Assessing Significance of Connectivity and Conservation in Protein Interaction Networks

Mehmet Koyutürk, Ananth Grama, Wojciech Szpankowski
Department of Computer Science, Purdue University

April 2, 2006

Acknowledgment: Funding for this work was provided by the National Institutes of Health Grant # R01 GM068959-01.

Outline

- Motivation
  - Connectivity, conservation, and modularity
  - Algorithmic approaches
- Existing techniques for assessing statistical significance
- Our approach: Analytical assessment of statistical significance
  - Largest dense subgraph problem
  - $G(n, p)$ model
  - Piecewise $G(n, p)$ model
  - Plugging significance into algorithms: Modified HCS algorithm
- Results
- Conclusion & Ongoing work

Connectivity and Conservation in Biological Networks

- Modular processes are likely to manifest themselves in terms of dense interactions in a particular network and conservation of these interactions across networks (Tornow and Mewes, *NAR*, 2003)
  - Many algorithmic approaches have successfully provided novel biological insights based on connectivity and conservation
- Identification of topological modules
  - MCODE (Bader and Hogue, *BMC Bioinformatics*, 2003)
  - TribeMCL (Pereira-Leal et al., *Proteins*, 2004)
  - Kernel-based clustering (Brun et al., *BMC Bioinformatics*, 2004)
- Identification of conserved subgraphs
  - MULE (Koyutürk et al., *ISMB*, 2004), MaWIsH (Koyutürk et al., *JCB*, 2006)
  - PathBlast (Kelley et al., *PNAS*, 2003), NetworkBlast (Sharan et al., *PNAS*, 2004)
  - CODENSE (Hu et al., *ISMB*, 2005)
  - NUKE (Novak et al., *Genome Informatics*, 2005)

Statistical Significance: 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
  - Correction for multiple testing + large space $\Rightarrow$ Underestimation
- Alternate analytical approaches
  - Sharan et al. (*RECOMB*, 2004) compare likelihood of null and conserved complex models
  - Itzkovitz et al. (*Physical Review*, 2003) estimate expectation analytically for given topological motifs
  - Our approach: Asymptotic analysis of behavior of largest pattern for general, but interesting pattern topology

Largest Dense Subgraph Problem

- A subnet of r proteins is said to be ρ-dense if $F(r) \geq \rho r^2$, where $F(r)$ is the number of interactions between these r proteins
  - Maximum-clique is a special case with $\rho = 1$
- Largest ρ-dense subgraph: $R_\rho = \max_{U \subseteq V(G): \delta(U) \geq \rho} |U|$
- What is the typical size of the largest ρ-dense subgraph in a random graph?
  - Any ρ-dense subgraph with larger size is statistically significant!
  - Typical size of maximum clique is $O(\log_{1/p} n)$ (Bollobás, *Random Graphs*, 2001)

Modeling Biological Networks

- Interaction networks generally exhibit power-law property (or exponential, geometric, etc.)
- Analysis simplified through independence assumption (Itzkovitz et al., *Physical Review*, 2003)
- 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
  - Models multi-graphs accurately rather than simple graphs
- Analytical techniques based on simplified models
  - Rigorous analysis on $G(n, p)$ model
  - It can be easily shown that, setting $p = d_{\max}/n$, largest dense subgraph in $G(n, p)$ stochastically dominates that for arbitrary degree distribution
  - Extension to piecewise $G(n, p)$ to capture network characteristics more accurately

Largest Dense Subgraph on $G(n,p)$

Theorem 1. *If G is a random graph with n vertices, where every edge exists with probability p, then*

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

where

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

Here, $H_p(\rho)$ denotes weighted entropy. More precisely,

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

where

$$r_0 = \frac{\log n - \log\log n + \log\kappa(p,\rho) - \log e + 1}{\kappa(p,\rho)}$$

for large n.

Proof

- $X_{r,\rho}$: number of subgraphs of size r with density at least ρ
  - $X_{r,\rho} = |\{U \subseteq V(G) : |U| = r \wedge |F(U)| \geq \rho r^2\}|$
- Y_r : number of edges induced by a set U of r vertices
  - $\mathbf{E}[X_r] = \binom{n}{r} P(Y_r \geq \rho r^2)$
  - $P(Y_r \geq \rho r^2)\binom{r^2}{\rho r^2}(r^2 - \rho r^2)p^{\rho r^2}(1-p)^{r^2-\rho r^2}$
- Upper bound: From first moment method, we have $P(R_\rho \geq r) \leq P(X_{r,\rho} \geq 1) \leq \mathbf{E}[X_{r,\rho}]$
  - Plug in Stirling's formula for appropriate regimes
- Lower bound: To use second moment method, we have to account for dependencies in terms of nodes and existing edges
  - Use Stirling's formula, plug in continuous variables for range of dependencies

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

- Random graph G with node set $V(G)$ that is composed of two disjoint subsets $V_h \subset V(G)$ and $V_l = V(G) \setminus V_h$, where $n_h = |V_h| \ll |V_l| = n_l$ and $n_h + n_l = n = |V(G)|$

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

  - Here, $p_l < p_b < p_h$.

