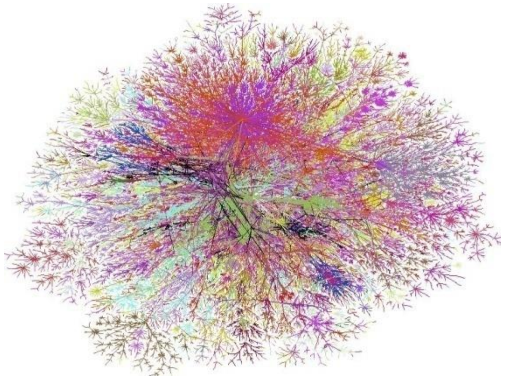
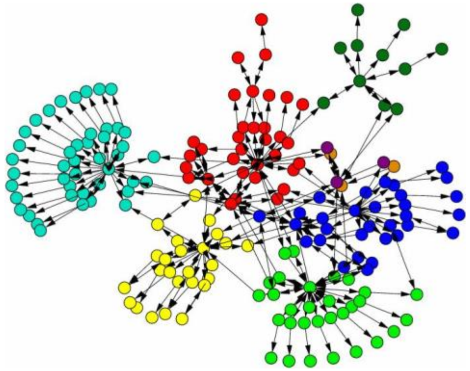
# Age of Networks

Jennifer Chayes
Managing Director
Microsoft Research New England
Microsoft Research New York City

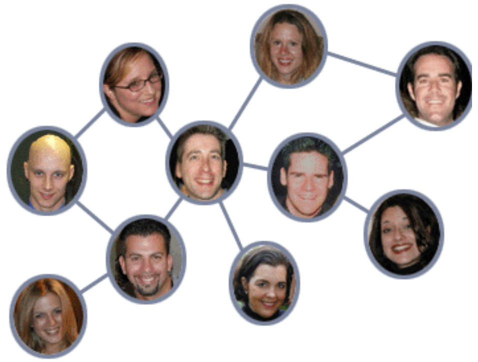
# Motivation: The Age of Networks



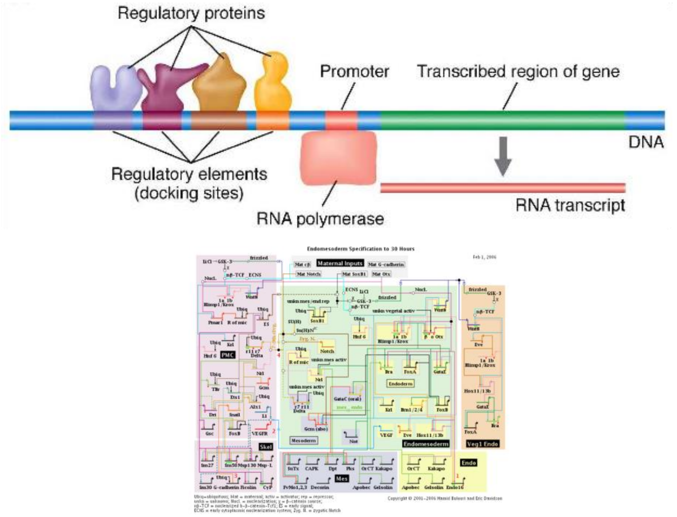
Internet



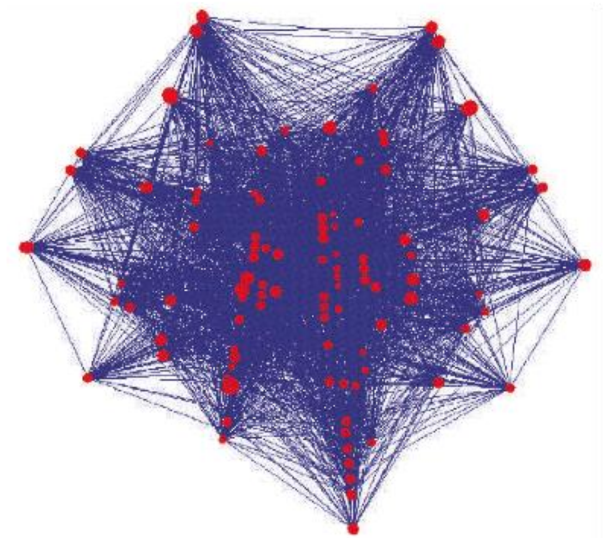
WWW



social networks


Regulatory proteins
Promoter
Transcribed region of gene
DNA
Regulatory elements (docking sites)
RNA polymerase
RNA transcript
Endomesoderm Specification to 30 Hours

gene regulatory networks



resource allocation networks
( = constraint satisfaction networks)

# Outline of the talk:

- “Observed” Networks
- Mathematical and Algorithmic Problems on Networks
- A Specific Class of Problems and Results

# Outline of the talk:

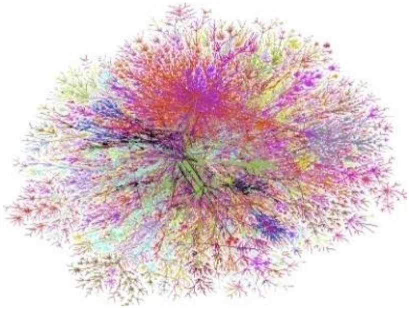
- “Observed” Networks
  - Technological networks
  - Social networks
  - Economic networks
  - Biological networks

# 1. Technological networks

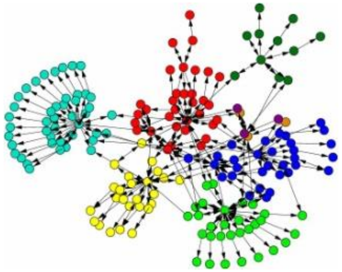
Note: we model these networks as graphs:

$$G = (V, E)$$

- AS (Autonomous System) Internet
  - V autonomous systems (AOL, MSN, Yahoo!, etc.)
  - E connections



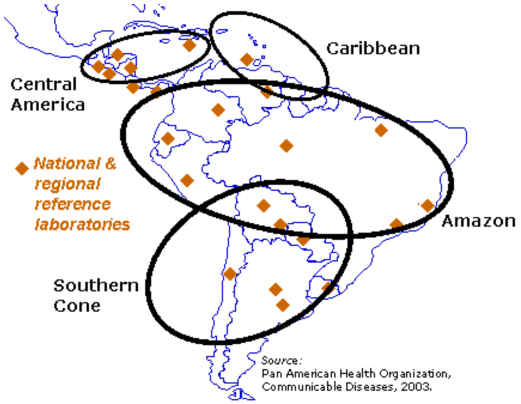
- WWW
  - V webpages
  - E hyperlinks (directed)



- Cloud (data center) networks …

# 2. Social networks

- Offline
  - e.g., epidemiological networks



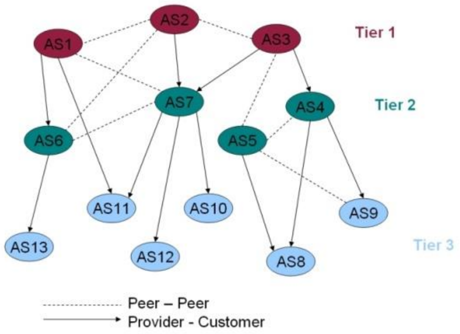
Source: Pan American Health Organization, Communicable Diseases, 2003.

- Online
  - Online social networks

    e.g., Facebook, LinkedIn
  - Mobile phone networks
  - Instant messaging (IM) networks
  - Twitter (microblogging) network

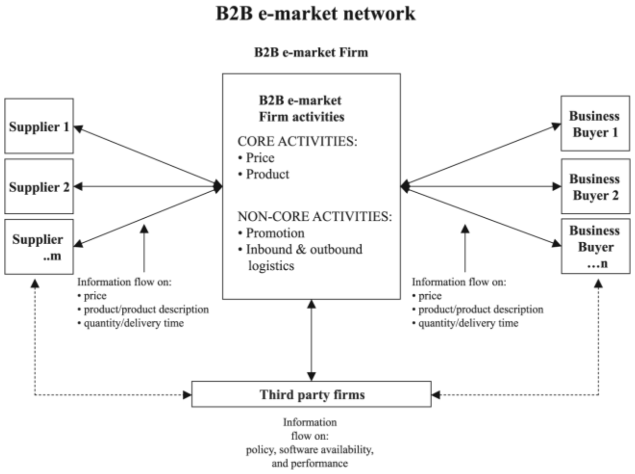
# 3. Economic networks

- Peering agreement networks



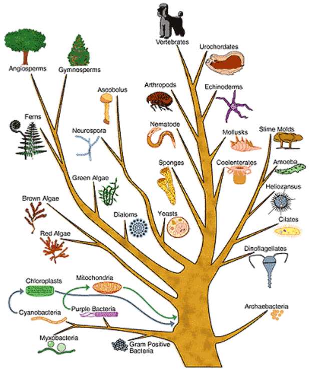
- Bipartite graphs of buyers and sellers

