Comparative Analysis of Networks

Ananth Grama
Coordinated Systems Lab and,
Computer Science Department,
Purdue University

http://www.cs.purdue.edu/people/ayg

This work is supported by National Institutes of Health, National Science Foundation, Department of Energy, Intel and Microsoft.
Outline

• Interaction Networks
  – Modeling, evolution, problems, practical implications

• Algorithms for Analyzing Interaction Networks
  – Analyzing biological networks for conserved interaction patterns
  – Pairwise Alignment of networks
  – Probabilistic models/analyses for assessing statistical significance

• Computational Synthesis of Interaction Networks
  – Inferring function from domain co-evolution
Conserved Interaction Patterns

- Given a collection of interaction networks, find sub-networks that are common to an interesting subset of these networks (Koyutürk, Grama, & Szpankowski, ISMB, 2004)
  - A sub-network is a group of interactions that are tied to each other (connected)
  - Frequency: The number of networks that contain a sub-network, is a coarse measure of statistical significance

- Computational challenges
  - How to relate nodes (proteins) in different networks (organisms)?
  - Requires solution of the intractable subgraph isomorphism problem
  - Must be scalable to potentially large number of networks
  - Networks are large (in the range of $10^K$ edges)
Graph Analysis

Network database

Interaction patterns that are common to all networks
Problem Statement

• Given a set of proteins $V$, a set of interactions $E$, and a many-to-many mapping from $V$ to a set of ortholog groups $\mathcal{L} = \{l_1, l_2, ..., l_n\}$, the corresponding interaction network is a labeled graph $G = (V, E, \mathcal{L})$.

  - $v \in V(G)$ is associated with a set of ortholog groups $L(v) \subseteq \mathcal{L}$.
  - $uv \in E(G)$ represents an interaction between $u$ and $v$.

• $S$ is a sub-network of $G$, i.e., $S \subseteq G$ if there is an injective mapping $\phi : V(S) \to V(G)$ such that for all $v \in V(S)$, $L(v) \subseteq L(\phi(v))$ and for all $uv \in E(S)$, $\phi(u)\phi(v) \in E(G)$.
Computational Problem

- **Conserved sub-network discovery**
  
  - **Instance:** A set of interaction networks \( \mathcal{G} = \{G_1 = (V_1, E_1, \mathcal{L}), G_2 = (V_2, E_2, \mathcal{L}), \ldots, G_m = (V_m, E_m, \mathcal{L})\} \), each belonging to a different organism, and a frequency threshold \( \sigma^* \).
  
  - **Problem:** Let \( H(S) = \{G_i : S \subseteq G_i\} \) be the occurrence set of graph \( S \). Find all connected subgraphs \( S \) such that \( |H(S)| \geq \sigma^* \), i.e., \( S \) is a frequent subgraph in \( \mathcal{G} \) and for all \( S' \subseteq S \), \( H(S') \neq H(S') \), i.e., \( S \) is maximal.
Algorithmic Insight: Ortholog Contraction

- Contract orthologous nodes into a single node
- No subgraph isomorphism
  - Graphs are uniquely identified by their edge sets
- Key observation: Frequent sub-networks are preserved ⇒ No information loss
  - Sub-networks that are frequent in general graphs are also frequent in their ortholog-contracted representation
  - Ortholog contraction is a powerful pruning heuristic
Example: Ortholog Contraction in Metabolic Pathways

- Directed hypergraph $\rightarrow$ uniquely-labeled directed graph
  - Nodes represent enzymes
  - Global labeling by enzyme nomenclature (EC numbers)
  - A directed edge from one enzyme to the other implies that the second consumes a product of the first
Example: Ortholog Contraction in Protein Interaction Networks

- Interaction between proteins $\rightarrow$ Interaction between ortholog groups or protein families
Preservation of Sub-networks

Theorem: Let $\tilde{G}$ be the ortholog-contracted graph obtained by contracting the orthologous nodes of network $G$. Then, if $S$ is a subgraph of $G$, $\tilde{S}$ is a subgraph of $\tilde{G}$.

Corollary: The ortholog-contracted representation of any frequent sub-network is also frequent in the set of ortholog-contracted graphs.
Simplifying the Graph Analysis Problem

- **Observation:** An ortholog-contracted graph is uniquely determined by the set of its edges.

