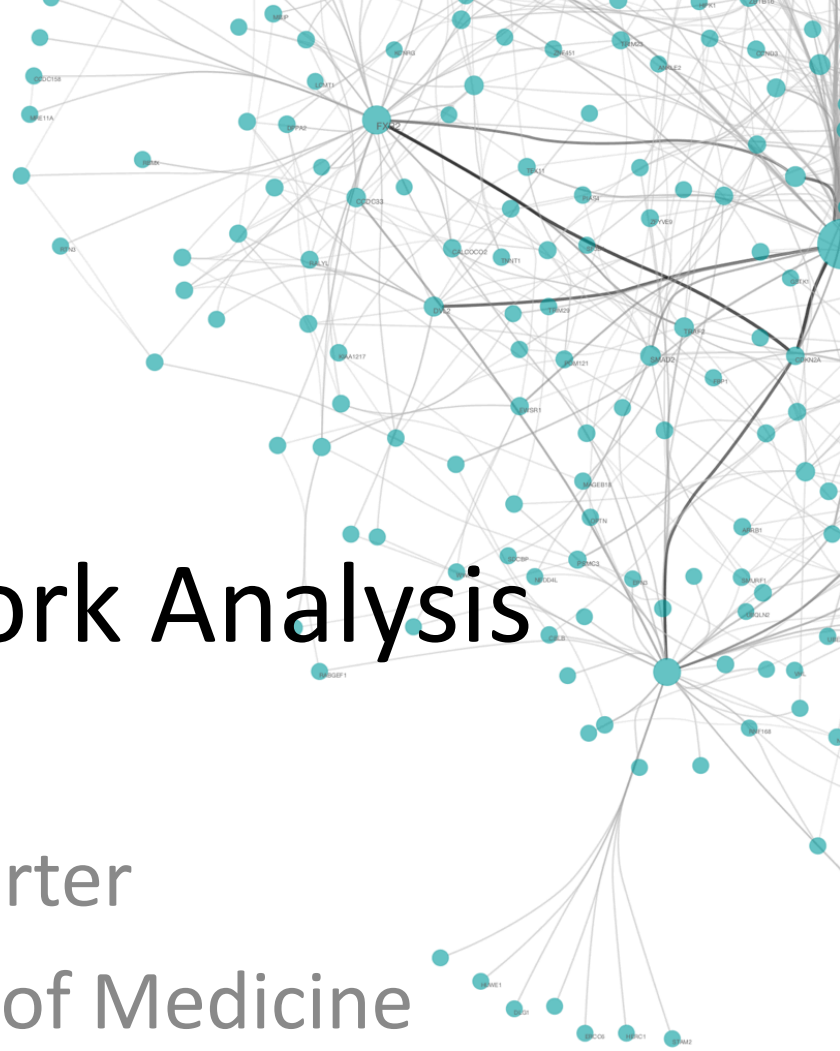


Biological Network Analysis

Hannah Carter

UCSD Department of Medicine

01/16/2018

Overview

- **Day 1: Network analysis**
  - Why networks?
  - Network basics
  - Examples of biological networks
  - Using networks to study biological systems

- Day 2: Network analysis applied
  - Exercises: Network basics
    - Creating, manipulating and analyzing networks with R
  - Introduction to Cytoscape and Ndex
    - Loading and visualizing networks

Class Survey

- Familiarity with biological network analysis?

- Familiarity with R and Jupyter Notebook

Network Module Goals

- Understand how networks can be used to represent, analyze and interpret biological systems including:
  - Examples of biological networks
  - Common network statistics used for biological inference
  - How to generate a random distribution over a network
- Understand how networks can be created from data
- Provide a basic computational toolkit for network analysis

Please ask questions!

Why Networks?

Why Networks?

Reductionist Biology

- Parts list
- Study each part individually

Systems biology

- Study behavior of the system
- Behavior arises from interactions among components

Networks as a tool for systems biology

- Explicitly model interactions
- Can be used to study complex behaviors of systems
- Are convenient tools for analyzing relationships in data generated from high throughput technologies

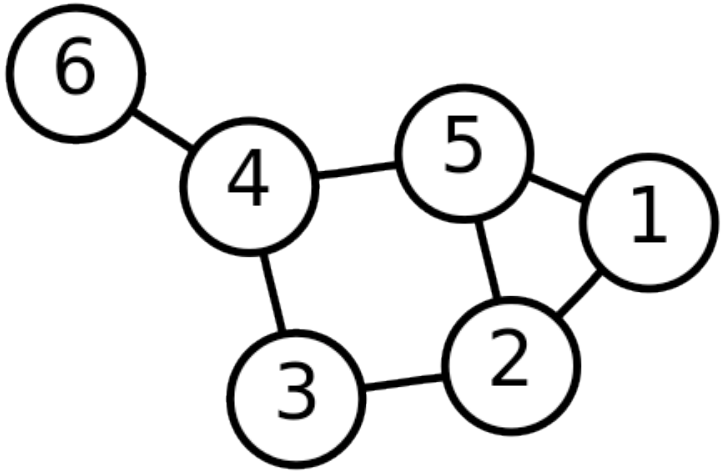
Network Theory

- Study of graphs as a representation of the relationships between discrete objects
  - Genes, proteins, enzymes, metabolites, drugs, diseases, …
- Objects represented as nodes, relationships represented as edges
  - Protein-protein interactions, enzyme-substrate relationships, co-expressed mRNAs …

Networks are modeled by Graphs

G=(V,E)

- V = vertices (nodes)
- E = edges



V = {1, 2, 3, 4, 5, 6}

E = {{1, 2}, {1, 5}, {2, 3}, {2, 5}, {3, 4}, {4, 5 }, {4, 6}}

Subgraph – connected subset of G

G’ = (V’,E’)

$V' \subseteq V$

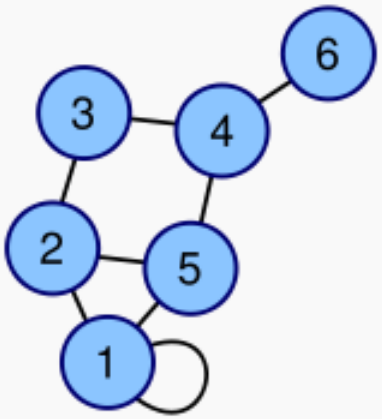
$E' \subseteq E \wedge ((v_1, v_2) \in E' \rightarrow v_1, v_2 \in V')$

Creating Graphs from Data

- Adjacency matrix
  - Can include edge weights in adjacency matrix

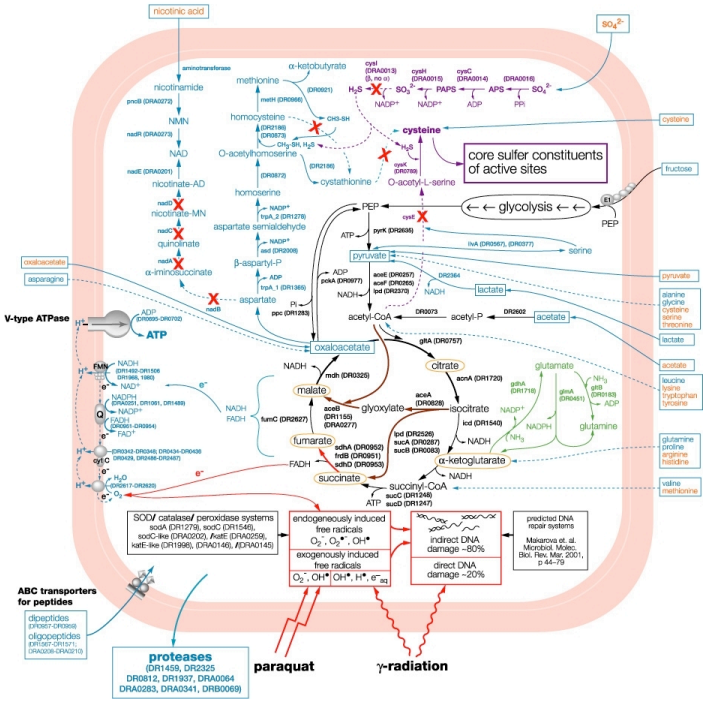
- SIF format

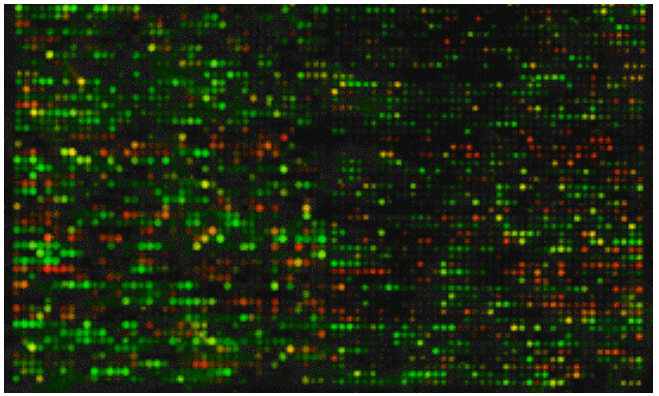
```
1  1
1  2
1  5
2  3
2  5
3  4
4  5
4  6
```

Labeled graph	Adjacency matrix
	$\begin{matrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{matrix}$

Creating Graphs from Data

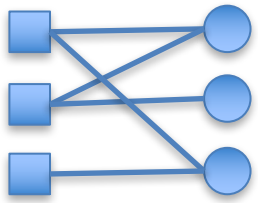
- Graphs that represent the **known structure** of biological systems
  - PPI networks
  - Gene regulatory networks from TF binding motifs
  - Metabolic networks
- Graphs that are inferred from **measured relationships** between biological variables
  - Gene co-expression networks
  - Co-morbidity maps





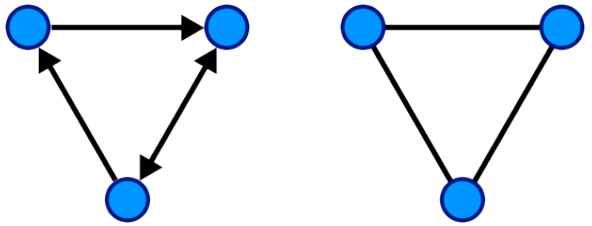
Versatility: Graphs can encode different types of information

- Nodes can represent one type of entity or more
  - Bipartite graph



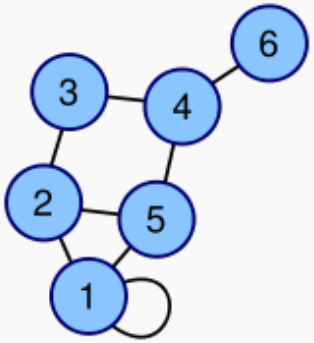
Example:
Enzymes and substrates

- Edges can be directed or undirected (or mixed)
  - Directed when the direction of action or information flow is known
  - Different types of relationship

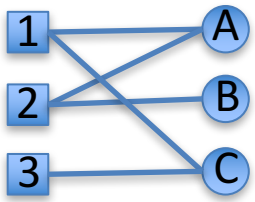


Examples:
Transcription factors and targets
Protein-protein interactions

Could you encode a directed network with an adjacency matrix?

Labeled graph	Adjacency matrix
	$\begin{matrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{matrix}$

Could you encode a bi-partite network with an adjacency matrix?

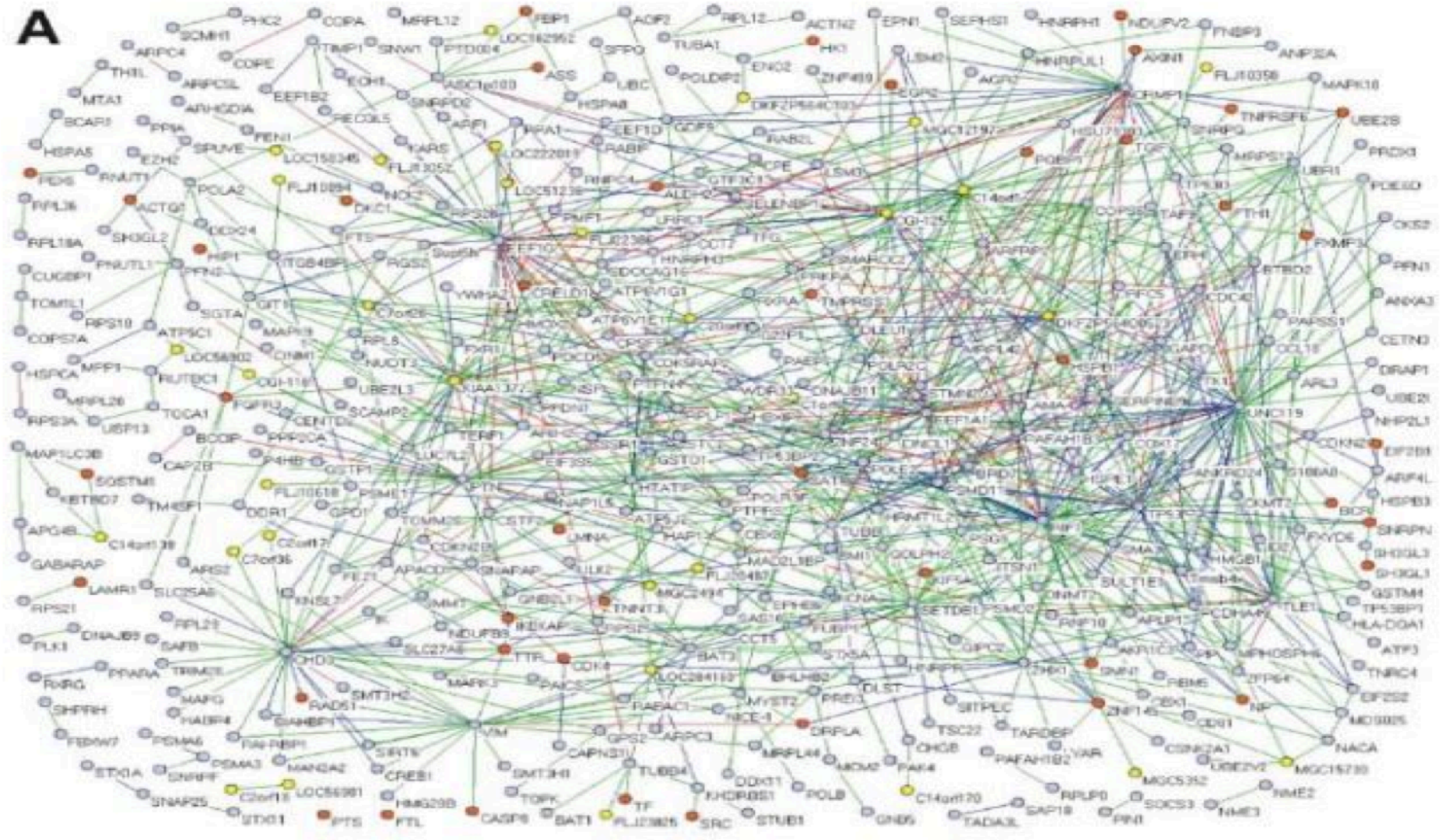


	A	B	C
1	1	0	1
2	1	1	0
3	0	0	1

Examples of Biological Networks

- Protein-protein interaction
- Metabolic
- Co-expression
- Regulatory
- Genetic Interaction
- Difference Networks
- Abstract

