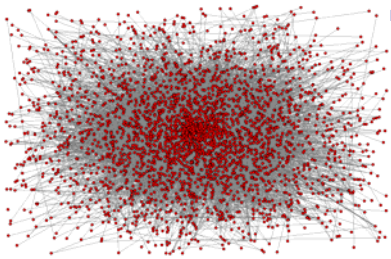
Protein interaction networks

- Large scale (genome wide networks):



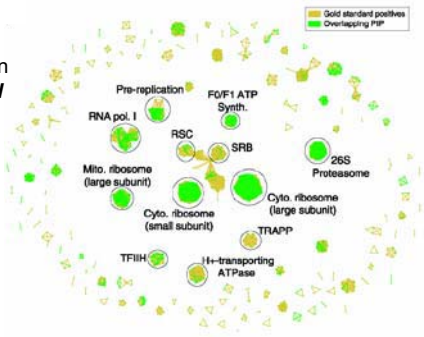
ProNet (Asthana et al.)
Yeast
3,112 nodes
12,594 edges

Analyzing Protein Networks

- Predict members of a partially known protein complex/pathway.
- Infer individual genes' functions on the basis of linked neighbors.
- Find strongly connected components, clusters to reveal unknown complexes.
- Find the best interaction path between a source and a target gene.

Simple analysis

The network can be ***thresholded*** to reveal clusters of interacting proteins



Complex/Pathway membership problem

- E.g.,
  - *C. elegans* cell death (apoptosis) pathway
  - Identified ~50 genes involved in the pathway.
  - Are there other genes involved in the pathway? Biologists would like to know:
    - Which genes (out of ~15K genes) should be tested in the RNAi screens next?

Complex/pathway membership problem

- Given a a set of proteins identified as the core complex (query), rank the remaining proteins in the network according to the probability that they "connect" to the core complex.
- This problem is very similar to the "network reliability" problem in communication networks.

Network reliability

- Two terminal network reliability problem:
  - Given a graph of connections between terminals:
    - Each connection weighted by the probability that the corresponding wire is functioning at a given time
  - What is the probability that some path of functioning wires connects two terminals at a given time?

Exact solution: NP-hard
Several approximation methods exist

Monte Carlo simulation

- Monte Carlo simulation (ProNet: Asthana *et al.* 2004)
  - Create a sample of **N** binary networks from the probabilistic network (according to a Bernoulli trial on each edge based on its probability).
- Use breadth-first search to determine the existence of a path between the nodes (i.e., the two terminals).
- The fraction of sampled networks in which there exists a path between the two nodes is an approximation to the exact network reliability.

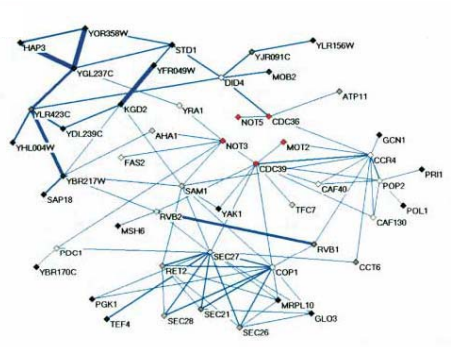
Parameters

- Number of binary networks (samples) to be sampled from the probabilistic network
  - 1000, 5000, 10000 ?
- The depth of the breadth-first search: complexity increases as you search for the existence of a path to a distant node.
  - 4, 10, 20 ?

ProNet

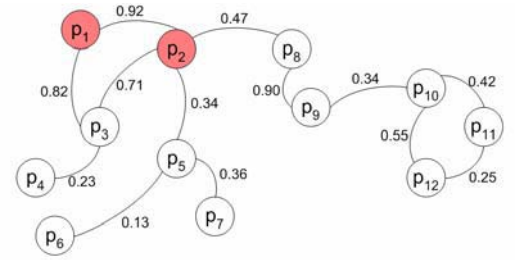
- Generate 10,000 binary networks from a probabilistic network (according to a Bernoulli trial on each edge based on its probability)
- Use breadth-first search to determine the existence of a path between two nodes
  - Limit the maximum depth to 4 to reduce computation
- For each protein *i* in the network, count the fraction C_i of sampled networks in which there exists a path between *i* and the core complex.
- Report proteins ranked by C_i