  - Conserved Sub-network Discovery Problem $\rightarrow$ Frequent Edge set Discovery Problem

\[
\begin{align*}
F_1 &= \{ab, ac, de\} \\
F_2 &= \{ab, ac, bc, de, ea\} \\
F_3 &= \{ab, ac, bc, ea\} \\
F_4 &= \{ab, ce, de, ea\}
\end{align*}
\]
Extending Frequent Itemset Mining to Graph Analysis

- Given a set of transactions, find sets of items that are frequent in these transactions
  - Extensively studied in data mining literature

- Algorithms exploit **downward closure** property
  - An edge set is frequent only if all of its subsets are frequent
  - Generate edge sets (sub-networks) from small to large, **pruning** supersets of infrequent sets

- No **redundancy**

- No **subgraph enumeration**
**MULE: Analyzing Ortholog-Contracted Networks**

Sample run of MULE for identifying maximal sub-networks that are common to at least 3 organisms.
Case Study: Analyzing Protein Interaction Networks

- PPI networks for 9 eukaryotic organisms derived from BIND and DIP
  - # of proteins ranges from 288 (Arabidopsis) to 8577 (fruit fly)
  - # of interactions ranges from 340 (rice) to 28829 (fruit fly)

- Ortholog contraction
  - Group proteins according to existing COG ortholog clusters
  - Merge Homologene groups into COG clusters
  - Cluster remaining proteins via BLASTCLUST
  - Ortholog-contracted fruit fly network contains 11088 interactions between 2849 ortholog groups

- MULE is available at
  http://www.cs.purdue.edu/pdsl/
Conserved Protein Interaction Patterns

Small nuclear ribonucleoprotein complex \( (p < 2e^{-43}) \)
Conserved Protein Interaction Patterns

Actin-related protein Arp2/3 complex \( (p < 9e^{-11}) \)
Conserved Protein Interaction Patterns

Endosomal sorting \( (p < 1e^{-78}) \)
Runtime Characteristics

Comparison with isomorphism-based algorithms

**FSG** (Kuramochi & Karypis, *IEEE TKDE*, 2004), **gSpan** (Yan & Han, *KDD*, 2003)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Minimum Support (%)</th>
<th>Runtime (secs.)</th>
<th>Largest pattern</th>
<th>Number of patterns</th>
<th>Runtime (secs.)</th>
<th>Largest pattern</th>
<th>Number of patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FSG</td>
<td></td>
<td></td>
<td>MULE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamate</td>
<td>20</td>
<td>0.2</td>
<td>9</td>
<td>12</td>
<td>0.01</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.7</td>
<td>10</td>
<td>14</td>
<td>0.01</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>5.1</td>
<td>13</td>
<td>39</td>
<td>0.10</td>
<td>13</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>22.7</td>
<td>16</td>
<td>34</td>
<td>0.29</td>
<td>15</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>138.9</td>
<td>16</td>
<td>56</td>
<td>0.99</td>
<td>15</td>
<td>56</td>
</tr>
<tr>
<td>Alanine</td>
<td>24</td>
<td>0.1</td>
<td>8</td>
<td>11</td>
<td>0.01</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.5</td>
<td>11</td>
<td>15</td>
<td>0.02</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>4.0</td>
<td>12</td>
<td>21</td>
<td>0.06</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>112.7</td>
<td>17</td>
<td>25</td>
<td>1.06</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>215.1</td>
<td>17</td>
<td>34</td>
<td>1.72</td>
<td>16</td>
<td>34</td>
</tr>
</tbody>
</table>

**Extraction of contracted patterns**

<table>
<thead>
<tr>
<th>Glutamate metabolism, ( \sigma = 8% )</th>
<th>Alanine metabolism, ( \sigma = 10% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of contracted pattern</td>
<td>Extraction time (secs.)</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>FN</td>
<td>gSpan</td>
</tr>
<tr>
<td>15</td>
<td>10.8</td>
</tr>
<tr>
<td>14</td>
<td>12.8</td>
</tr>
<tr>
<td>13</td>
<td>1.7</td>
</tr>
<tr>
<td>12</td>
<td>0.9</td>
</tr>
<tr>
<td>11</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Total number of patterns: 56
Total runtime of FSG alone: 138.9 secs.
Total runtime of MULE+FSG: 0.99+100.5 secs.

