Network Science meets Tissue-specific Biology

Shahin Mohammadi and Ananth Grama

Department of Computer Science
Purdue University

October, 2016

PURDUE UNIVERSITY

Outline

1. **Part 1**
   - Datasets
   - Measuring similarity of cells
   - Identifying cell types
   - Identifying cell type specific markers

2. **Part 2**
   - Datasets
   - Constructing tissue-specific network
   - Identifying differential protein complexes/pathways
   - Prioritizing disease genes
   - Identifying disease-related pathways

PURDUE UNIVERSITY

Outline

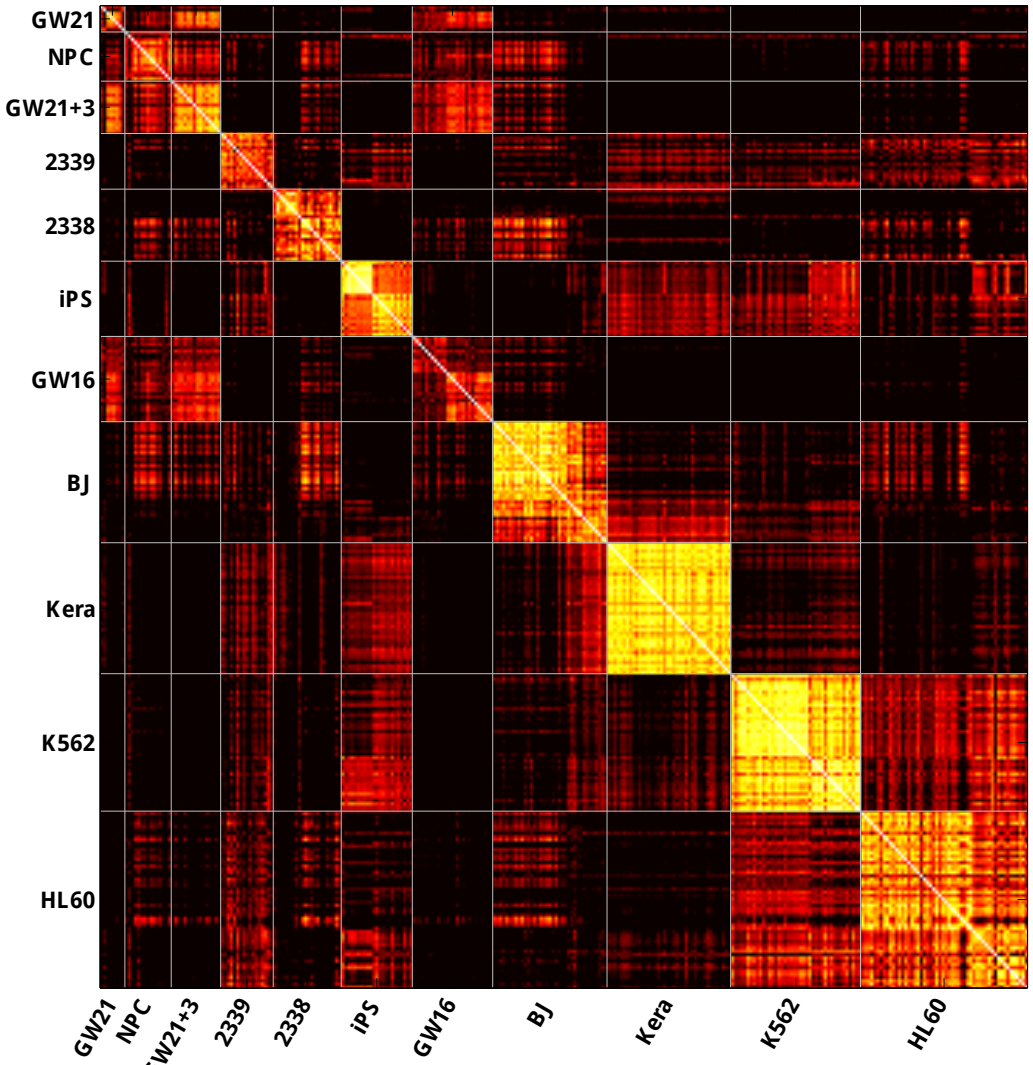
PURDUE UNIVERSITY

Databases

GEO (US): https://www.ncbi.nlm.nih.gov/geo
Example: 430 cells from 5 different primary glioblastomas (GSE57872)

PURDUE UNIVERSITY

Single cell gene expression profiles



- 301 cells
- 11 cell types
- Pubmed: `https://www.ncbi.nlm.nih.gov/pubmed/25086649`

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Identifying the shared subspace among a group of tissues

Low-rank decomposition

$$A = U_r \Sigma_r V_r = \sum_{i=1}^{r} \sigma_i u_i {v_i}^T,$$

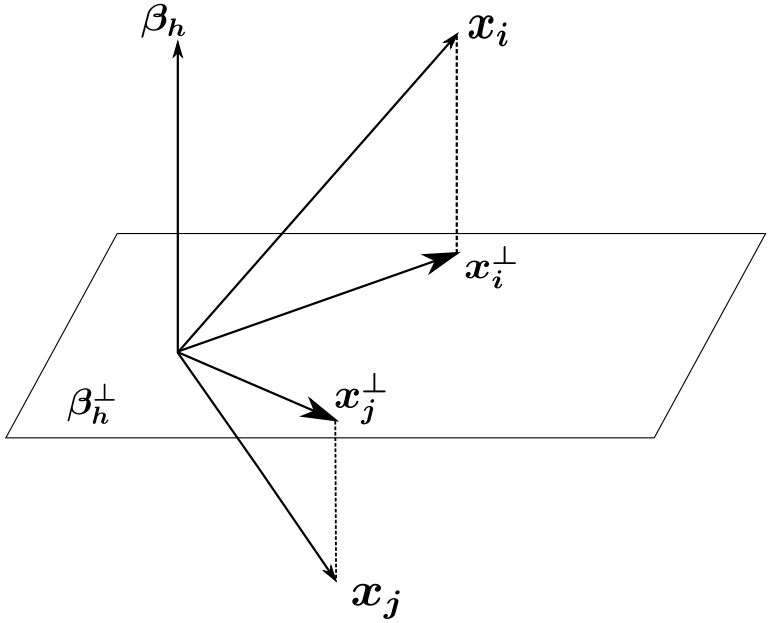
- $T \in \mathbb{R}^{n_g \times n_t}$
- n_g rows correspond to genes
- n_t columns correspond to various tissues
- $r \leq \min(n_g, n_t)$
- SVD or NMU

