



# An Efficient Computational Method to Identify a Protein Community from a Seed Protein

---

Daniel Wu & Xiaohua Hu

College of Information Science & Technology

Drexel University





# Outline

---

- Introduction
- Related work
- The algorithm - CommBuilder
- Experimental results
- Conclusions & discussions
- Q&A



# Introduction

---

- Graph Theory
  - A mathematical formalism
  - Topology
    - Global properties: diameter, clustering, degree distribution, and community structure
    - Local properties: motif and graphlet
  - Models – random, small-world, scale-free, random geometric
  - Modeling real-world phenomena



# Introduction

---

- Biological Networks
  - Modeling biological systems
  - Genetic networks
    - Gene association and expression
  - Protein networks
    - Protein structure and interactions
  - Metabolic pathways
  - Challenging
    - Non-trivial and irregular
    - Incomplete and noisy



# Introduction

---

- Protein–Protein Interaction (PPI) Networks
  - Proteins are executor of genetic program and rarely act alone
  - PPI models
    - Scale-free (Barabasi & Albert, 1999)
    - Geometric random (Przulj et al, 2004)
  - Tolerant to random errors and Fragile against the removal of the most connected nodes
  - Modularity and community structure



# Introduction

---

- Community structure in networks
  - Gathering of vertices into groups such that the connections within groups are denser than between groups (Girvan & Newman, 2002)
  - An important property of PPI networks
    - Delineation of functional groups/processes
    - Transfer of information



# Introduction

---

- What is the community to which a given protein (or proteins) belongs?
  - Desirable and computationally more feasible to study a community containing one or a few proteins of interest



# Related Work

---

- Finding communities
  - The GN algorithm (Girvan & Newman, 2002)
    - Based on betweenness
    - High computational cost
    - Well adopted
      - Metabolic networks (Holme et al, 2003)
        - Functional units
      - Gene networks (Wilkinson & Huberman, 2004)
        - Related genes



# Related Work

---

- Finding complexes and modules
  - MCODE (Bader & Hogue, 2003)
    - Vertex weighting based on local density
  - Monte Carlo simulation (Spirin & Mirny, 2003)
  - Clustering purifications (Krause et al, 2003)
  - Bayesian approach (Sharan et al, 2004)
  - Spectral-based clustering (Bu et al, 2003, Hu et al, 2004, and Hu et al 2005)
  - Probabilistic network reliability (Asthana et al, 2004)



# Related Work

---

- Growing network from seeds
  - Genetic regulatory networks (Hashimoto et al, 2004)
    - Based on probabilistic Boolean networks
  - Web (Flake et al, 2002)
    - Based on the self-organization of the network structure and a maximum flow method



# The Algorithm

---

## □ Notation

- PPI networks are modeled as undirected, unweighted, and simple graph
- An undirected graph,  $G(V, E)$ , has two sets, vertices  $V$  and edges  $E$
- In-community degree for vertex  $i$ ,  $k_{in}(i, G')$ : number of edges connecting vertex  $i$  to other vertices within the same subgraph  $G'$
- Out-community degree,  $k_{out}(i, G')$ : number of edges connecting vertex  $i$  to other vertices outside subgraph  $G'$ .



# The Algorithm

---

## □ Notation

A subgraph  $G'$  is a community

- In a strong sense if  $k_{in}(i, G') > k_{out}(i, G')$  for each vertex  $i$  in  $G'$
- In a weak sense if the sum of all degrees within  $G'$  is greater than the sum of all degrees from  $G'$  to the rest of the graph



# The Algorithm

---

---

## Algorithm 1 CommBuilder( $G, s, f$ )

---

- 1:  $G(V, E)$  is the input graph with vertex set  $V$  and edge set  $E$ .
  - 2:  $s$  is the seed vertex,  $f$  is the affinity threshold.
  - 3:  $N \leftarrow \{\text{Adjacency list of } s\} \cup \{s\}$
  - 4:  $C \leftarrow \text{FindCore}(N)$
  - 5:  $C' \leftarrow \text{ExpandCore}(C, f)$
  - 6: **return**  $C'$
-