- Market networks

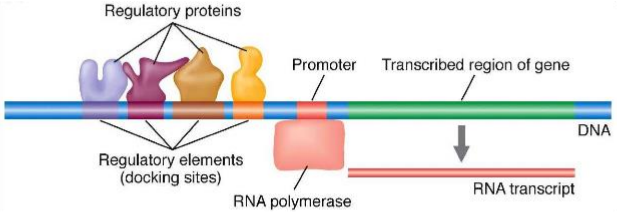


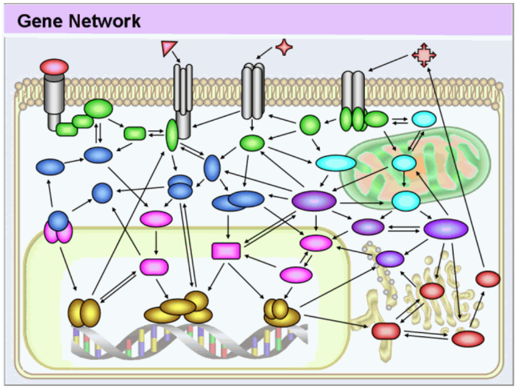
# 4. Biological networks

- Phylogenetic trees



- Gene regulatory networks


Regulatory proteins
Promoter
Transcribed region of gene
DNA
Regulatory elements (docking sites)
RNA polymerase
RNA transcript



- (Real) neural networks . . .

# Outline of the talk:

- “Observed” Networks
- Mathematical and Algorithmic Problems on Networks
- A Specific Class of Problems and Results

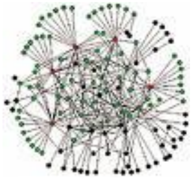
# Mathematical and Algorithmic Problems on Networks

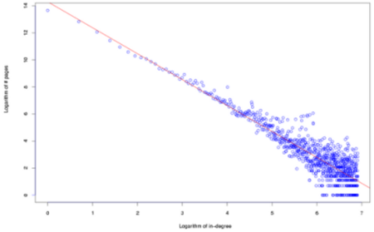
- Modeling networks
- Sampling from large networks
- Processes on networks
- Algorithms on networks
- Network reconstruction algorithms

# 1. Modeling networks

- Observations of tech and social networks
  - Small diameter
    ~ 6 degrees of separation: 1929
    Frigyes Karinthy's short story, "Chains"

  - Power-law degree distribution





  - Aging of vertices
    - On both the AS Internet and the WWW, older vertices tend to be more highly connected
    - This is why web-spammers (a.k.a. Search Engine Optimizers) like to buy old domain names – they are highly connected, and the spammers can use these connections to artificially enhance the connectedness, and hence Pagerank, of commercial sites

# Interlude: Search engines and graph theory

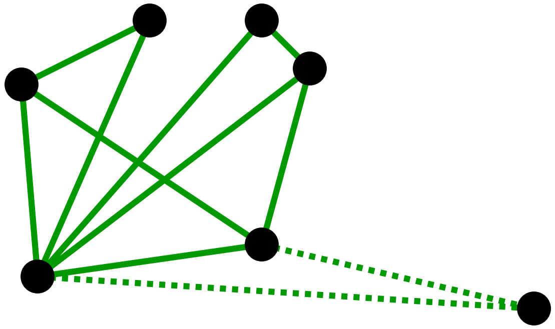
- Early search engines used semantics (i.e., content and language) to find the most relevant webpages
- Later search engines (e.g., Google, Bing) used the structure of the web graph (i.e., graph theory and algorithms) to find the most relevant webpages
- Pagerank: Do a random walk on the web graph, following the hyperlinks, restarting every say 7 steps. The relative weight of webpage in the stationary distribution of this walk is its Pagerank

Later ranking algorithms detect and avoid anomalies to downgrade rankings of web-spammers

(Andersen, Borgs, Chayes, Hopcroft, Mirrokni, Teng ‘07)

# 1. Modeling networks (cont)

- First model:
  - Barabási–Albert 1999
    - At each time step, a new vertex is created and attaches to m old vertices
    - The probability that the new vertex attaches to an old vertex i is proportional to the degree $d_i$ of vertex i.



preferential attachment model

- First rigorous work:
  - Bollobás–Riordan 2000

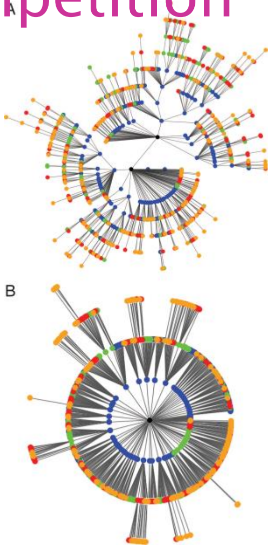
# 1. Modeling networks (cont.)

- Other types of models:
  - Variants of preferential attachment:
    - E.g., Preferential attachment with fitness

      (Bianconi-Barabási '01; rig. work: Borgs, Chayes, Daskalakis, Roch '07)

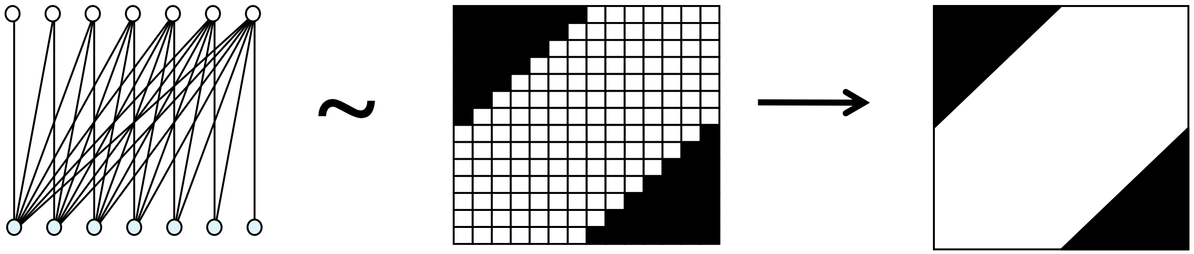
      Observation: There are many exceptions to the fact that older vertices tend to be more highly connected (e.g., Google vs. Alta-Vista).

      Model: Vertices are born with a distribution of fitnesses → phase transitions.
  - Competition models: Optimization models in which the choice of the next vertex is determined by a competition between different factors.
    - E.g., Competition-induced preferential attachment
      - (Berger, Borgs, Chayes, D'Souza, R. Kleinberg '04, '08)
  - Fully game theoretic models
    - E.g., Borgs, Chayes, Ding, Lucier '11



# 2. Sampling from large networks

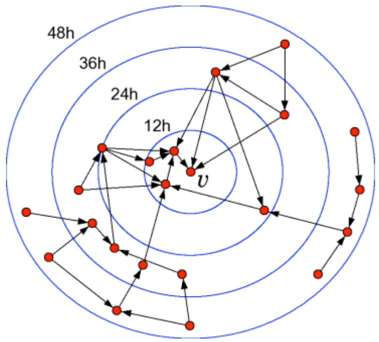
- The WWW is very large (order of a trillion static sites) and growing.
- How do we sample from it, e.g., to calculate pagerank?
- To deal with this, we developed a theory of graph limits and testing:
  - Borgs, Chayes, Lovász, Sós, Vesztergombi ‘06 – ‘12



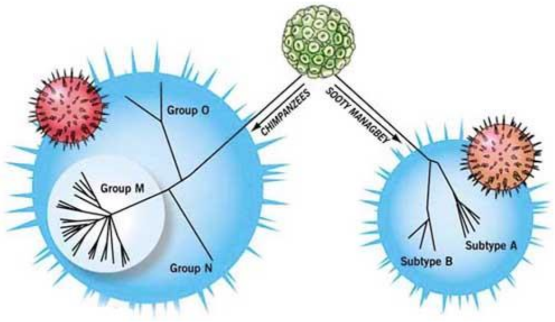
New Result: Graph limits for sparse graphs with power-law tails (Borgs, Chayes, Cohn, Zhao ‘14)

# 3. Processes on networks

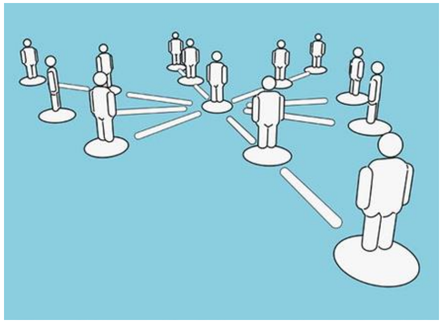
- Flow of information
  - J. Kleinberg *et al.*



- Spread of epidemics
  - Berger, Borgs, Chayes, Saberi ’05, ’13



- Viral marketing
  - Kempe, J. Kleinberg, E. Tardos



# 4. Algorithms on networks

- Ranking algorithms for web search (e.g., Sublinear Time Pagerank: Brautbar, Borgs, Chayes, Teng '12)