PURDUE UNIVERSITY

Adjusting transcriptional signatures

- Vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ represent normalized transcriptional signatures of a tissues/cell types i and j
- Vector $\boldsymbol{\beta}_h$ represent the normalized housekeeping signature
- Projection to the orthogonal subspace of $\boldsymbol{\beta}_h$:

$$\begin{aligned} \mathbf{x}_i^{\perp} &= \mathbf{P}^{\perp}\mathbf{x}_i \\ &= (\mathbf{I} - \mathbf{P})\mathbf{x}_i \\ &= \left(\mathbf{I} - \frac{\boldsymbol{\beta}_h \boldsymbol{\beta}_h^T}{\| \boldsymbol{\beta}_h \|_2}\right)\mathbf{x}_i \end{aligned}$$



PURDUE UNIVERSITY

Adjusted transcriptional similarities

Let $\mathbf{X}^{\perp}$ represent the adjusted signature of tissues/cell types. Let $\mathbf{Z} = \mathbf{zscore}(\mathbf{X}^{\perp})$ be its normalized version. Then, we can defined the following kernel:

Adjusted similarity scores

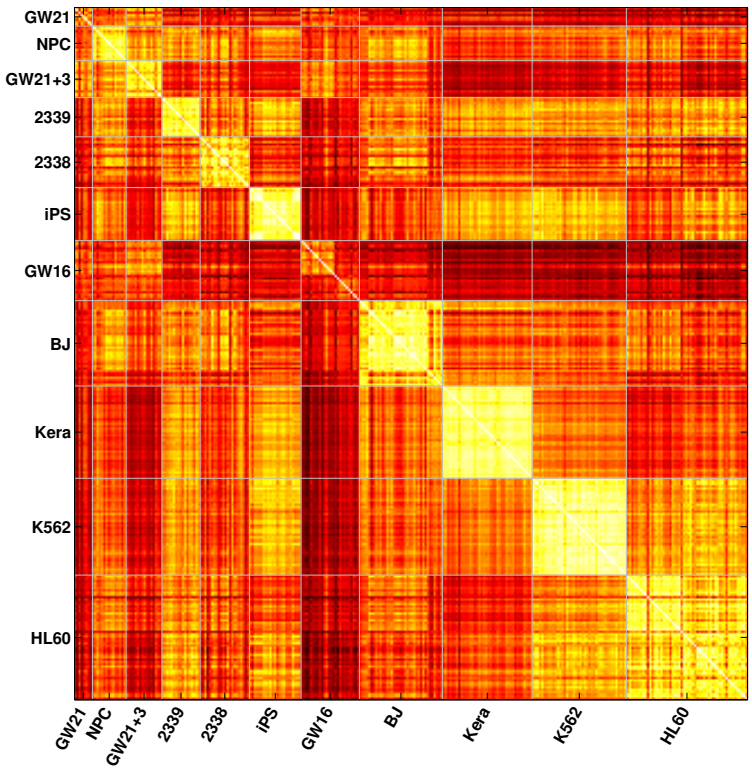
$$\mathbf{K} = \mathbf{Z}^{T} * \mathbf{Z}$$

$\rightarrow$ This is closely related to the partial correlation score after correcting for the effect of housekeeping genes