# The Algorithm

---

---

7: *FindCore*( $N$ )

---

8: **for each**  $v \in N$

9:     calculate  $k_v^{in}(N)$

10: **end for**

11:  $Kmin \leftarrow \min \{ k_v^{in}(N), v \in N \}$

12:  $Kmax \leftarrow \max \{ k_v^{in}(N), v \in N \}$

13: **if**  $Kmin = Kmax$  **then return**  $N$

14: **else return** *FindCore*( $N - \{v\}, k_v^{in}(N) = Kmin$ )

---



# The Algorithm

---

15: *ExpandCore*( $C, f$ )

---

16:  $D \leftarrow \bigcup_{(v,w) \in E, v \in C, w \notin C} \{v, w\}$

17:  $C' \leftarrow C$

18: **for each**  $t \in D$  **and**  $t \notin C$

19:     calculate  $k_t^{in}(D)$

20:     calculate  $k_t^{out}(D)$

21:     **if**  $k_t^{in}(D) > k_t^{out}(D)$  **or**  $k_t^{in}(D)/|D| > f$  **then**

$C' \leftarrow C' \cup \{t\}$

22:     **end for**

23: **if**  $C' = C$  **then return**  $C$

---



# Experimental Results

---

- Data sets
  - Yeast PPI from the General Repository for Interaction Datasets (GRID)
    - 4,907 proteins
    - 17,598 interactions
- Average run time for finding a community of 50 members from the network: 20 ms



# Experimental Results

---

- The SAGA/SRB community
  - Seed: TAF6
  - 39 members
  - 14/16 SAGA and 14/21 SRB (MIPS)
  - 27/50 (Asthana et al, 2004)
  - All members are related to transcription regulation



