Role of Analytical and Statistical Modeling in Data Analytics

Ananth Grama

Center for Science of Information, and
Department of Computer Science
Purdue University

With lots of help from Shahin Mohammadi, Abram Magner, Mehmet Koyuturk.

July, 2016

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

- p-Value: the probability of obtaining at least as extreme results given that the null hypothesis is true. Stated otherwise, this is the probability that a random selection (null hypothesis) yields an observation as extreme or more.
  Example: We have 30 people in this room. All 30 of you are under 40 years of age. The p-value of this observation is the probability that a randomly drawn sample of 30 people from the general population only contains people who are no more than 40 years old. *The lower this probability, the more surprising is the observation.*
- Statistical Signifiance: is attained when a p-value is less than a prescribed significance level.
- Statistical significance is typically used for hypothesis testing. However, we use this to estimate the "surprisingness" of observations.

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Can I predict the Coffee Drinkers – a classification problem?
If I build a classifier with 90% precision and 90% recall, would you be happy?

Can I predict the Coffee Drinkers – a classification problem?
If I build a classifier with 90% precision and 90% recall, would you be happy?

What if I now told you that 90% of all people are coffee drinkers – the Null Hypothesis!!

PURDUE UNIVERSITY

Can I predict the Coffee Drinkers – a classification problem?
If I build a classifier with 90% precision and 90% recall, would you be happy?

What if I now told you that 90% of all people are coffee drinkers – the Null Hypothesis!!

In the context of the null hypothesis, a classifier with 90% precision and 90% recall has no statistical significance at all!!

PURDUE UNIVERSITY

How can I infer rules of the kind: P(Bread) $\rightarrow$ P(Butter)? That is, people who buy bread are also likely to buy butter.

The association rule mining algorithm works in two steps – in the first step, all frequent sets are identified, and in the second step, the conditional probability for these frequent sets is computed to identify association rules.

Frequent sets are themselves computed using downward closure – all subsets of frequent subsets must themselves be frequent.

PURDUE UNIVERSITY

Motivation: Some Examples: Associations

- In reality, it does not work the way it should – the method mostly identifies things that are obvious – nothing interesting.
- This is because frequency is a poor measure of statistical significance (if at all).
- Even for the simplest priors (all items purchase probabilities are i.i.d), frequent sets are not "statistically significant sets".

PURDUE UNIVERSITY

I run a set of uniformly distributed points through a standard clustering algorithm, say, k-means. What happens?

I get a set of completely meaningless clusters! Should I be impressed?

How do I find the significance of a given set of clusters under a prior?

PURDUE UNIVERSITY

Motivation: Some more questions to ponder.

- You run a graph clustering algorithm on facebook and you find a group of 300 people who are (almost) completely connected. Should you be excited?
- 70% of your friends on Facebook are male, 60% of your friends are computer scientists, and 80% of your friends are Indian. Do these numbers (and their overlaps) tell you anything?
- 75% of everyone's friends on Facebook live within 50 miles of them. Can I use distance as an indicator of friendship (topology completion)?
- Can I trace the flow of a contagion through a network (its lineage), in the form of a directed acyclic graph?

PURDUE UNIVERSITY

Motivation: Most probable explanations.

- In each of the questions above, I am looking for "surprisingness" of observations. In other words, I am trying to find observations that minimize the p-value.
- In yet another class of problems, we can look to maximize the probability, with respect to a prior:
  - Given a dynamic network (say the spread of an infection), who was patient 0?
  - Who was the first person on facebook?
- In each of these questions, we are looking for the explanation most consistent with our understanding of driving processes (priors). These maximize probability w.r.t. the prior.

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

What kinds of answers are we looking for?

When posed with an optimization problem (or an analytics problem) we can look for different kinds of answers:

- The optimal solution (as defined by a suitable optimization problem). This solution is typically infeasible at scale, and often, an overkill, because of noise and missing data.
- The obvious solution. This solution is the one that you get from most analytics algorithms. However, this is rarely useful.
- The most statistically significant solution. This is similar to the optimal solution. It is generally infeasible at scale, but if feasible, it is typically the most desirable.
- Any solution with high statistical significance. This trades off feasibility and utility. In most cases, this is the solution of choice.
- The most probable solution. Generally computationally infeasible, however, proofs of near-optimality are feasible.

PURDUE UNIVERSITY

How do we develop solutions?

- Minimize statistical significance explicitly (i.e., use minimum p-value as an optimization criteria). This is generally hard to formulate, and even harder to realize in an efficient algorithm.
- Separate the algorithm from the quantification of p-values. Design heuristic algorithms and show that the results are statistically significant.

PURDUE UNIVERSITY

Formulating and computing p-values?

- A p-value for an observation can be formulated in different ways. For example, when formulating a p-value for the facebook dense subgraph example, we can ignore the graph and simply look at the hypergeometric p-value (probability of k successes in n trials, drawn from a sample of N objects without replacement).
- Desirable formulations of p-values provide discriminating power.
- Defining suitable priors are critical for p values. A prior that is very distant from the data will lead to very low p-values for all observations. This is not useful.
- Priors may themselves be non-parametric.
- p-values may be analytical (often very hard to derive), or empirical (often expensive to compute because of number of trials required).

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

You run a graph clustering algorithm on facebook and you find a group of 300 people who are (almost) completely connected. Should you be excited?

PURDUE UNIVERSITY

Analytical Assessment of Statistical Significance of Clusters

- What is the significance of a dense component in a network?
- What is the significance of a conserved component in multiple networks?

PURDUE UNIVERSITY

- Interaction networks generally exhibit power-law property (or exponential, geometric, etc.)
- Analysis simplified through independence assumption
- Independence assumption may cause problems for networks with arbitrary degree distribution
- $P(uv \in E) = d_u d_v / |E|$, where d_u is expected degree of u, but generally $d_{\max}^2 > |E|$ for PPI networks
- Rigorous analysis on $G(n, p)$ model
- Extension to piecewise $G(n, p)$ to capture network characteristics more accurately

PURDUE UNIVERSITY

Significance of Dense Subgraphs

- A subgraph of r nodes is said to be ρ-dense if $F(r) \geq \rho r^2$, where $F(r)$ is the number of interactions between these r nodes
- What is the expected size of the largest ρ-dense subgraph in a random graph?
- Any ρ-dense subgraph with larger size is statistically significant!

PURDUE UNIVERSITY

Significance of Dense Subgraphs

- $G(n,p)$ model
  - n nodes, each interaction occurs with probability p
  - Simple enough to facilitate rigorous analysis
  - If we let $p = d_{\max}/n$, largest ρ-dense subgraph in $G(n,p)$ stochastically dominates that in a graph with arbitrary degree distribution
- Piecewise $G(n,p)$ model
  - Few nodes with many interacting partners, many nodes with few interacting partners
  - Captures the basic characteristics of many networks
  - Analysis of $G(n,p)$ model generalizes to this model

PURDUE UNIVERSITY

- Theorem: If G is a random graph with n nodes, where every edge exists with probability p, then

$$\lim_{n\to\infty} \frac{R_\rho}{\log n} = \frac{1}{\kappa(p,\rho)} \qquad (pr.), \tag{1}$$

where

$$\kappa(p,\rho) = \rho \log \frac{\rho}{p} + (1-\rho)\log \frac{1-\rho}{1-p}. \tag{2}$$

More precisely,

$$P(R_\rho \geq r_0) \leq O\left(\frac{\log n}{n^{1/\kappa(p,\rho)}}\right), \tag{3}$$

where

$$r_0 = \frac{\log n - \log\log n + \log \kappa(p,\rho)}{\kappa(p,\rho)} \tag{4}$$

for large n.

PURDUE UNIVERSITY

Largest Dense Subgraph

Plugging in values of n, p, and setting $\rho = 0.9$, the p-value of a 300 person ρ dense component in facebook is about 10^{-4}!

PURDUE UNIVERSITY

Piecewise $G(n,p)$ model

- The size of largest dense subgraph is still proportional to $\log n/\kappa$ with a constant factor depending on number of hubs
- Model:

$$P(uv \in E(G)) = \begin{cases} p_h & \text{if } u,v \in V_h \\ p_l & \text{if } u,v \in V_l \\ p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h \end{cases}$$

- Result:
  Let $n_h = |V_h|$. If $n_h = O(1)$, then $P(R_n(\rho) \geq r_1) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right)$, where

$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log \kappa(p_l,\rho) - \log e + 1}{\kappa(p_l,\rho)}$$

and $B = \frac{p_b q_l}{p_l} + q_b$, where $q_b = 1 - p_b$ and $q_l = 1 - p_l$.