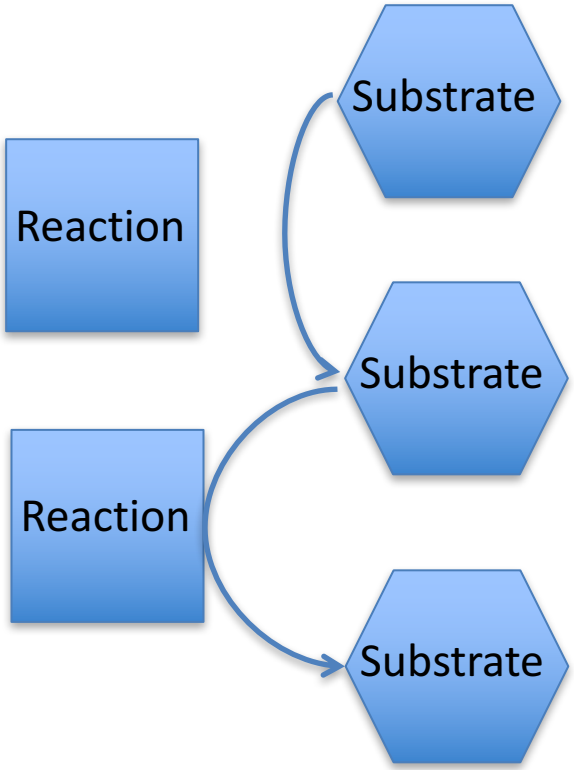
Protein Protein Interaction Network

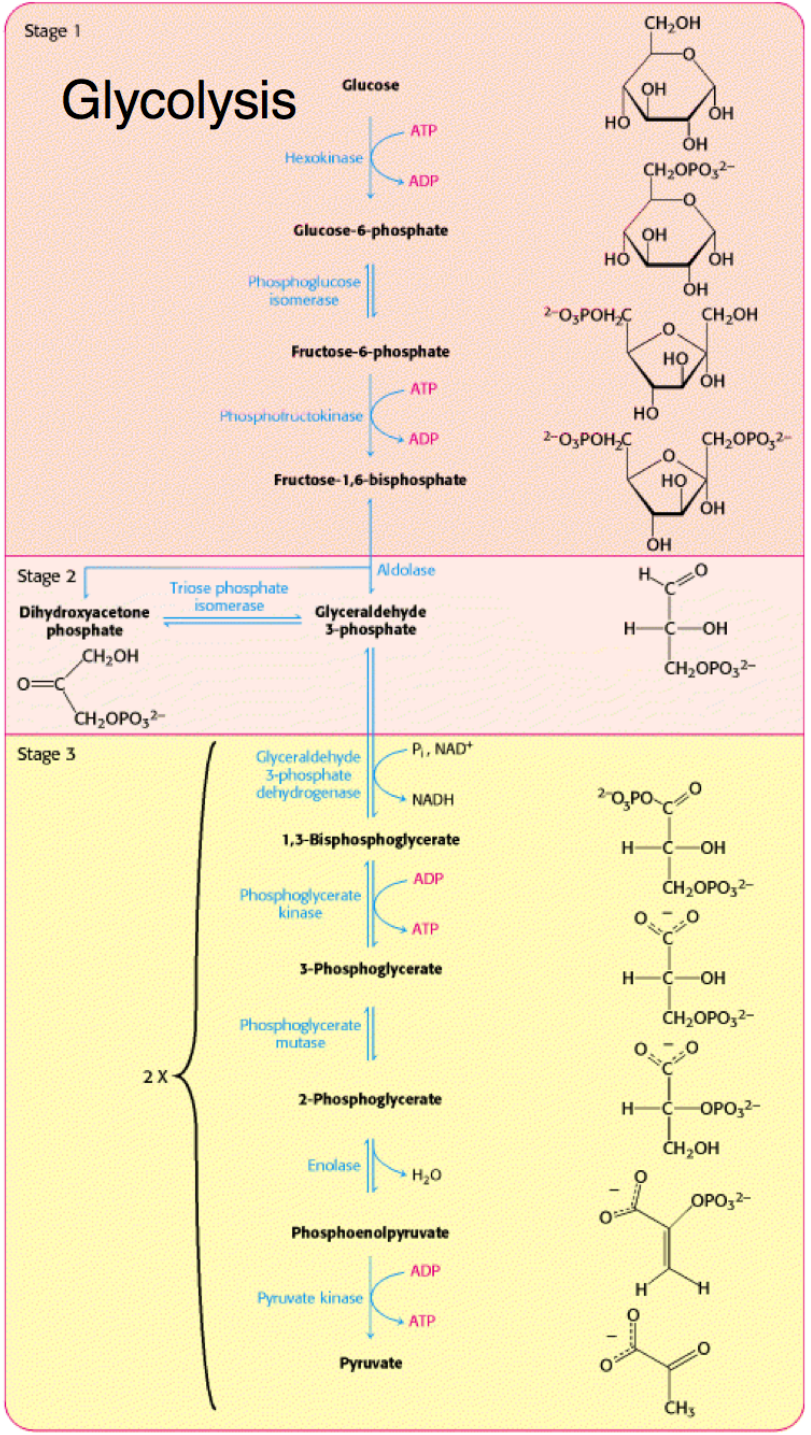


Generated from Y2H data

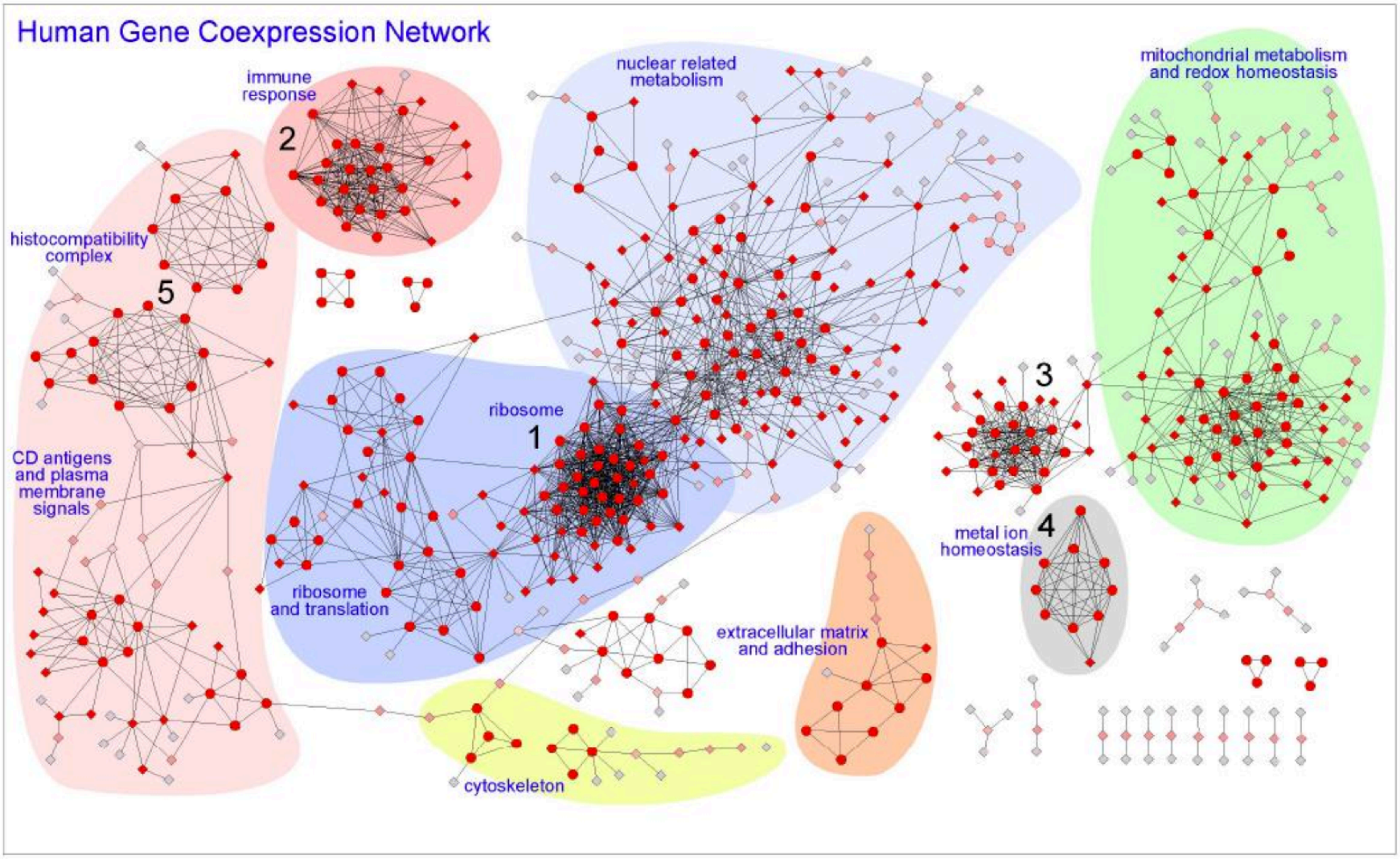
MacArthur et al., *PLoS ONE* 3: e3086 (2008)

Metabolic Networks





Hi-Fi Human Co-expression Network



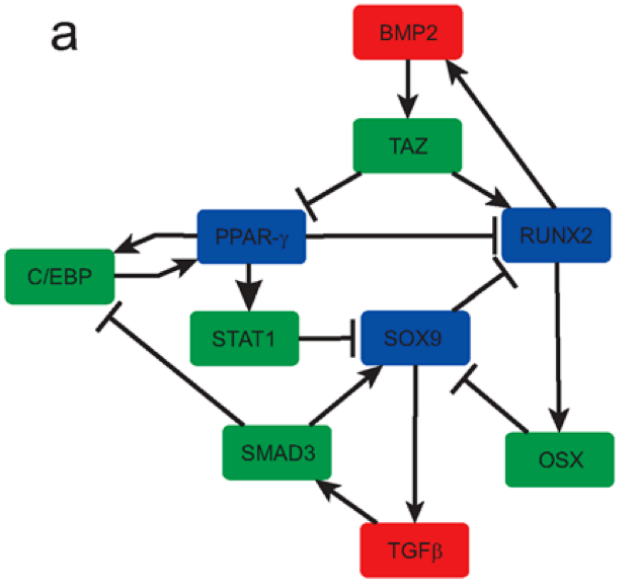
Generated from microarrays on a panel of healthy normal tissues

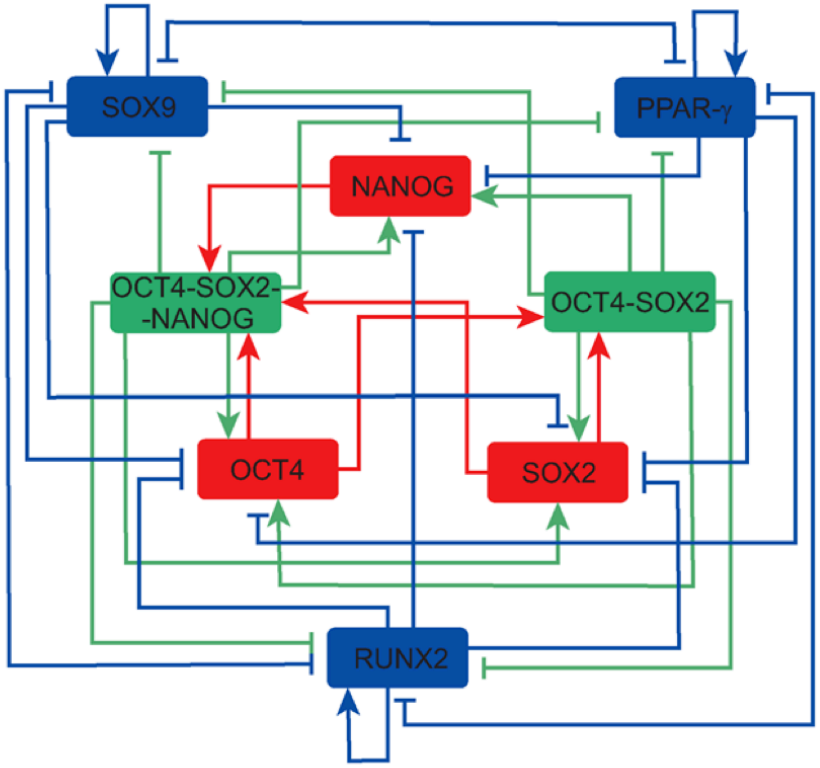
3327 gene-nodes and 15841 coexpression-links

PLoS One. 2008

Gene Regulatory Network

Stem cell differentiation regulation

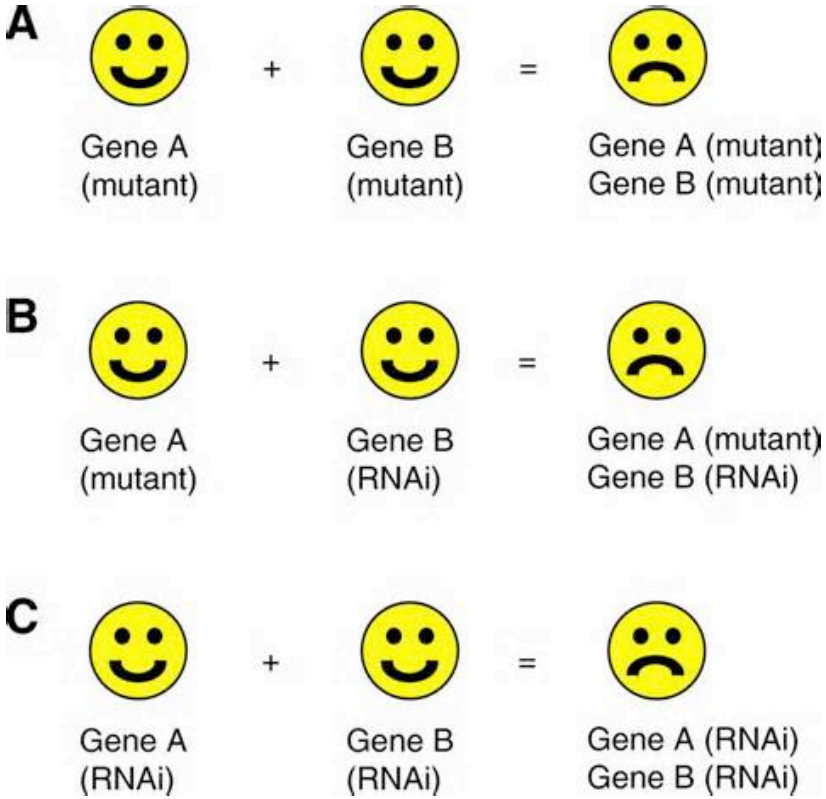


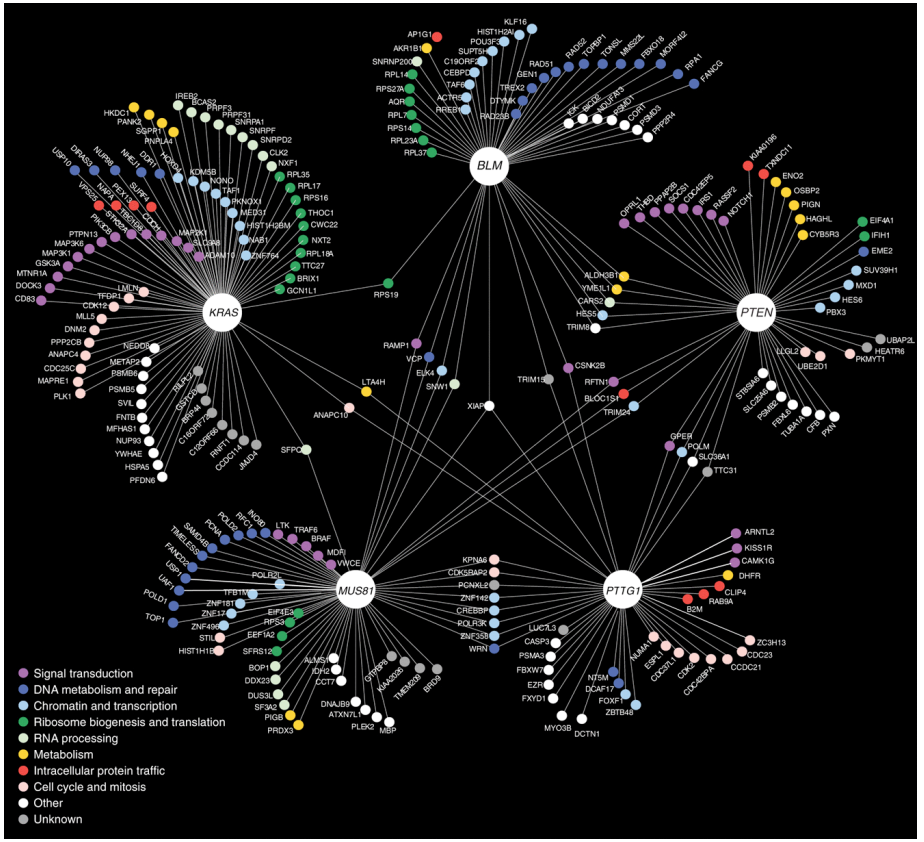


- Nodes are genes and transcription factors
- Interactions can be directional or bidirectional
- Interactions can be activation or inhibition

MacArthur et al., *PLoS ONE* 3: e3086 (2008)

Genetic Interaction Network

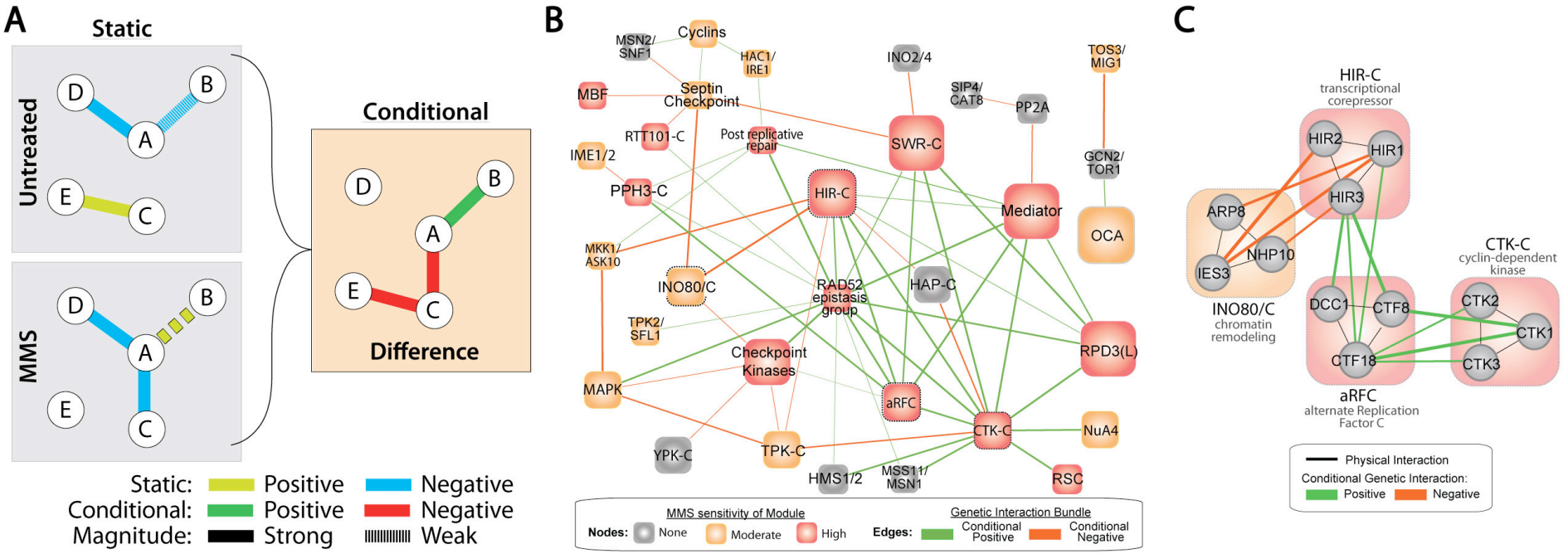




Vizeacoumar et al Molecular Systems Biology (2013)