# Experimental Results

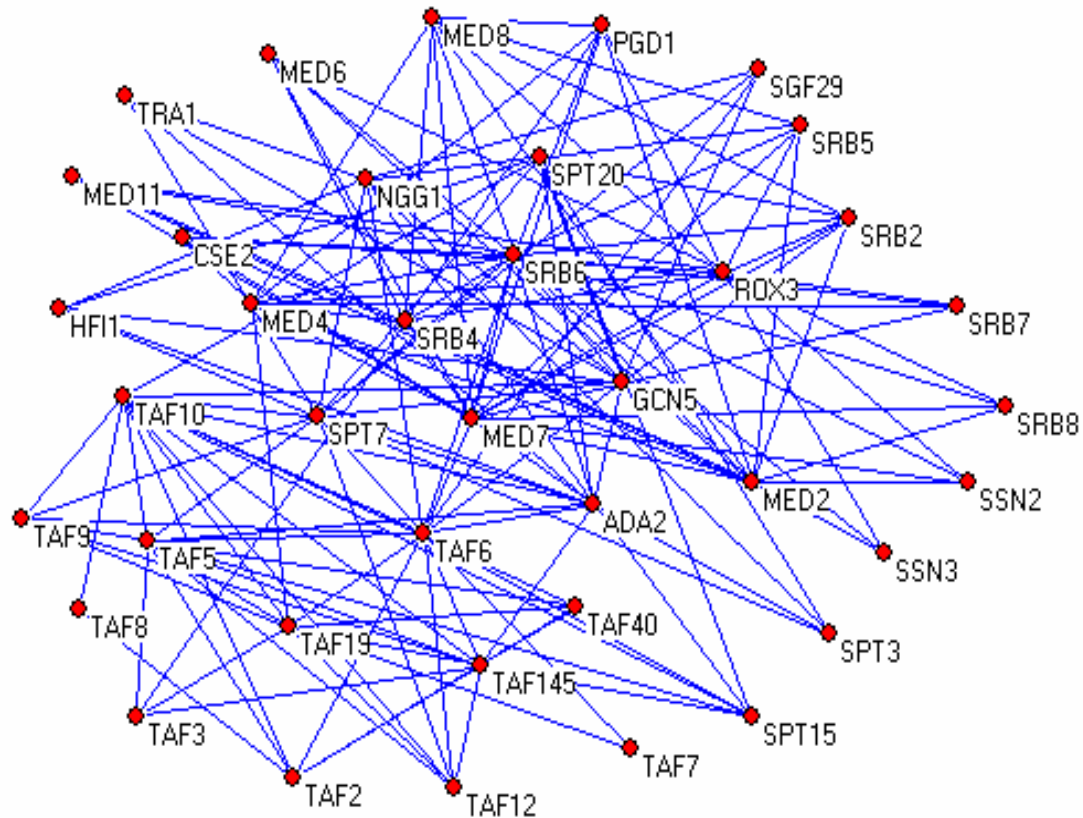


Fig. 1. The SAGA-SRB community.



# Experimental Results

---

- The CCR4-NOT community
  - Seed: NOT3
  - 40 members
  - 5/5 NOT and 11/13 CCR4 (MIPS)
  - 27/50 (Asthana et al, 2004)
  - Related to transcription and DNA/chromatin maintenance



# Experimental Results

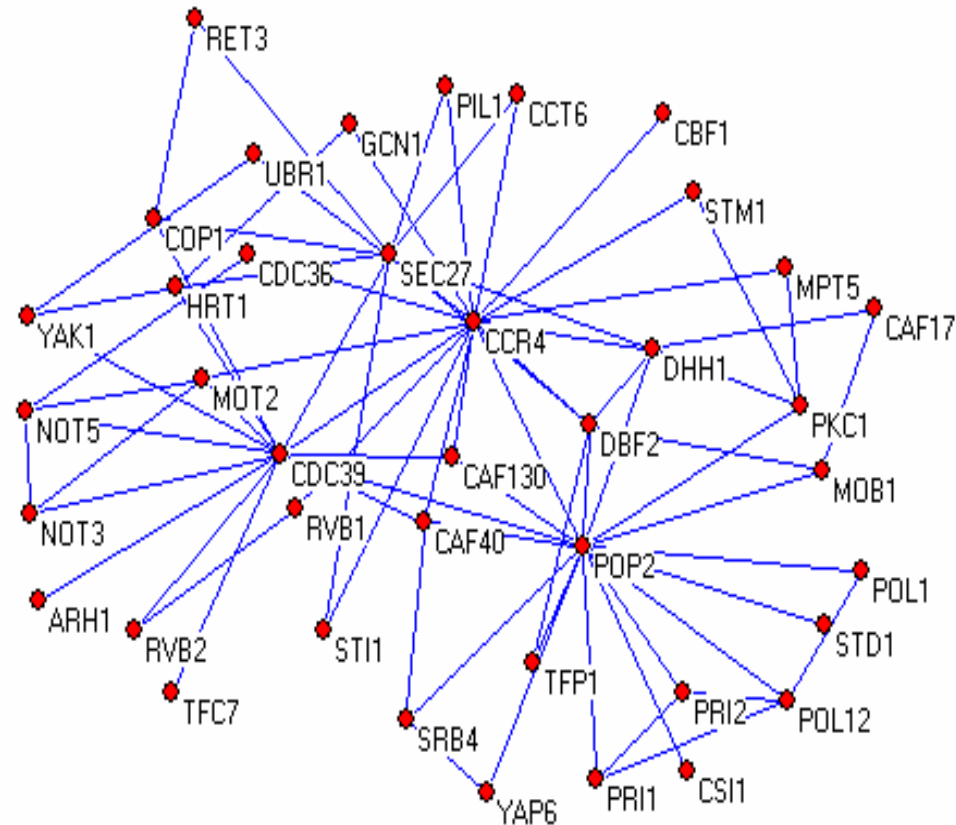


Fig. 2. The CCR4-NOT Community.



