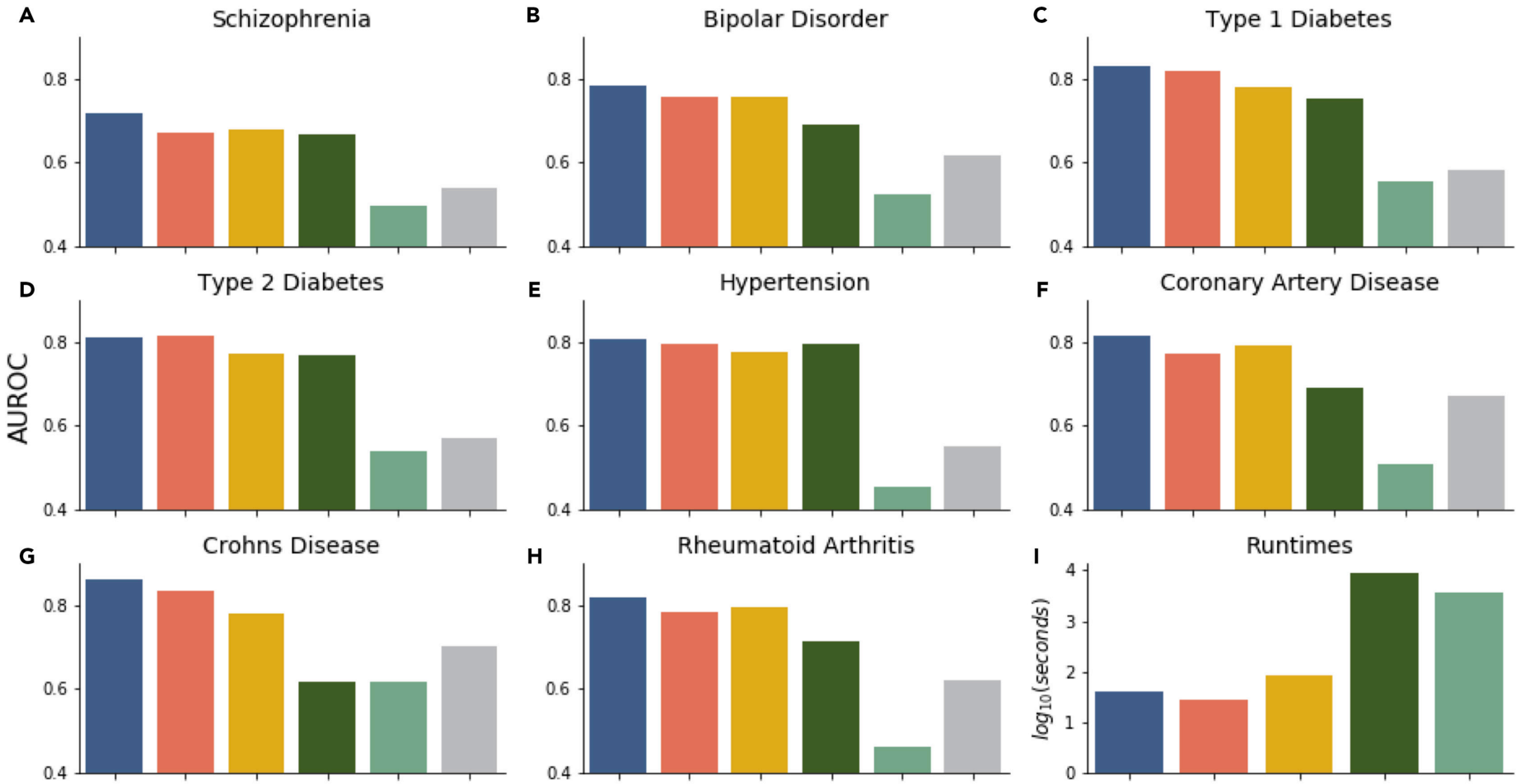
G = STRING protein interaction network

Vertex scores X_v = MutSIg2CV z-scores computed based on frequency of somatic mutations in TCGA tumor samples

Note: "Scores-only" has good performance – how helpful is interaction network?