Application: Identify possible synthetic lethal interactions

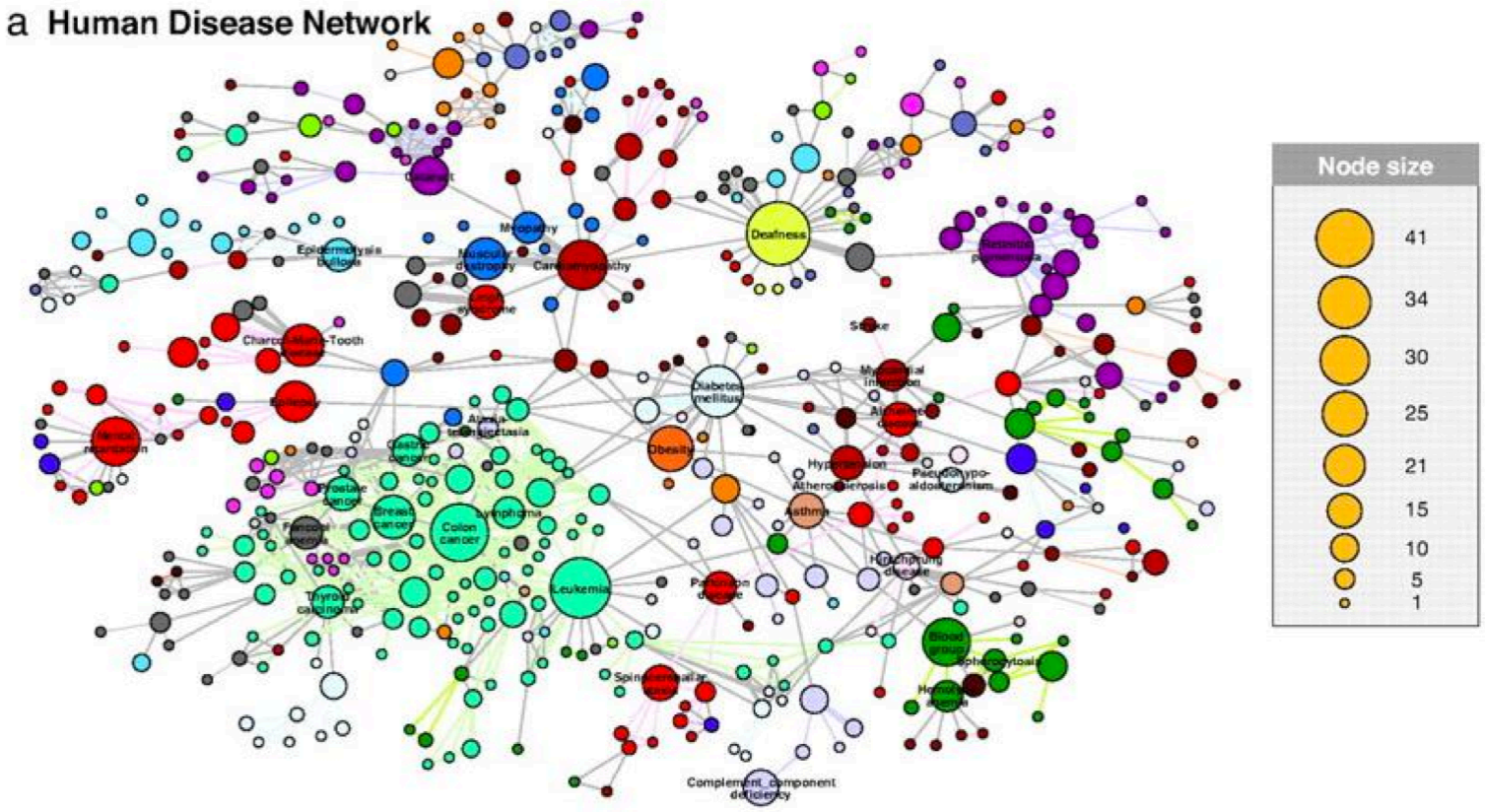
Difference Networks



Edges represent changes between different experimental conditions

Application: Compare how two distant species respond to a common stimulus

Diseaseome

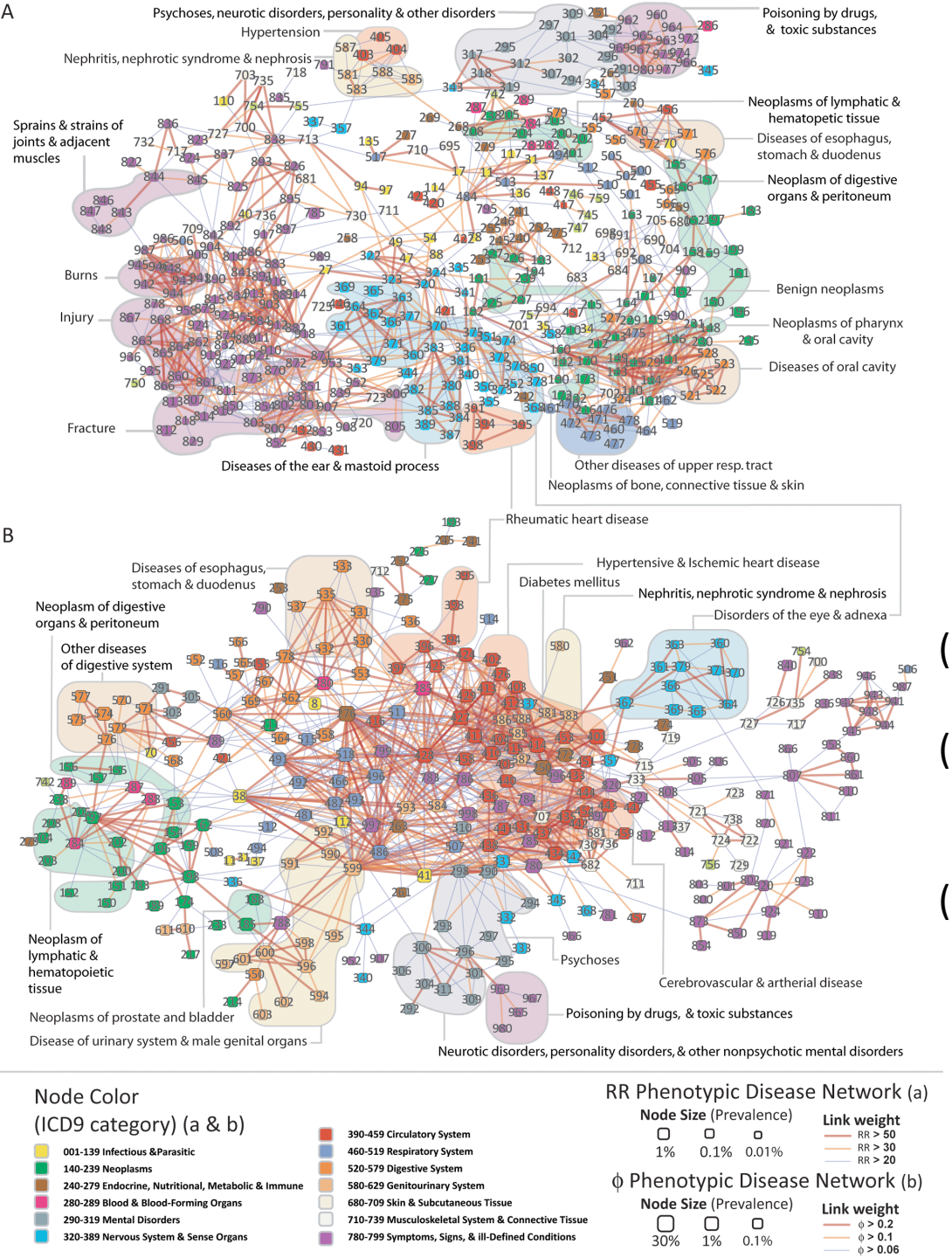


Goh et al. Proc Natl Acad Sci USA. (2007) 104:8685-90

Started from a bipartite graph

- nodes represent genes and diseases
- edges connect genes to diseases

Phenotypic Disease Networks (PDNs).



correlations obtained from the disease history of more than 30 million patients

(1) patients develop diseases close in the network to those they already have
(2) the progression of disease is different for patients of different genders and ethnicities;
(3) patients diagnosed with more highly connected diseases tend to die sooner

Hidalgo *et al*. PLoS Comp Bio 2009

At first glance

- Cyclic vs acyclic
  - Trees are graphs with no cycles
  - Nodes can sometimes have self loops

- Connected components
  - Set of connected nodes
  - Can have multiple connected components
  - Giant component – a component that includes the majority of nodes in the graph

Properties of the graph may have implications for various graph algorithms

Why use graphs to analyze biology?

Graph theory can provide insights into the evolution and function of biological systems

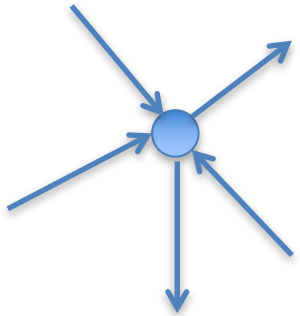
Characteristics of biological networks

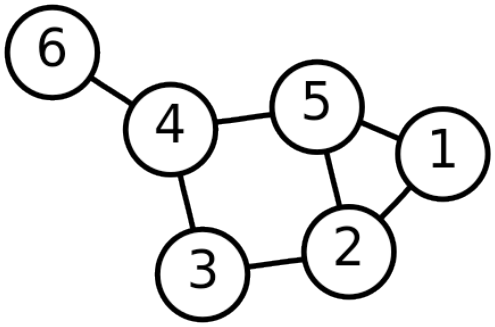
- Scale-free
  - degree distribution follows a power law

- Small-world
  - most nodes can be reached from all other nodes by traversing a small number of edges

Relevant graph variables

- Degree
  - The number of edges incident on a node
  - In a directed graph can be divided into in-degree and out-degree
  - High degree nodes called ‘hubs’





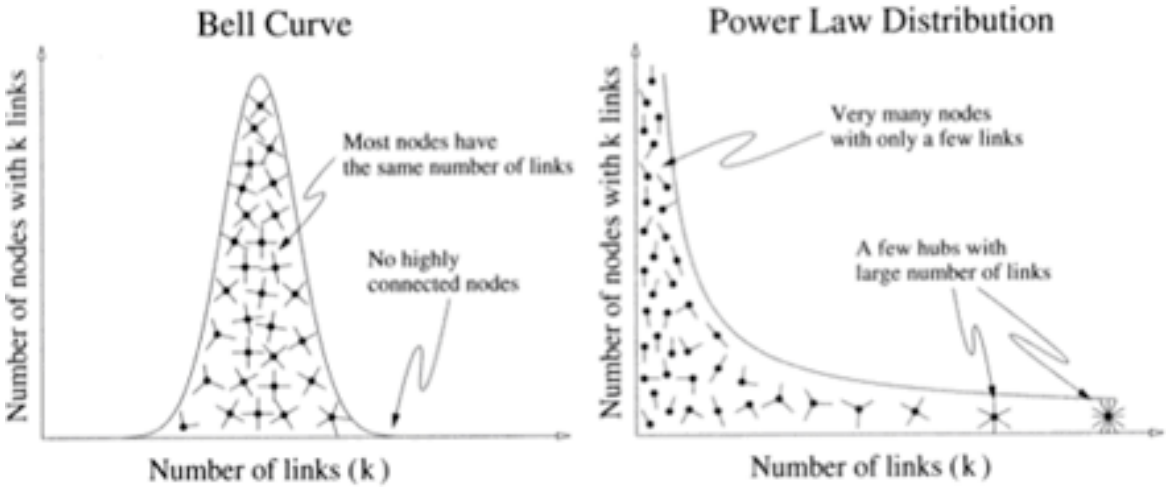
Scale Free Networks

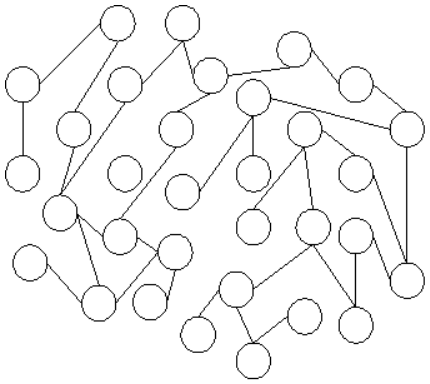
Implications for error tolerance:

P(deg=k) - k more likely to be higher in random network

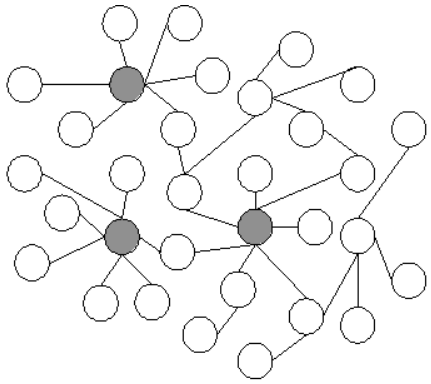
More tolerant to random failures

“Attacks” targeting higher degree nodes more likely to be disruptive to function





(a) Random network



(b) Scale-free network

Biological Networks have Hubs

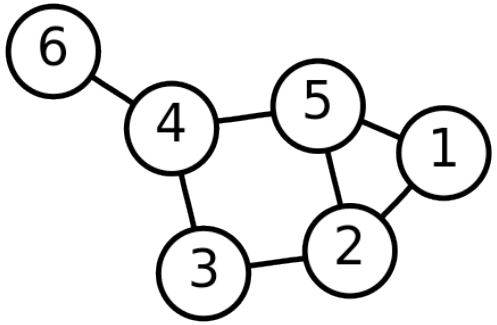
- PPI networks, regulatory networks and many other biological networks have hubs (Nodes of high degree)

- PPI network hubs are older[1] and more essential[2] than random proteins in the network

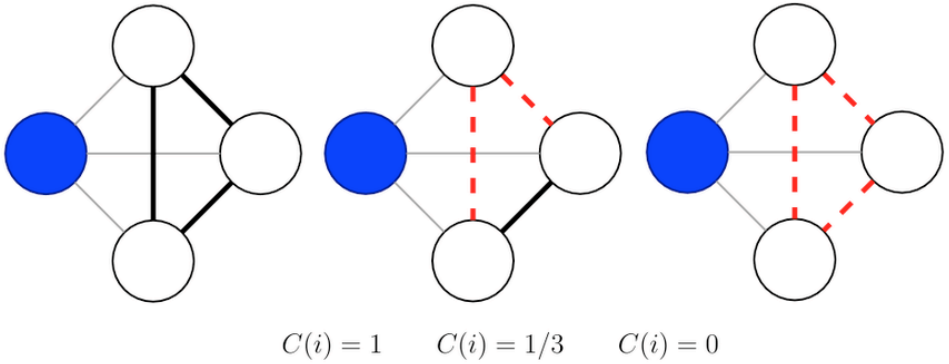
1 Fraser et al, Science (2002)
2 Jeong et al, Nature (2001)

Relevant graph variables

- Diameter
  - Longest shortest path
- Path
  - A walk in which no vertex occurs more than once
  - Shortest path



- Clustering Coefficient
  - A measure of how densely connected a graph is
  - Can calculate node-wise or graph-wise CC



Relevant graph variables

- Clustering Coefficient
  - How connected are my neighbors?
  - Count triangles

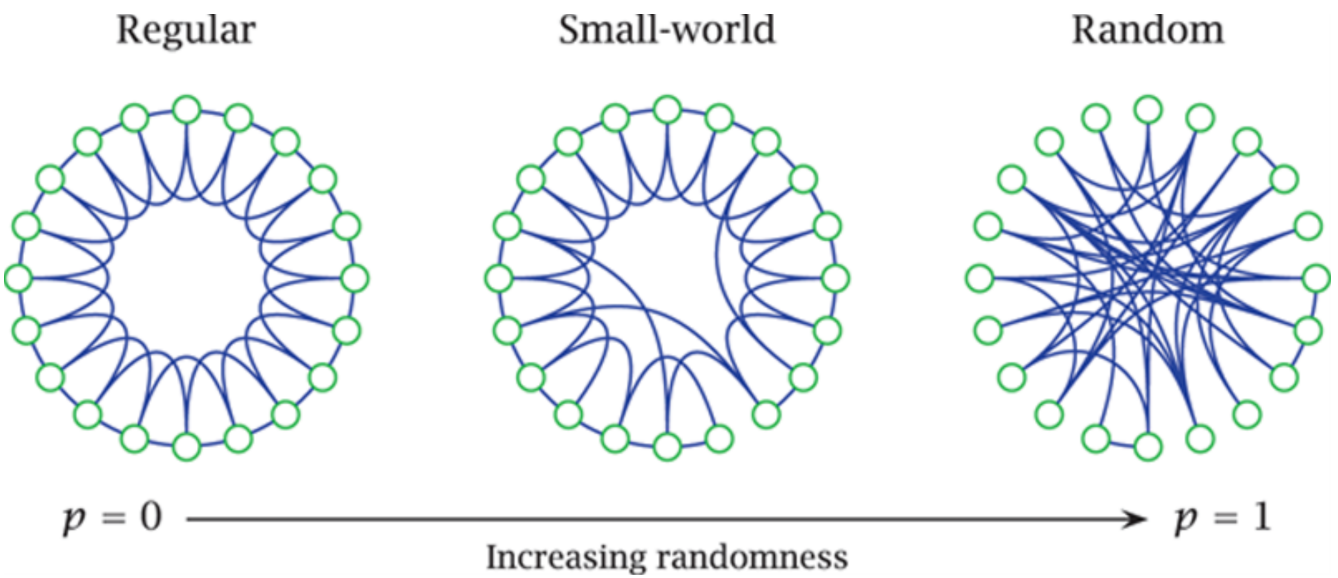
$$C(v) = \frac{\#\,triangles - containing - v}{\deg ree(v) * \frac{(\deg ree(v) - 1)}{2}}$$