ProNet: example



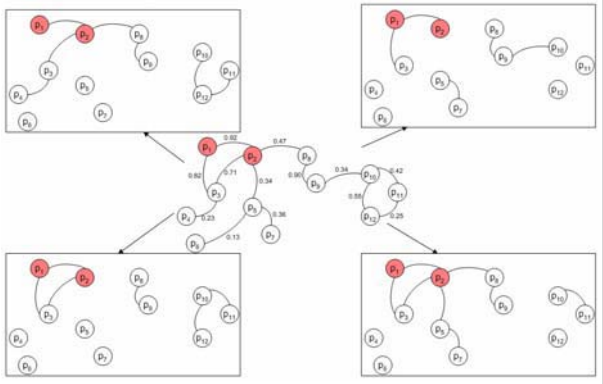
Example

- Complex nodes: p_1 and p_2



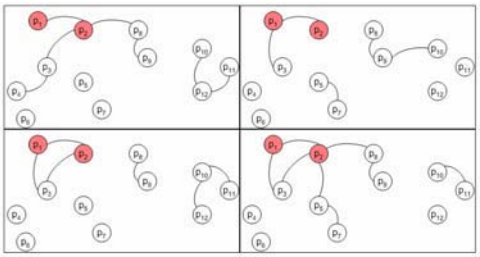
Example

- Sample size: 4, maximum search depth: 3



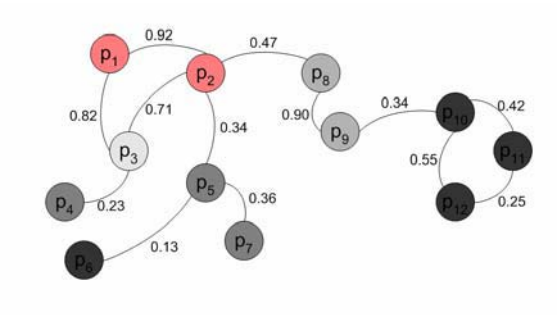
Example

- Sample size: 4, maximum search depth: 3

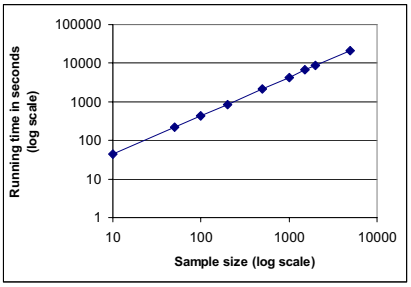


C_{p3} = 4/4 = 1.0
C_{p4} = 1/4 = 0.25
C_{p5} = 1/4 = 0.25
C_{p6} = 0/4 = 0.0
C_{p7} = 1/4 = 0.25
C_{p8} = 2/4 = 0.5
C_{p9} = 2/4 = 0.5
C_{p10} = 0/4 = 0.0
C_{p11} = 0/4 = 0.0
C_{p12} = 0/4 = 0.0

Results



Running time vs. sample size



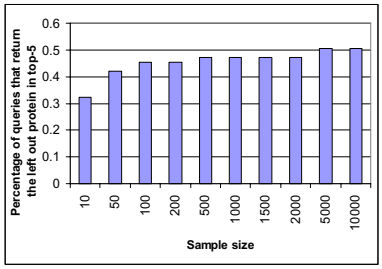
What about accuracy of the technique? Is it able to give a good ranking for the nodes of the network, based on their closeness to the core?

Leave-one-out benchmark

- Use known complexes to evaluate the accuracy of the method
- Leave one member (in turn) from each complex/pathway.
- Use the rest of the complex/pathway as the starting, i.e., query, set.
- Examine the rank of the left-out protein.
  - What do we expect from a good technique?

Accuracy vs. sample size

- How does the sample size effect returned results?



Monte Carlo simulation

- Disadvantages:
  - What is the best choice for the number of samples?
  - What should be the maximum depth for breadth-first search? (Need a cutoff to decrease running time)
  - Scalability issues: May need a lot of computation time for large networks

Assignment #4

- Implement the “Network Reliability by Monte Carlo Sampling” technique
- Run your program on the two test networks (small network - for debugging - and the yeast network)
- Sample size 5000, search depth 4
- Perform leave one out experiments for the 3 yeast complexes given and report “the percentage of the experiments in which the left-out protein ranks in top-5.”