Results: GWAS

Recent study by Carlin et al (iScience 2019) – evaluates how well methods identify known disease reference genes

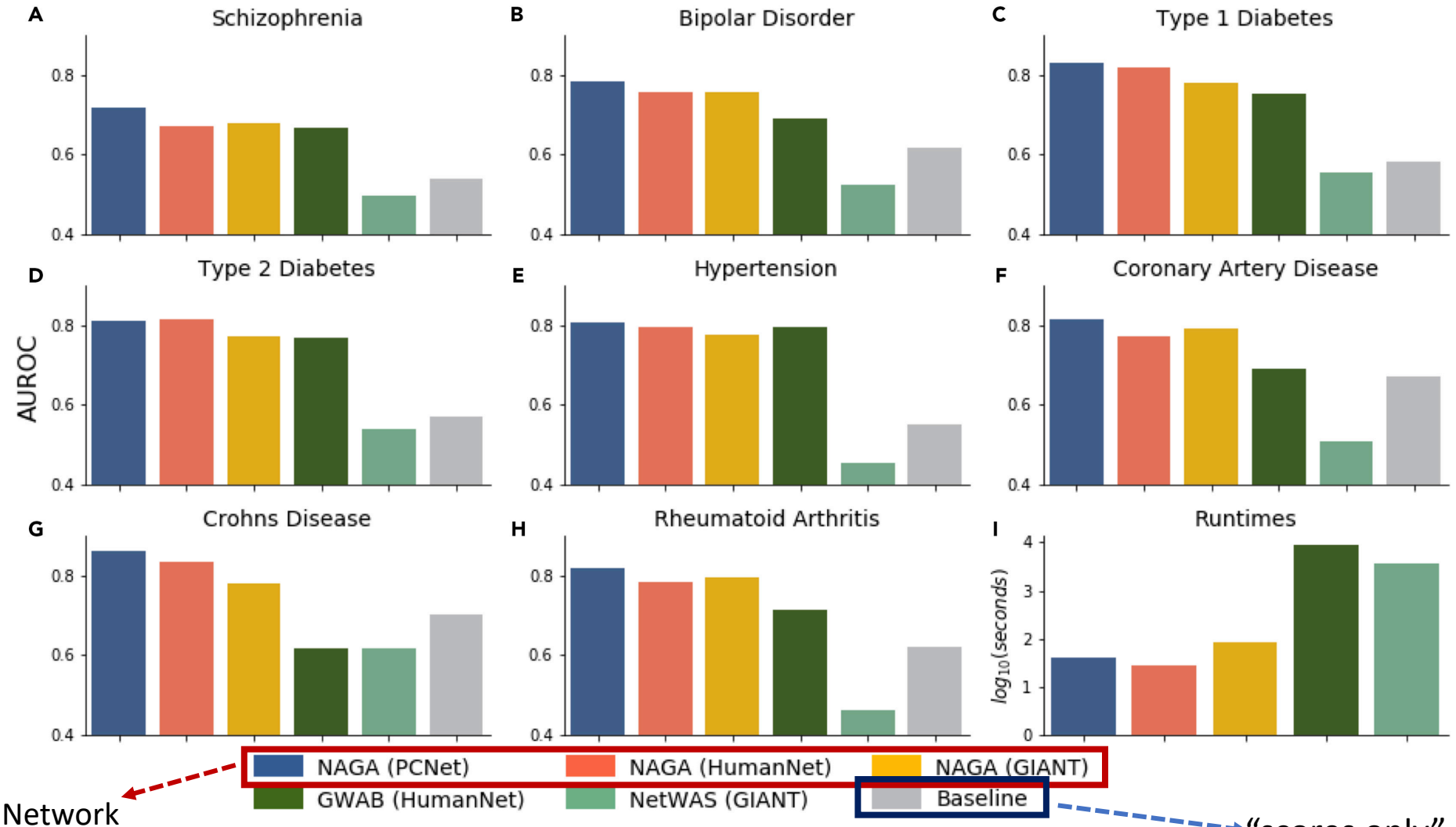


Network propagation

"scores only"

Results: GWAS

Recent study by Carlin et al (iScience 2019) – evaluates how well methods identify known disease reference genes