Denominator enumerates the possible ways a node can connect to two of its neighbors

$$C(G) = \frac{\sum_{v \in V} C(v)}{|V|}$$

Small-world Networks

- A compromise between modularity and average distance to other network nodes



High clustering coefficient
High average path length

Low clustering coefficient
Low average path length

Biological Networks often have Scale-free and Small-world Architectures

- Scale-free and small-world characteristics are derived from graph theory
- Biological systems tend to have these characteristics
- This provides clues as to selective pressures acting on biological systems architectures
  - Error tolerance
  - Compromise between vulnerability and effective information flow

Implications for network architecture for disease

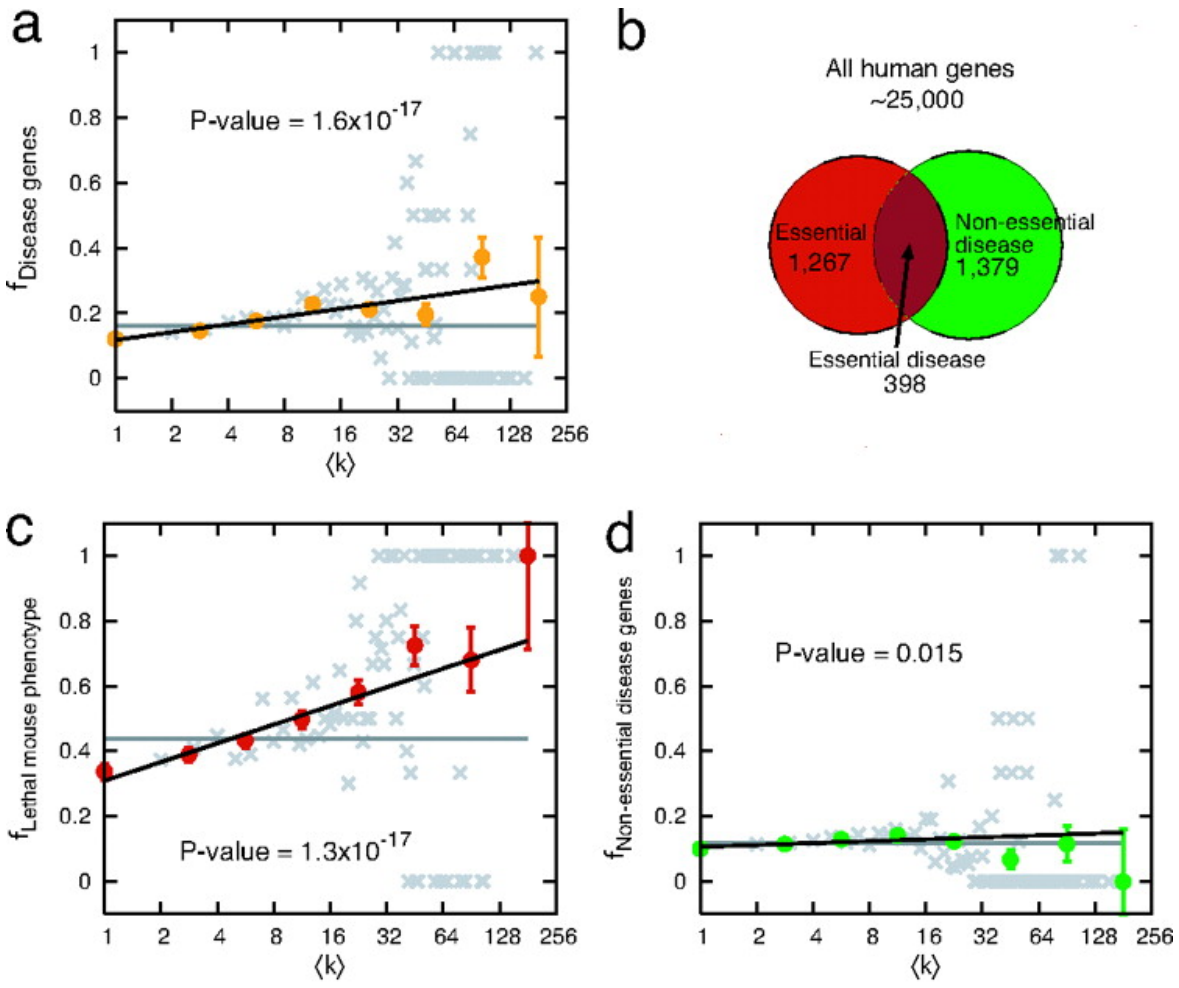
- The architecture of a graph determines the severity of a perturbation to a node or edge

- Can node/edge level features inform the potential for disease?

Degree of Disease Genes

High degree proteins more likely disease associated

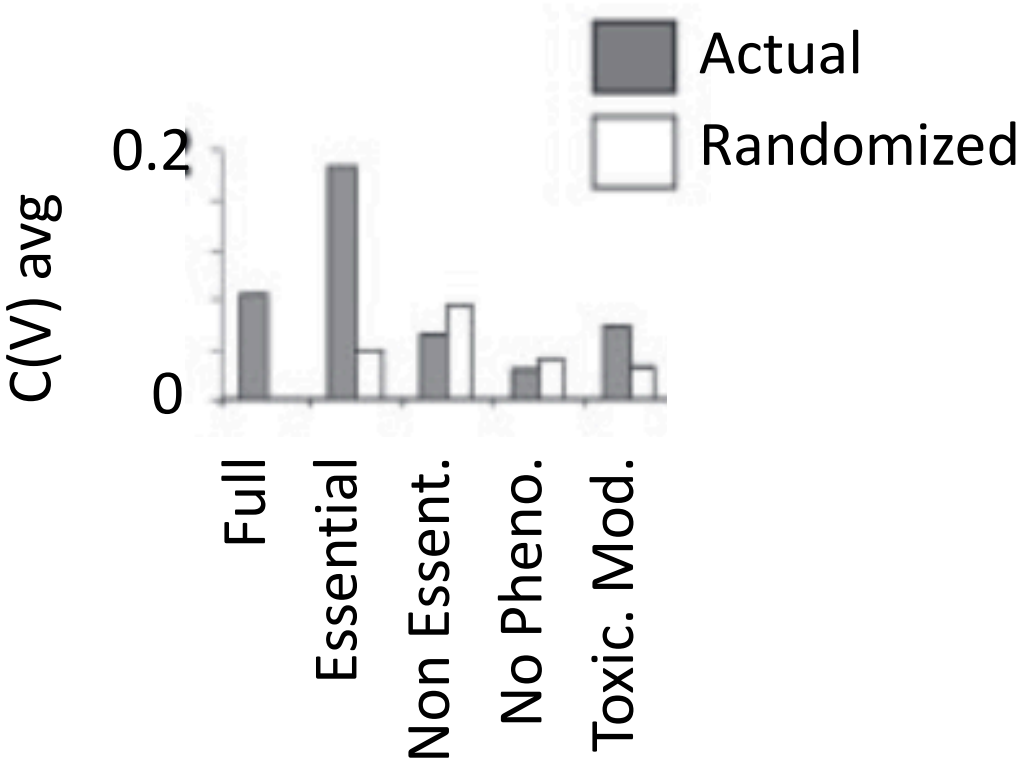
Essential genes are much more likely to be hubs



Kwang-Il Goh et al. PNAS 2007;104:8685-8690

Node clustering coefficient

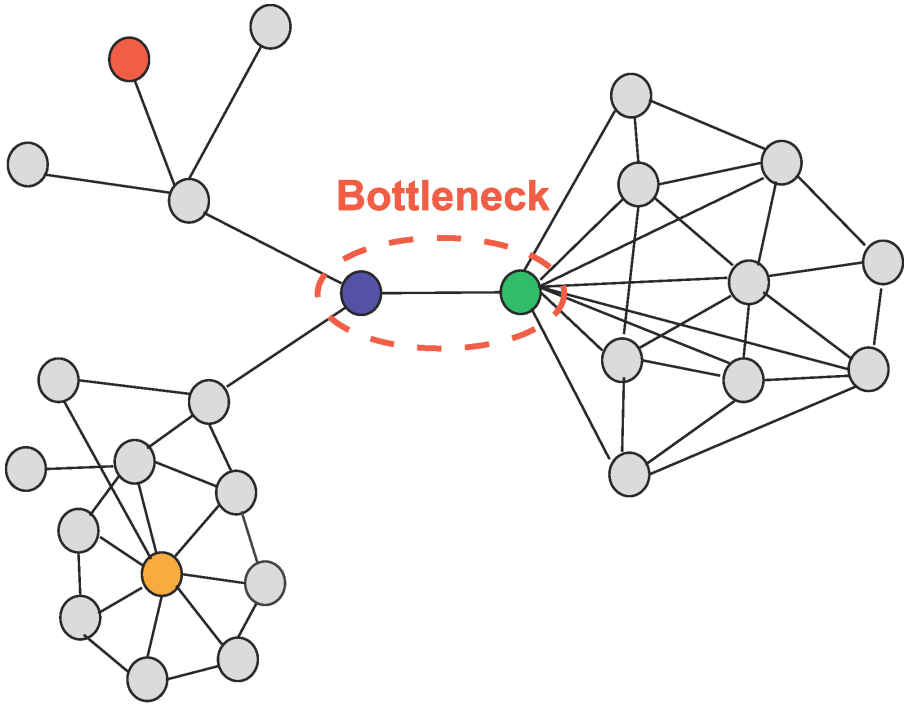
- Clustering Coefficient
  - Essential genes have higher C(v)



Said *et al* PNAS 2004

Node centrality

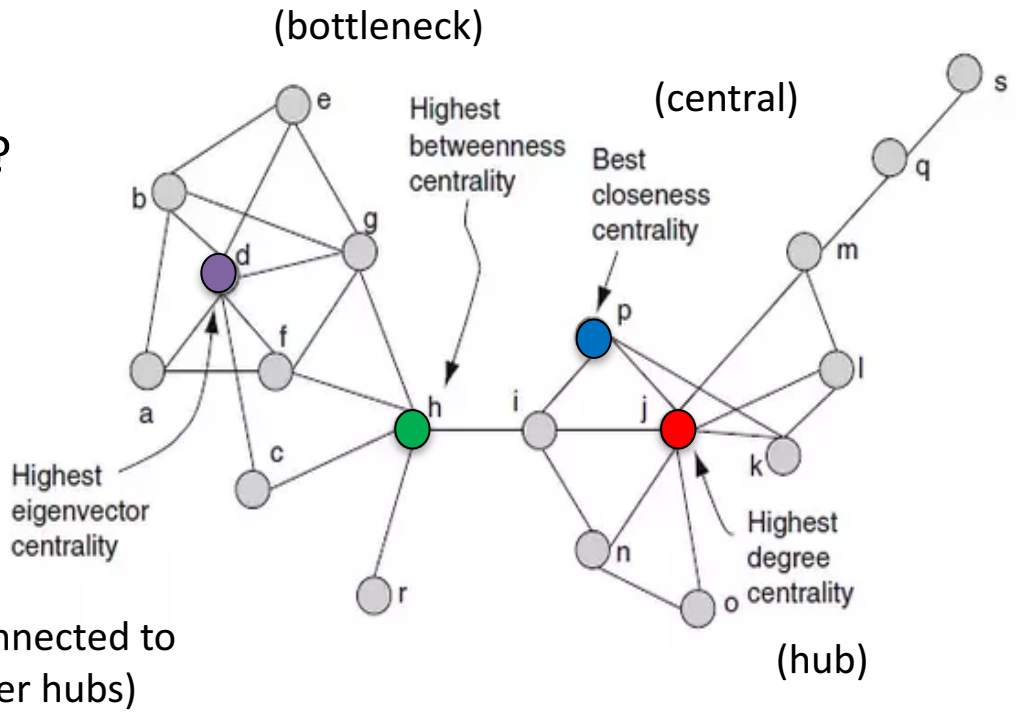
- Bottlenecks: nodes with high centrality
  - Can have low or high degree
- Centrality-Lethality Hypothesis
  - The more central a gene/protein is, the more likely it is essential



Yu et al Plos Comp Bio 2007

Network centrality features

- **Degree**: How many nodes can this node reach directly?
- **Betweenness**: How likely is this node to be the most direct route between two nodes in the network?
- **Closeness**: How fast can this node reach all nodes in the network?
- **Eigenvector**: How well is this node connected to other well-connected nodes?



Node Centrality

- Measures of centrality
  - Degree centrality
    - normalized

$$\frac{\deg(u)}{|V|-1}$$

  - Closeness centrality

$$\frac{|V|-1}{\sum_{v \neq u} length(shortest-path(u,v))}$$

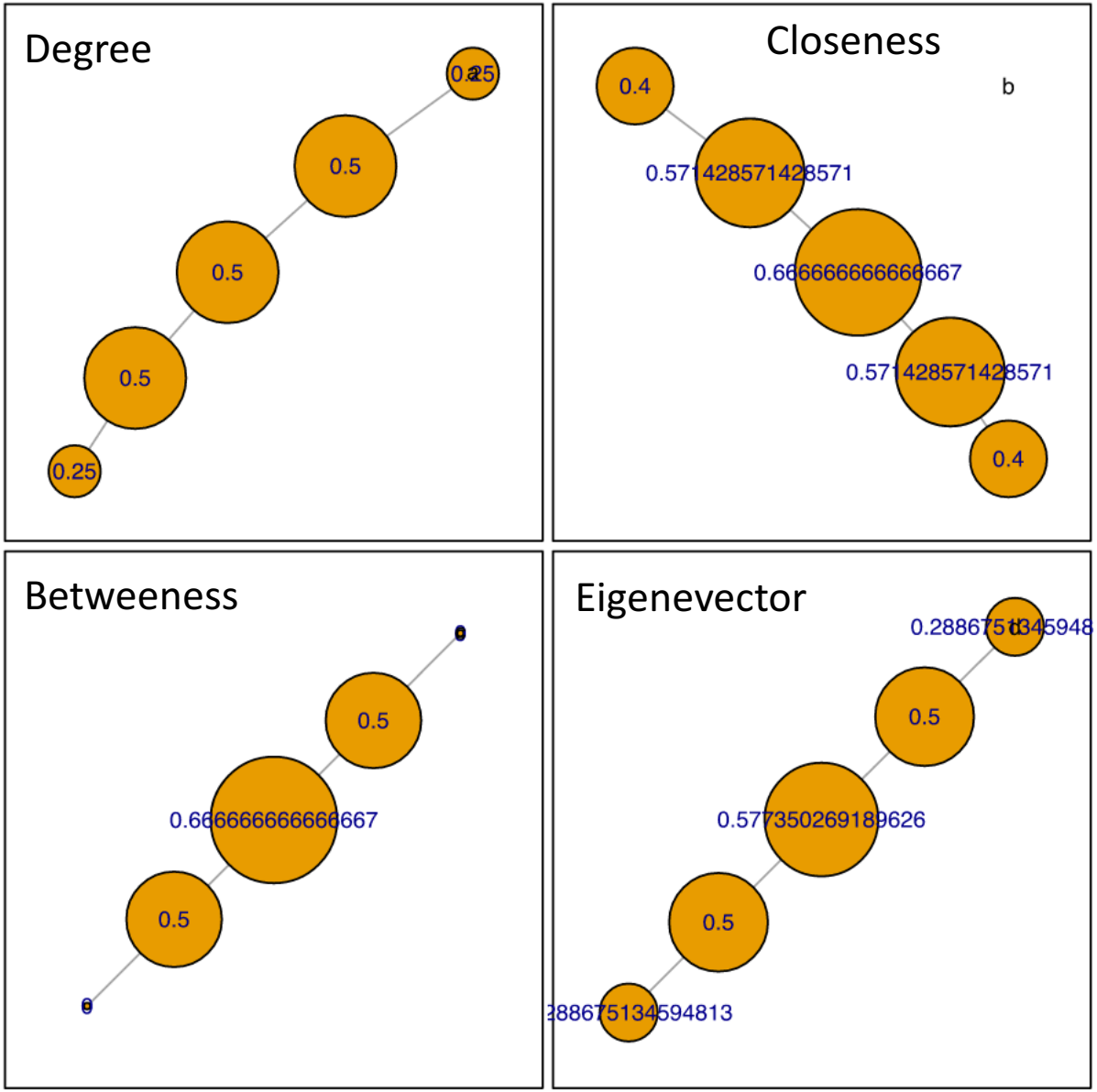
  - Eigenvector centrality
    - normalized eigenvector of first eigenvalue of the adjacency matrix

  - Between-ness centrality

$$\sum_{v,w \neq u} \frac{\#\, shortest-paths(v,w) through(u)}{\#\, shortest-paths(v,w)}$$