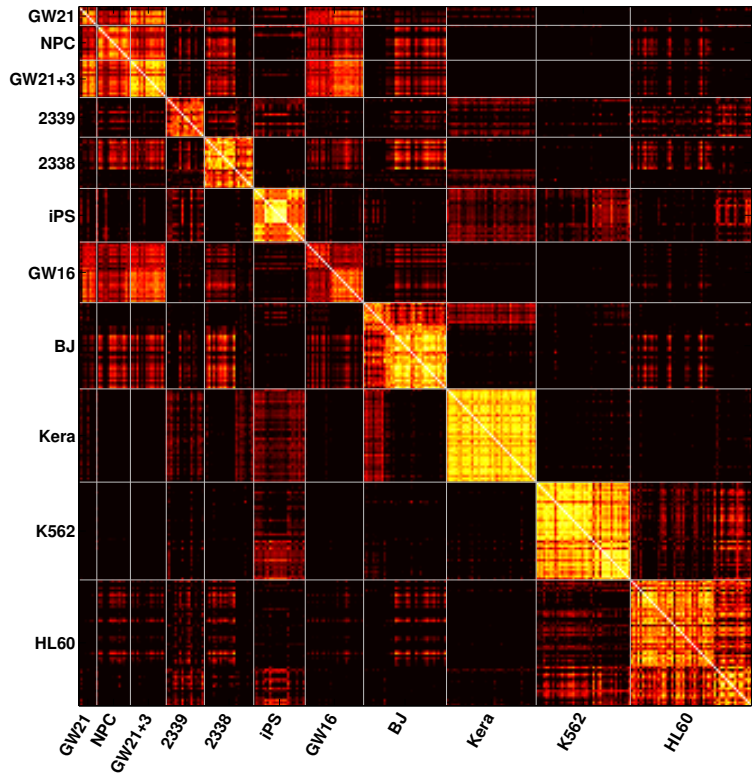
PURDUE UNIVERSITY

Effect of adjustment

Before adjustment

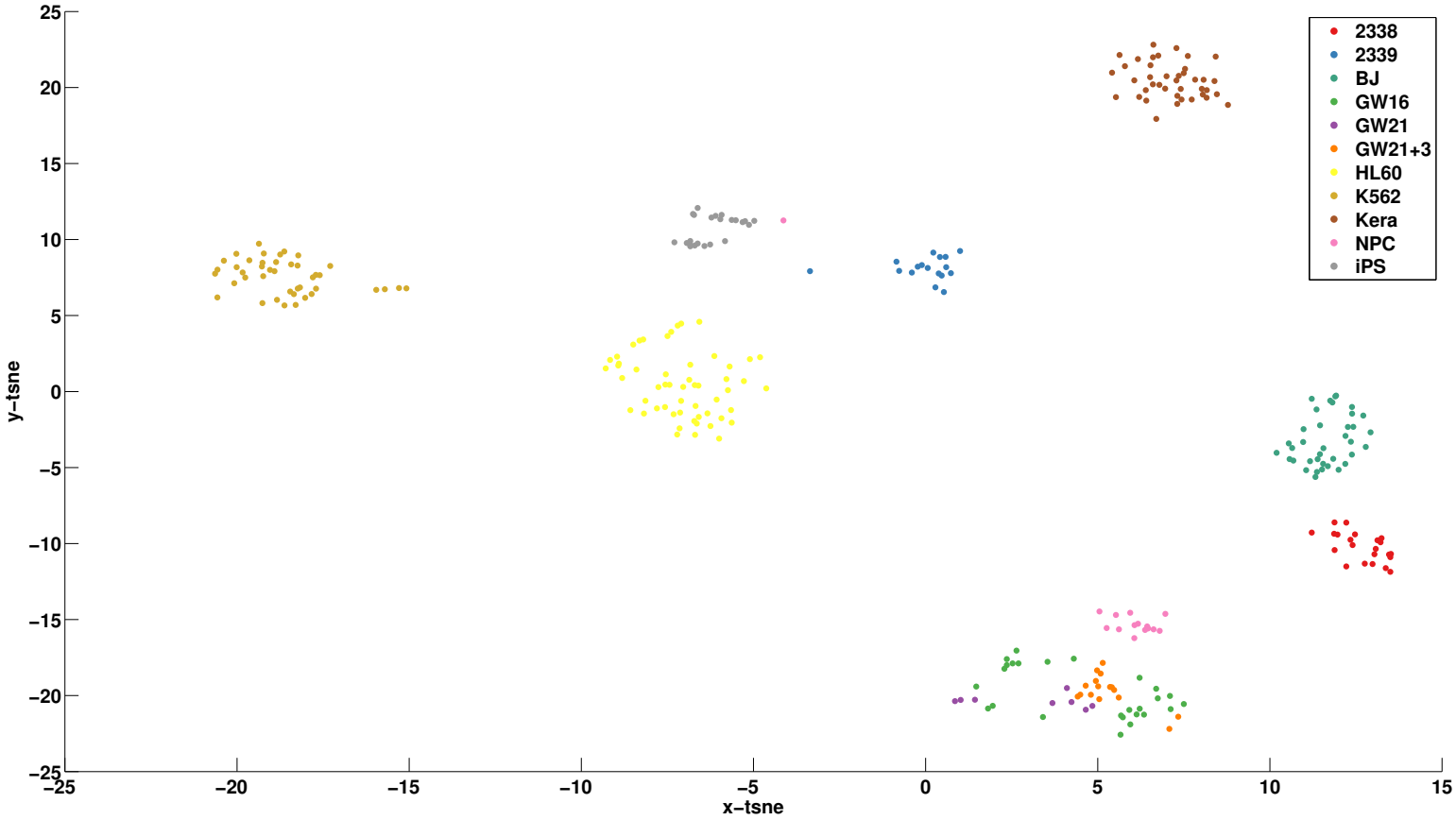


After adjustment



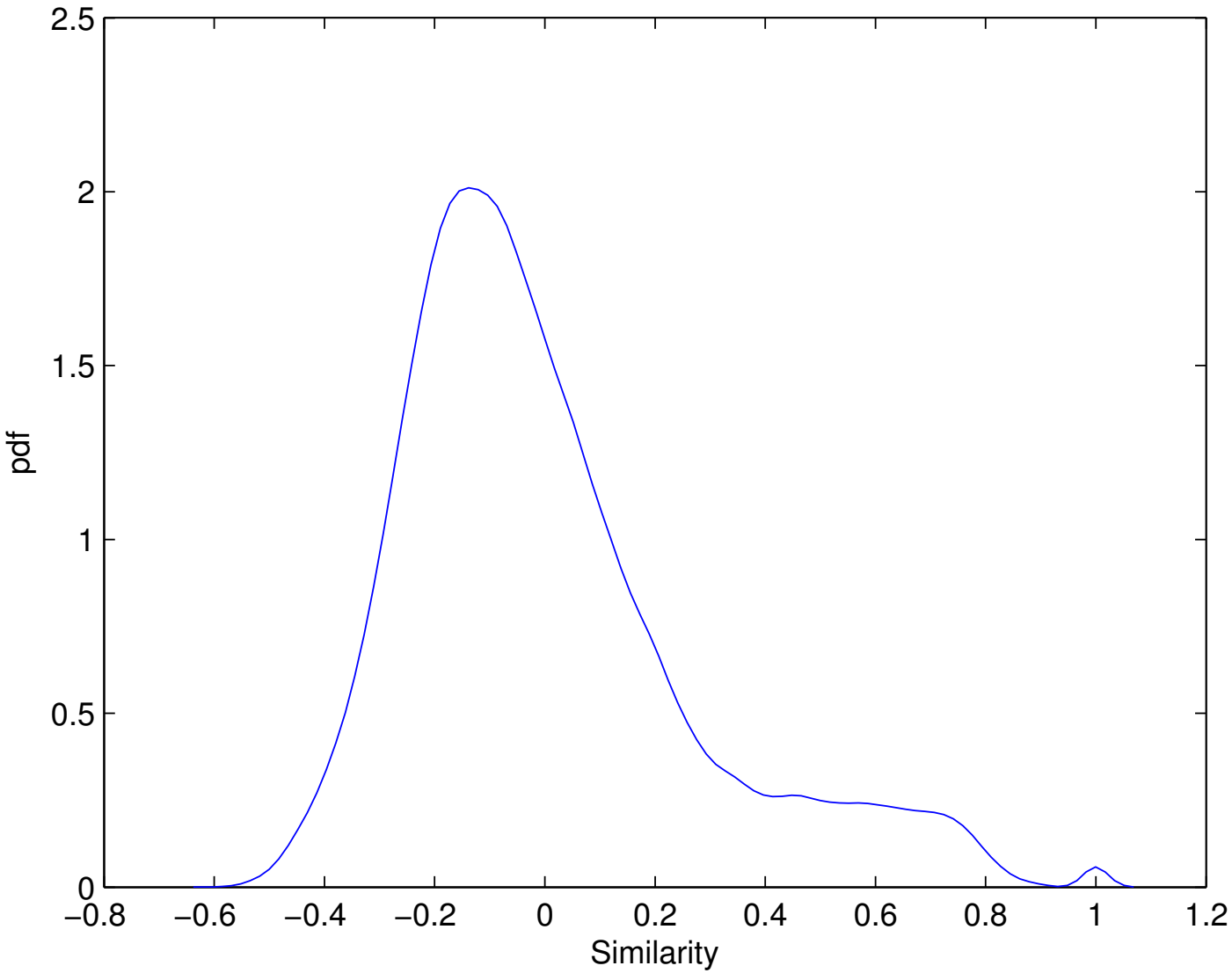
PURDUE UNIVERSITY

t-distributed Stochastic Neighbor Embedding (t-SNE)