Total number of patterns: 34
Total runtime of FSG alone: 215.1 secs.
Total runtime of MULE+FSG: 1.72+160.6 secs.
Total runtime of MULE+gSpan: 1.72+31.0 secs.
Alignment of Networks

• Given two networks, identify sub-networks that are similar to each other
  - Biological implications
  - Mathematical modeling

• Existing algorithms
  - PathBLAST aligns pathways (linear chains) to simplify the problem while maintaining biological meaning (Kelley et al., PNAS, 2004)
  - NetworkBLAST compares conserved complex model with null model to identify significantly conserved subnets (Sharan et al., J. Comp. Biol., 2005)

• Our approach (Koyutürk et al., RECOMB, 2005) (Koyutürk et al., J. Comp. Biol., 2006)
  - Guided by models of evolution
  - Scores evolutionary events
  - Identifies sets of proteins that induce high-scoring sub-network pairs
Many networks evolve through the process of Duplication/ and Divergence. 
- Interactions of duplicated nodes are also duplicated 
- Duplicated nodes rapidly lose interactions through mutations

Allows defining and scoring evolutionary events as graph-theoretical concepts

Duplication  Elimination  Emergence
Evolutionary events as graph-theoretic concepts

- A match $\in M$ corresponds to two pairs of homolog nodes from each network such that both pairs interact in both networks. A match is associated with score $\mu$.
- A mismatch $\in N$ corresponds to two pairs of homolog nodes from each graph such that only one pair is interacting. A mismatch is associated with penalty $\nu$.
- A duplication $\in D$ corresponds to a pair of homolog nodes that are in the same network. A duplication is associated with score $\delta$. 

```
G:     H:
  u_1  v_1
  u_2  
  u_3  v_3
  u_4  v_2
```
Scoring Matches, Mismatches and Duplications

- **Quantizing similarity between two nodes**
  - Confidence in two nodes being orthologous
  - E-value: \( S(u, v) = \log_{10} \frac{p(u,v)}{p_{\text{random}}} \)
  - Ortholog clustering: \( S(u, v) = c(u)c(v) \)

- **Match score**
  - \( \mu(uu', vv') = \bar{\mu} \min\{S(u, v), S(u', v')\} \)

- **Mismatch penalty**
  - \( \nu(uu', vv') = \bar{\nu} \min\{S(u, v), S(u', v')\} \)

- **Duplication score**
  - \( \delta(u, u') = \bar{\delta}(\hat{\delta} - S(u, u')) \)
  - \( \hat{\delta} \) specifies threshold for sequence similarity to be considered functionally conserved
Pairwise Alignment of Networks as an Optimization Problem

- **Alignment score:**
  \[ \sigma(A(P)) = \sum_{M \in M} \mu(M) - \sum_{N \in N} \nu(N) + \sum_{D \in D} \delta(D) \]
  - Matches are rewarded for conservation of interactions
  - Duplications are rewarded/penalized for functional conservation/differentiation after split
  - Mismatches are penalized for functional divergence (what about experimental error?)

- Scores are functions of similarity between associated proteins

- **Problem:** Find all protein subset pairs with significant alignment score
  - High scoring protein subsets are likely to correspond to conserved modules

- A graph equivalent to BLAST
Weighted Alignment Graph

- \( G(V, E) : V \) consists of all pairs of homolog proteins \( v = \{ u \in U, v \in V \} \)

- An edge \( vv' = \{uv\}\{u'v'\} \) in \( E \) is a
  - match edge if \( uu' \in E \) and \( vv' \in V \), with weight \( w(vv') = \mu(uv, u'v') \)
  - mismatch edge if \( uu' \in E \) and \( vv' \notin V \) or vice versa, with weight \( w(vv') = -\nu(uv, u'v') \)
  - duplication edge if \( S(u, u') > 0 \) or \( S(v, v') > 0 \), with weight \( w(vv') = \delta(u, u') \) or \( w(vv') = \delta(v, v') \)
Maximum Weight Induced Subgraph Problem

• **Definition**: \((\text{MAWISH})\)
  - Given graph \(G(\mathcal{V}, \mathcal{E})\) and a constant \(\epsilon\), find \(\tilde{\mathcal{V}} \subseteq \mathcal{V}\) such that
  \[
  \sum_{v,u \in \tilde{\mathcal{V}}} w(vu) \geq \epsilon.
  \]
  - NP-complete by reduction from Maximum-Clique

• **Theorem**: \((\text{MAWISH} \equiv \text{Pairwise alignment})\)
  - If \(\tilde{\mathcal{V}}\) is a solution for the MAWISH problem on \(G(\mathcal{V}, \mathcal{E})\), then \(P = \{\tilde{U}, \tilde{V}\}\) induces an alignment \(A(P)\) with \(\sigma(A) \geq \epsilon\), where \(\tilde{\mathcal{V}} = \tilde{U} \times \tilde{V}\).