Node Centrality

- Centrality measures
  - Subtle differences



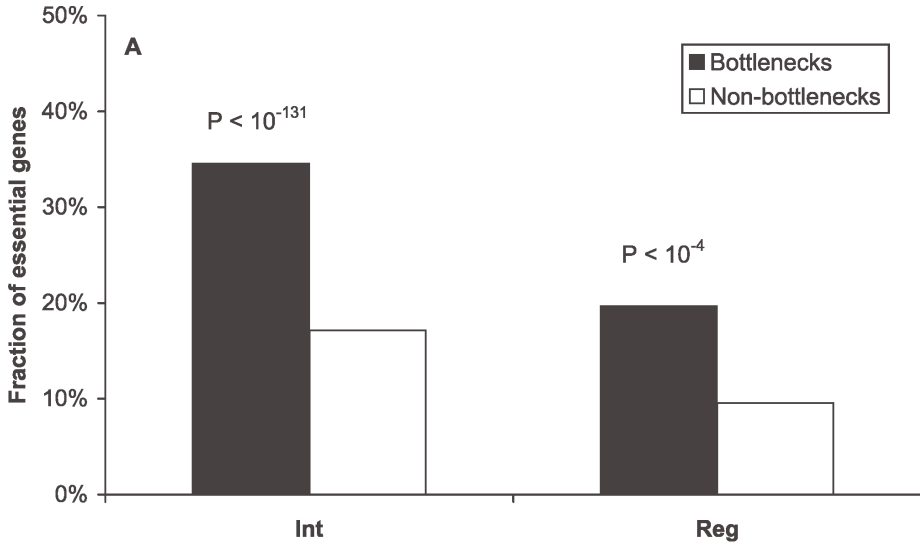
Node centrality in biology

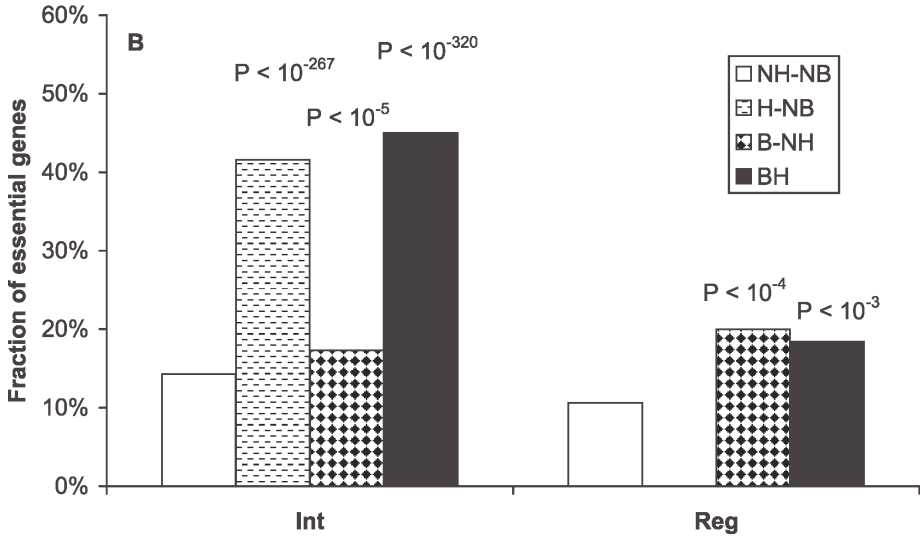
Bottlenecks: nodes with high centrality

- Can have low or high degree

Centrality-Lethality Hypothesis

- The more central a gene/protein is, the more likely it is essential

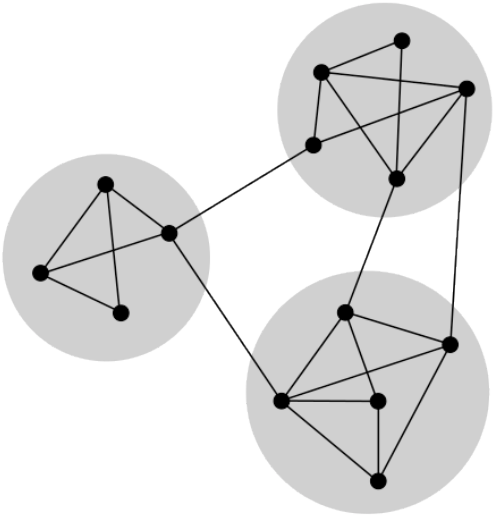




Yu et al Plos Comp Bio 2007

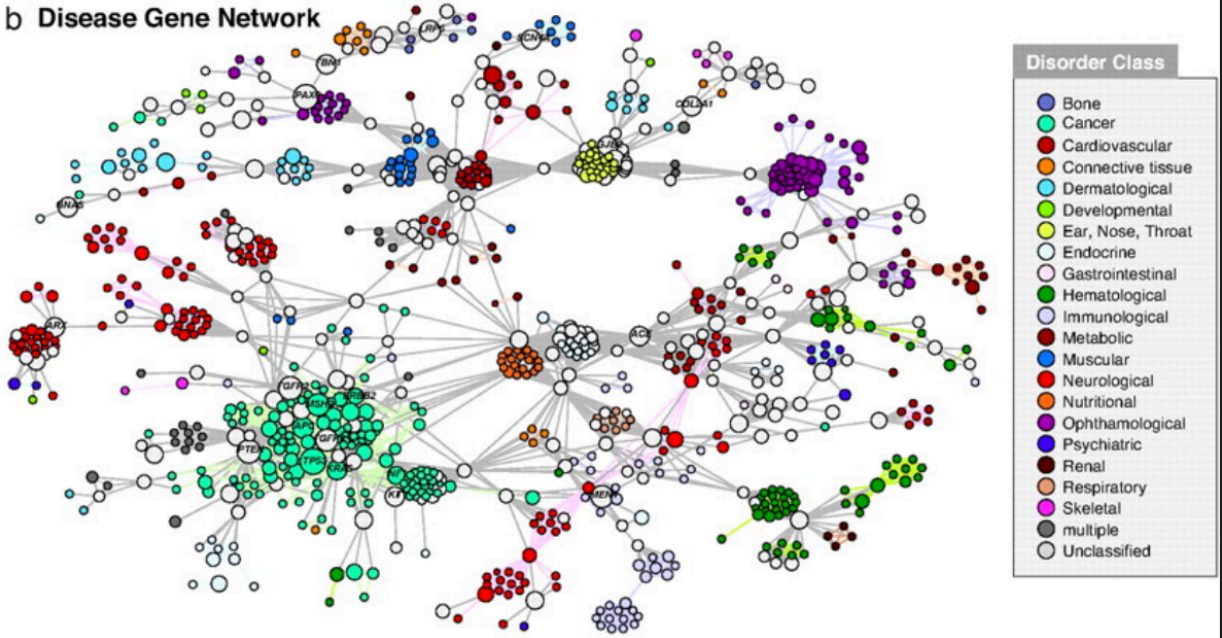
Modularity in Biological Networks

Groups of more highly connected nodes

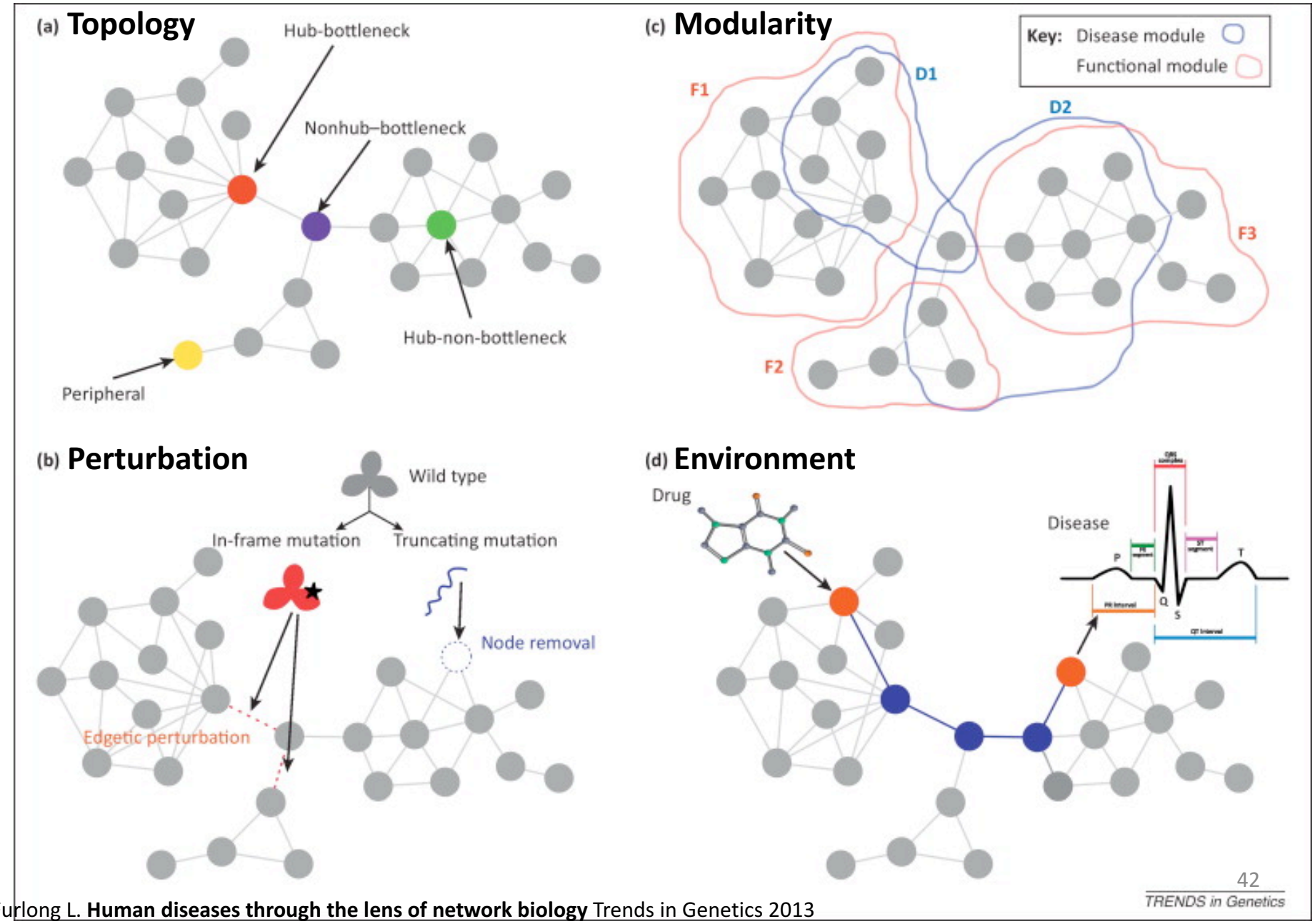


Tend to be more functionally related

Guilt-by-association



Network properties of human disease genes

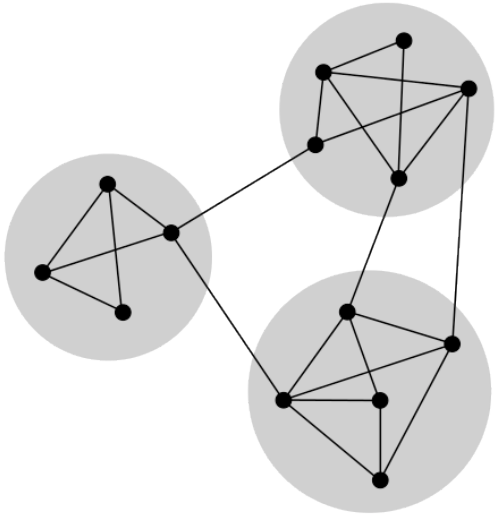


Furlong L. **Human diseases through the lens of network biology** Trends in Genetics 2013

Algorithmic and Statistical Tasks Involving Graphical Models

Algorithmic Problems on Graphs

- Clique/motif counting
- Module/community finding
- Shortest path in a weighted or directed graph
- Graph partitioning



Statistical Problems on Graphs

- Hypothesis testing with networks
  - Determine if motif is over-represented
  - Compare behavior of groups of nodes
  - Test randomness of spatial distributions on the network
  - Compare properties of two networks

- Theoretical null models often not available
  - Need background of random expectation
  - "observations" or scores in network data are not "independent" samplings from populations
  - Violates assumptions of many statistical tests

Random Graphs

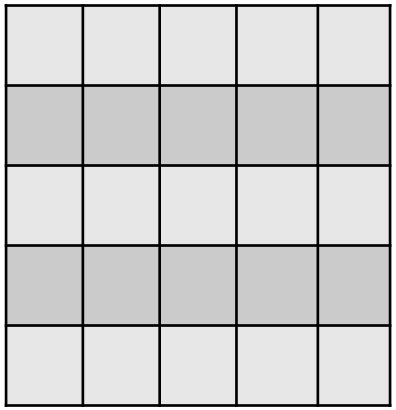
- Comparing biological networks to random networks can provide biological insight

- Original Erdos-Renyi (not widely used)
  - Given n vertices, all possible graphs with k edges – choose a graph at random
  - Exactly k edges

- Erdos-Renyi (1960) / Gilbert (1959)
  - Given n vertices, two vertices are connected by an edge with a probability p
  - On average k edges, with the actual number following a probability distribution

Random graphs: Preserving underlying sub-structure

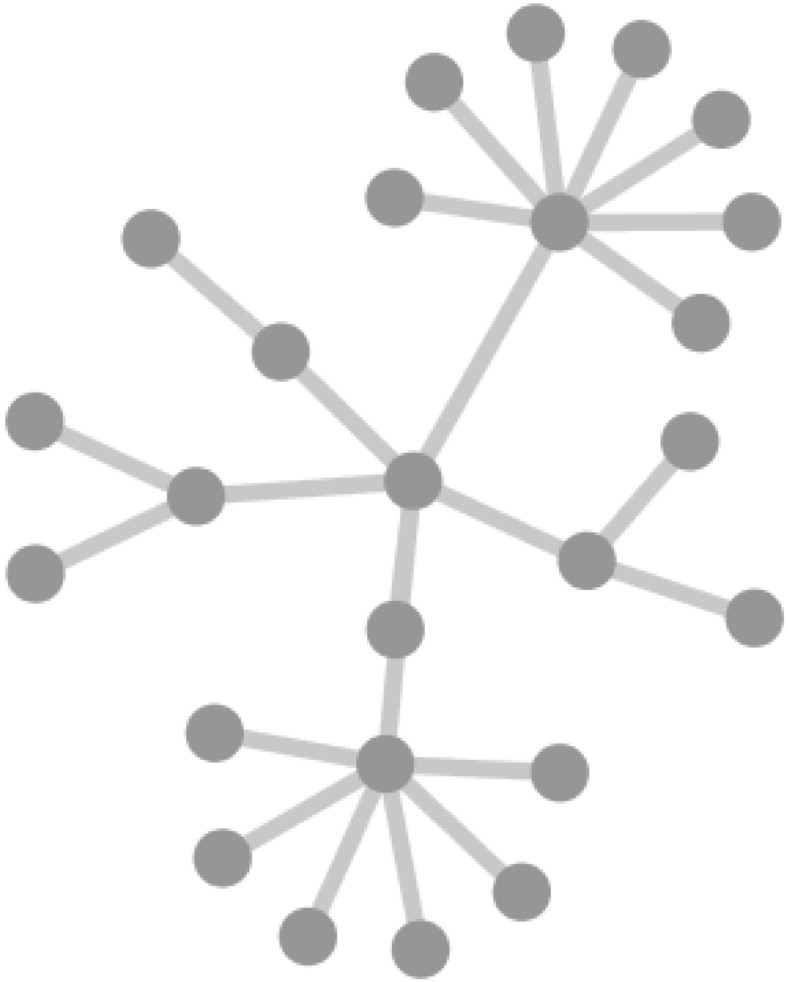
- Degree-preserving Permutation
  - Would you expect to see a statistic as extreme as *t* in a random network WITH THE SAME PROPERTIES?



Corresponds to an adjacency matrix where the row and column sums are the same.