# Experimental Results

---

- The RFC community
  - Seed: RFC2
  - 17 members
  - 5/5 RFC (MIPS)
  - 16/17 in the functional category of DNA recombination/repair and cell cycle checkpoints (MIPS)
  - 9/50 (8/8) (Asthana et al, 2004)



# Experimental Results

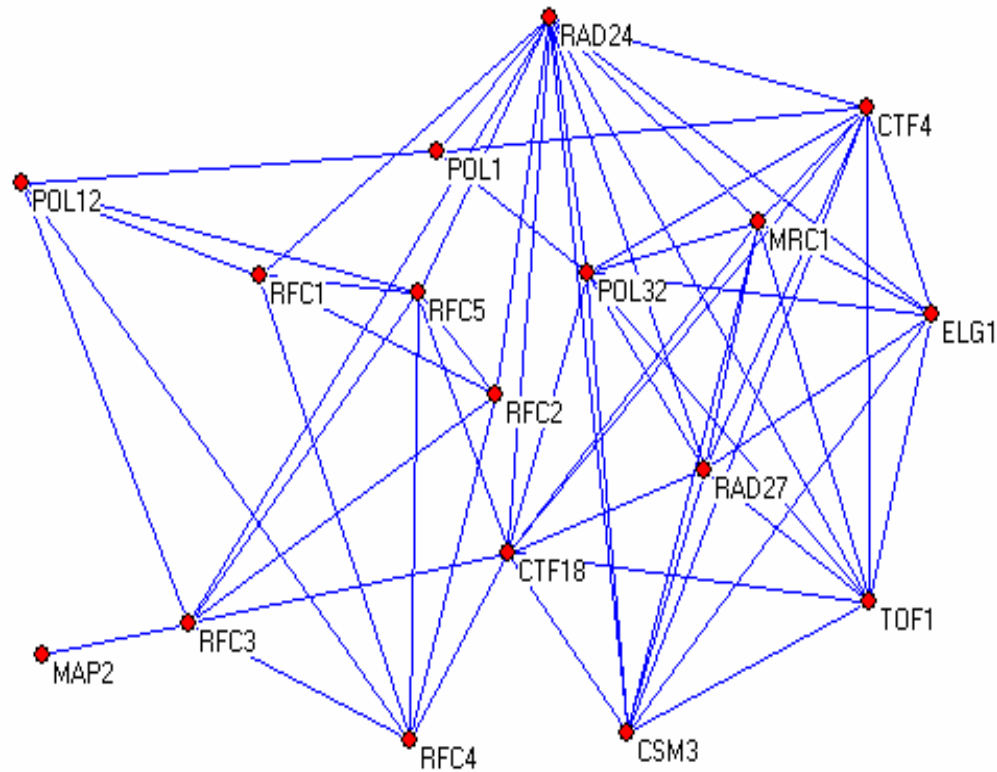


Fig. 3. The RFC Community.

Wu & Hu (2005)



# Experimental Results

---

- The Arp2/3 community
  - Seed: ARP3
  - 20 members
  - 7/7 Arp2/3 (MIPS)
  - 14 in the functional category of budding, cell polarity, and filament formation (MIPS)
  - Top 6 (Asthana et al, 2004)