Largest Dense Subgraph on Piecewise $G(n,p)$

Theorem 2. *Let G be a random graph with piecewise degree distribution. If $n_h = O(1)$, then*

$$P(R_\rho \geq r_1) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right),$$

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 $q_l = 1 - p_l$.

Note: For power-law graphs, $n_h = \sum_{d=(n/\zeta(\gamma))^{1/\gamma}}^{\infty} nd^{-\gamma}/\zeta(\gamma)$ is bounded, provided the series converges.

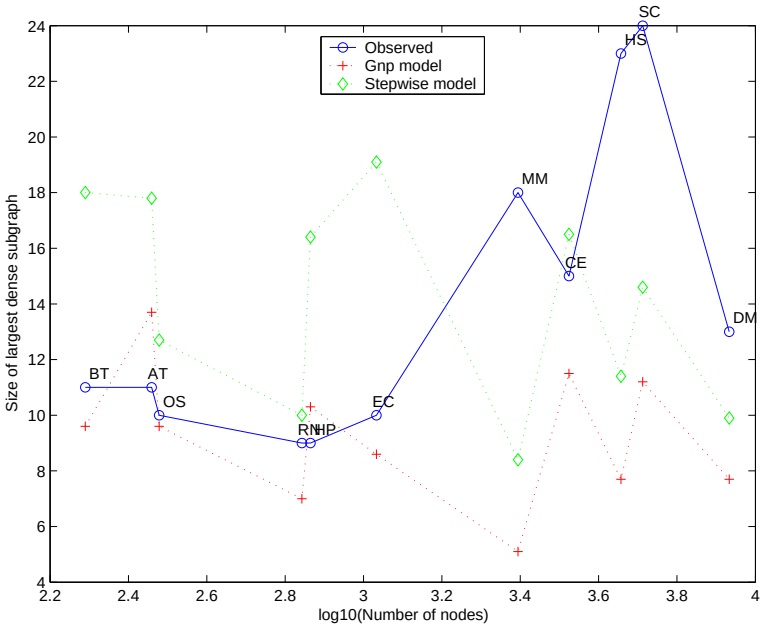
Proof

- Graph can be divided into three disjoint graphs
  - $G = G_l \cup G_h \cup G_b$
  - G_l and G_h are $G(n_h, p_h)$ and $G(n_l, p_l)$, respectively
  - G_b, is a random bipartite graph with node sets V_l, V_h, where each edge occurs with probability p_b
- $E(G) = E(G_l) \cup E(G_h) \cup E(G_b)$
- We emphasize on $\mathbf{E}[X^b_{r,\rho}]$
  - Using $2n_h r << \rho r^2$, we obtain
    $\mathbf{E}[X^b_{r,\rho}] = \mathbf{E}[X^l_{r,\rho}] \sum_{l=0}^{2n_h r} \binom{2n_h r}{l} \left(\frac{p_b q_l}{p_l}\right)^l q_b^{2n_h r - l}$
- Summation term contributes an additive factor of $2n_h \log B$ to the exponent, which is less than $\log n$ for large n

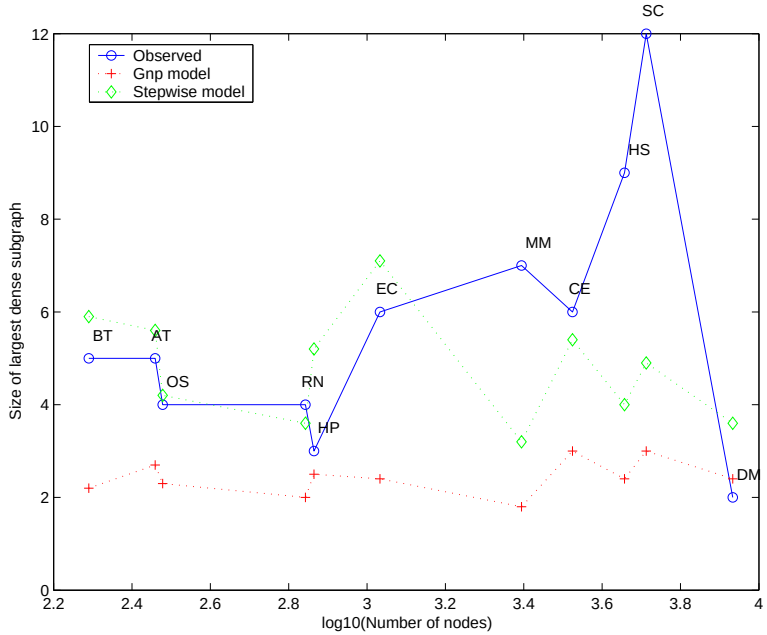
Algorithms Based on Statistical Significance

