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
1 Background
   - Basic Concepts
   - Motivating Examples
   - Analytical Frameworks

2 Analytics in Action
   - Statistical Significance of Clusters
   - Statistical Significance of Overlap of Clusters
   - Clusters and Lineage
   - Design of (Nonparametric) Priors

3 Some Open Problems
1. **Background**
   - Basic Concepts
   - Motivating Examples
   - Analytical Frameworks

2. **Analytics in Action**
   - Statistical Significance of Clusters
   - Statistical Significance of Overlap of Clusters
   - Clusters and Lineage
   - Design of (Nonparametric) Priors

3. **Some Open Problems**
Basic Concepts

- **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 Significance**: 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.
Outline

1 Background
   - Basic Concepts
   - Motivating Examples
   - Analytical Frameworks

2 Analytics in Action
   - Statistical Significance of Clusters
   - Statistical Significance of Overlap of Clusters
   - Clusters and Lineage
   - Design of (Nonparametric) Priors

3 Some Open Problems
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!!
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!!
How can I infer rules of the kind: $P(\text{Bread}) \rightarrow P(\text{Butter})$? 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.
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”.
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?
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?
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.
1. **Background**
   - Basic Concepts
   - Motivating Examples
   - Analytical Frameworks

2. **Analytics in Action**
   - Statistical Significance of Clusters
   - Statistical Significance of Overlap of Clusters
   - Clusters and Lineage
   - Design of (Nonparametric) Priors

3. **Some Open Problems**
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.
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.
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).
Outline

1 Background
   • Basic Concepts
   • Motivating Examples
   • Analytical Frameworks

2 Analytics in Action
   • Statistical Significance of Clusters
   • Statistical Significance of Overlap of Clusters
   • Clusters and Lineage
   • Design of (Nonparametric) Priors

3 Some Open Problems
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?
What is the significance of a dense component in a network?
What is the significance of a conserved component in multiple networks?
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

\[ P(uv \in E) = \frac{d_u d_v}{|E|}, \]
where \( d_u \) is expected degree of \( u \), but generally \( d_{\text{max}}^2 > |E| \) for PPI networks

Rigorous analysis on \( G(n, p) \) model

Extension to piecewise \( G(n, p) \) to capture network characteristics more accurately
A subgraph of \( r \) nodes is said to be \( \rho \)-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 \( \rho \)-dense subgraph in a random graph?

Any \( \rho \)-dense subgraph with larger size is statistically significant!
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 $p = \frac{d_{\text{max}}}{n}$, largest $\rho$-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
Theorem: If $G$ is a random graph with $n$ nodes, where every edge exists with probability $p$, then

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

(1)

where

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

More precisely,

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

(3)

where

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

for large $n$. 
Plugging in values of $n$, $p$, and setting $\rho = 0.9$, the $p$-value of a 300 person $\rho$ dense component in facebook is about $10^{-4}$!
**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 $q_l = 1 - p_l$. 
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
SiDeS Algorithm

SiDeS is available at http://www.cs.purdue.edu/pdsl
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 $= 100 \times \frac{n_{CT}}{n_C}$
- Sensitivity $= 100 \times \frac{n_{CT}}{n_T}$

<table>
<thead>
<tr>
<th></th>
<th>SiDeS</th>
<th></th>
<th></th>
<th></th>
<th>MCODE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity (%)</td>
<td>43.0</td>
<td>100.0</td>
<td>91.2</td>
<td>0.0</td>
<td>100.0</td>
<td>77.8</td>
<td></td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>2.0</td>
<td>100.0</td>
<td>55.8</td>
<td>0.0</td>
<td>100.0</td>
<td>47.6</td>
<td></td>
</tr>
</tbody>
</table>

Comparison of SiDeS with MCODE
Performance of SiDeS

Size vs Significance

Sensitivity vs Specificity
Performance of \textsc{SiDeS}

Size vs Specificity

Size vs Sensitivity
Outline

1 Background
- Basic Concepts
- Motivating Examples
- Analytical Frameworks

2 Analytics in Action
- Statistical Significance of Clusters
- Statistical Significance of Overlap of Clusters
- Clusters and Lineage
- Design of (Nonparametric) Priors

3 Some Open Problems
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?
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”
Combing Density and Overlap (CoDO)