- Clustering algorithms for collaborative filtering on bipartite graphs (if you like . . . then you'll also like . . . )

amazon.com

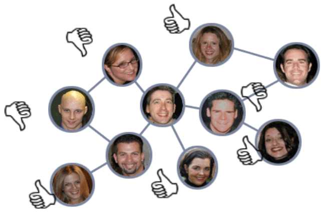
NETFLIX

- Algorithms for multicasting (belief propagation Steiner tree algorithm by Bayati, Borgs, Braunstein, Chayes, Zecchina '08) and for web hosting (Leighton, Lewin)

Akamai

- Fast (sublinear) algorithms for identifying influential sites (Brautbar, Borgs, Chayes, Lucier '13)

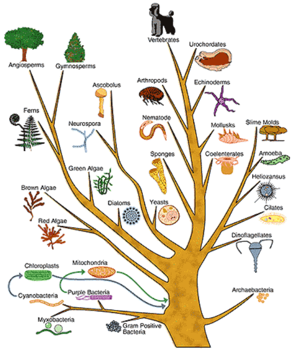
- Algorithms for recommendation systems on online trust networks (Andersen, Borgs, Feige, Flaxman, Kalai, Malekian, Mirrokni, Tennenholtz '08, '10)



# 5. Network reconstruction algorithms

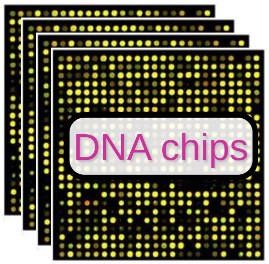
- Phylogenetic network reconstruction

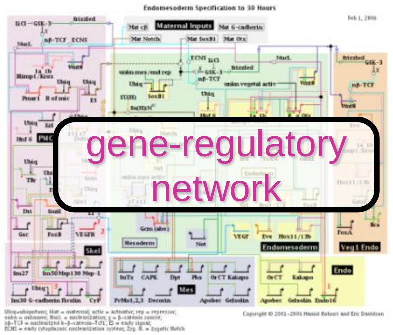
  (also used for linguistic evolution reconstruction)



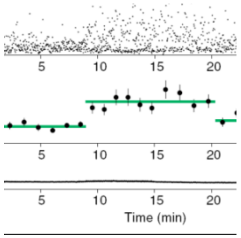
- Gene regulatory network reconstruction

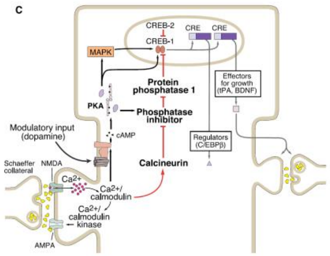
Specific Class of Problems





- Reconstruction of learning processes in networks of synapses





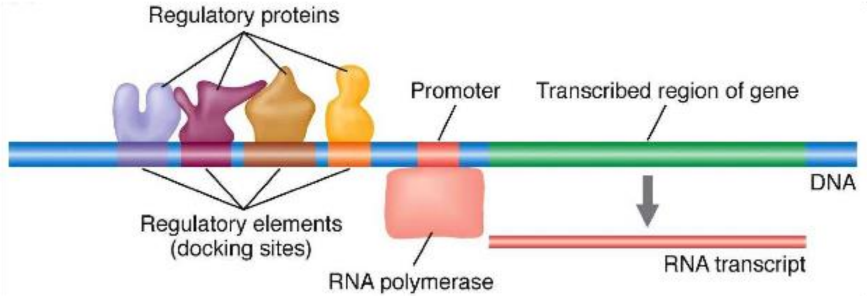
# Outline of the talk:

- “Observed” Networks
- Mathematical and Algorithmic Problems on Networks
- A Specific Class of Problems and Results:

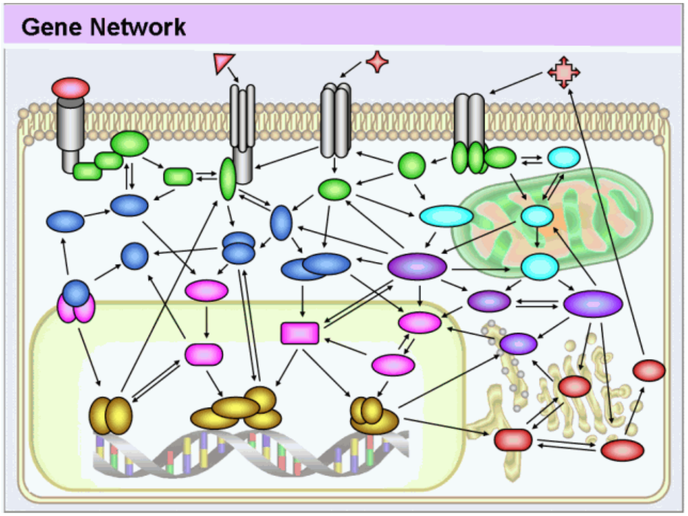
  Reconstruction of gene regulatory networks

# Reconstruction of Gene Regulatory Networks

- Standard Dogma:  DNA → RNA → Proteins



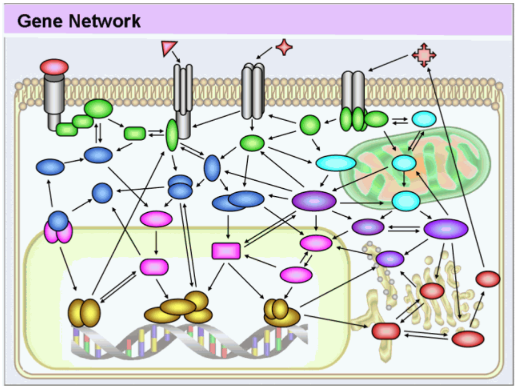
⇒ Gene Regulatory Network



Protein Interactome

# Gene Regulation and Disease

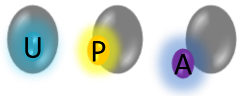
- Problems with the gene regulatory network are the sources of many diseases
- How do we infer the network structure from partial data?
- Can we identify particular nodes on the network responsible for disregulation in certain diseases and individuals?
- Are one or more nodes in combination viable drug targets?

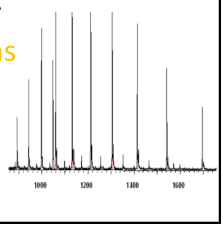


# Drug Discovery Paradigm

Mass spectrometry

Protein Modifications

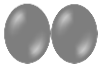




Yeast two-hybrid

Affinity capture mass-spec

Protein-protein interactions

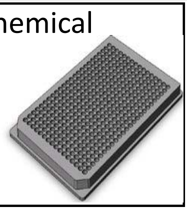


ChIP-Seq, Dnase-Seq, …

Protein-DNA interactions

Genetic/Chemical

Screens

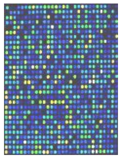


Microarrays

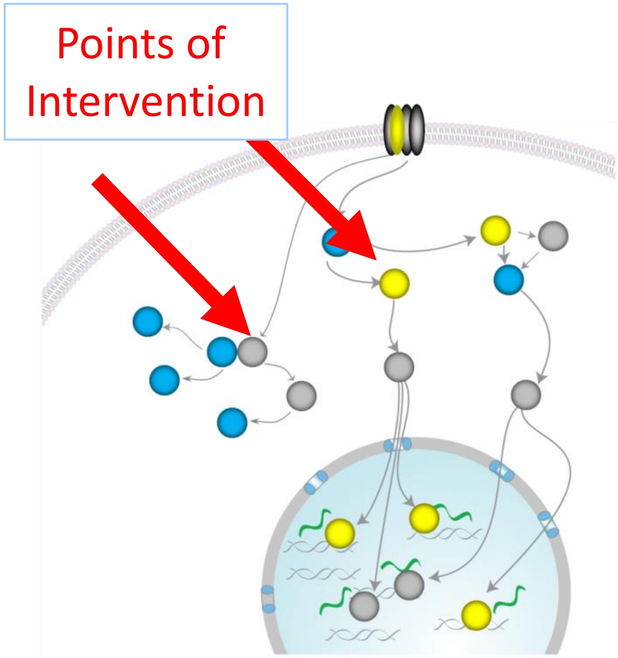
RNA-Seq

mRNA

```
AAATAGCCATTATACGTA
CCTAATACTGAAGAGTCA
TTCCTAGTAAAGCATGCT
ACTTTTCAGTATATTCCA
TTATATTTTTAACTACAA
GCGGCGCAGAAACCAGAG
```

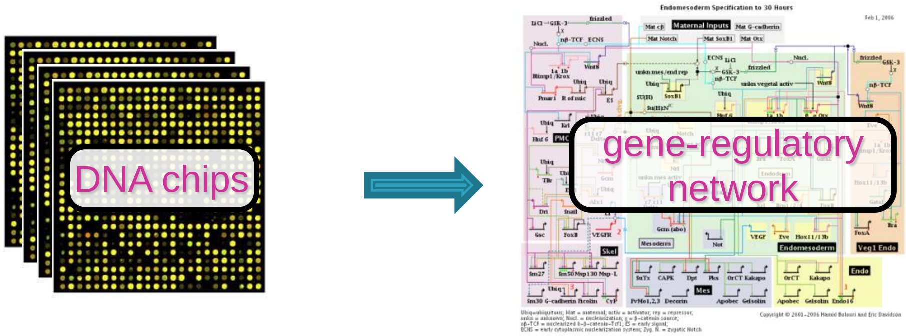


Computational Models



Points of Intervention

# Gene Expression Data



- Microarrays tell us which gene is expressed in the presence of which other gene under a particular set of conditions
- From the differential expression of a particular gene, we infer the node weight of the corresponding (transcription factor) protein
- To get edge weights between two proteins, we use the probability of interaction of these two proteins inferred from (properly weighted) databases of known interactions for the given organism

Question: How do we determine the network most likely to have produced this data?