- Identification of topological modules
- Use statistical significance as a stopping criterion for graph clustering heuristics
- HCS Algorithm (Hartuv & Shamir, *Inf. Proc. Let.*, 2000)
  - 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
- 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

Behavior of Largest Dense Subgraph Across Species



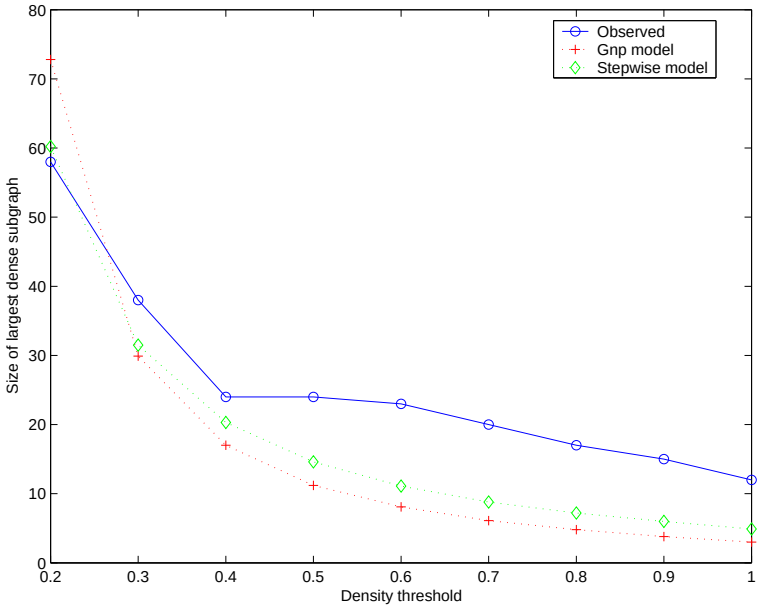
$\rho = 0.5$



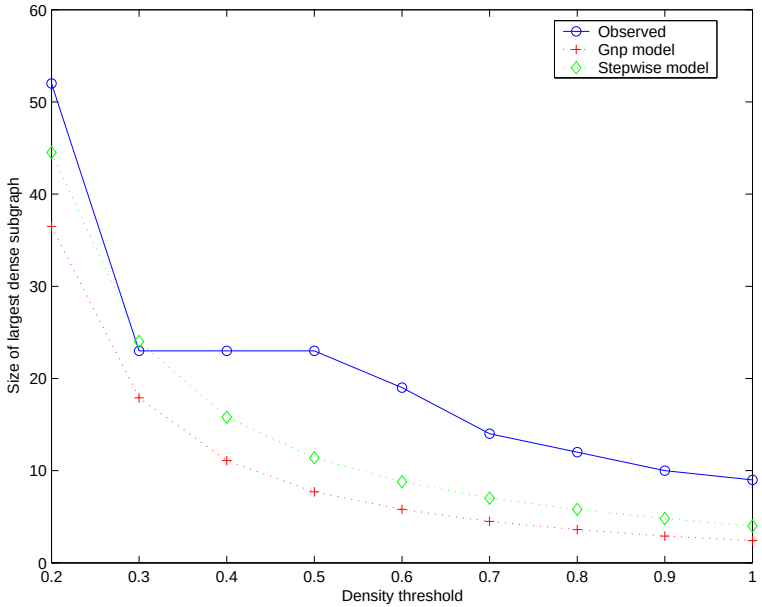
$\rho = 1.0$

Number of nodes vs. Size of largest dense subgraph
for PPI networks belonging to 9 Eukaryotic species