Formal Statement

Definition

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

where $Z$ is the set of vertices in the overlap subgraph and $\delta()$ measures the density of a graph, i.e. $\frac{|E|}{C(|V|,2)}$. 
By conditioning on the size of the overlap, we can get an explicit formula for this $p$-value in terms of hypergeometric tails:

$$p_{\text{CoDO}} = \min\{|\hat{A}|,|\hat{B}|\} \sum_{j=|Z|} \Pr[|\hat{A} \cap \hat{B}| = j] \cdot \Pr[\delta(\hat{A} \cap \hat{B}) \geq \delta(Z)| |\hat{A} \cap \hat{B}| = j]$$
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
Combing Density and Overlap (CoDO)
Application– Social networks

Definition

Ego Net is the induced subgraph among friends (alters) of a given user (ego)
Overlap among KEGG pathways is an indicator of pathway cross-talk.
1 Background
- Basic Concepts
- Motivating Examples
- Analytical Frameworks

2 Analytics in Action
- Statistical Significance of Clusters
- Statistical Significance of Overlap of Clusters
- Clusters and Lineage
- Design of (Nonparametric) Priors

3 Some Open Problems
Say, your favorite clustering algorithm gives you the following clusters:

Can we say anything about the lineage of these 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.
Tracking Lineage of Clusters

Segmentation:
- Cut Y
- Cut X
- Cut X and Y

Significant overlap?
- Yes
- No

Relationship among sets:
- Overlap
- Embedding

Insignificant Overlap:
- Significant union?
  - Yes
  - No

Independent
- A
- B

Union
- A+B

Are sets embedded?
- Yes
- No
1 Background
   - Basic Concepts
   - Motivating Examples
   - Analytical Frameworks

2 Analytics in Action
   - Statistical Significance of Clusters
   - Statistical Significance of Overlap of Clusters
   - Clusters and Lineage
   - Design of (Nonparametric) Priors

3 Some Open Problems
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?
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 housekeeping 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 \( V_{\mathcal{R}} = V_U \cup V_S \), with \( V_S \) being random set of vertices of size \( n_T - n_U \) selected from \( V_G \setminus V_U \)
Significance of network alignment(s)

Definition

- **Original alignment:** \( W = w^T x, O = \frac{1}{2} x^T S x \)
- **Monte-Carlo simulation:** Let \( W_R \) and \( O_R \) be the random vectors representing the weight and overlap of aligning \( k_R \) random tissue-specific networks with yeast.
- **Positive/Negative cases:** \( k_P \) is the number of random cases with both \( W_R \leq W \) and \( O_R \leq O \). \( k_N \) is defined as the size of complement set.
- **p-value bounds:**
  \[
  \delta_R = \frac{k_P}{k_R} \leq \text{alignment p-value} \leq 1 - \frac{k_N}{k_R} = \Delta_R
  \]
- **Alignment p-value:**
  \[
  p - \text{value} = \text{Prob}(\alpha \ast O + \beta \ast W \leq OW_R)
  \]
Definition

Selectivity $p$-value— Given a cluster of homogenous tissues:

\[
\text{p-value}(X = c_n) = \text{Prob}(c_n \leq X) = HGT(c_n|N, n, c_N)
\]

\[
= \min(c_N, n) \sum_{x=c_n}^{\min(c_N, n)} \frac{C(c_N, x)C(N - c_N, n - x)}{C(N, n)}
\]

$N$: 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.
### 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
Core genes—The most conserved subset of housekeeping genes
Functional enrichment of HK genes
Core subset

- Ribosome biogenesis
- Translation
- Protein targeting
- RNA splicing
- mRNA surveillance
Functional enrichment of HK genes
Human-specific subset

- Anatomical structure development
- Paracrine signaling
- NADH dehydrogenase (mitochondrial Complex I)
The most similar tissues to yeast