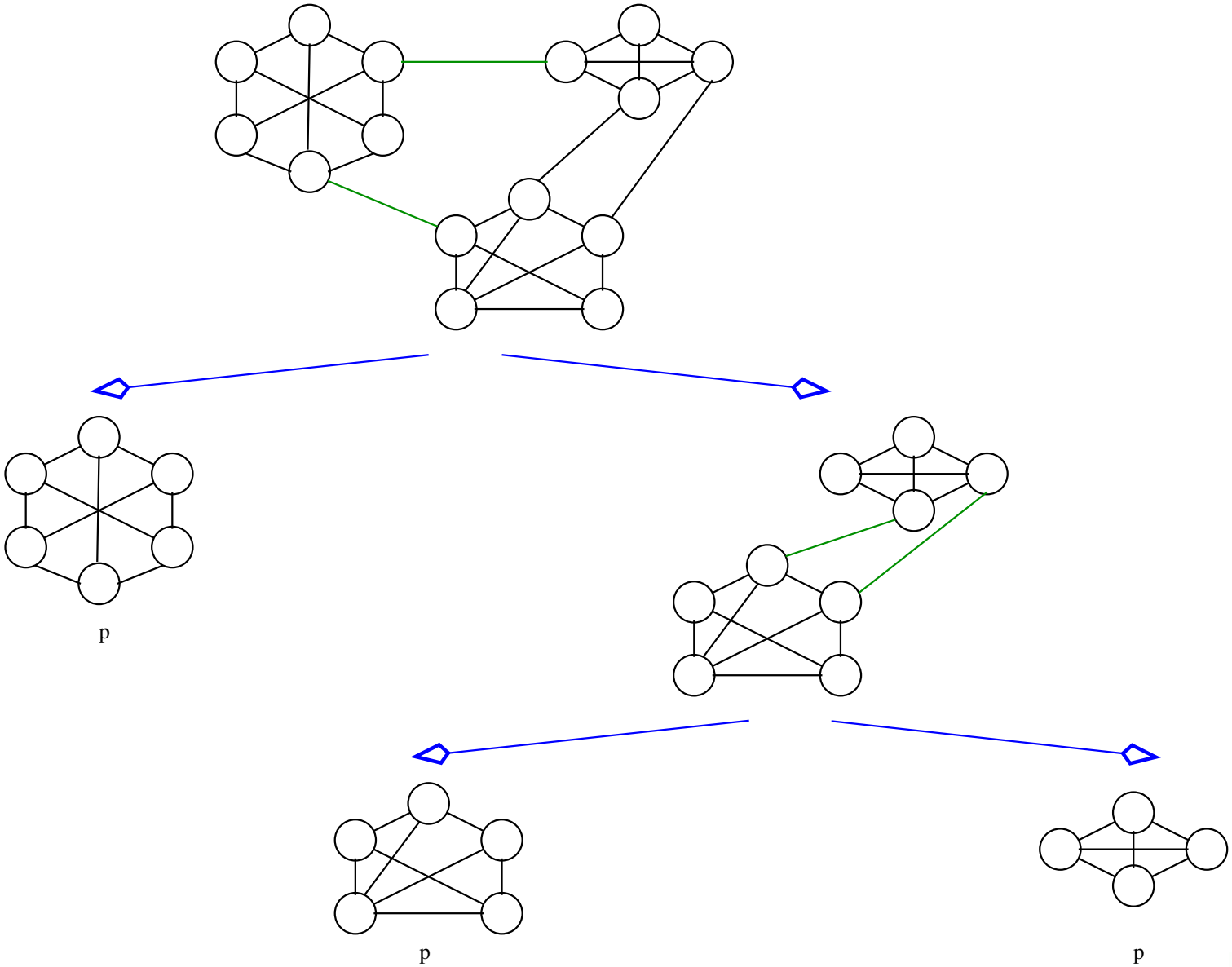
PURDUE UNIVERSITY

Algorithms Based on Statistical Significance

- Identification of topological modules
- Use statistical significance as a stopping criterion for graph clustering heuristics
- Find a minimum-cut bipartitioning of the network
- If any of the parts is dense enough, record it as a dense cluster of proteins
- Else, further partition them recursively

PURDUE UNIVERSITY

SiDeS Algorithm



SiDeS is available at `http://www.cs.purdue.edu/pdsl`

PURDUE UNIVERSITY

Performance of SiDeS

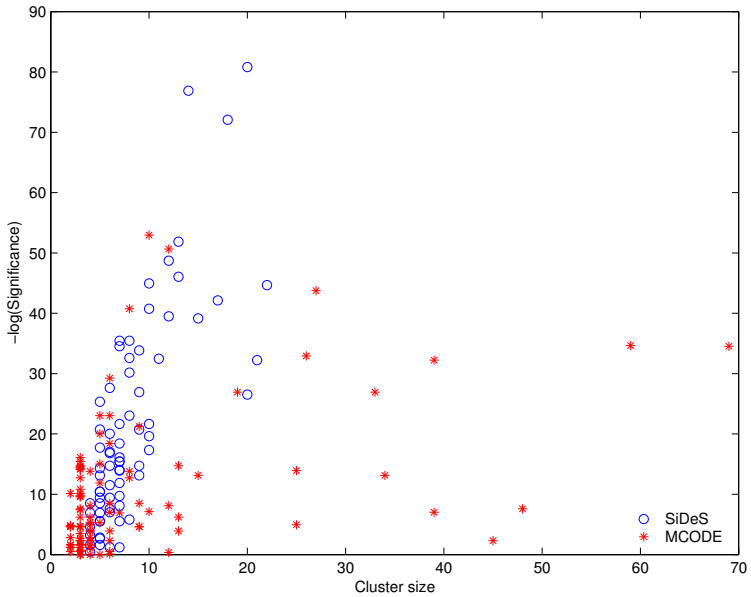
- Biological relevance of identified clusters is assessed with respect to Gene Ontology (GO)
- Estimate the statistical significance of the enrichment of each GO term in the cluster
- Quality of the clusters with respect to GO annotations
- Assume cluster C containing n_C genes is associated with term T that is attached to n_T genes and n_{CT} of genes in C are attached to T
- specificity $= 100 \times n_{CT}/n_C$
- sensitivity $= 100 \times n_{CT}/n_T$

	SiDeS			MCODE		
	Min.	Max.	Avg.	Min.	Max.	Avg.
Specificity (%)	43.0	100.0	91.2	0.0	100.0	77.8
Sensitivity (%)	2.0	100.0	55.8	0.0	100.0	47.6