PURDUE UNIVERSITY

Distribution of similarity scores



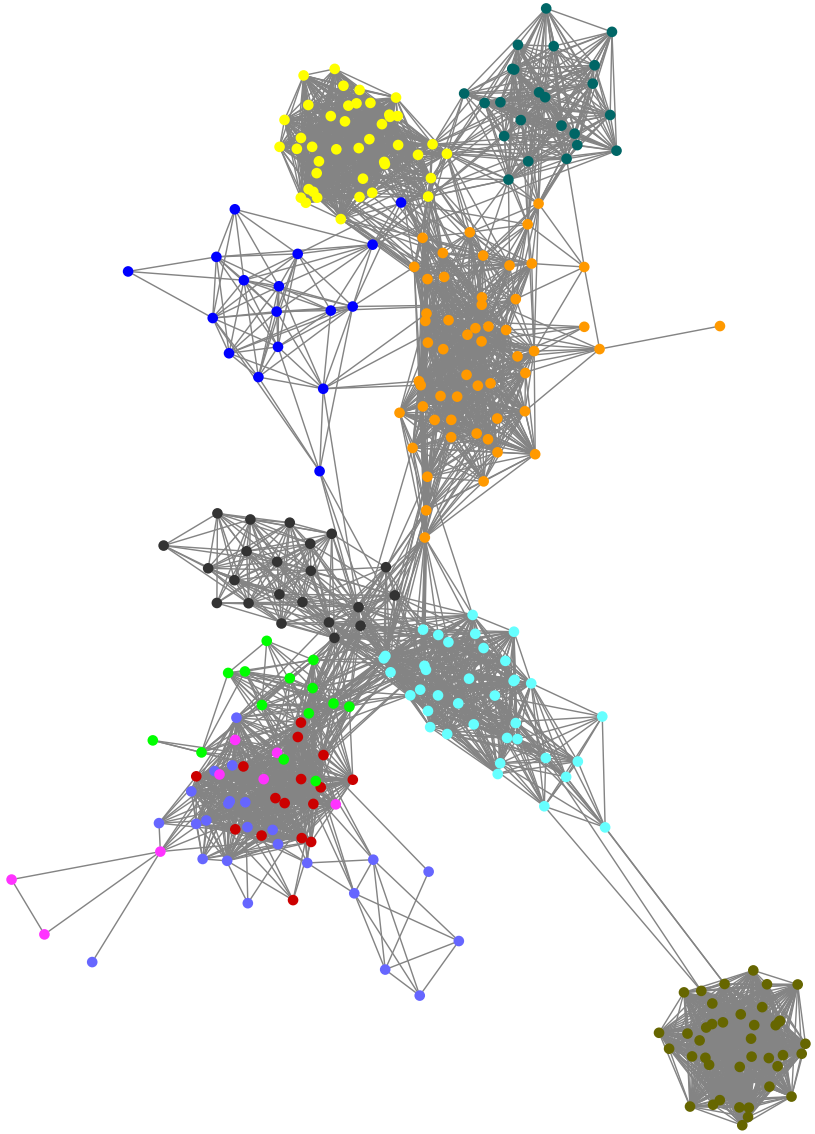
PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Cell-cell similarity network