<table>
<thead>
<tr>
<th>Name</th>
<th>pval lower bound</th>
<th>overall pval</th>
<th>pval upper bound</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myeloid Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Monocytes</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Dendritic Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>NK Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>T-Helper Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Cytotoxic T-Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>B-Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Endothelial</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Hematopoietic Stem Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>MOLT-4</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>B Lymphoblasts</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>HL-60</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>K-562</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Early Erythroid</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>1</td>
</tr>
<tr>
<td>Bronchial Epithelial Cells</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>0.0002</td>
<td>0.9998</td>
</tr>
<tr>
<td>Colorectal Adenocarcinoma</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>0.0004</td>
<td>0.9996</td>
</tr>
<tr>
<td>Daudi</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>0.0009</td>
<td>0.9991</td>
</tr>
<tr>
<td>Testis Seminiferous Tubule</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>0.0012</td>
<td>0.9988</td>
</tr>
<tr>
<td>Smooth Muscle</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>0.0016</td>
<td>0.9984</td>
</tr>
<tr>
<td>Blood (Whole)</td>
<td>&lt; 1.00e-04</td>
<td>&lt; 1.00e-04</td>
<td>0.0053</td>
<td>0.9947</td>
</tr>
<tr>
<td>Thymus</td>
<td>&lt; 1.00e-04</td>
<td>0.0001</td>
<td>0.0062</td>
<td>0.9938</td>
</tr>
<tr>
<td>Testis Interstitial</td>
<td>&lt; 1.00e-04</td>
<td>0.0004</td>
<td>0.0086</td>
<td>0.9914</td>
</tr>
</tbody>
</table>
## The least similar tissues to yeast

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

Figures: Blood cells, Brain tissues, Testis tissues, Ganglion tissues
Blood cells

(h) Conserved

(i) Human-specific

Figure: Enrichment map of unique blood-selective functions.
Figure: Enrichment map of unique brain-selective functions.
### Enriched disease classes

<table>
<thead>
<tr>
<th>Disease class</th>
<th>Conserved genes</th>
<th>Human-specific genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blood cells</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>2.85 * 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>Immune</td>
<td>1.88 * 10^{-9}</td>
<td></td>
</tr>
<tr>
<td>Infection</td>
<td>1.00 * 10^{-2}</td>
<td></td>
</tr>
<tr>
<td><strong>Brain tissues</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psych</td>
<td>3.59 * 10^{-4}</td>
<td></td>
</tr>
<tr>
<td>Chemdependency</td>
<td>2.60 * 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>Pharmacogenomic</td>
<td>9.74 * 10^{-2}</td>
<td></td>
</tr>
<tr>
<td>Psych</td>
<td>5.70 * 10^{-8}</td>
<td></td>
</tr>
<tr>
<td>Neurological</td>
<td>2.97 * 10^{-2}</td>
<td></td>
</tr>
</tbody>
</table>
## Comparative analysis of brain-specific pathologies

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Conserved genes</th>
<th>Human-specific genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>schizophrenia</td>
<td>0.008573</td>
<td>8.4905E-06</td>
</tr>
<tr>
<td>autism</td>
<td>0.048288</td>
<td>0.00077448</td>
</tr>
<tr>
<td>dementia</td>
<td>0.0014356</td>
<td>-</td>
</tr>
<tr>
<td>schizophrenia; schizoaffective disorder; bipolar disorder</td>
<td>-</td>
<td>0.0021433</td>
</tr>
<tr>
<td>myocardial infarct; cholesterol, HDL; triglycerides; atherosclerosis, coronary; macular degeneration; colorectal cancer</td>
<td>0.0051617</td>
<td>-</td>
</tr>
<tr>
<td>epilepsy</td>
<td>0.071562</td>
<td>0.0064716</td>
</tr>
<tr>
<td>seizures</td>
<td>-</td>
<td>0.020381</td>
</tr>
<tr>
<td>bipolar disorder</td>
<td>0.048288</td>
<td>0.022016</td>
</tr>
<tr>
<td>attention deficit disorder conduct disorder oppositional defiant disorder</td>
<td>0.032444</td>
<td>0.023865</td>
</tr>
</tbody>
</table>
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 rigorous, IMHO) formulation would order documents by the statistical significance of keyword occurrence.

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).
Consider your favorite social network and construct a mapping from node attributes 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 $c_1, \ldots, c_j$ in a subgraph $G'$ of $G$ of diameter $d$ or less?
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 don’t have to!
Arrival Sequence of Nodes in Dynamic Graphs

Figure: Probability of randomly selected feasible arrival sequences.
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.
Many thanks to the organizers for giving me this opportunity to talk, and to you for engaging with me!