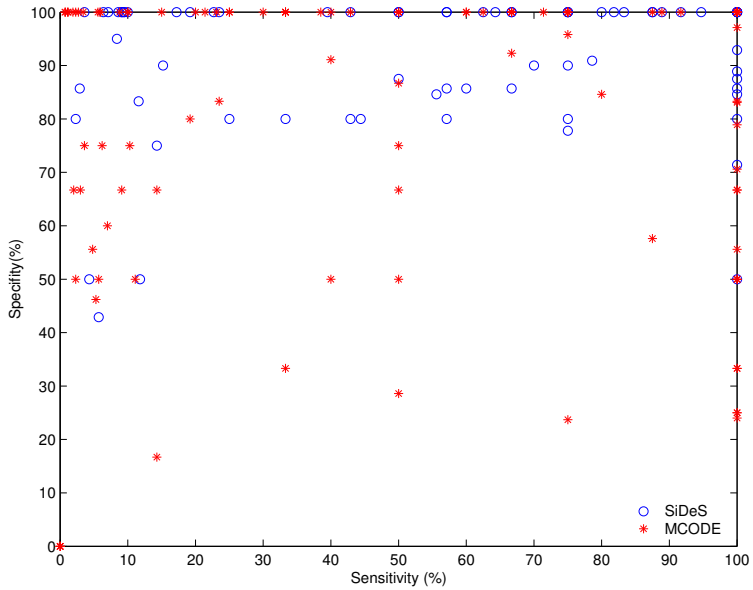
Comparison of SiDeS with MCODE

PURDUE UNIVERSITY

Performance of SiDeS



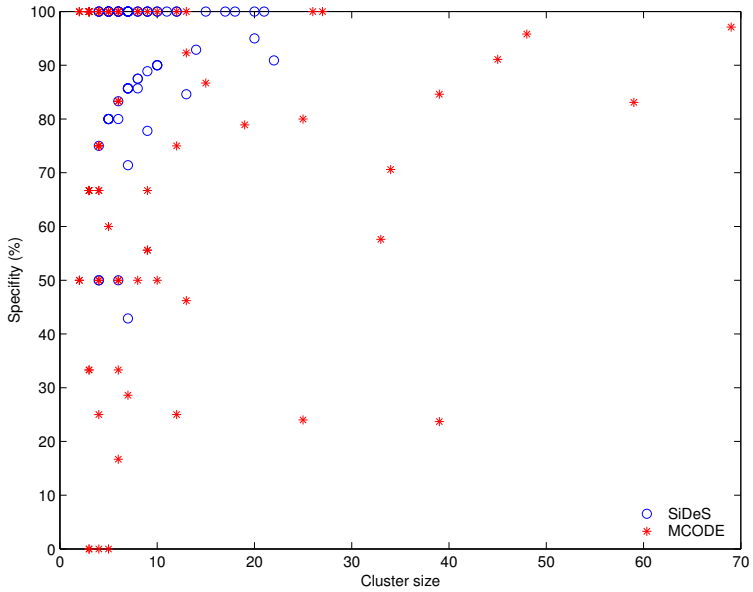
Size vs Significance



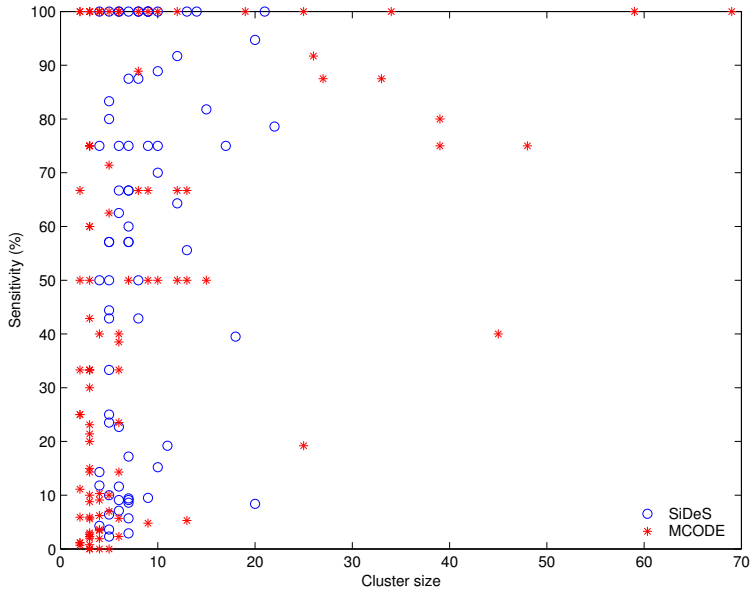
Sensitivity vs Specifity

PURDUE UNIVERSITY

Performance of SiDeS



Size vs Specificity



Size vs Sensitivity

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

70% of your friends on Facebook are male, 60% of your friends are computer scientists, and 80% of your friends are Indian. Do these numbers (and their overlaps) tell you anything?

PURDUE UNIVERSITY

Combining Density and Overlap (CoDO) of Clusters in Graphs

An overlap is significant if ...

1. It is at least as dense as the constituting graphs
2. It is "large enough"

PURDUE UNIVERSITY

Definition

$$p_{CoDO} = \Pr[|\hat{A} \cap \hat{B}| \geq |Z| \cap \delta(\hat{A} \cap \hat{B}) \geq \delta(Z)]$$

where Z is the set of vertices in the overlap subgraph and $\delta()$ measures the density of a graph, i.e. $\frac{|E|}{C(|V|,2)}$.

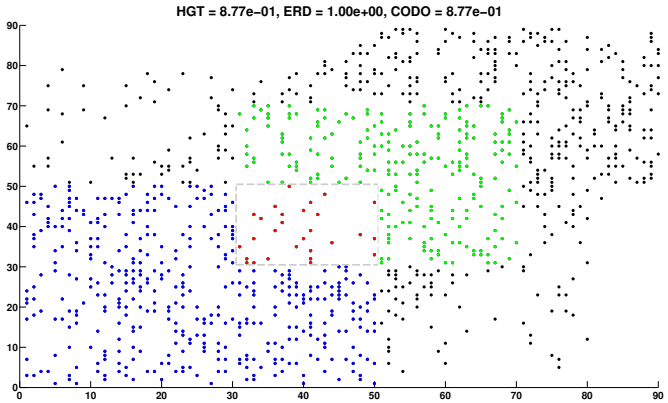
PURDUE UNIVERSITY

By conditioning on the size of the overlap, we can get an explicit formula for this p-value in terms of hypergeometric tails:

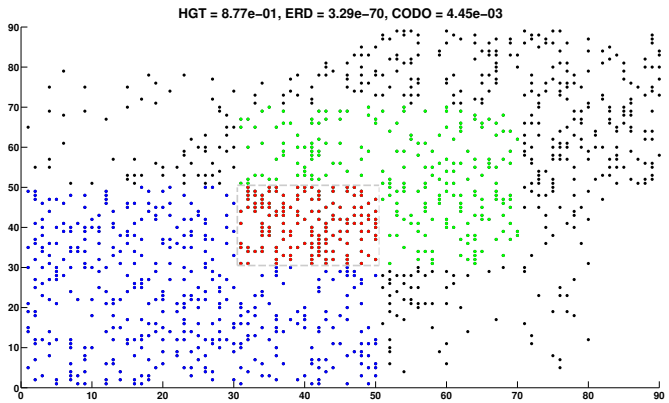
$$p_{CoDO} = \sum_{j=|Z|}^{\min\{|\hat{A}|,|\hat{B}|\}} \Pr\left[|\hat{A}\cap\hat{B}| = j\right] \cdot \Pr\left[\delta(\hat{A}\cap\hat{B}) \geq \delta(Z) \middle| |\hat{A}\cap\hat{B}| = j\right]$$

Combing Density and Overlap (CoDO)

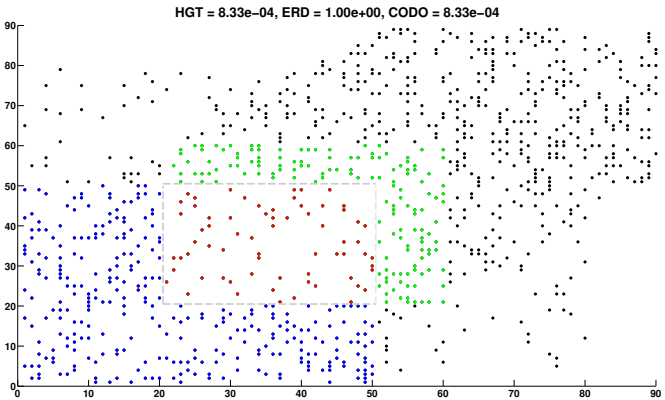
Example



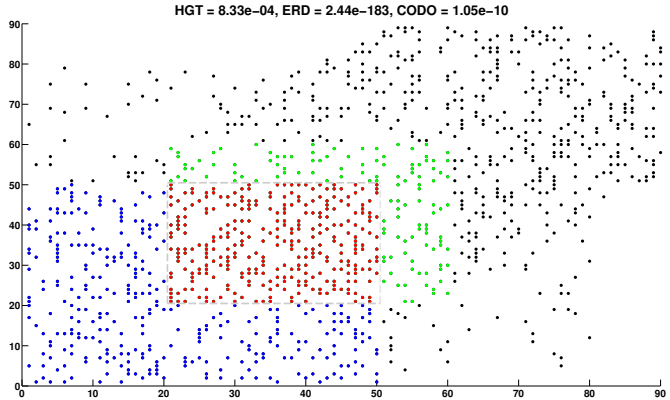
(a) Insignificant Overlap/Insignificant Density