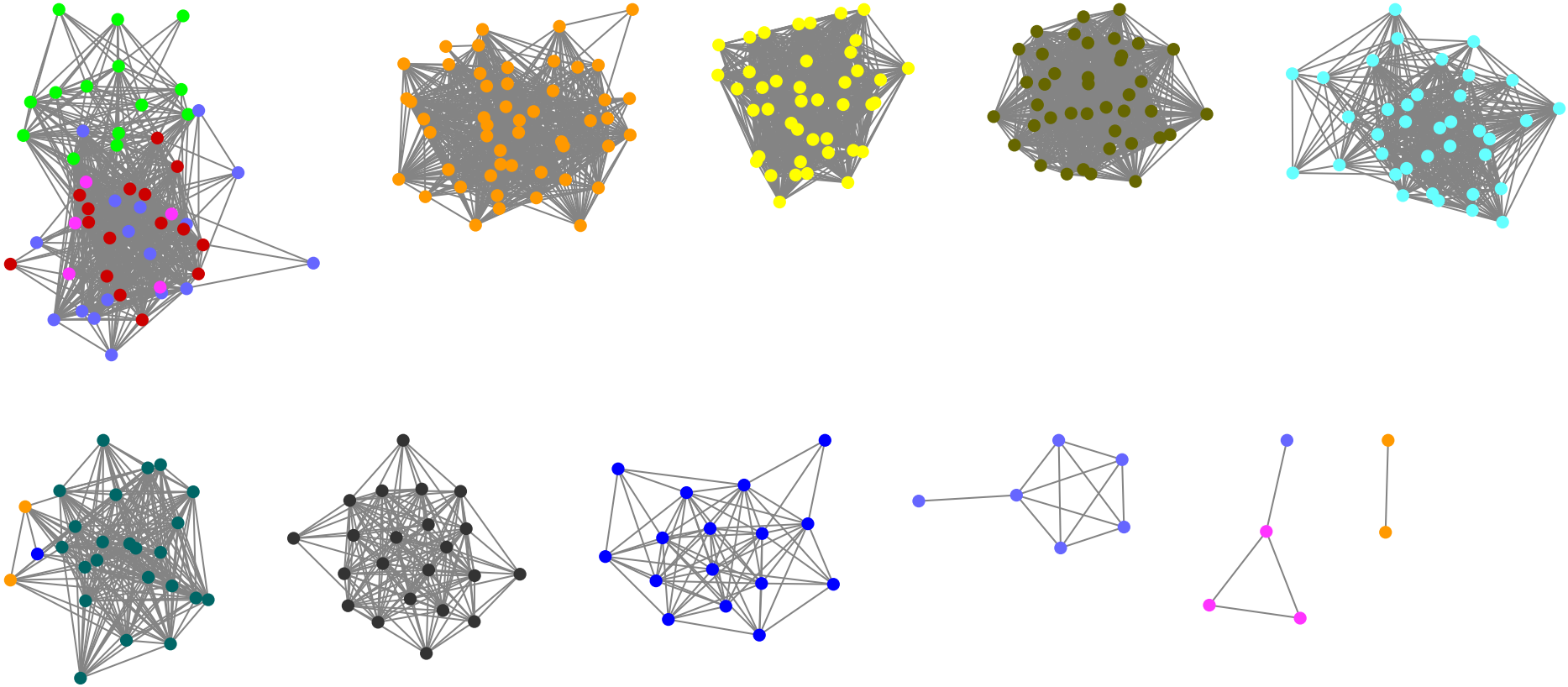
Visualization using Cytoscape



PURDUE UNIVERSITY

Cell-cell similarity network

Clustering



Validation

Our method

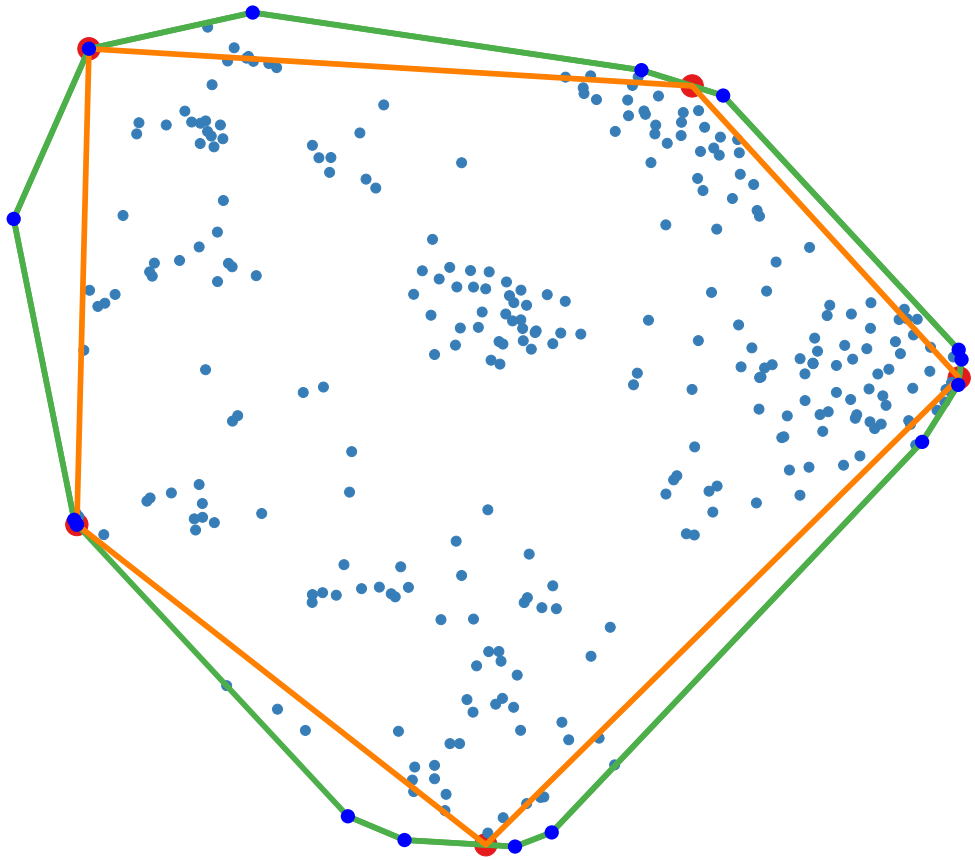
- NMI: 0.86
- ARI: 0.76

SNN_Cliq

- NMI: 0.75
- ARI: 0.55

PURDUE UNIVERSITY

Archetypal-analysis for Cell type identificaTION (ACTION)



A geometric approach to identify cell types

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Hypothesis 1

$$H_N : \quad \mu_c = \quad \mu_{\mathcal{C}\setminus c}$$
$$H_A : \quad otherwise$$

Hypothesis 2

$$H_N : \quad \mu_c = \mu_{c'} \quad \forall c' \in \mathcal{C} \setminus c$$
$$H_A : \quad otherwise$$

PURDUE UNIVERSITY

Table : $p - val \leq 10^{-3}$ **and** $2 \leq fold \rightarrow$ 130 genes

p-value	Term
3.0E-11	nervous system development
2.3E-10	neuron projection morphogenesis
2.5E-08	cytoskeletal protein binding
4.7E-08	neuron projection
8.7E-08	axon guidance

PURDUE UNIVERSITY

Table : $p - val \leq 10^{-3}$ **and** $2 \leq fold \rightarrow$ 52 genes

p-value	Term
7.1E-06	nervous system development
6.3E-04	neuron part
1.0E-03	neuron projection
2.1E-03	neuron projection development
2.7E-03	cytoskeletal protein binding

PURDUE UNIVERSITY

Table : $100 < factor \rightarrow 584$ genes

p-value	Term
2.3E-42	nervous system development
2.9E-31	neuron projection development
1.4E-20	neuron projection
1.7E-17	regulation of neurogenesis
1.1E-14	cytoskeletal protein binding

PURDUE UNIVERSITY

Outline

1. Part 1
   - Datasets
   - Measuring similarity of cells
   - Identifying cell types
   - Identifying cell type specific markers

2. Part 2
   - Datasets
   - Constructing tissue-specific network
   - Identifying differential protein complexes/pathways
   - Prioritizing disease genes
   - Identifying disease-related pathways