# The Steiner Tree Problem

- Given
  - Graph $G = (V, E)$
  - Costs $\{c_{ij}\}_{ij \in E}$, $c_{ij} \geq 0$
  - Set of "terminals" $S \subseteq V$
- Problem: Find a tree $T \subseteq G$ containing all terminals, i.e. all nodes in $S$, which minimizes the cost $\sum_{ij \in E(T)} c_{ij}$
- Solution: In general, the minimizing tree contains other nodes in addition to the terminals. These additional nodes are called Steiner nodes.
- Computational issues: Bayati, Borgs, Braunstein, Chayes, Ramezanpour, Zecchina PRL'08 found a new representation of the Steiner tree problem which allowed it to be solved very quickly with belief propagation algorithms.

# Biological Problem Formulation: The Prize-Collecting Steiner Tree

- Given
  - Graph $G = (V, E)$
  - Costs $\{c_{ij}\}_{ij \in E}$, $c_{ij} \geq 0$
  - Set of "prize terminals" $S \subseteq V$ with prizes $\{\pi_i\}_{i \in S}$, $\pi_i > 0$
  - Parameter $\lambda > 0$
- Problem: Find a tree $T \subseteq G$ which minimizes the cost:

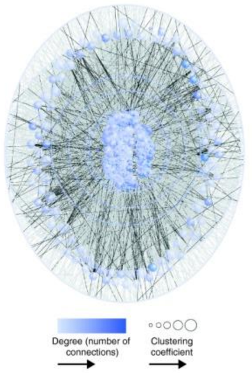
$$C(T) = \sum_{ij \in E(T)} c_{ij} - \lambda \sum_{i \in V(T)} \pi_i$$

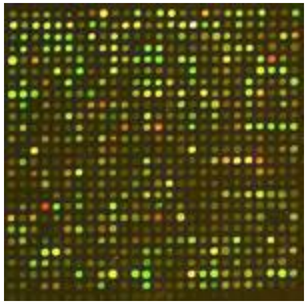
- Note: As $\lambda \to \infty$, this turns into the standard Steiner tree problem with terminals $S = \{i | \pi_i > 0\}$.

# Mapping to Biological Data

- Find the tree which minimizes

$$C(T) = \sum_{ij \in E(T)} c_{ij} - \lambda \sum_{i \in V(T)} \pi_i$$





$$c_{ij} = -\log \mathrm{prob}(ij\ \mathrm{exists})$$

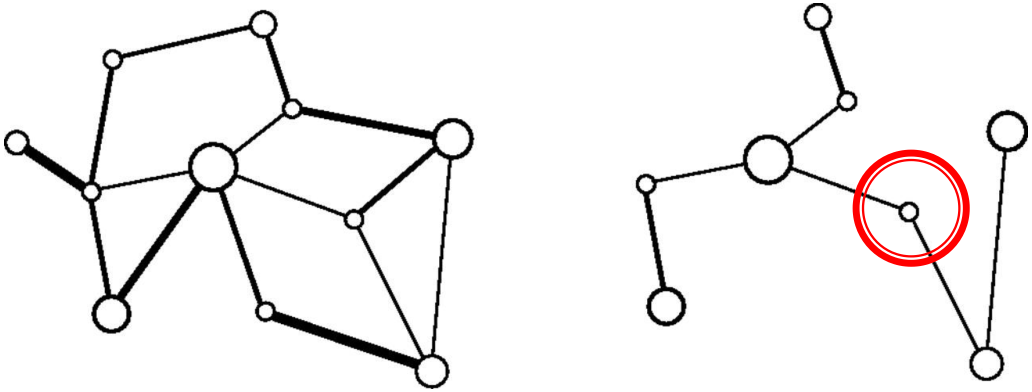
where $\mathrm{prob}(ij\ \mathrm{exists})$is the probability that proteins $i$ and $j$ interact in the given organism (from organism databases)

$$\pi_i = -\log p_{\mathrm{value}}(i)$$

where $p_{\mathrm{value}}(i)$ is the p–value of the differential expression of the gene corresponding to protein $i$, in the given experiment

# Steiner Nodes

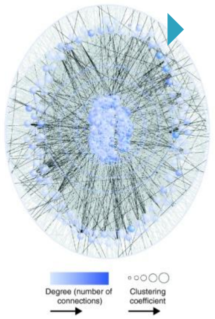
- In the standard Steiner tree problem, nodes which are included in the minimizing solution but which are not terminals, i.e. not in the set $S$, are called Steiner nodes
- Similarly, in the PCST, nodes which have zero (or low) prizes but which are included in the minimizing solution are called Steiner nodes



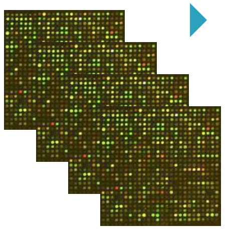
- In the context of the gene regulatory networks, Steiner nodes correspond to proteins whose genes which are not differentially expressed, but which nevertheless seem likely to participate in the network ⇒ identification of proteins not previously know to participate in the pathway

# Example 1: Yeast Pheromone Response Pathway

(Bailley-Bechet, Borgs, Braunstein, Chayes, Dagkessamanskaia, Francois, Zecchina: PNAS '11)



- Yeast protein signal transduction network:
  - 4689 Proteins
  - 14928 Protein-Protein interactions
  - Gives set of weights $\{c_{ij}\}$ for relevant proteins in pheromone response pathway

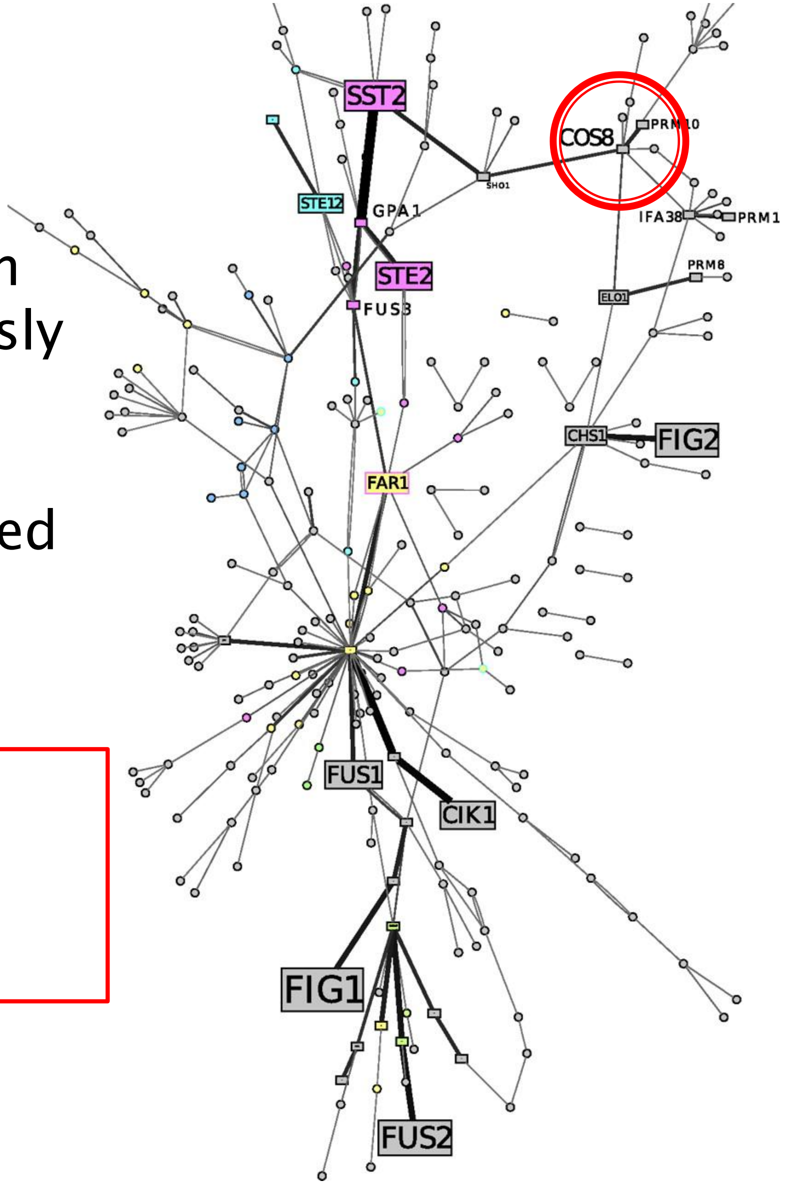


- Considered 56 large-scale gene expression data sets used to reconstruct the yeast pheromone pathway. For each data set
  - Get set of prizes $\{\pi_i\}$
- Construct 56 solutions to bounded-D PCST problem
- "Merge solutions" to get one network

# Results: Pathway identified

- Two types of proteins on network
  - Proteins differentially expressed in pheromone response and previously discovered by transcriptomic studies (terminals)
  - Proteins not differentially expressed but bridging between different subnetworks ("Steiner proteins")

Question: Are the Steiner proteins important in the pheromone response pathway?