# Experimental Results

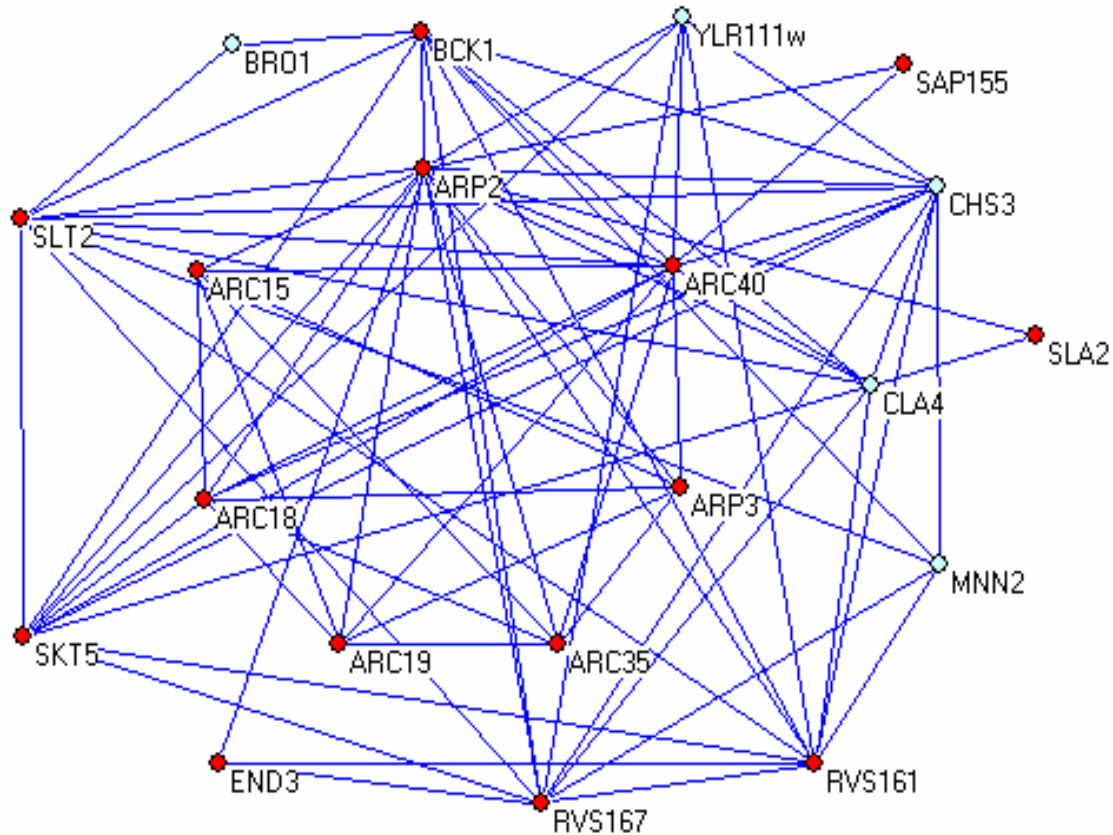


Fig. 4. The ARP2/ARP3 Community.

Wu & Hu (2005)



# Conclusions & Discussions

---

- ❑ Efficiently build a community from a seed protein
- ❑ The communities built reveal strong structural and functional relationships among member proteins



# References

---

- ❑ Barabasi, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science* 286:509-512.
- ❑ Przulj, N., et al (2004) Modeling interactome: scale-free or geometric? *Bioinformatics* 20(18): 3508-3515
- ❑ Girvan, M. & Newman, M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99: 7821-7826.
- ❑ Donetti, L. and Munoz, M.A. (2004). Detecting Network Communities: a new systematic and efficient algorithm. *J. Stat. Mech.* P10012.
- ❑ White, S. and Smyth, P. (2005). A Spectral Clustering Approach to Finding Communities in Graphs. *SIAM International Conference on Data Mining 2005*, Newport Beach, CA, USA.
- ❑ Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 19(4): 532-538.
- ❑ Wilkinson, D. and Huberman, B.A (2004). A Method for Finding Communities of Related Genes. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl 1): 5241-5248.
- ❑ Hashimoto, R.F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M.L., and Dougherty, E.R. (2004). Growing genetic regulatory networks from seed genes. *Bioinformatics* 20(8): 1241–1247.
- ❑ Flake, G. W., Lawrence, S. R., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of Web communities, *IEEE Computer* 35: 66-71.
- ❑ Jansen, R., Lan, N., Qian, J., and Gerstein, M. (2002). Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Functional Genomics* 2: 71–81.
- ❑ Bader, G.D. and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4: 2.
- ❑ Asthana, S., King, O.D., Gibbons, F.D., and Roth, F.P. (2004). Predicting Protein Complex Membership Using Probabilistic Network Reliability. *Genome Res.* 14: 1170-1175