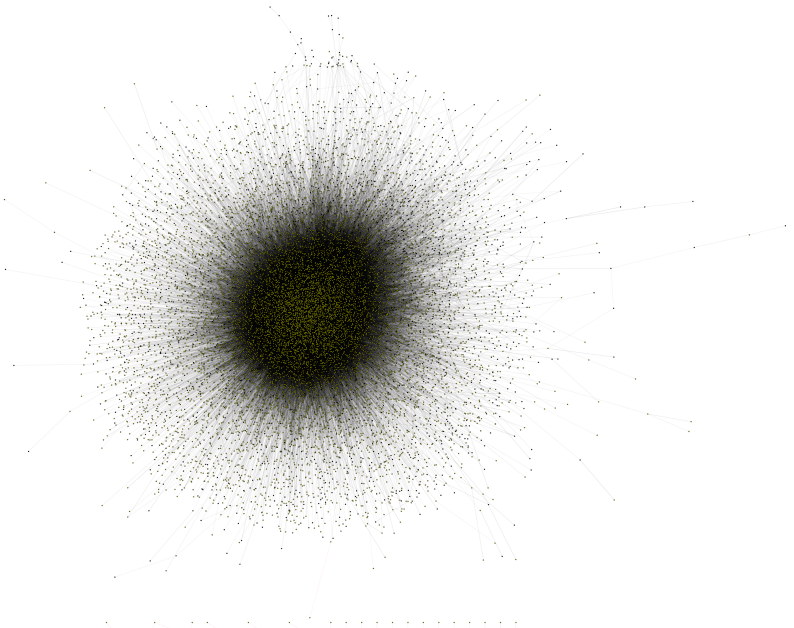
PURDUE UNIVERSITY

Databases

BioGRID: http://thebiogrid.org/
iRefIndex: http://irefindex.org/wiki/index.php

PURDUE UNIVERSITY

Single cell gene expression profiles



Before pruning:

- 14,658 nodes
- 147,444 edges

After pruning:

- 7,651 nodes
- 82,097 edges

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Motivation

Global human interactome is a superset of all possible physical interactions that can take places in the cell. It does not provide any information as to which one of these interactions do take place in a given tissue/cell-type context.

PURDUE UNIVERSITY

Statement

Available data sources:

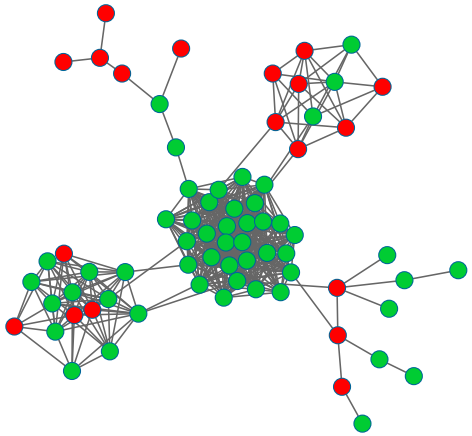
1. A global interactome, which contains the set of *possible* interacting pairs.
2. A tissue-specific measurement of gene/protein activity within each tissue/cell type.

Problem

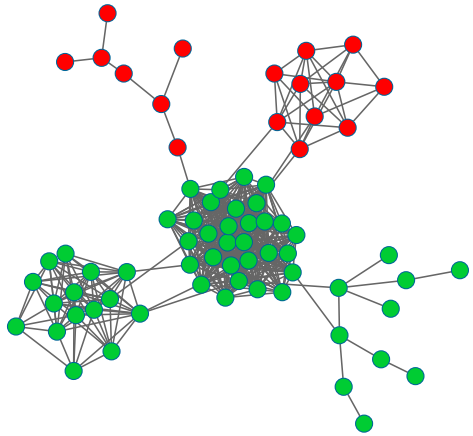
How can we optimally utilize transcriptional activity of gene products to construct the most informative tissue-specific sub-network?

PURDUE UNIVERSITY

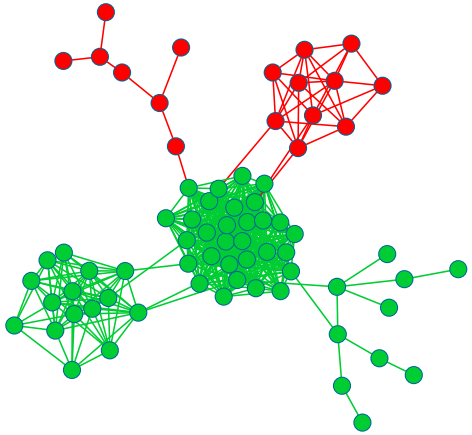
Toy Example



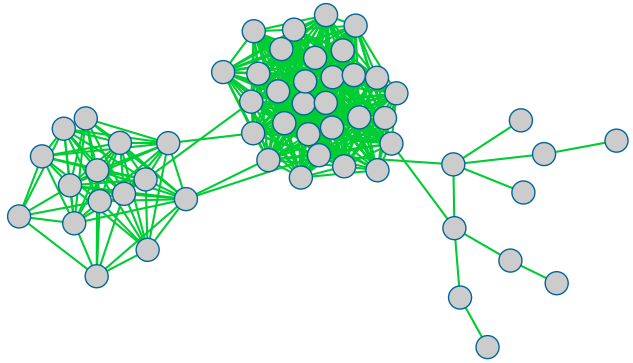
(a) Original



(b) Diffusion



(c) Projection



(d) Pruning

PURDUE UNIVERSITY

Minimal number of changes that smooths transcriptional activities over adjacent nodes in the network:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\mathbf{x}\mathbf{L}\mathbf{x} + \alpha \parallel \mathbf{x} - \mathbf{z} \parallel_1 \right\}$$

$$\text{Subject to: } \begin{cases} \mathbf{1}^T\mathbf{x} = 1 \\ 0 \leq \mathbf{x} \end{cases}$$

- Vector $\mathbf{z}$ initial value of transcriptional activities estimated by UPC
- Matrix $\mathbf{L}$ is the Laplacian matrix, defined as $\mathbf{A} - \mathbf{D}$, where d_{ii} is the weighted degree of i^{th} vertex in the global interactome.
- Parameter α controls the weight of regularization

PURDUE UNIVERSITY

- The first term defines a diffusion kernel that propagates activity of genes through network links.
- We can expand it as $\sum_{i,j} w_{i,j}(x_i - x_j)^2$, which is the accumulated difference of values between adjacent nodes scaled by the weight of the edge connecting them.
- The Laplacian operator $\mathbf{L}$ acts on a given function defined over vertices of a graph, such as $\mathbf{x}$, and computes the smoothness of $\mathbf{x}$ over adjacent vertices.

PURDUE UNIVERSITY

- The second term is a regularizer which penalizes changes or deviations
- We can expand it as $\sum_i |x_i - z_i|$, where x_i and z_i are the (inferred) functional and the transcriptional activity of gene i, respectively.
- It enforces sparsity over the vector of differences between *transcriptional* and *functional* activities.