• **Solution**: **Local** graph expansion
  - Greedy graph growing + iterative refinement
  - Linear-time heuristic

• **Source code available at**
### Case Study: Alignment of Yeast and Fruit Fly Networks

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>z-score</th>
<th># Proteins</th>
<th># Matches</th>
<th># Mismatches</th>
<th># Dups.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.97</td>
<td>6.6</td>
<td>18 (16, 5)</td>
<td>28</td>
<td>6</td>
<td>(4, 0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>protein amino acid phosphorylation (69%)</td>
<td>JAK-STAT cascade (40%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13.93</td>
<td>3.7</td>
<td>13 (8, 7)</td>
<td>25</td>
<td>7</td>
<td>(3, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>endocytosis (50%) / calcium-mediated signaling (50%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8.22</td>
<td>13.5</td>
<td>9 (5, 3)</td>
<td>19</td>
<td>11</td>
<td>(1, 0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>invasive growth (sensu Saccharomyces) (100%)</td>
<td>oxygen and reactive oxygen species metabolism (33%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8.05</td>
<td>7.6</td>
<td>8 (5, 3)</td>
<td>12</td>
<td>2</td>
<td>(0, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ubiquitin-dependent protein catabolism (100%)</td>
<td>mitosis (67%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>4.36</td>
<td>6.2</td>
<td>9 (5, 4)</td>
<td>18</td>
<td>13</td>
<td>(0, 5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cytokinesis (100%, 50%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>3.76</td>
<td>39.6</td>
<td>6 (3, 5)</td>
<td>5</td>
<td>1</td>
<td>(0, 6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DNA replication initiation (100%, 80%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Subnets Conserved in Yeast and Fruit Fly

Proteosome regulatory particle subnet

Calcium-dependent stress-activated signaling pathway
Analytical Assessment of Statistical Significance

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

Existing techniques
- Mostly computational (e.g., Monte-Carlo simulations)
- Compute probability that the pattern exists rather than a pattern with the property (e.g., size, density) exists
- Overestimation of significance
Random Graph Models

- Interaction networks generally exhibit power-law property (or exponential, geometric, etc.)


- 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 $d_{\text{max}}^2 > |E|$ for PPI networks

- Analytical techniques based on simplified models (Koyutürk, Grama, Szpankowski, RECOMB, 2006)

  - Rigorous analysis on $G(n, p)$ model
  - Extension to piecewise $G(n, p)$ to capture network characteristics more accurately
Significance of Dense Subgraphs

- A subnet of \( r \) proteins is said to be \( \rho \)-dense if \( F(r) \geq \rho r^2 \), where \( F(r) \) is the number of interactions between these \( r \) proteins.

- 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!

- \( G(n, p) \) model
  - \( n \) proteins, each interaction occurs with probability \( p \)
  - Simple enough to facilitate rigorous analysis
  - If we let \( p = d_{\text{max}}/n \), largest \( \rho \)-dense subgraph in \( G(n, p) \) stochastically dominates that in a graph with arbitrary degree distribution.

- Piecewise \( G(n, p) \) model
  - Few proteins with many interacting partners, many proteins with few interacting partners
  - Captures the basic characteristics of PPI networks
  - Analysis of \( G(n, p) \) model immediately generalized to this model.
Largest Dense Subgraph

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

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

where

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

More precisely,

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

where

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

for large $n$. 
**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 $q_l = 1 - p_l$. 
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

- **SiDES**: Use statistical significance to determine whether a subgraph is sufficiently dense
  - For given number of proteins and interactions between them, we can determine whether those proteins induce a significantly dense subnet
SiDES Algorithm

SiDES is available at [http://www.cs.purdue.edu/pdsl](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>SiDES</th>
<th></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>
<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>
<td></td>
</tr>
</tbody>
</table>

Comparison of SiDES with MCODE *(Bader & Hogue, BMC Bioinformatics, 2003)*
Performance of SiDES

Size vs Significance

Correlation

SiDES: 0.76
MCODE: 0.43
Performance of SIDES

Size vs Specificity

Size vs Sensitivity

Correlation

SIDES: 0.22
MCODE: -0.02

SIDES: 0.27
MCODE: 0.36
Functional Annotation of Pathways: From Molecules to Systems

- Annotation is at the node level
- Map networks to function space (can generate a library of annotated modular (sub-) networks)

Network of Gene Ontology terms based on significance of pairwise interactions in yeast synthetic gene array (SGA) (Tong et al., Science, 2004)
Narada Functionality
Narada Network Annotation