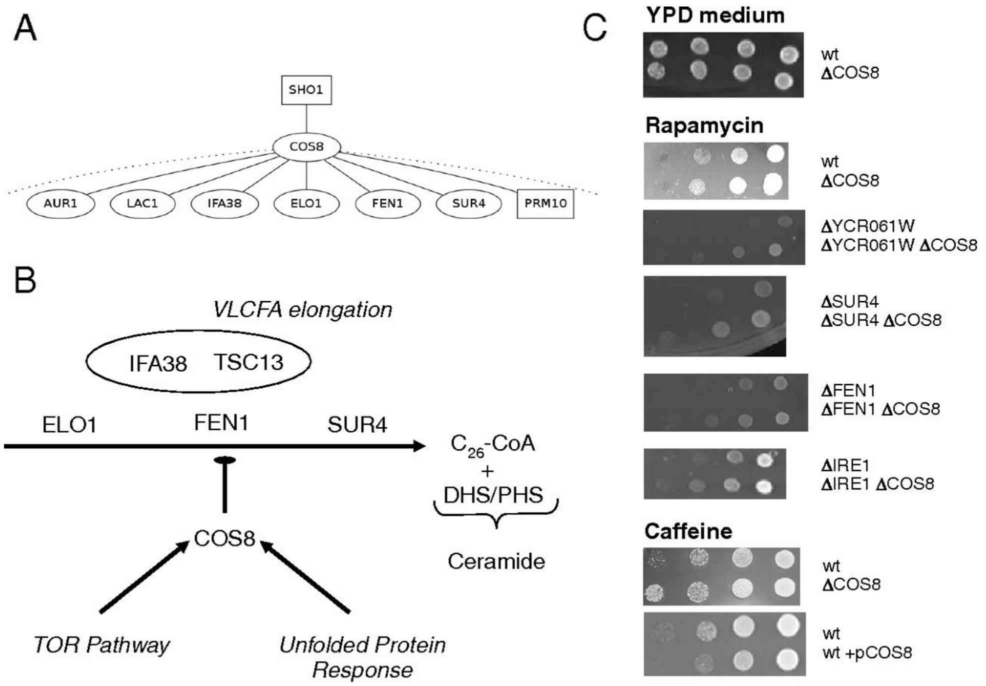
# Testing a Steiner Node

- Did an experiment to knock out the gene corresponding to COS8

⇒

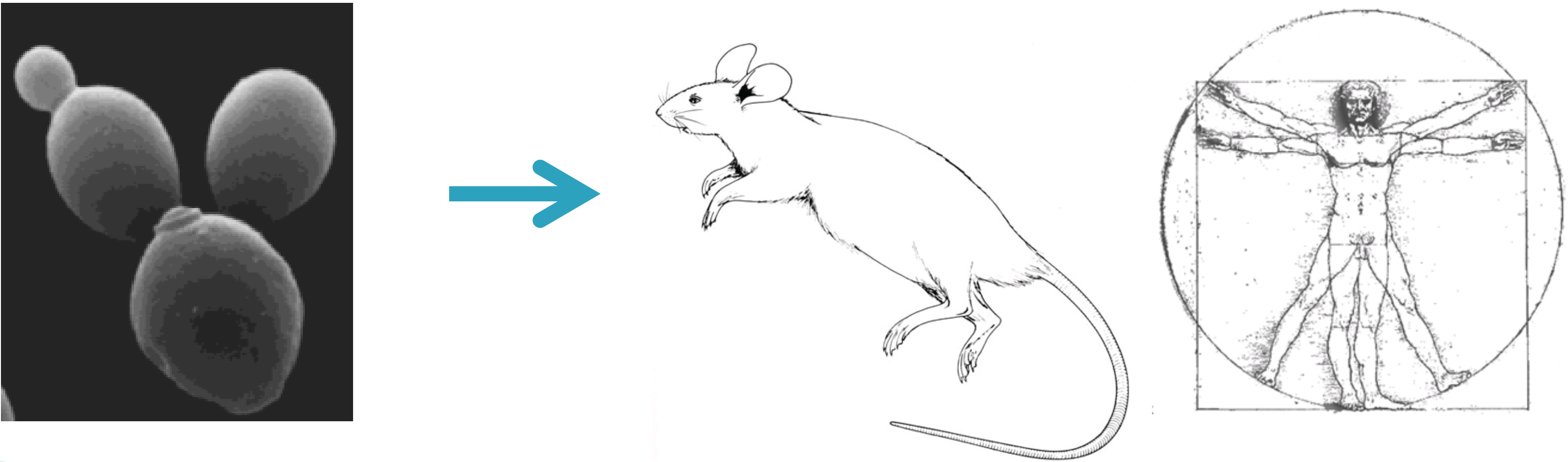
Pheromone response pathway failed.

"Experimental proof" of the importance of the Steiner node



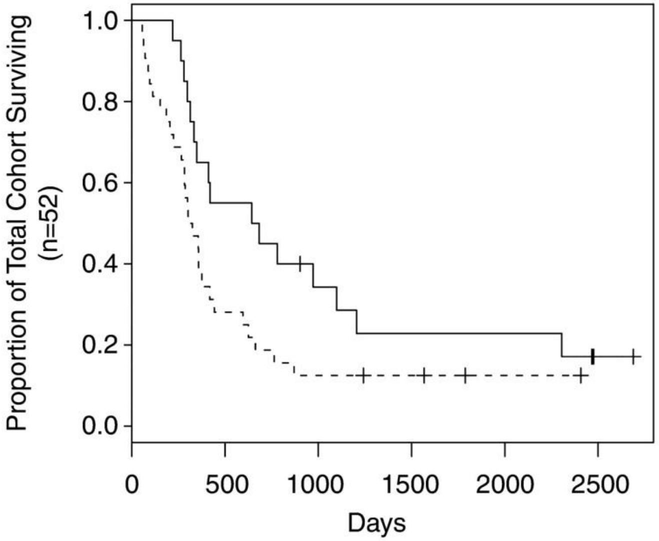
# From Yeast to Mammals

- Problems (mammals relative to yeast):
  - Incomplete interactome data
  - Ten times as many transcription factors
  - Huge intergenic regions
- Need fast algorithms

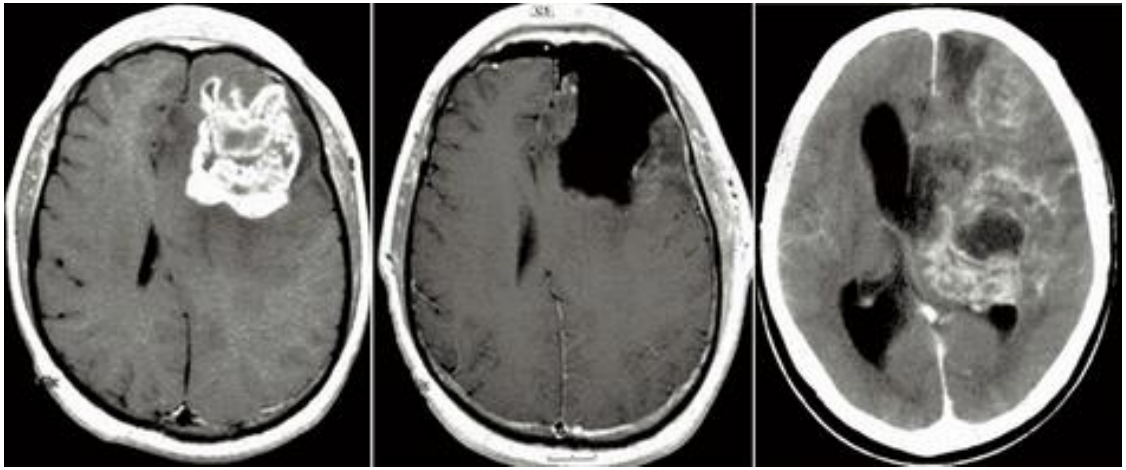


# Example 2: Glioblastoma Pathways

- Glioblastoma:
  - particular form of brain cancer
  - the human cancer with the worst outcome
  - much more common in men than women



Pope W B et al. Radiology 2008;249:268-277

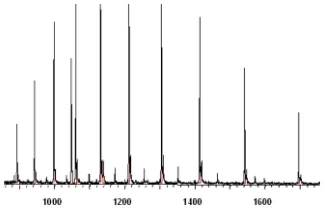


Presentation Post-op Recurrence

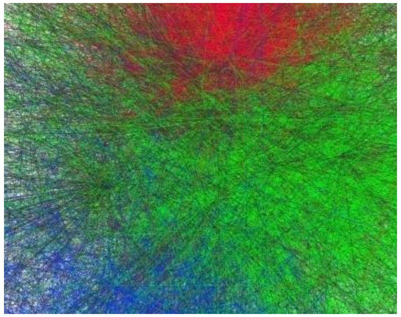
Weil RJ (2006) PLoS Med 3(1): e31.