(b) Insignificant Overlap/Significant Density

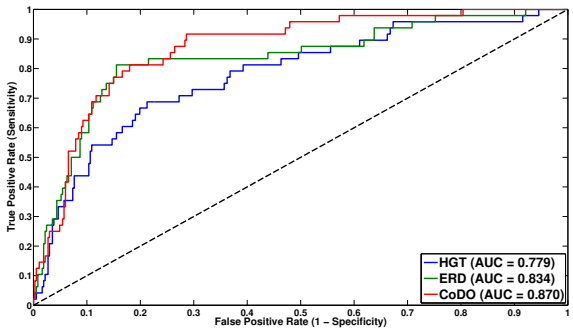


(c) Significant Overlap/Insignificant Density

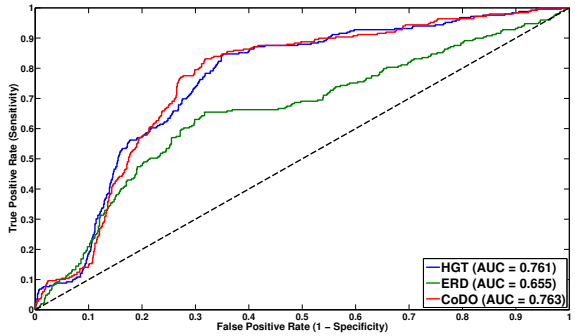


(d) Significant Overlap/Significant Density

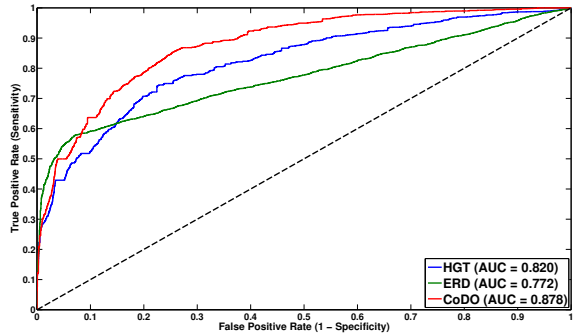
PURDUE UNIVERSITY



(e) Facebook



(f) Google+



(g) Twitter

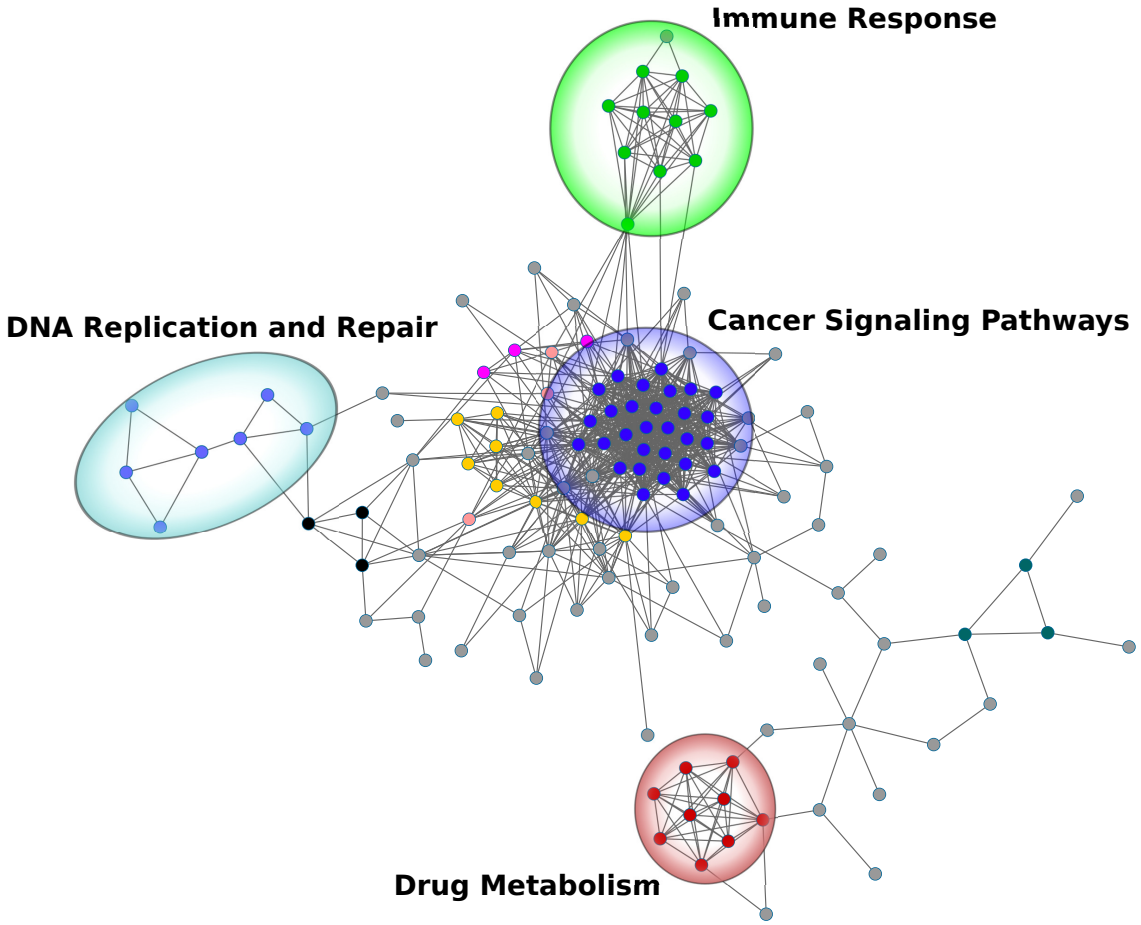
Definition

Ego Net is the induced subgraph among friends (alters) of a given user (ego)

PURDUE UNIVERSITY

Combing Density and Overlap (CoDO)

Application– Biological networks



Overlap among KEGG pathways is an indicator of pathway cross-talk

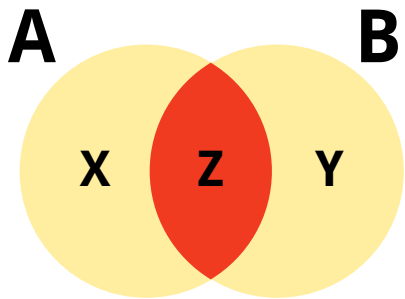
PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Tracking Lineage of Clusters

Say, your favorite clustering algorithm gives you the following clusters:



Can we say anything about the lineage of these clusters?