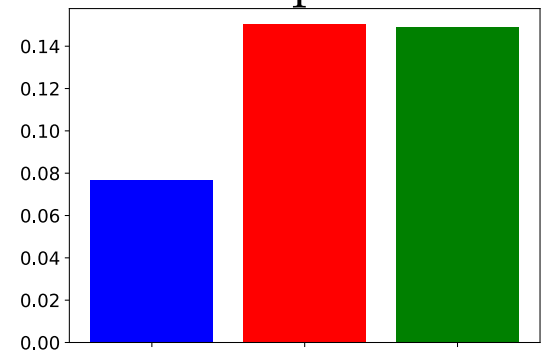
Network propagation

"scores only"

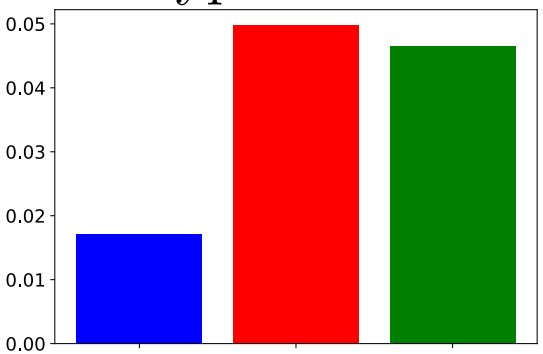
Issue: AUROC is poor metric for small reference sets! (<1% of 15,000 genes)