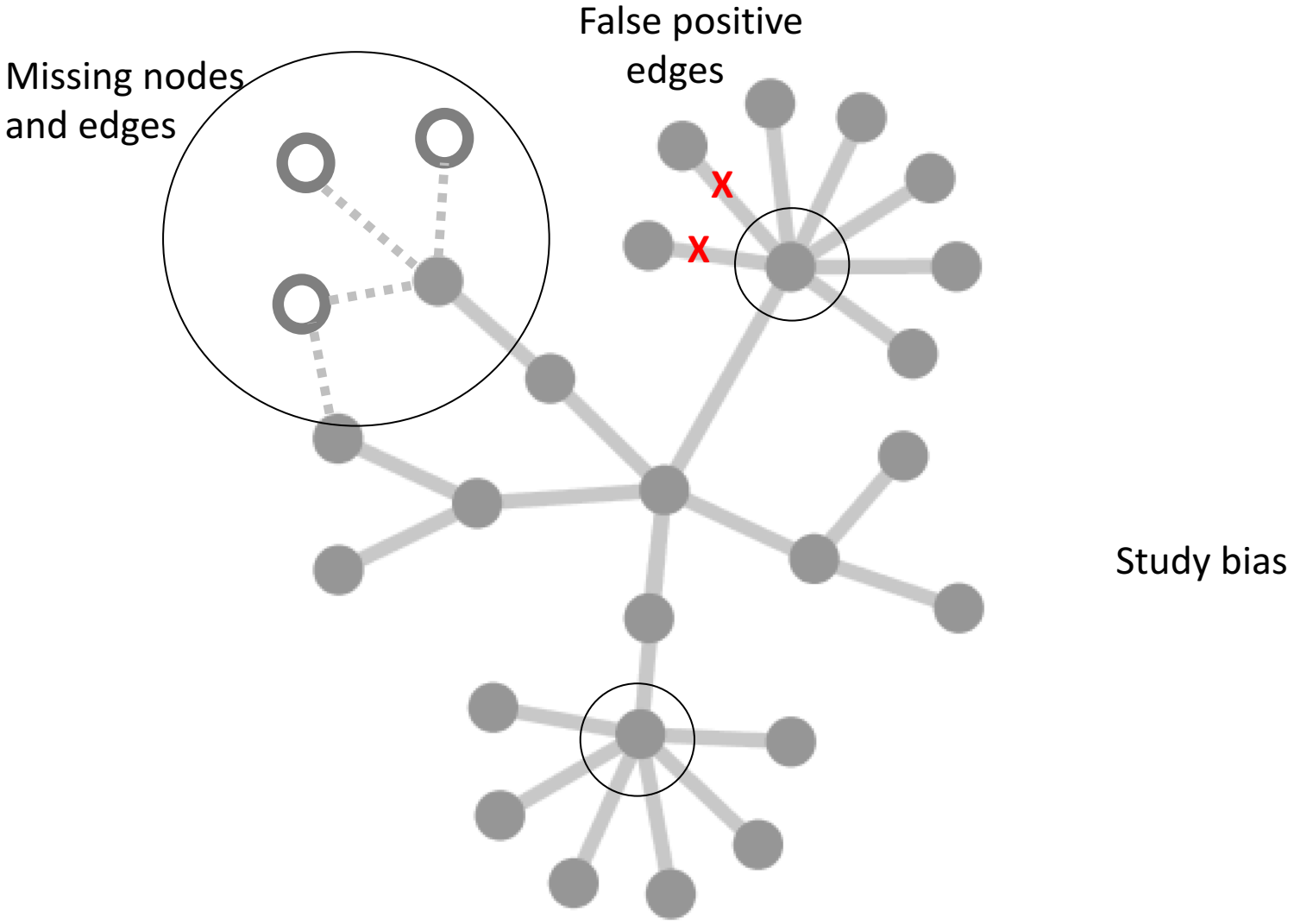
Discussion of Biological Networks in Practice

Challenges for Network Inference

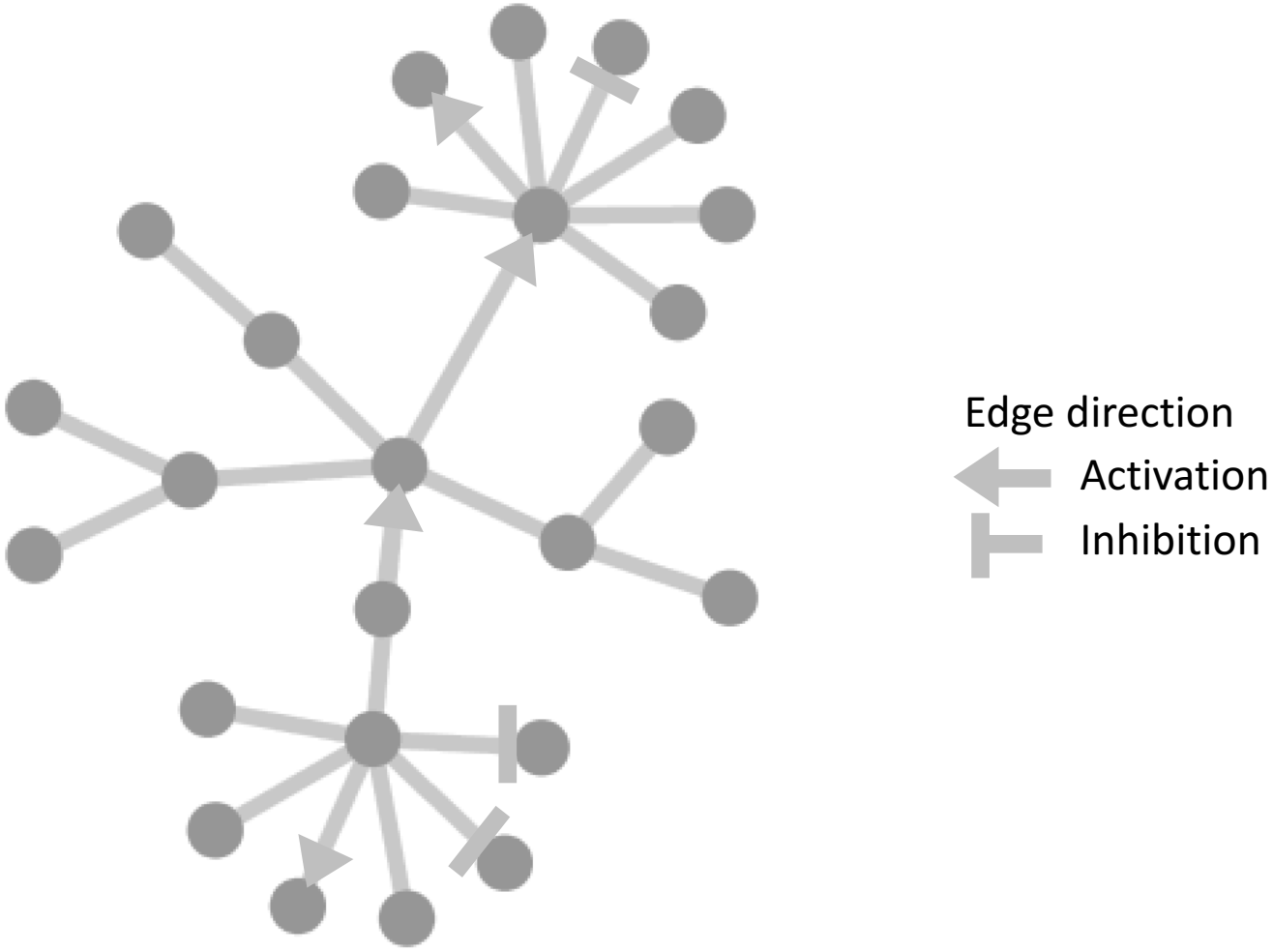


Network Type?
Protein interactions
Regulatory interactions
Genetic interactions
Metabolic interactions
PTMs …