# Can we find GBM pathways using the PCST?

(Fraenkel Lab, MIT, using our PCST algorithm; more recent joint work)

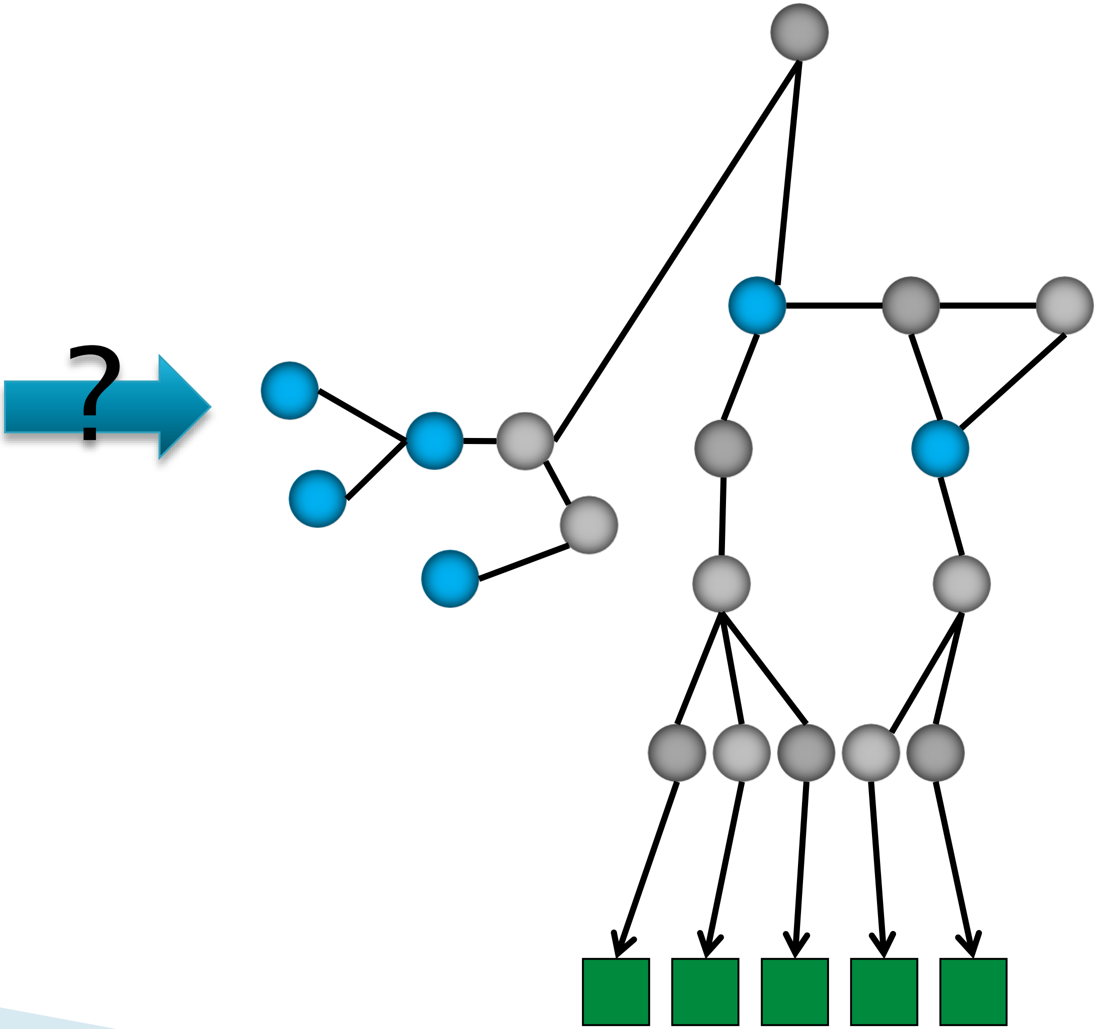


Mass spectrometry



Interactome

Expression/Epigenomics

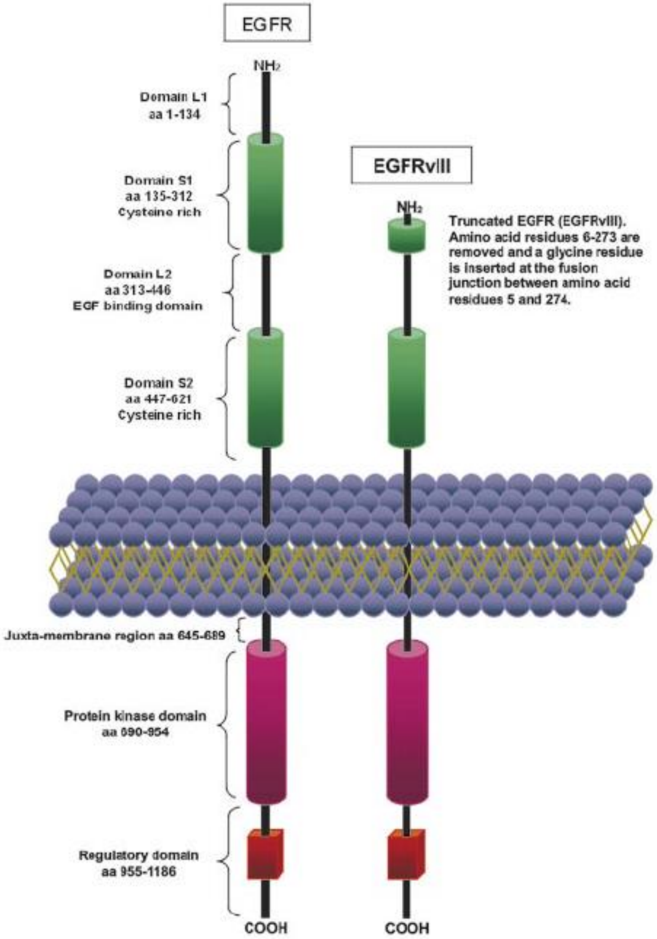


# How to choose the root of the PCST?

Always good to choose receptor proteins since these often begin signaling pathways