PURDUE UNIVERSITY

Updating edges

$$\hat{\mathbf{A}} = \mathbf{diag}(\mathbf{x}^*) * \mathbf{A} * \mathbf{diag}(\mathbf{x}^*)$$

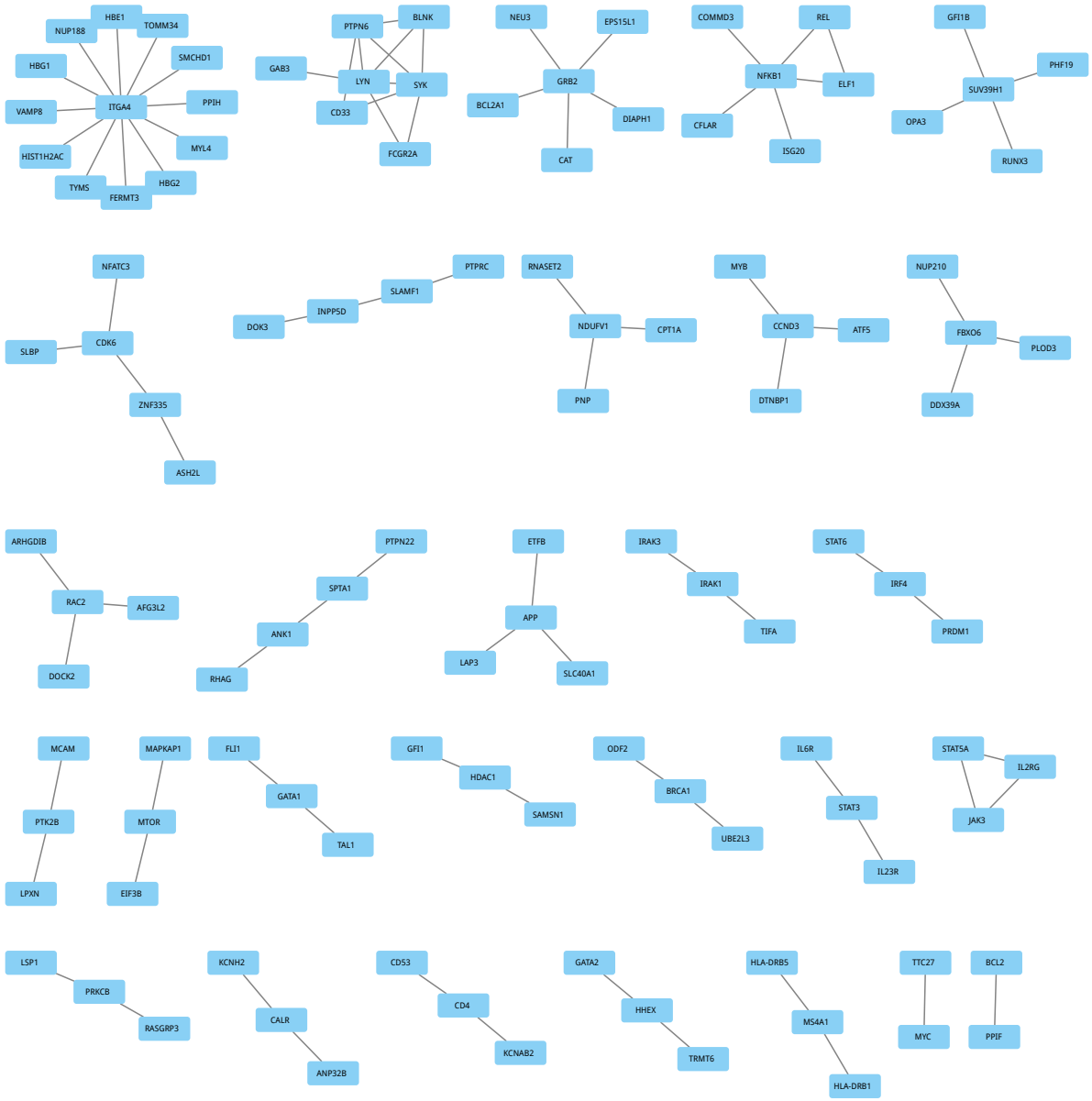
- $\mathbf{x}^*$ is the solution of optimization problem
- It represents functional activity of genes
- Functional activities are inferred from the global network context
- We update each edge according to the functional activity of its end-points

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Differential blood complexes



PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Experimental design

- 28 Leukemia-related genes
- Use each gene as seed, rank the rest of genes using random walk with restart (RWR)
- Compute p-value for each ranking
- Combine p-values to compute a meta p-value

PURDUE UNIVERSITY

Definition

Random walk on a graph G, initiated from vertex v, is the sequence of transitions among vertices, starting from v. At each step, the random walker randomly chooses the next vertex from among the neighbors of the current node.

It is a Markov chain with the transition matrix P, where $p_{ij} = Prob(S_{n+1} = v_i | S_n = v_j)$ and random variable S_n represents the state of the random walk at the time step n.

PURDUE UNIVERSITY

Definition

Random walk with restart (RWR) is a modified Markov chain in which, at each step, a random walker has the choice of either continuing along its path, with probability α, or jump (teleport) back to the initial vertex, with probability $1 - \alpha$.

The transition matrix of the modified chain, M, can be computed as $M = \alpha P + (1 - \alpha)\mathbf{e}_v \mathbf{1}^T$, where $\mathbf{e}_v$ is a stochastic vector of size n having zeros everywhere, except at index v, and $\mathbf{1}$ is a vector of all ones.

PURDUE UNIVERSITY

The portion of time spent on each node in an infinite random walk with restart initiated at node v, with parameter α.

Definition

Stationary distribution of the modified chain

$$\begin{aligned} \boldsymbol{\pi}_v(\alpha) &= M\boldsymbol{\pi}_v(\alpha) \\ &= (\alpha P + (1-\alpha)\mathbf{e}_v\mathbf{1}^T)\boldsymbol{\pi}_v(\alpha) \end{aligned}$$

Enforcing a unit norm on the dominant eigenvector to ensure its stochastic property, $\| \boldsymbol{\pi}_v(\alpha) \|_1 = \mathbf{1}^T\boldsymbol{\pi}_v = 1$, we will have:

PURDUE UNIVERSITY

Definition

Iterative form of the information flow process:

$$\boldsymbol{\pi}_v(\alpha) = \alpha P \boldsymbol{\pi}_v(\alpha) + (1-\alpha)\mathbf{e}_v,$$

Definition

Explicit (direct) formulation of the information flow process:

$$\boldsymbol{\pi}_v(\alpha) = \underbrace{(1-\alpha)(I-\alpha P)^{-1}}_{Q}\mathbf{e}_v,$$