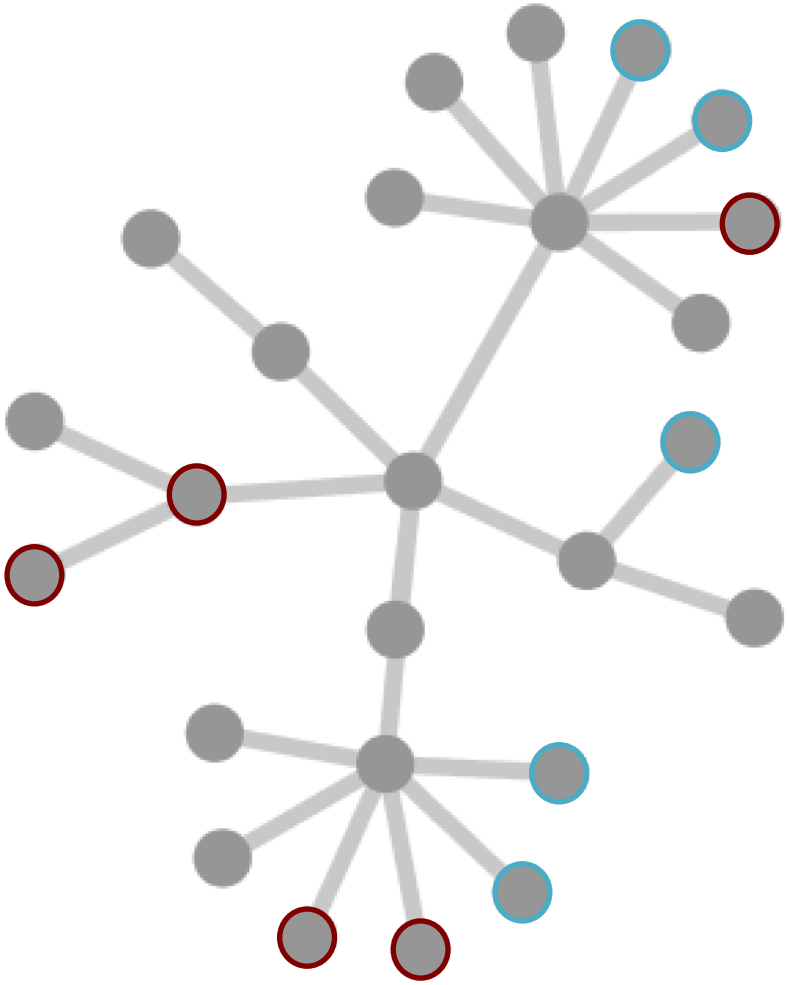
Challenges for Network Inference



Challenges for Network Inference



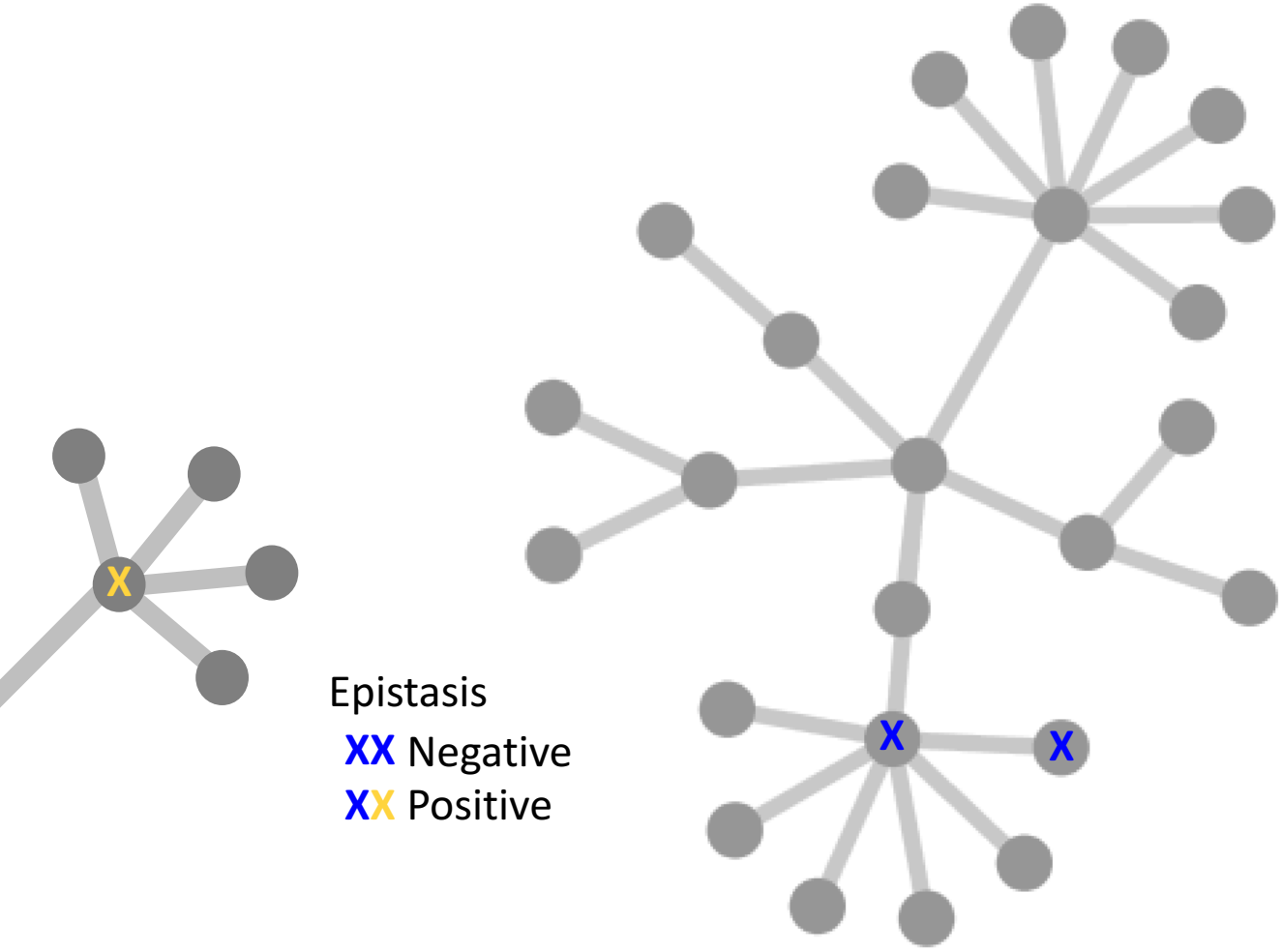
Challenges for Network Inference



Tissue specific expression

- Ubiquitous
- Tissue A
- Tissue B
- ? Cancer

Challenges for Network Inference

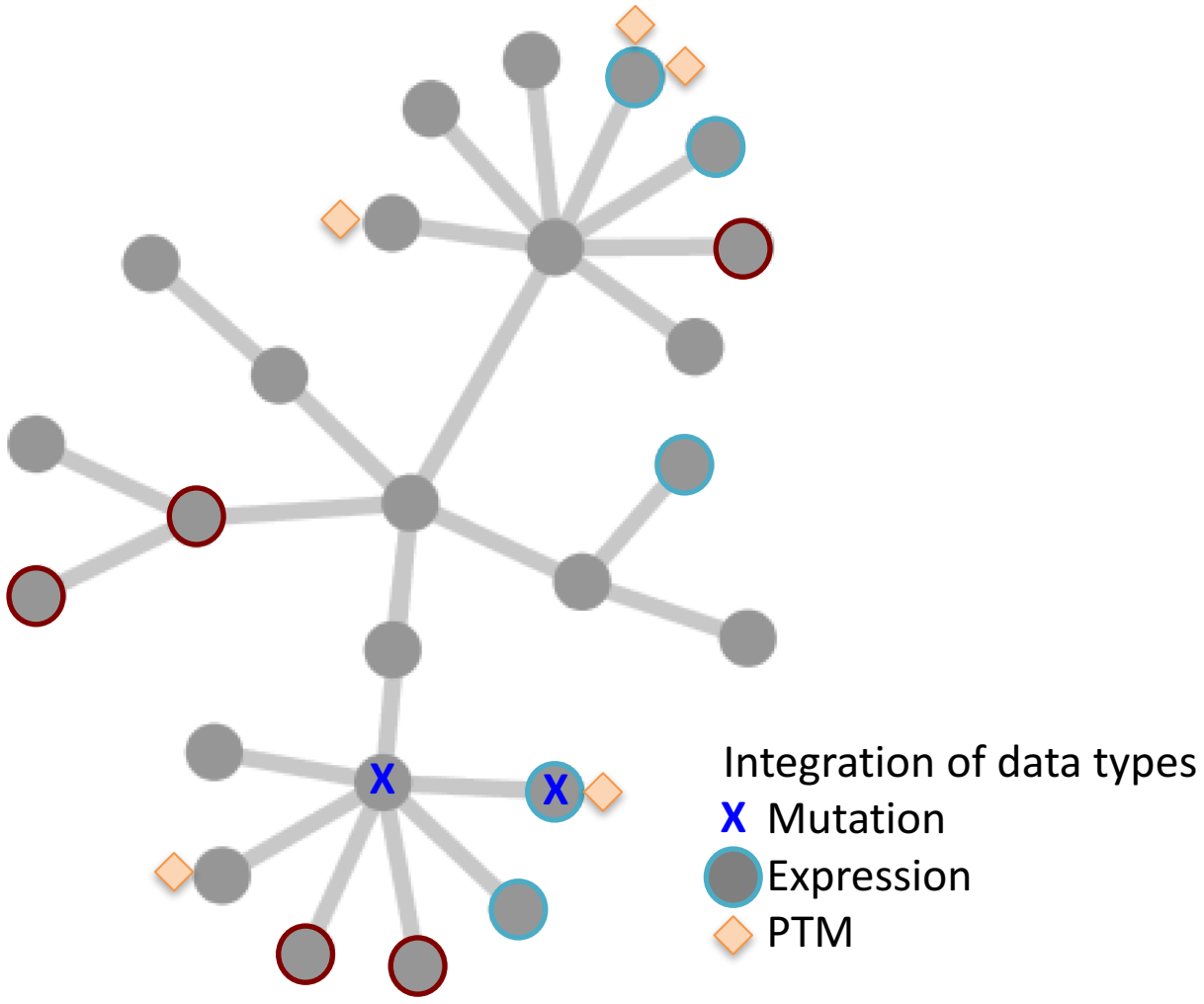


Epistasis

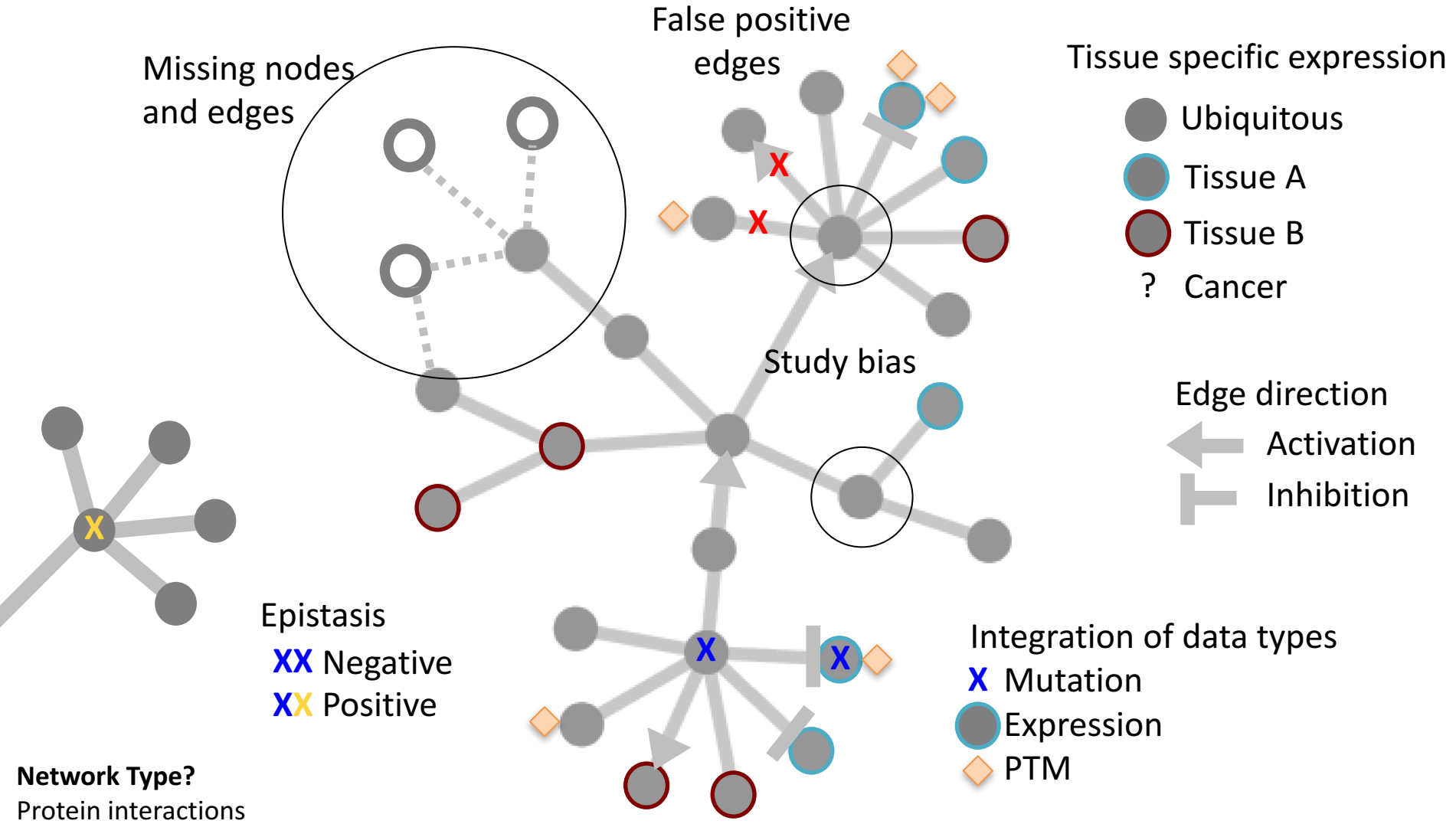
XX Negative

XX Positive

Challenges for Network Inference

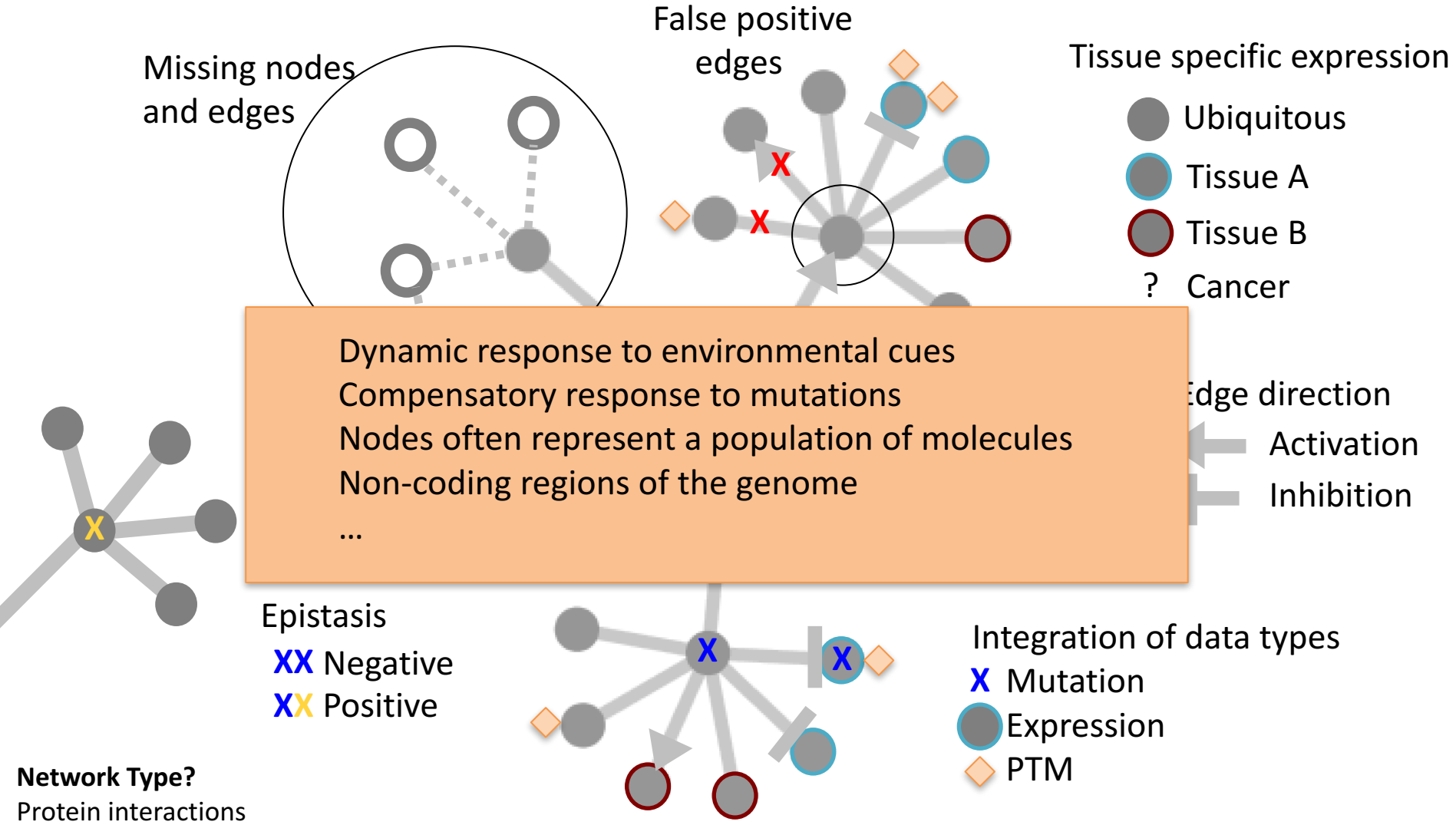


Challenges for Network Inference



Network Type?
Protein interactions
Regulatory interactions
Genetic interactions
Metabolic interactions
PTMs …

Challenges for Network Inference

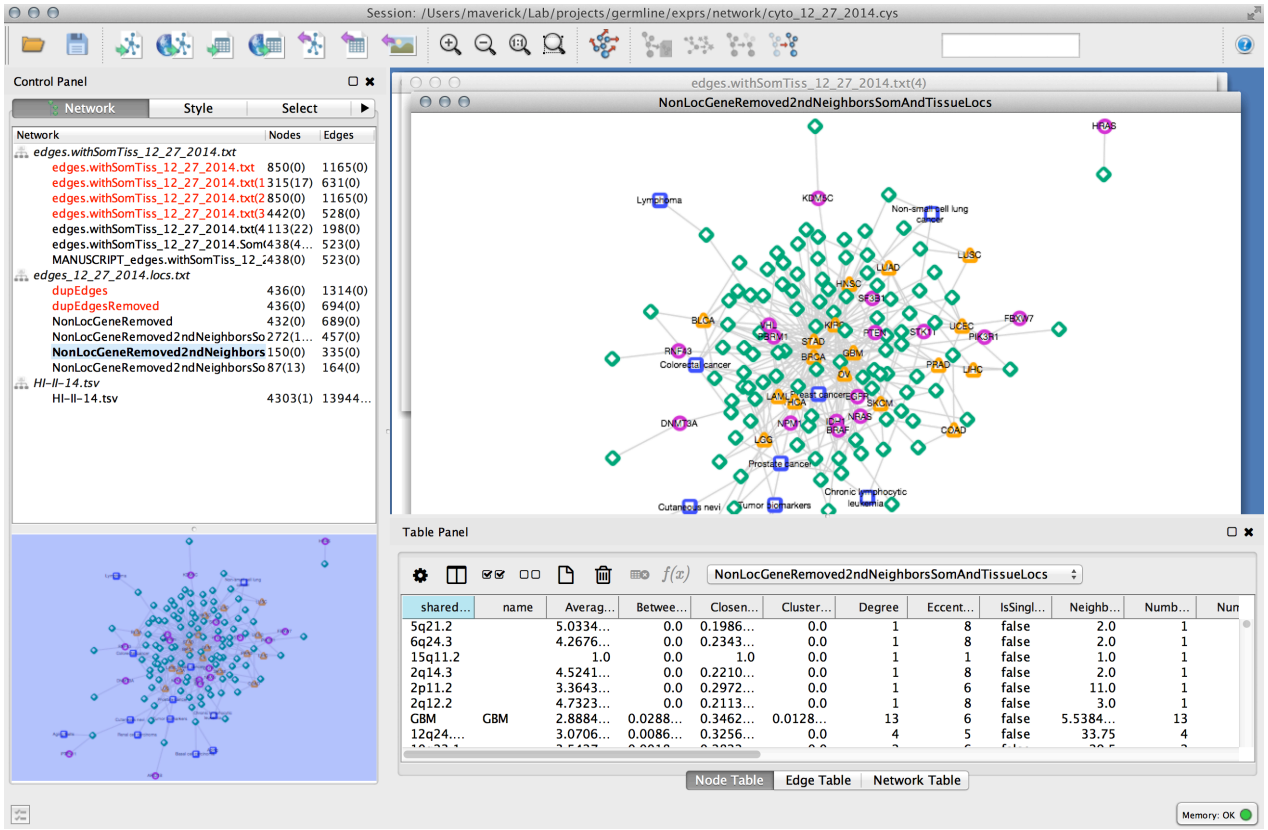


Network Type?
Protein interactions
Regulatory interactions
Genetic interactions
Metabolic interactions
PTMs …

Sources for Biological Network Data

- Literature
- Interaction databases
  - STRING
  - MINT
  - HPRD
  - MIPS
  - BIND
  - GRID
  - DIP
  - HumanNet / YeastNet
- Pathway Databases
  - Pathway Commons
  - Biogrid
  - NCBI PID
  - Reactome
  - Gene Ontology
- High-throughput experiments
  - Y2H – binary interactions
  - APMS – protein complexes

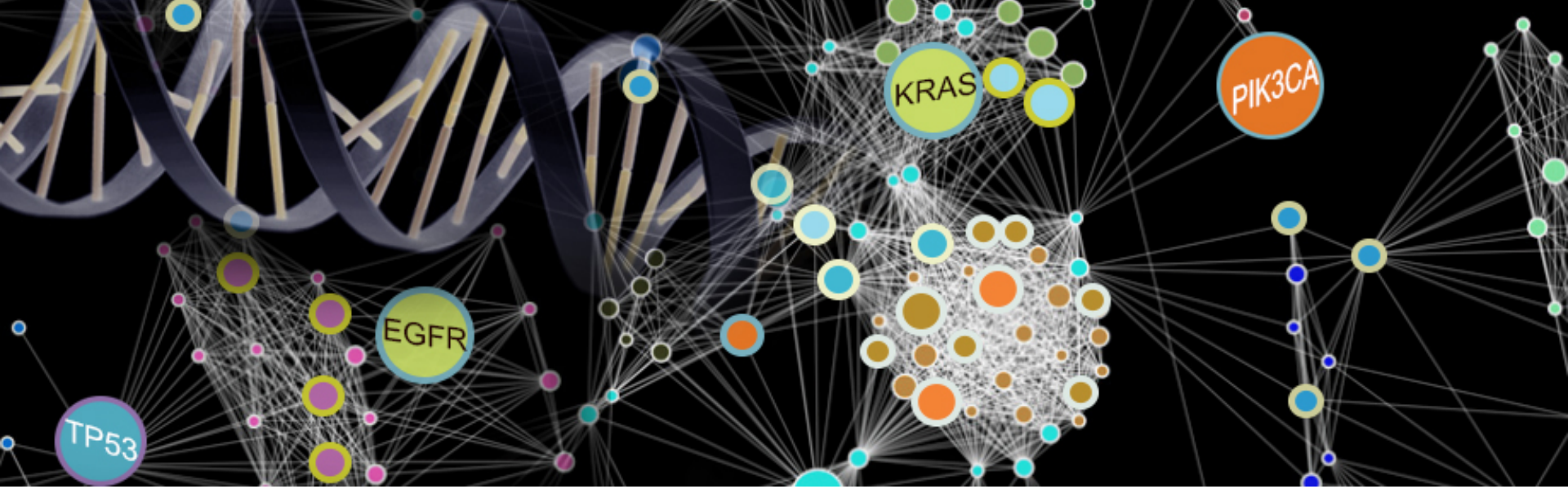
Visualizing / Analyzing Graphs



- iGraph (R)
- NetworkX (Python)

Conclusions

- Properties of biological networks can lead to new insights about the function/evolution of biological systems
- Need to carefully consider what you can and can’t infer and given the limitations of the data
- A growing area of research – new technologies enable construction of better networks



THANK YOU!

Extra material

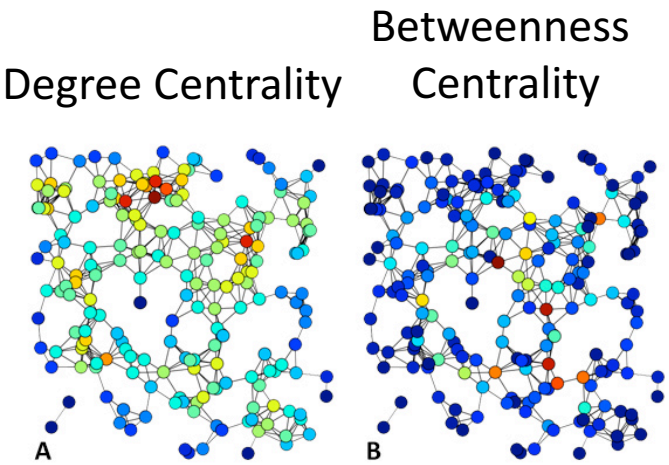
Excercises

- Jupyter notebook: Network basics
  - Reset kernel and clear all output
- Questions / comments / suggestions
  - hkcarter@ucsd.edu

Properties of Graphs

- Density
  - the extent to which a graph is complete (all possible edges present)

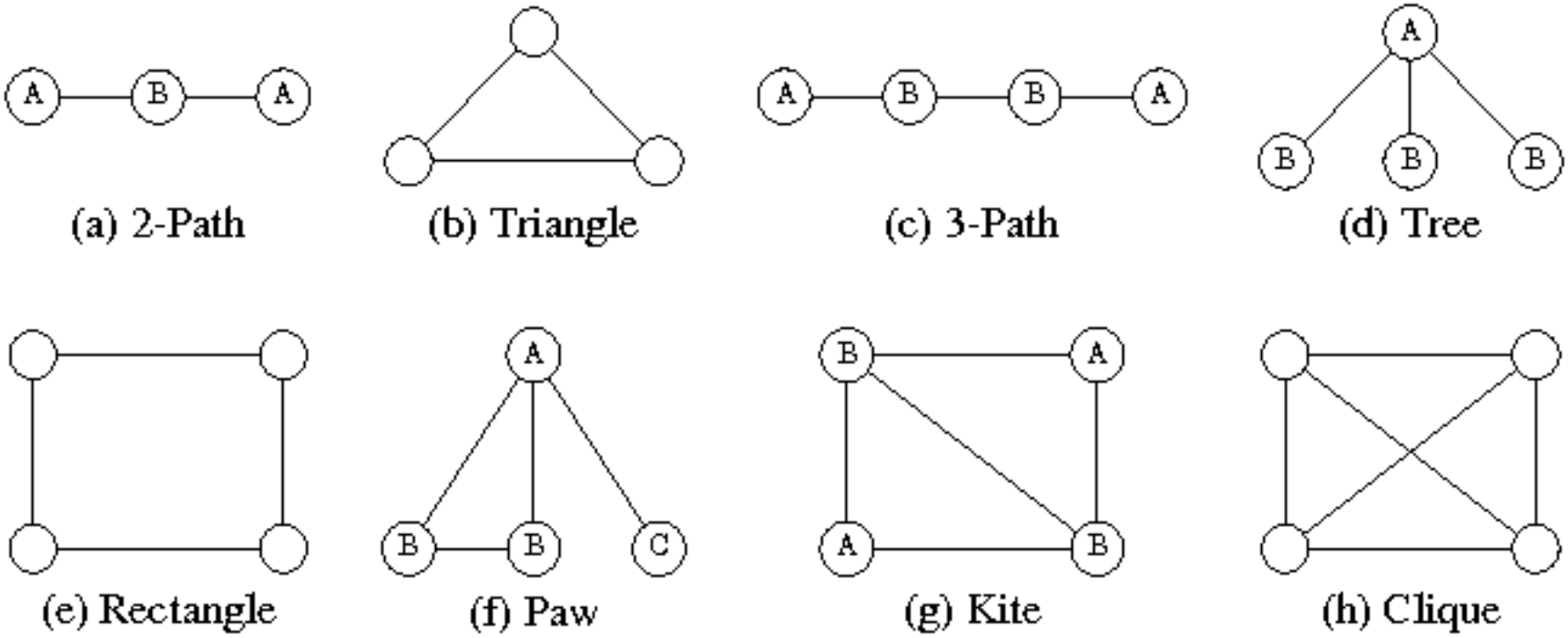
- Centrality – identifies the most "important" nodes in a graph
  - Various measures: degree centrality, betweenness centrality, eigenvector, …