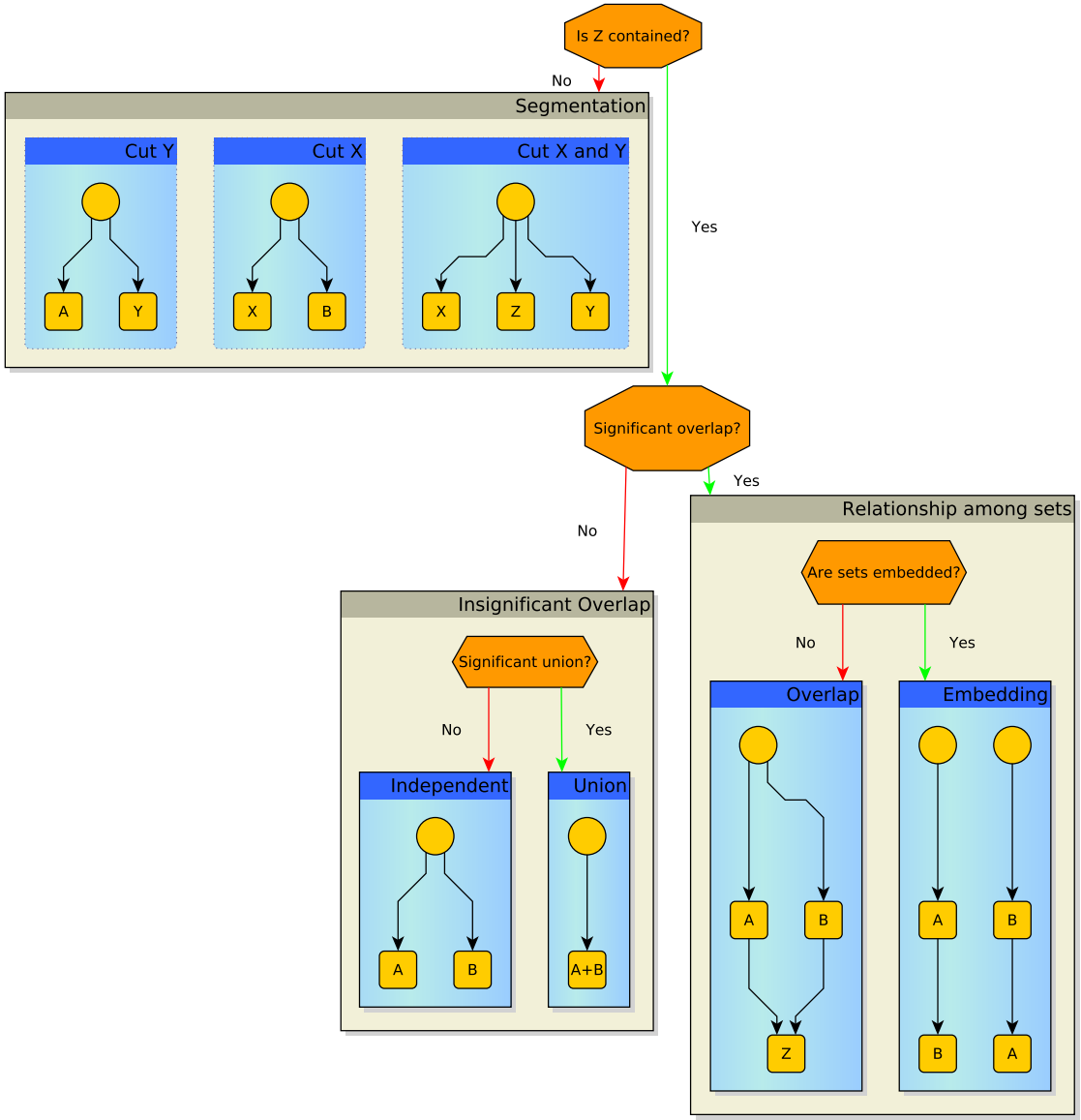
PURDUE UNIVERSITY

Tracking Lineage of Clusters

- We need to make sure that each of the clusters, x, y, and z are statistically significant themselves.
- We need to make sure that one of the clusters is not contained in the other (statistical significance of edge cut).
- We need to assess the statistical significance of the overlap of the two clusters.

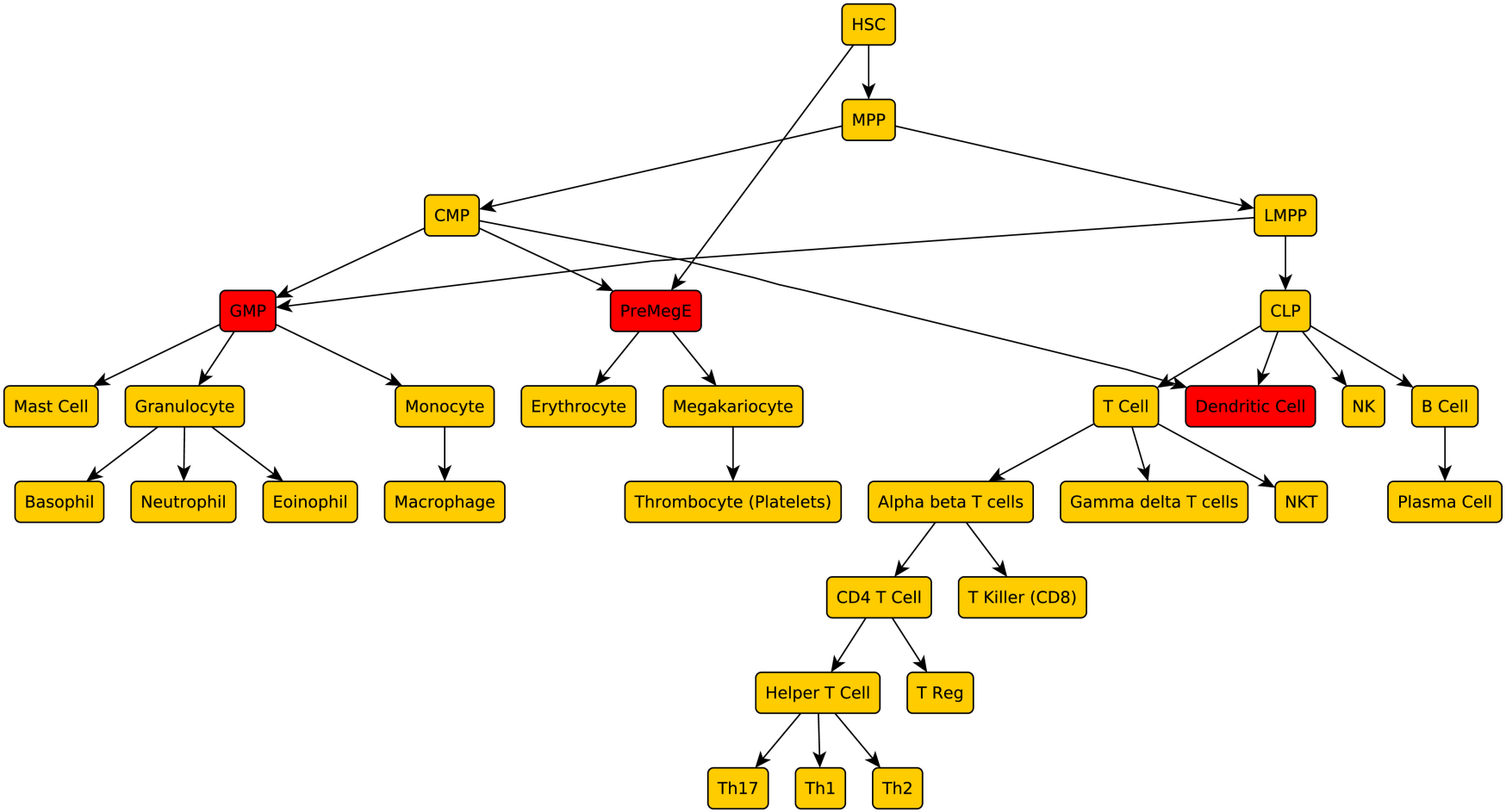
PURDUE UNIVERSITY

Tracking Lineage of Clusters



PURDUE UNIVERSITY

Tracking Lineage of Clusters



PURDUE UNIVERSITY

Outline

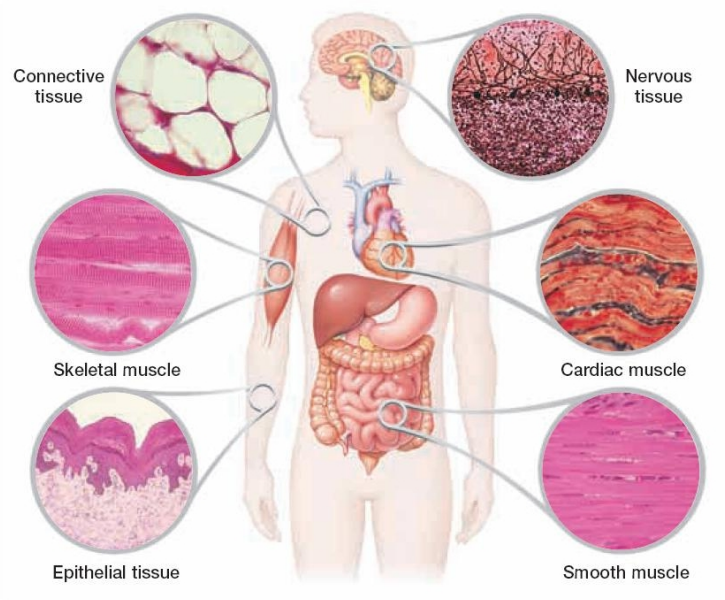
PURDUE UNIVERSITY

Constructing suitable background distributions – the case for non-parametric priors.

Aligning a set of tissue-specific interaction networks to the interactions in yeast with the goal of identifying the most statistically significant alignments.

Problem statement

For which tissues is yeast a good model organism?



Different human tissues, while inheriting a similar genetic code, exhibit unique anatomical and physiological properties.

What are the shared/ missing functional components in yeast, compared to human tissues?