PURDUE UNIVERSITY

Definition

Expansion using the Neumann series:

$$\boldsymbol{\pi}_v(\alpha) \quad = \quad (1-\alpha)\sum_{i=0}^{\infty}(\alpha P)^i \mathbf{e}_v$$

Thus, $\boldsymbol{\pi}_v(\alpha)$ is a function of:

- Distance to source node (v)
- Multiplicity of paths

PURDUE UNIVERSITY

Assigning p-value to each prioritization using mHG

Hypergeometric pvalue– fixed cut

$$
\begin{aligned}
p\text{-value}(Z = b_l(\lambda)) &= Prob(b_l(\lambda) \leq Z) \\
&= HGT(b_l(\lambda)|m, T, l) \\
&= \sum_{x=b_l(\lambda)}^{min(T,l)} \frac{\binom{T}{x}\binom{m-T}{l-x}}{\binom{m}{l}}
\end{aligned}
$$

PURDUE UNIVERSITY

Assigning p-value to each prioritization using mHG

minimum Hypergeometric (mHG) score

$$mHG(\lambda) \quad = \quad min_{1 \leq l \leq m} p\text{-value}(Z = b_l(\lambda))$$

minimum Hypergeometric (mHG) p-value is computed from mHG score using dynamic programing

PURDUE UNIVERSITY

Computing meta p-value

This method gathers a statistic $\mathcal{S} = \sum_{i=1}^{k} p_i$ for a set of k given p-values, and computes the meta p-value by assigning significance to $\mathcal{S}$ as:

Edgington method

$$\sum_{j=0}^{\lfloor \mathcal{S} \rfloor} -1^j \binom{k}{j} \frac{(\mathcal{S} - j)^k}{k!}$$

PURDUE UNIVERSITY

Moral of story!

Global interactome

- p-value $= 1.6 \times 10^{-199}$

Leukemia-specific interactome

- p-value $= 1.7 \times 10^{-244}$

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Prize-Collecting Steiner Tree (PCST)

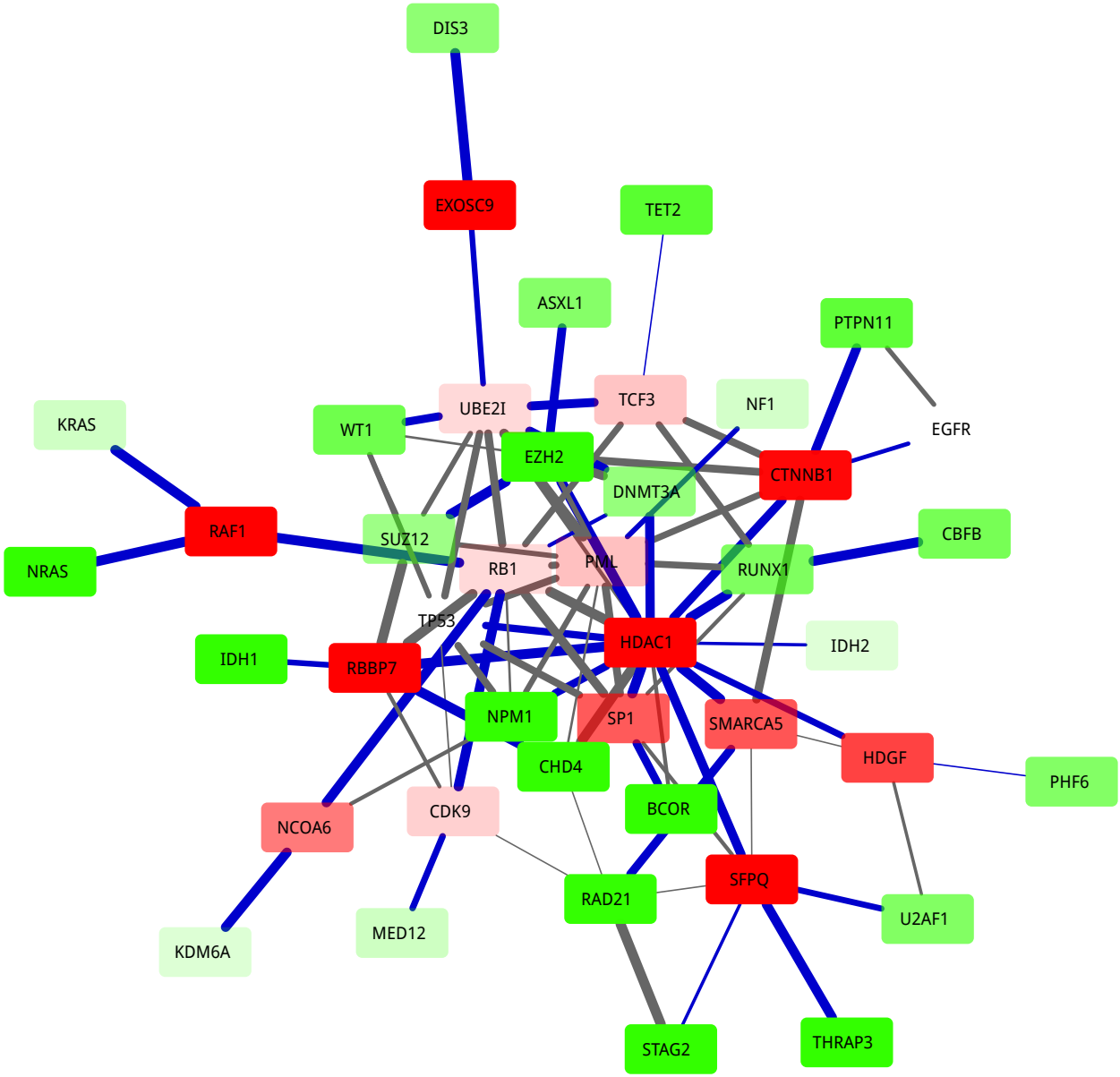
Objective

$$\underset{<v,e>\in T}{\operatorname{argmin}} \left\{ \sum_{e} c_e - \lambda \sum_{v} b_v \right\},$$

T is an induced tree of the given graph, v and e are the set of vertices and edges in T, respectively, c_e is the cost of choosing edge e, and b_v is the reward/prize of collecting node v.

PURDUE UNIVERSITY

Leukemia



PURDUE UNIVERSITY