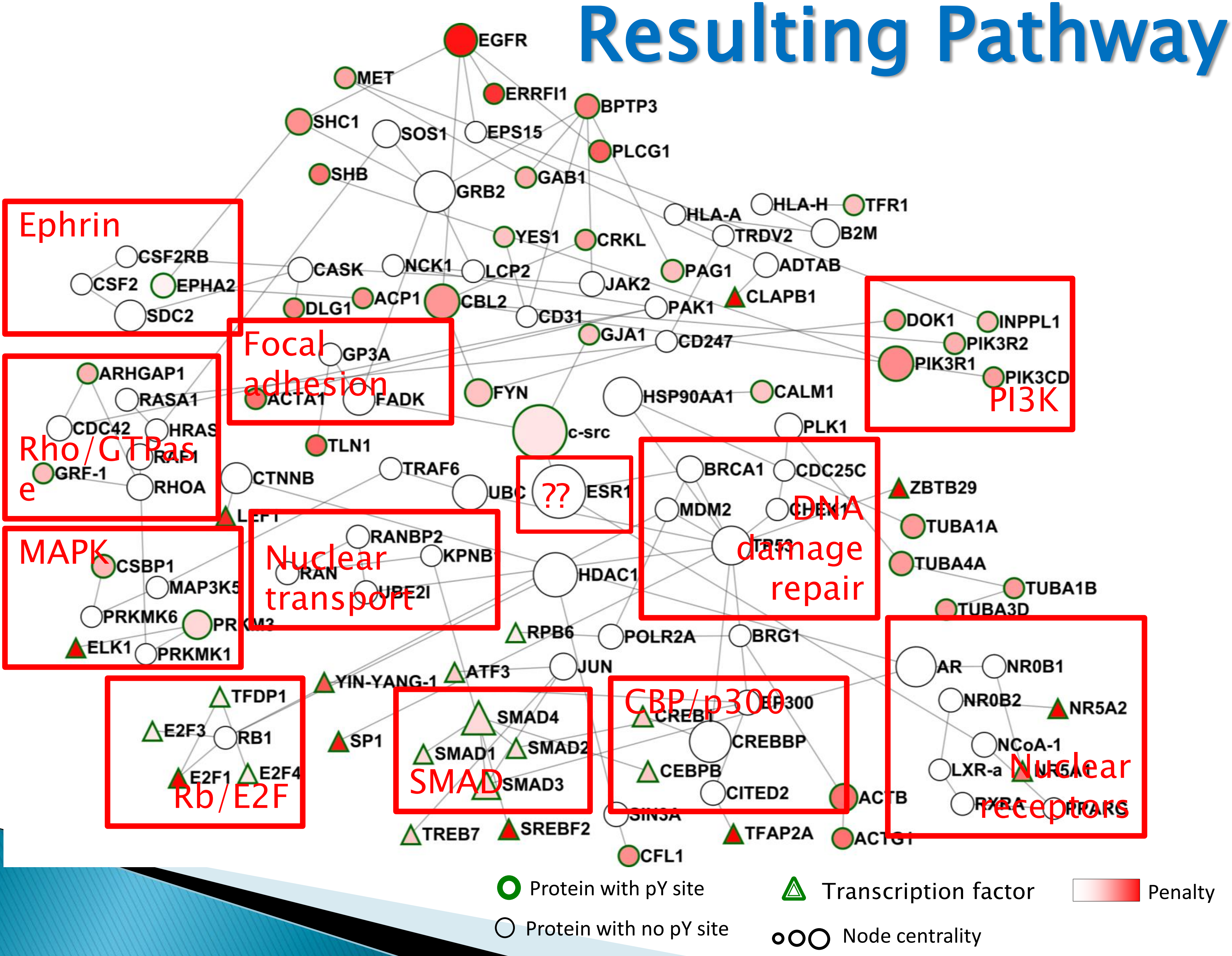
Try EGFR

- EGFR variant III mutation is most common EGFR mutation in human cancer
- Present in 60% of GBMs
- EGFRvIII expression correlates with shorter life expectancies



Truncated EGFR (EGFRvIII). Amino acid residues 6-273 are removed and a glycine residue is inserted at the fusion junction between amino acid residues 5 and 274.

# Resulting Pathway



# Identify interesting Steiner nodes

- Top 5 Nodes ranked by betweeness centrality*:
  SRC, ESR1, HDAC1, CREBBP, GRB2
- SRC well-known to be active in many types of cancer, and had relatively large “prize”
- What about ESR1?
  - No “prize” and not previously identified for Glioblastoma
  - What is ESR1?
  - This is the Estrogen Receptor
- **First pathway link between glioblastoma and gender!**
- Experimental test: EGFR inhibitor and Estrodiol together inhibit the growth of GBM cells in culture better than the EGFR inhibitor alone

  ⇒ ??? possible drug therapy for glioblastoma

*Relative percentage of shortest paths in graph through given node

# Multiple Signaling Pathways

(Tuncbag, Braunstein, Pagnani, Huang, Chayes, Borgs, Zecchina, Frankel ’12)

- How do we explain multiple signaling pathways altered in a particular condition?
- Use Prize-Collecting Steiner Forest (PCSF):
- Just like prize-collecting Steiner tree, but now we also specify that there be $k$ disjoint trees* $(=$ forest $F)$ as the minimizing solution of

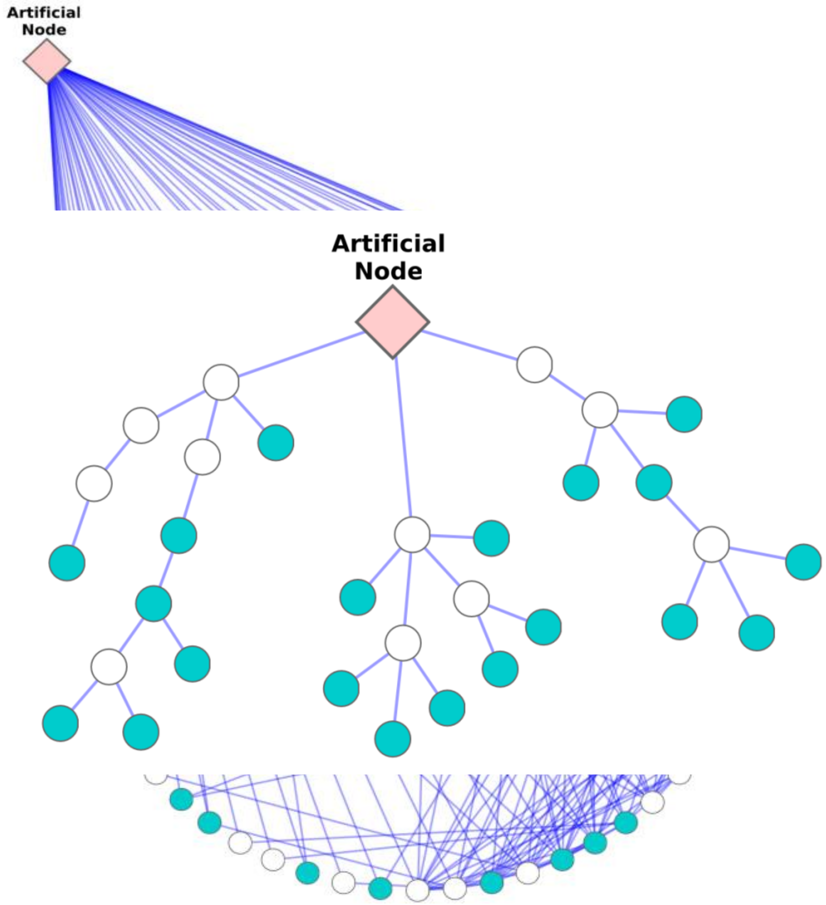
$$C(F) = \sum_{ij \in E(F)} c_{ij} - \lambda \sum_{i \in V(F)} \pi_i$$

- To implement PCSF, just add an “artificial node” $A$, connect every node $i$ to $A$ with strength $c_{iA} \Rightarrow$ new PCST with 1 more node and $|V|$ more edges