PURDUE UNIVERSITY

Random model for tissue-specific networks

Definition

- Global human interactome: All potential interactions between human proteins, represented by graph $G = (V_G, E_G)$
- Tissue-specific network(s): Vertex-induced subgraph(s) of the Global human interactome, represented by $G_T = (V_T, E_T)$ with $n_T = |V_T|$, $V_T \subset V_G$, and $E_T \subset E_G$
- Universal genes: Ubiquitously expressed subset of human genes corresponding to houskeeping functions, represented by $V_U \subset V_G$, and $n_U = |V_U|$
- Random tissue-specific network(s): Vertex-induced subgraphs of G, constructed from $V_{\mathcal{R}} = V_U \cup V_{\mathcal{S}}$, with $V_{\mathcal{S}}$ being random set of vertices of size $n_T - n_U$ selected from $V_G \setminus V_U$

PURDUE UNIVERSITY

Significance of network alignment(s)

Definition

- Original alignment: $\mathcal{W} = \boldsymbol{w}^T \boldsymbol{x}$, $\mathcal{O} = \frac{1}{2}\boldsymbol{x}^T \mathcal{S} \boldsymbol{x}$
- Monte-Carlo simulation: Let $\mathcal{W}_{\mathcal{R}}$ and $\mathcal{O}_{\mathcal{R}}$ be the random vectors representing the weight and overlap of aligning $k_{\mathcal{R}}$ random tissue-specific networks with yeast
- Positive/Negative cases: k_P is the number of random cases with both $\mathcal{W}_{\mathcal{R}} \leq \mathcal{W}$ and . $\mathcal{O}_{\mathcal{R}} \leq \mathcal{O}$. k_N is defined as the size of complement set.
- p-value bounds:

$$\delta_{\mathcal{R}} = \frac{k_P}{k_{\mathcal{R}}} \leq \text{alignment p-value} \leq 1 - \frac{k_N}{k_{\mathcal{R}}} = \Delta_{\mathcal{R}}$$

- Alignment p-value:

$$p - value = Prob(\alpha * \mathcal{O} + \beta * \mathcal{W} \leq \mathcal{O}\mathcal{W}_{\mathcal{R}})$$

PURDUE UNIVERSITY

Definition

Selectivity p-value– Given a cluster of homogenous tissues:

$$\begin{aligned} p\text{-value}(X = c_n) &= Prob(c_n \leq X) \\ &= HGT(c_n|N, n, c_N) \\ &= \sum_{x=c_n}^{min(c_N,n)} \frac{C(c_N, x)C(N - c_N, n - x)}{C(N, n)} \end{aligned}$$

N: total number of tissues, n: number of tissues in the cluster, c_N: number of tissues in which a given gene is expressed, c_n: number of tissue in the cluster that the given gene is expressed.

PURDUE UNIVERSITY

Human-specific or conserved?

Definition

Classification of human tissue-selective genes:

- Conserved: Subset of tissue-selective genes that are consistently aligned in the "majority" of aligned tissues in the given group
- Human-specific: Subset of tissue-selective genes that are consistently unaligned in the "majority" of tissues in the given group
- Unclassified: None of the above

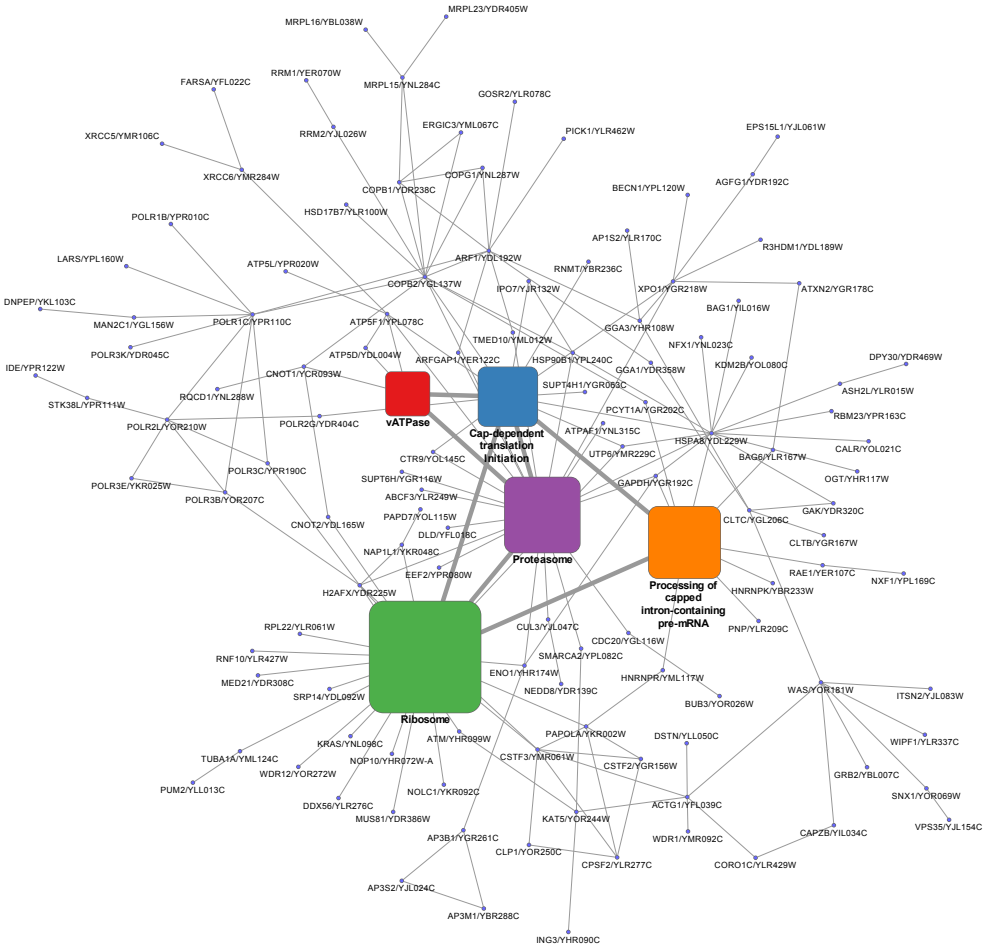
Definition

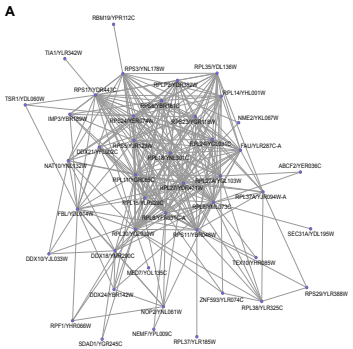
Majority voting:

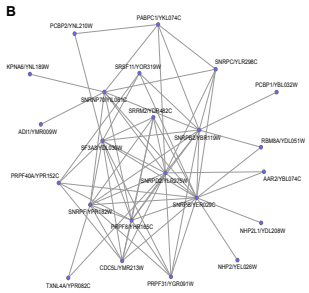
- Alignment consistency table: Yeast partner of each tissue-selective gene in the given cluster of tissues
- Consensus rate: Minimum percentage of tissues (columns) in each row of the alignment consistency table that have to agree to make a decision about conserved/human-specificity

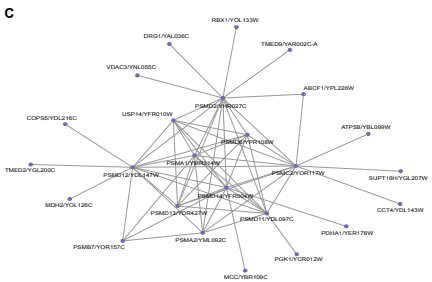
PURDUE UNIVERSITY

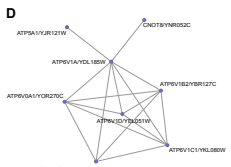
Core genes– The most conserved subset of housekeeping genes

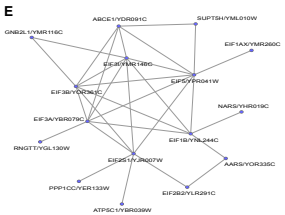












PURDUE UNIVERSITY

Functional enrichment of HK genes

Core subset

- Ribosome biogenesis
- Translation
- Protein targeting
- RNA splicing
- mRNA surveillance

PURDUE UNIVERSITY

- Anatomical structure development
- Paracrine signaling
- NADH dehydrogenase (mitochondrial Complex I)

The most similar tissues to yeast

Name	pval lower bound	overall pval	pval upper bound	confidence
Myeloid Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Monocytes	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Dentritic Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
NK Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
T-Helper Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Cytotoxic T-Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
B-Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Endothelial	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Hematopoietic Stem Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
MOLT-4	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
B Lymphoblasts	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
HL-60	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
K-562	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Early Erythroid	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Bronchial Epithelial Cells	< 1.00e-04	< 1.00e-04	0.0002	0.9998
Colorectal Adenocarcinoma	< 1.00e-04	< 1.00e-04	0.0004	0.9996
Daudi	< 1.00e-04	< 1.00e-04	0.0009	0.9991
Testis Seminiferous Tubule	< 1.00e-04	< 1.00e-04	0.0012	0.9988
Smooth Muscle	< 1.00e-04	< 1.00e-04	0.0016	0.9984
Blood (Whole)	< 1.00e-04	< 1.00e-04	0.0053	0.9947
Thymus	< 1.00e-04	0.0001	0.0062	0.9938
Testis Interstitial	< 1.00e-04	0.0004	0.0086	0.9914