Behavior of Largest Dense Subgraph w.r.t Density



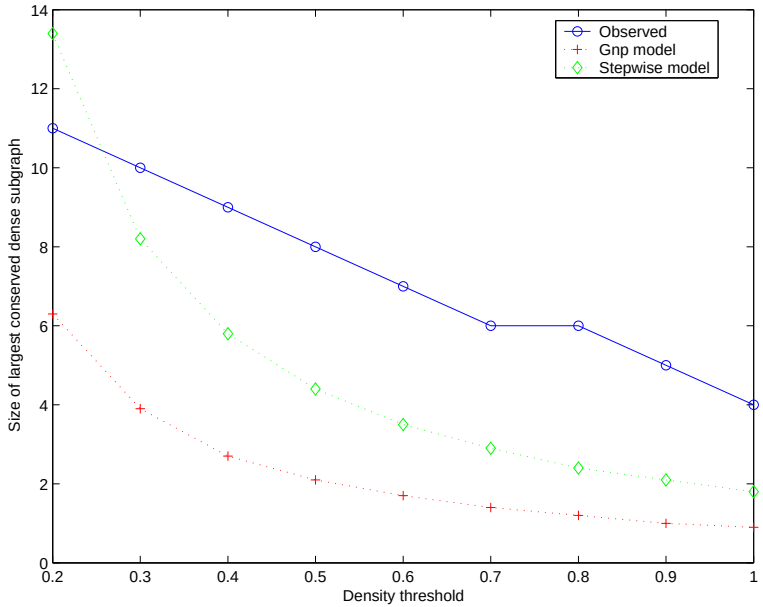
S. cerevisiae



H. sapiens

Density threshold vs. Size of largest dense subgraph
for Yeast and Human PPI networks

Behavior of Largest Dense Subgraph w.r.t Density



S. cerevisiae & H. sapiens

Density threshold vs. Size of largest conserved subgraph
for Yeast and Human PPI networks

Significantly Connected Subnets on Yeast

# Prot	# Int	$p <$	GO Annotation
24	165	10^{-175}	(C) nucleolus (54%, $p < 10^{-7}$)
20	138	10^{-187}	(P) ubiquitin-dep prot catabolism (80%, $p < 10^{-21}$) (F) endopeptidase activity (50%, $p < 10^{-11}$) (C) proteasome reg part, lid subcomp (40%, $p < 10^{-12}$)
16	104	10^{-174}	(P) histone acetylation (62%, $p < 10^{-15}$) (C) SAGA complex (56%, $p < 10^{-15}$) (P) chromatin modification (56%, $p < 10^{-14}$)
15	90	10^{-145}	(F) RNA binding (80%, $p < 10^{-12}$) (C) mRNA cleav & polyadenyl SFC (80%, $p < 10^{-24}$) (P) mRNA polyadenylylation (80%, $p < 10^{-21}$)
14	79	10^{-128}	(P) mRNA catabolism (71%, $p < 10^{-16}$) (F) RNA binding (64%, $p < 10^{-6}$) (P) nuclear mRNA splicing (57%, $p < 10^{-7}$)
10	45	10^{-200}	(P) ER to Golgi transport (90%, $p < 10^{-14}$) (C) TRAPP complex(90%, $p < 10^{-23}$)
7	20	10^{-30}	(C) mitochondrial OMTC (100%, $p < 10^{-20}$) (F) protein transporter activity (100%, $p < 10^{-14}$) (P) mitochondrial matrix prot import (100%, $p < 10^{-16}$)

Significantly Conserved Subnets on Yeast & Human

# Prot	# Cons Int	$p <$	COG Annotation
10	17	10^{-68}	RNA polymerase (100%)
11	11	10^{-26}	Mismatch repair (33%) RNA polym II TI/nucleo excision repair fac TFIIH (33%) Replication factor C (22%),
7	7	10^{-25}	Exosomal 3'-5' exoribonuclease complex (86%)
4	4	10^{-24}	Single-stranded DNA-binding repl prot A (50%) DNA repair protein (50%)
5	4	10^{-12}	Small nuclear ribonucleoprotein(80%) snRNP component (20%)
5	4	10^{-12}	Histone (40%) Histone transcription regulator (20%) Histone chaperone (20%)
3	3	10^{-9}	Vacuolar sorting protein (33%) RNA polymerase II TFC subunit (33%) Uncharacterized conserved protein (33%)

Conclusion & Ongoing Work

- Asymptotic analysis provides clear understanding of the behavior of interesting patterns
- These results provide solid bases for evaluating statistical significance
- More complicated models are necessary for distinguishing nature of network from biological significance
  - Can we approach more realistic skewed degree distributions by increasing the number of pieces?
  - How can we handle dependence?
  - What about growth models?