*Or let $k$ vary by adding another term to $C$

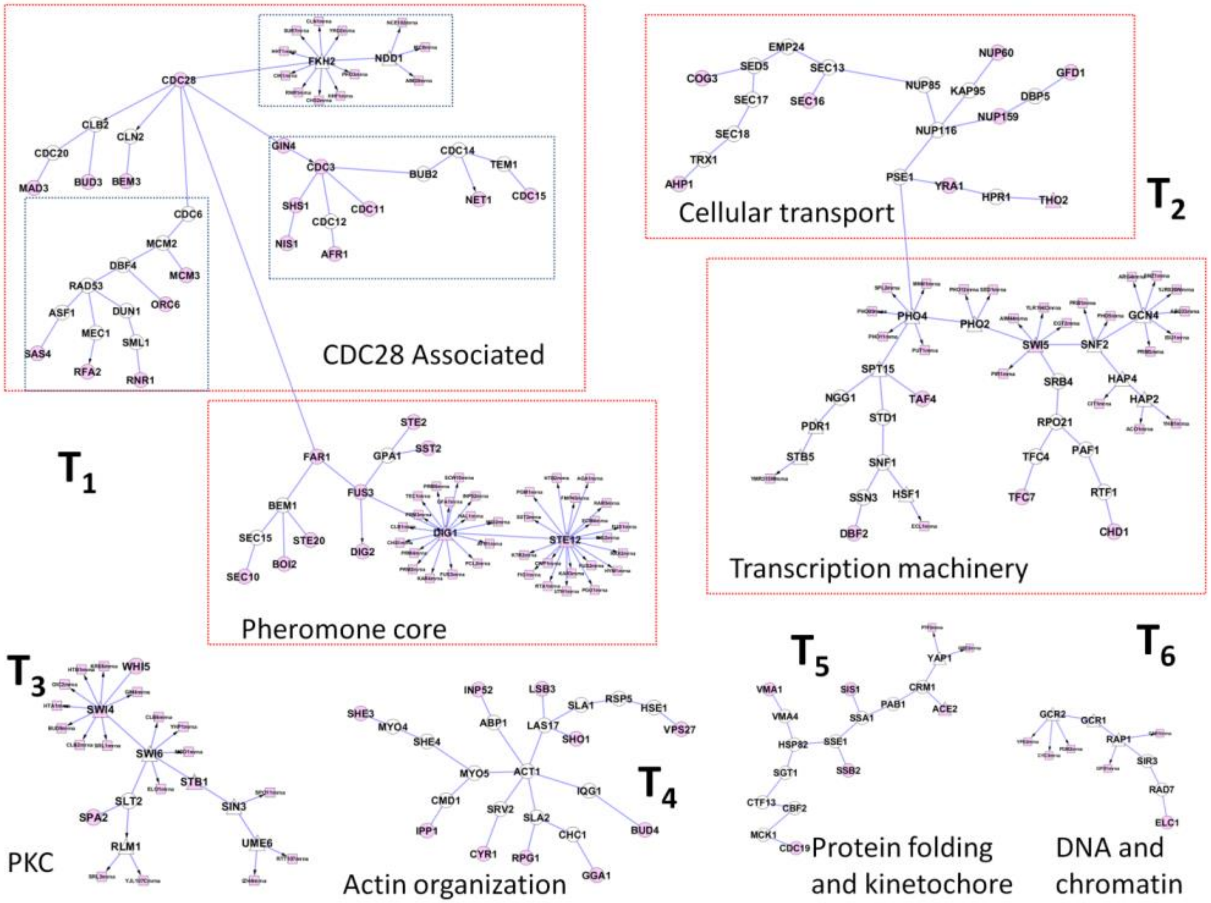
# Method

## Prize Collecting Steiner Forest

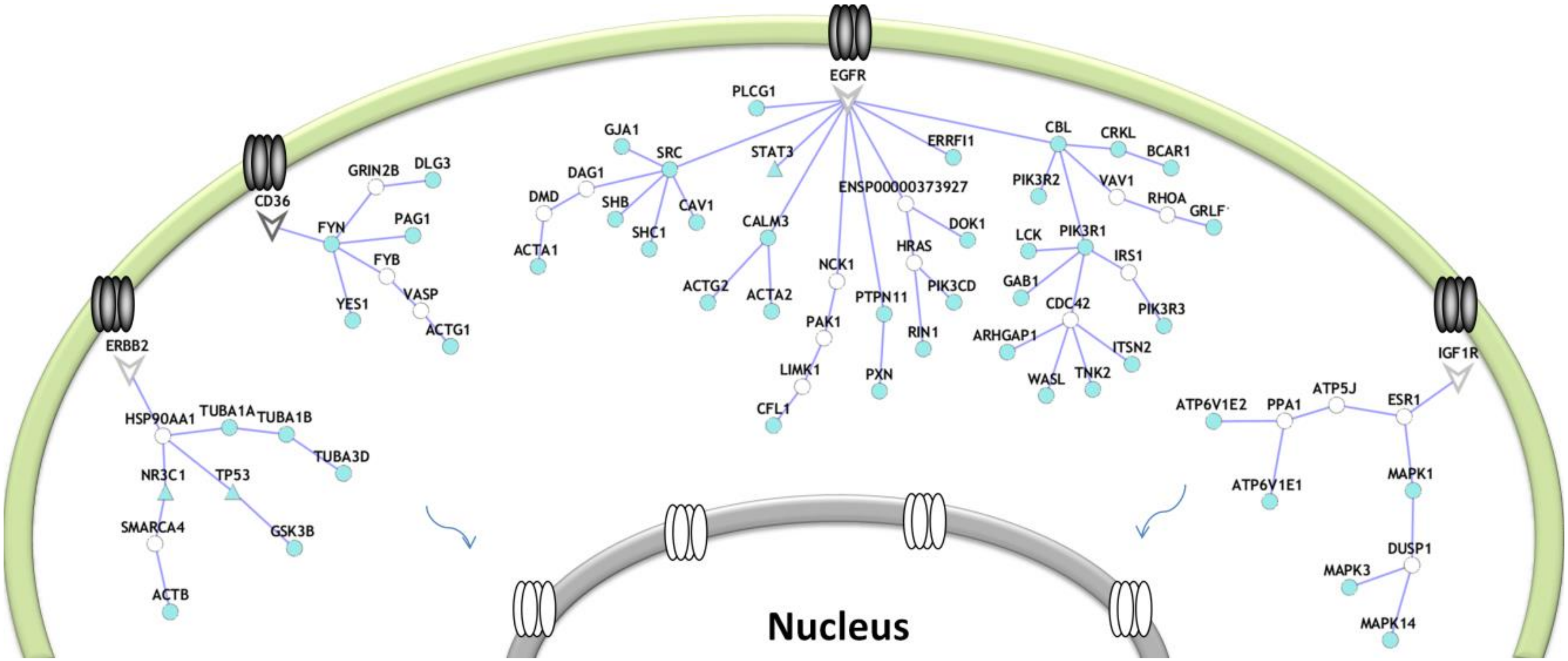


Reveals parallel working pathways, in addition to "hidden" (Steiner) individual proteins or genes

# Derived Forest: Yeast Pheromone Response Network



# Derived Forest: Human Glioblastoma Data Set



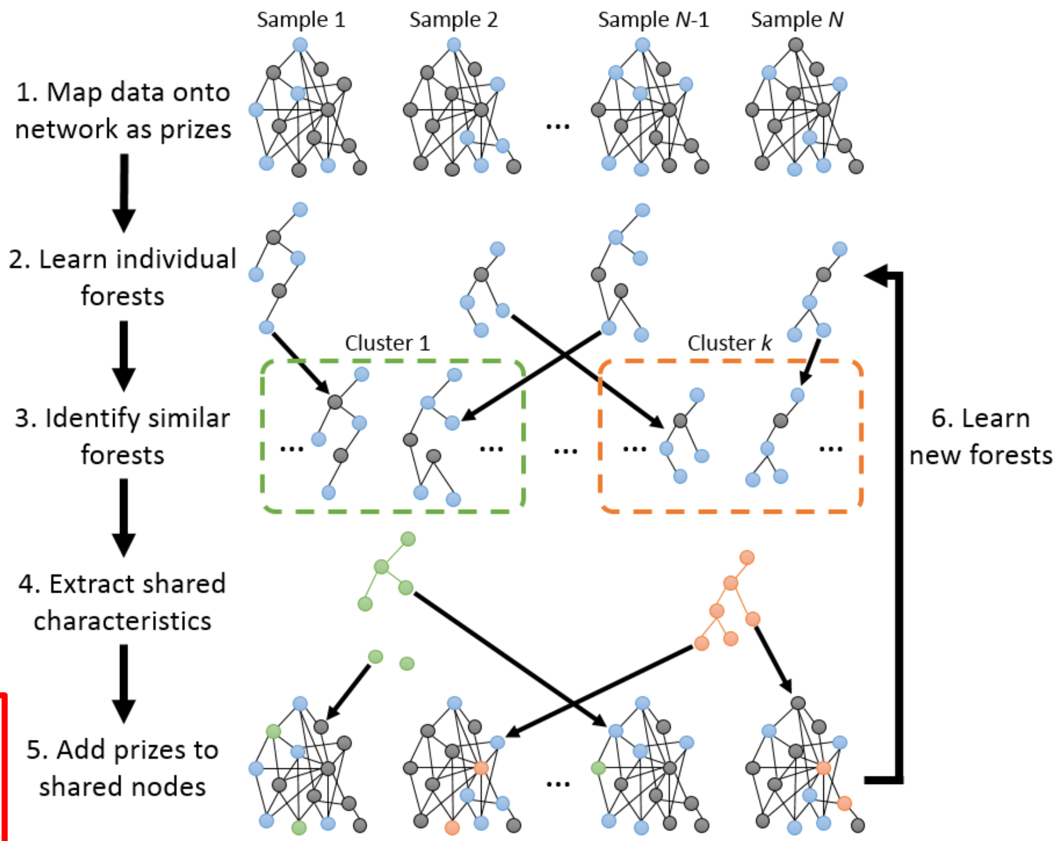
# Recent Extension to Reconstruction of Patient-Specific Networks (Multi-PCSF)

TCGA Breast Cancer Data:

Learn networks of individual breast cancer patients, extract shared features, & update algorithm for individual patients. Iterate.

- → Highly patient-specific networks, which have input from networks of other patients.

(E.g., found subclass whose Steiner nodes implied they might be treatable with drugs for KIT-positive gastrointestinal tumors)


Sample 1
Sample 2
Sample *N*-1
Sample *N*
1. Map data onto network as prizes
...
2. Learn individual forests
Cluster 1
Cluster *k*
3. Identify similar forests
...
4. Extract shared characteristics
5. Add prizes to shared nodes
6. Learn new forests

(Gitter, Braunstein, Pagnini, Baldassi, Borgs, Chayes, Zecchina, Fraenkel; PSB'14)

# Summary

- Everywhere we look, we see large-scale networks – technological, social, economic, biological
- Modeling and analysis of these networks uses approaches from graph theory, combinatorics, probability, game theory, algorithms
- Results include new theories, theorems, experimental predictions

  … even new business models

  … and possibly new (personalized) drug therapies

# Thanks for your attention