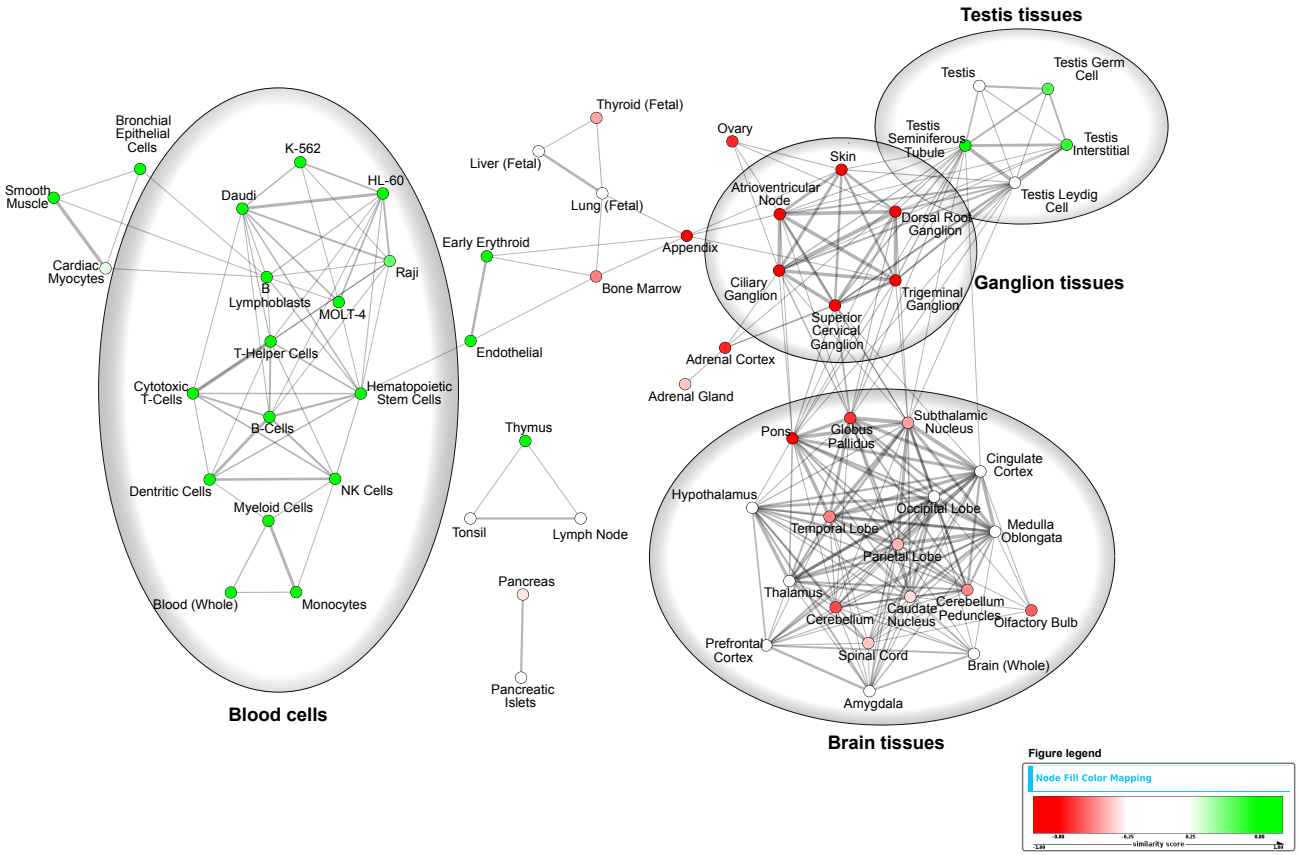
PURDUE UNIVERSITY

The least similar tissues to yeast

Name	pval lower bound	overall pval	pval upper bound	confidence
Trigeminal Ganglion	0.9947	0.9994	1	0.9947
Superior Cervical Ganglion	0.9847	0.9991	1	0.9847
Ciliary Ganglion	0.9407	0.9813	0.9964	0.9443
Atrioventricular Node	0.8746	0.9792	0.9921	0.8825
Skin	0.8355	0.9297	0.9809	0.8546
Heart	0.7934	0.9585	0.9815	0.8119
Appendix	0.7596	0.9371	0.973	0.7866
Dorsal Root Ganglion	0.7065	0.933	0.9717	0.7348
Skeletal Muscle	0.3994	0.5902	0.7866	0.6128
Uterus Corpus	0.233	0.7736	0.8769	0.3561
Lung	0.0771	0.3853	0.5544	0.5227
Pons	0.0674	0.5201	0.6983	0.3691
Salivary Gland	0.0639	0.3449	0.5173	0.5466
Liver	0.0600	0.6857	0.8519	0.2081
Ovary	0.0388	0.2735	0.4481	0.5907
Trachea	0.0259	0.2376	0.4146	0.6113
Globus Pallidus	0.0206	0.2471	0.4336	0.587
Cerebellum	0.0127	0.1950	0.3783	0.6344