AUPRC

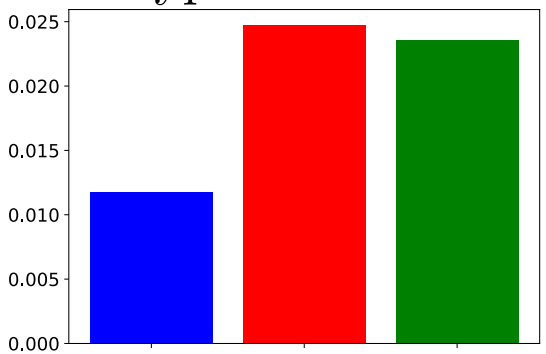
Schizophrenia



Hypertension

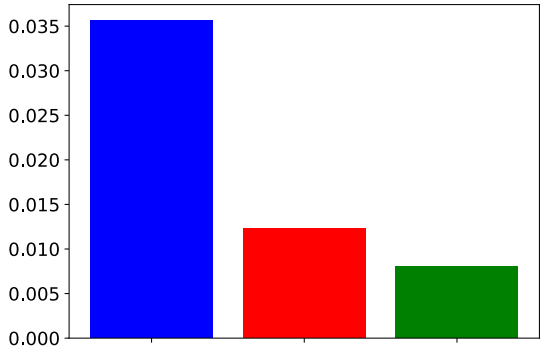


Type 1 Diabetes

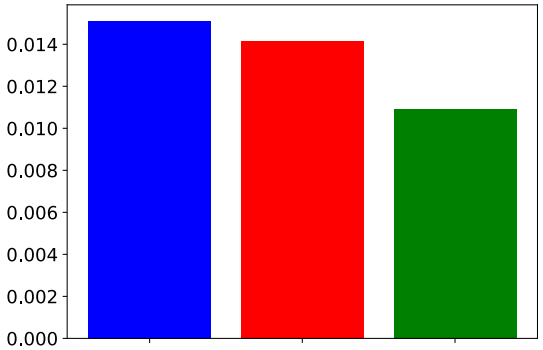


(N) Network alone is sufficient to identify reference genes

Crohn's Disease

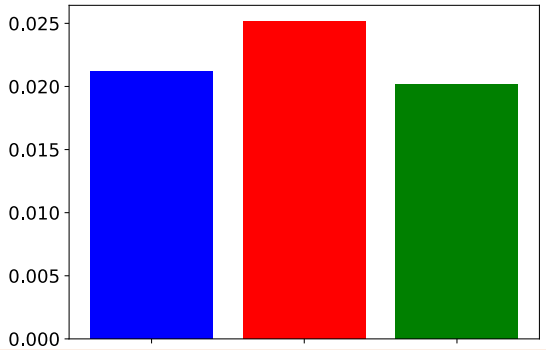


Coronary Artery Disease

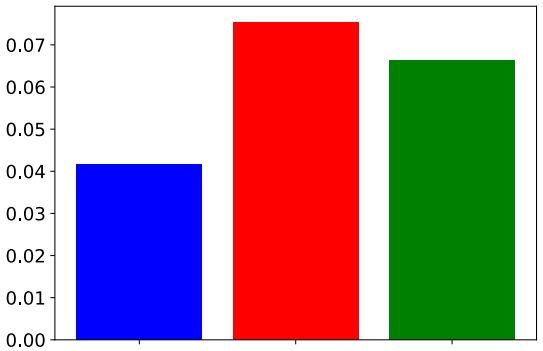


(S) Scores alone are sufficient to identify reference genes

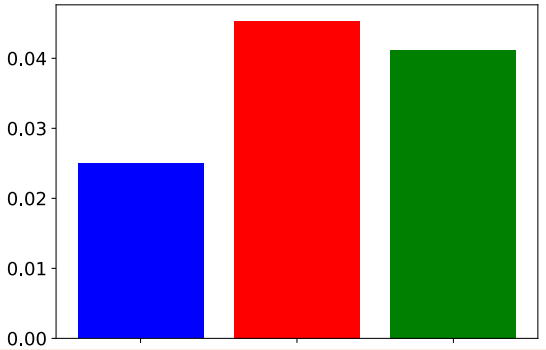
Rheumatoid Arthritis



Bipolar Disorder



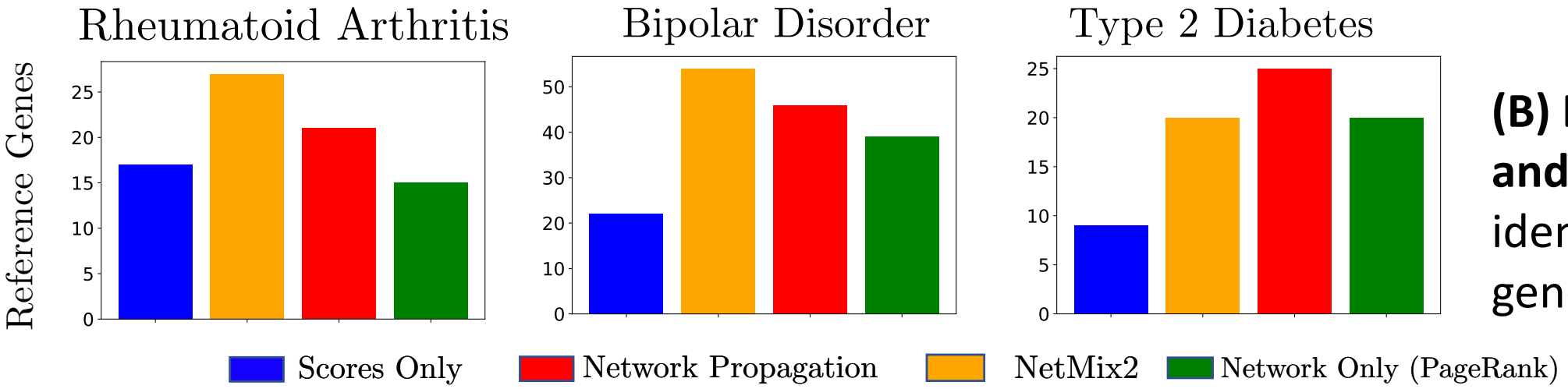
Type 2 Diabetes



(B) Both network and scores help identify reference genes

Scores Only Network Propagation Network Only (PageRank)

NetMix2 results on diseases where both network and scores help



(B) Both network and scores help identify reference genes

NetMix2 outperforms network propagation on 2/3 diseases

Summary

- **Generative model** for altered subnetworks from different subnetwork families
- **Propagation family** approximates subnetworks identified by network propagation
- **NetMix2** algorithm: principled network propagation approach for altered subnetwork identification
- Important to correctly benchmark network algorithms against simple “scores only” and “network only” baselines!

Acknowledgments

Tyler Park*

Raphael Group:

Ben Raphael

Tae Yoon (Tyler) Park

Matt Myers

Ahmed Shuaibi

Alexander Strzalkowski

Brian Arnold

Cong Ma

Henri Schmidt

Madelyne Xiao

Maya Gupta

Palash Sashittal

Xinhao Liu

Rishabh Rout

Ellen Su

Paper (bioRxiv)

Code: https://github.com/raphael-group/netmix2