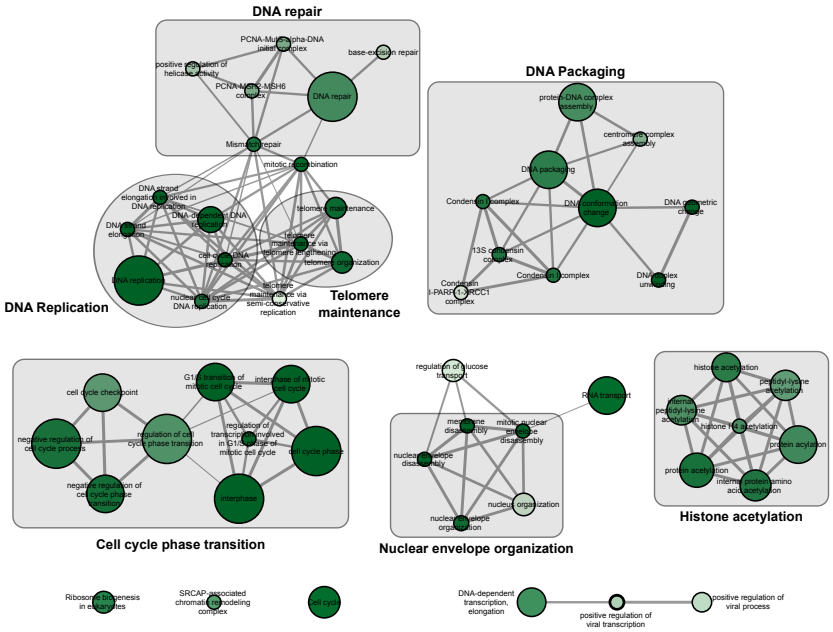
PURDUE UNIVERSITY

Tissue-tissue similarity network

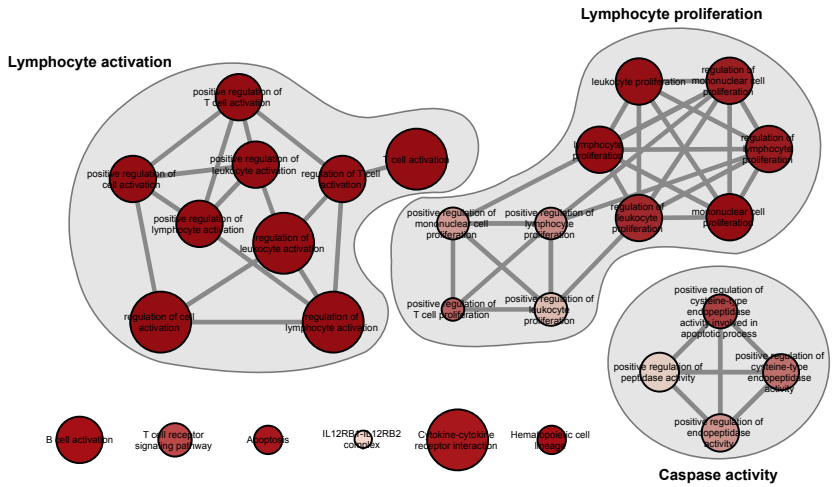


PURDUE UNIVERSITY

Blood cells



(h) Conserved

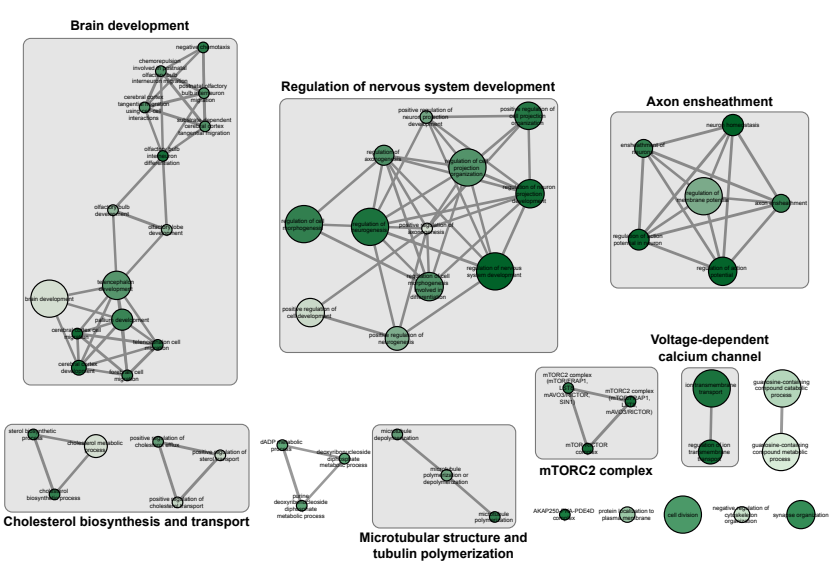


(i) Human-specific

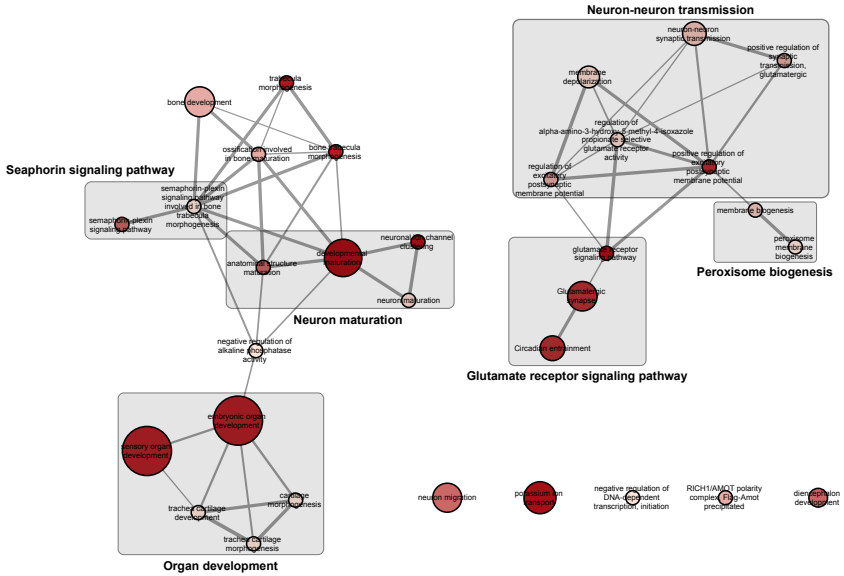
Figure: Enrichment map of unique blood-selective functions.

PURDUE UNIVERSITY

Brain tissues



(a) Conserved



(b) Human-specific

Figure: Enrichment map of unique brain-selective functions.

PURDUE UNIVERSITY

Enriched disease classes

	Conserved genes		Human-specific genes	
	Disease class	p-value	Disease class	p-value
Blood cells	Cancer	$2.85 * 10^{-3}$	Immune	$1.88 * 10^{-9}$
			Infection	$1.00 * 10^{-2}$
Brain tissues	Psych	$3.59 * 10^{-4}$	Psych	$5.70 * 10^{-8}$
	Chemdependency	$2.60 * 10^{-3}$	Neurological	$2.97 * 10^{-2}$
	Pharmacogenomic	$9.74 * 10^{-2}$		

PURDUE UNIVERSITY

Comparative analysis of brain-specific pathologies

Disorder	Conserved genes	Human-specific genes
schizophrenia	0.008573	8.4905E-06
autism	0.048288	0.00077448
dementia	0.0014356	-
schizophrenia; schizoaffective disorder; bipolar disorder	-	0.0021433
myocardial infarct; cholesterol, HDL; triglycerides; atherosclerosis, coronary; macular degeneration; colorectal cancer	0.0051617	-
epilepsy	0.071562	0.0064716
seizures	-	0.020381
bipolar disorder	0.048288	0.022016
attention deficit disorder conduct disorder oppositional defiant disorder	0.032444	0.023865

PURDUE UNIVERSITY

Some Interesting Questions

PURDUE UNIVERSITY

Significance of Keywords in Text

- Search engines typically search for keywords, and use proximity of keywords as one of the primary heuristics for ordering search results.
- An alternate (and more rogorous, IMHO) formulation would order documents by the statistical significance of keyword occurrance.
- Given a sequence $< S >$ (along with a generation model for $< S >$, preferably Markovian), what is the likelihood of observing a given set of keywords $k_1, ..., k_j$ within the shortest subsequence $< S' >$ of $< S >$ of length d or less.
- The lower this p-value, the more significant the match (i.e., rank this higher among returned results).

PURDUE UNIVERSITY

Significance of Snippets in Graphs

- Consider your favorite social network and construct a mapping from node attrubutes to a scalar (the color of the node). For instance, females, over 30, making more than \$200K are mapped to color Red; males over 30 with two children are mapped to color Green.
- Can we argue the statistical significance of tight subgraphs that contain prescribed colors – for instance, do we see overrepresentation of tight subgraphs containing Red, Green, and Blue (children, perhaps) nodes?
- Given a graph G (along with a generation model for G), what is the likelihood of observing a given set of query colors $c_1, ..., c_j$ in a subgraph G' of G of diameter d or less?

PURDUE UNIVERSITY

Arrival Sequence of Nodes in Dynamic Graphs

- For a dynamic network, how can we infer a likely arrival sequence for nodes?
- Given a graph G (along with a generation model for G), what is the likelihood of observing the graph G for a given arrival sequence $n_1, n_2, ..., n_k$. How do we determine this arrival sequence to maximize the probability?
- For a preferential attachment model, we can generate a set of precedence constraints. Any arrival sequence that satisfies all of these precedence constraints is a potential true sequence.
- How do we optimize within this (super)exponential space? We have some experimental evidence and theoretical justification to show that we dont have to!

PURDUE UNIVERSITY

Arrival Sequence of Nodes in Dynamic Graphs

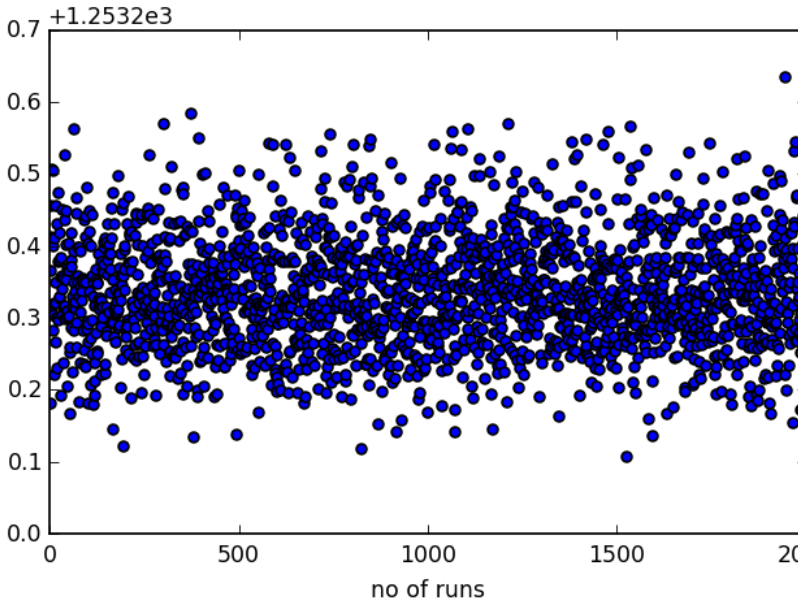


Figure: Probability of randomly selected feasible arrival sequences.

PURDUE UNIVERSITY

Some Parting Thoughts

- Significance if an essential aspect of quantifying the usefulness of an observation.
- Significance is an essential component of validating observations.
- Significance testing is hard; requiring solutions to complex analysis problems.
- Maximizing significance should be one of the primary goals of algorithms/ heuristics.

PURDUE UNIVERSITY

Thanks

Many thanks to the organizers for giving me this opportunity to talk, and to you for engaging with me!

PURDUE UNIVERSITY