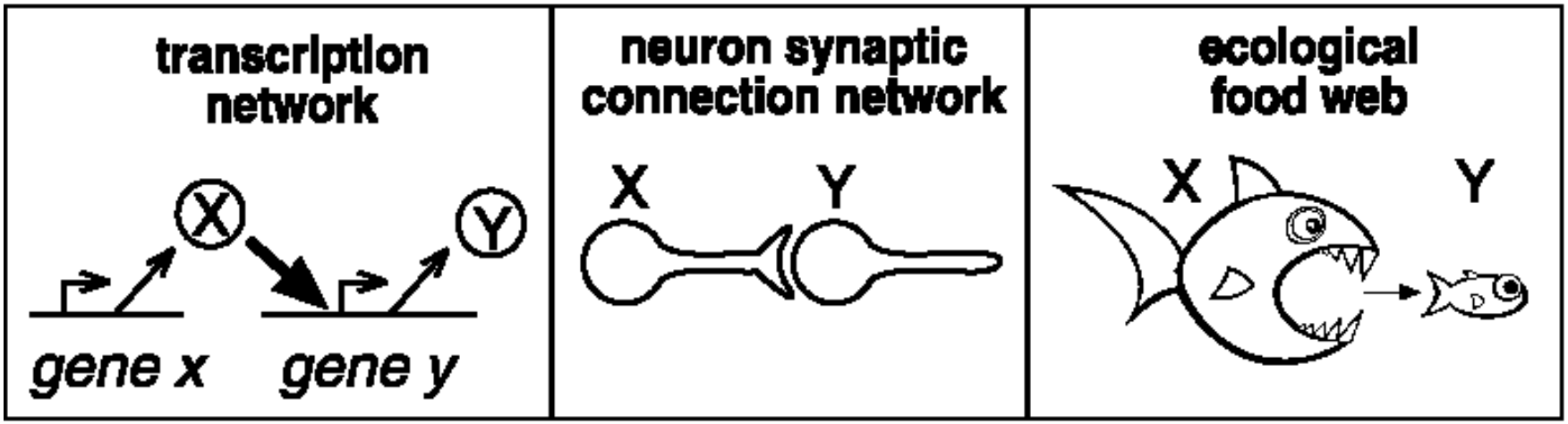
- Assortativity
  - the tendency of nodes with similar characteristics to interact

Motifs in graphs

- Motifs
  - Subgraphs that occur within a larger graph



Example: Motif searches in 3 different contexts

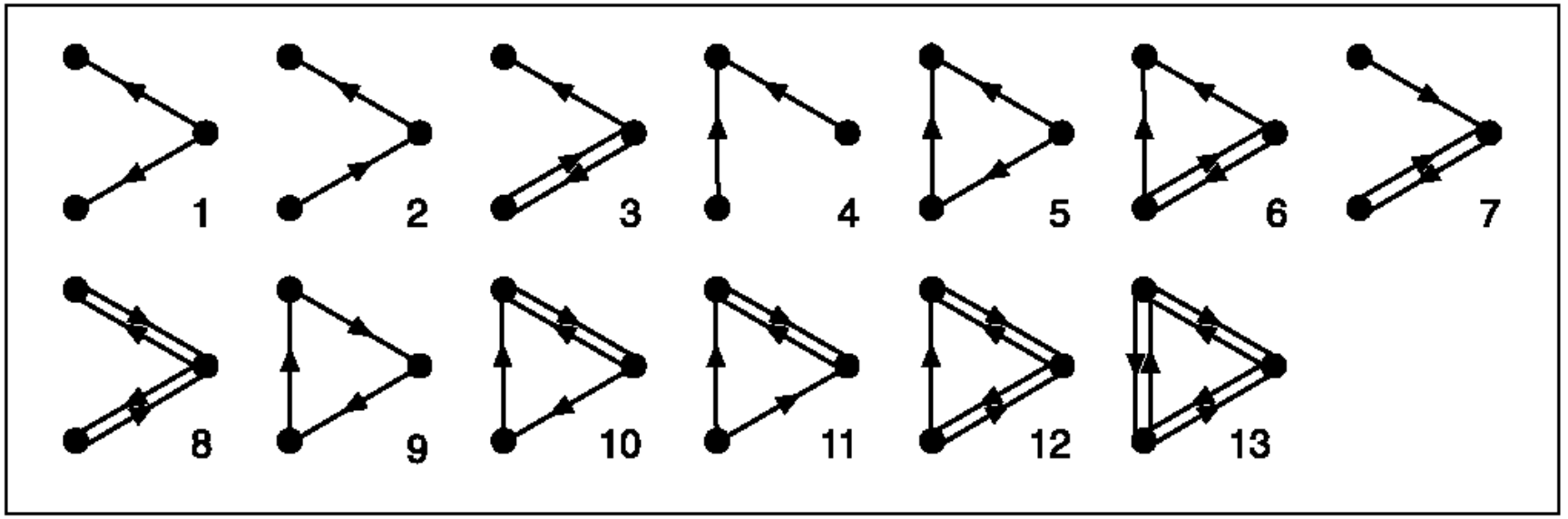


How many motifs (connected subgraph topologies) exist involving three nodes?

If the graph is undirected?

If the graph is directed?

All 3-node directed subgraphs



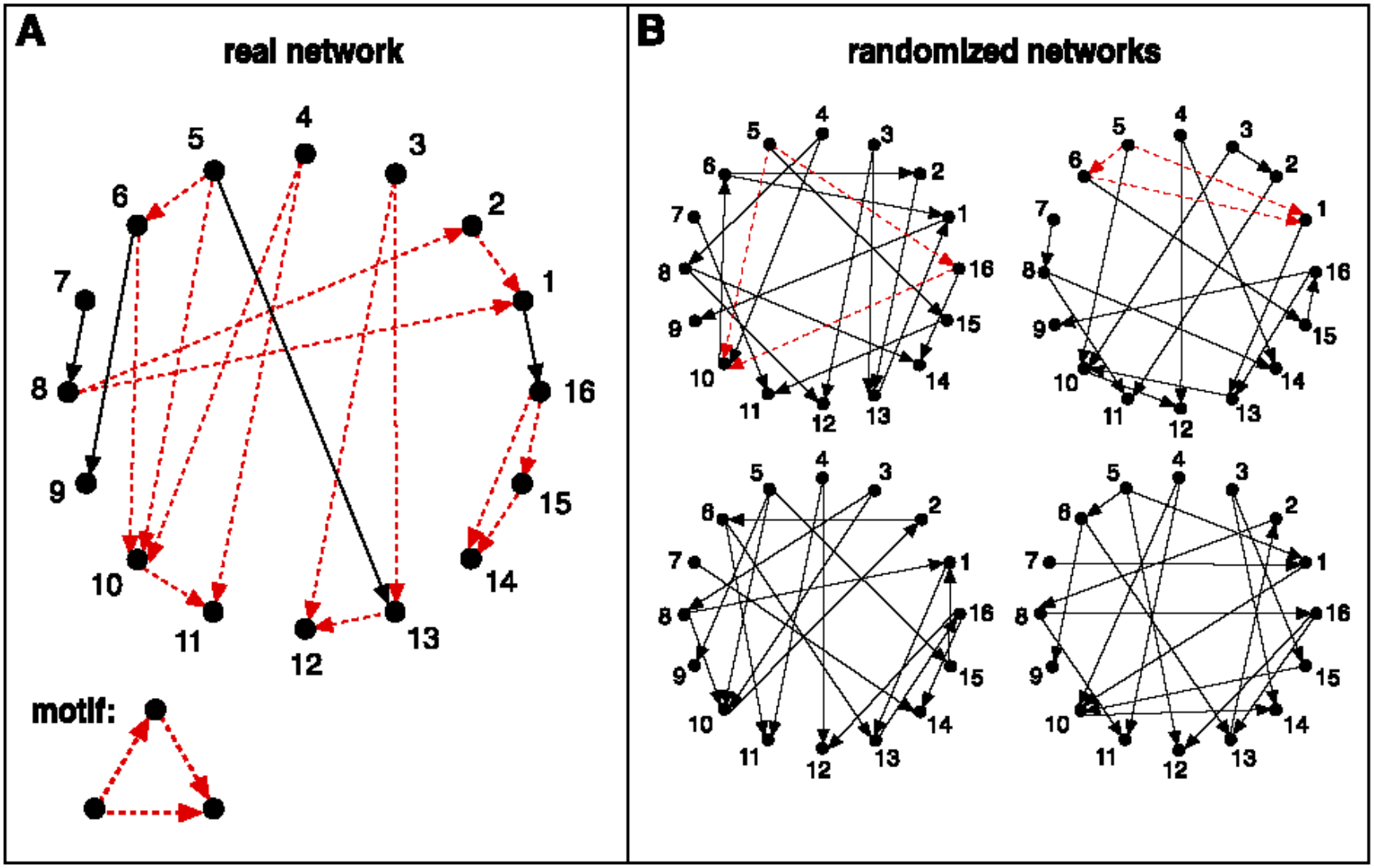
What is the frequency of each in the network?

Are any of these motifs overrepresented? (would suggest function)

Outline of the Approach

- Search network to identify all possible *n*-node connected subgraphs (here *n*=3 or 4)
- Get # occurrences of each subgraph type
- The significance for each type is determined using permutation testing, in which the above process is repeated for many randomized networks (preserving node degrees–why?)
- Use random distributions to compute a p-value for each subgraph type. The "network motifs" are subgraphs with $p < 0.001$ (why?)

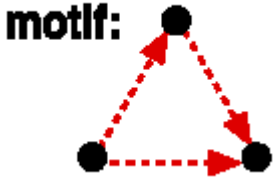
Schematic view of network motif detection

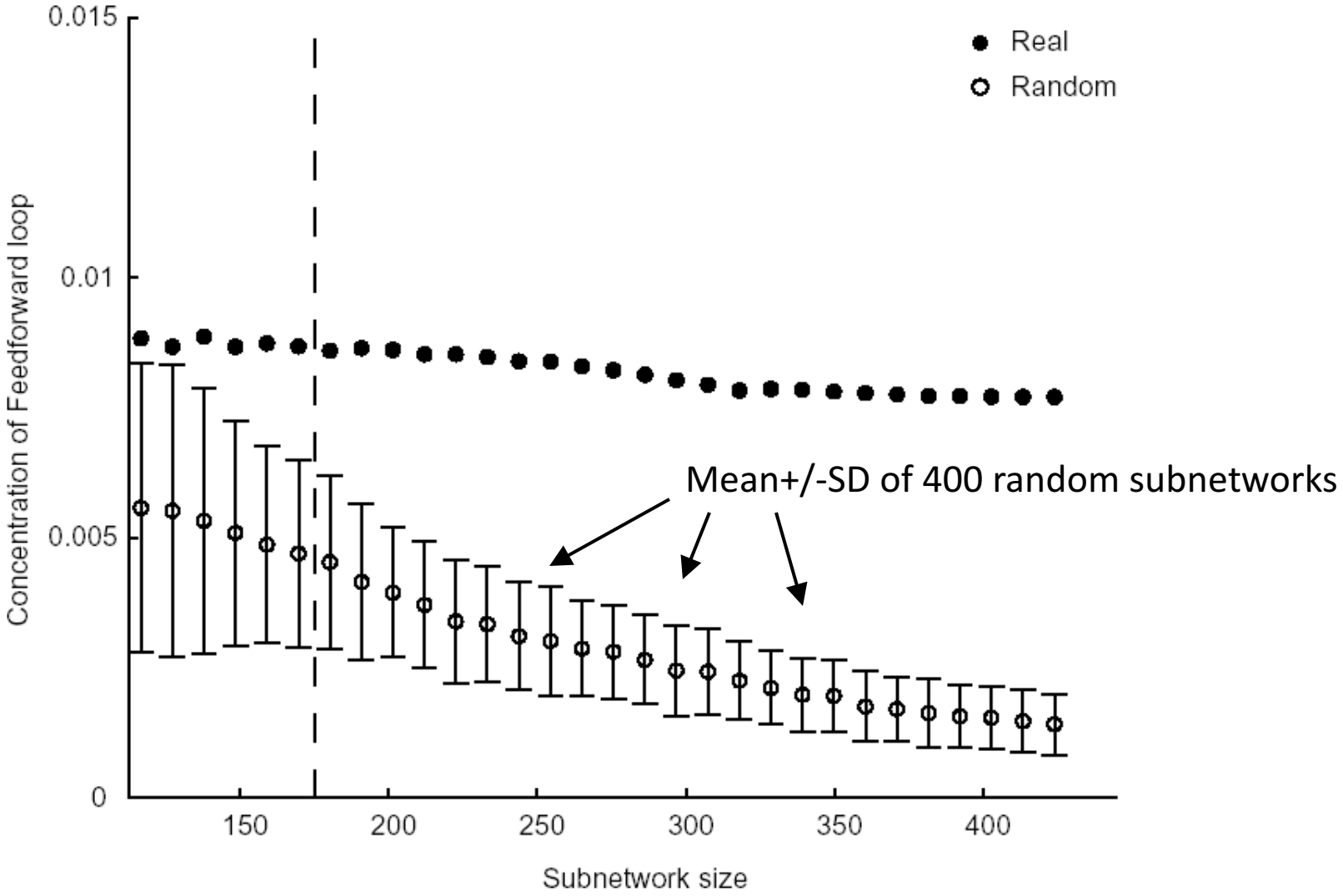


Networks are randomized preserving node degree

Concentration of feedforward motif:

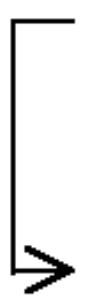
(Num. appearances of motif divided by
all 3 node connected subgraphs)



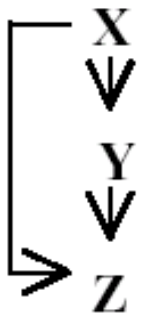
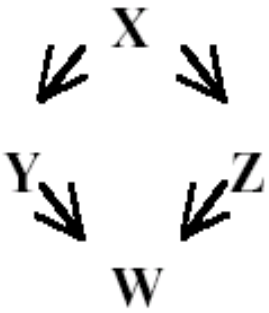


Transcriptional network results

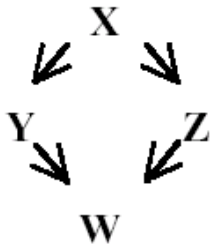
Network	Nodes	Edges
Gene regulation (transcription)		
E. coli	424	519
*S. cerevisiae**	685	1,052

N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
 **Feed-forward loop**			 **Bi-fan**		
40	7 ± 3	10	203	47 ± 12	13
70	11 ± 4	14	1812	300 ± 40	41

Neural networks

Network	Nodes	Edges	N_{real}	$N_{rand} \pm$ SD	Z score
Neurons					**Feed-forward loop**
C. elegans†	252	509	125	90 ± 10	3.7
		**Bi-fan**			**Bi-parallel**
127	55 ± 13	5.3	227	35 ± 10	20

Food webs

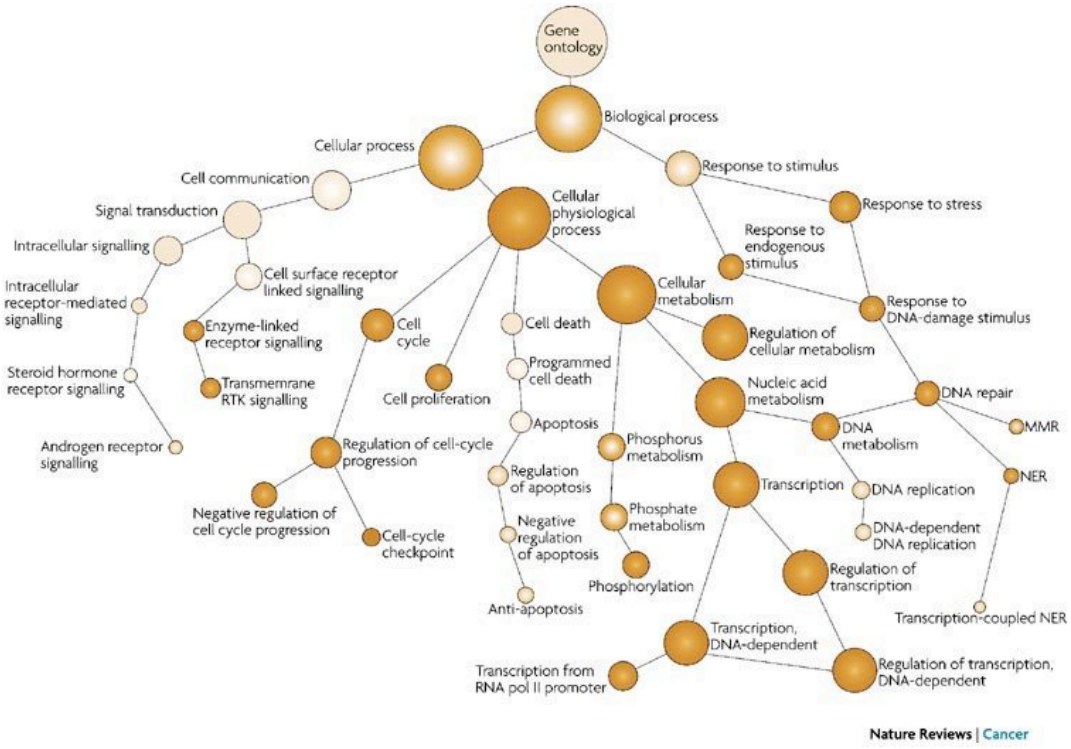
Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Food webs				 **Three chain**				**Bi-parallel**
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32

Interesting questions

- Which networks have motifs in common?
- Which networks have completely distinct motifs versus the others?
- Does this tell us anything about the design constraints on each network?
- E.g., the feedforward loop may function to activate output only if the input signal is persistent (i.e., reject noisy or transient signals) and to allow rapid deactivation when the input turns off
- E.g., food webs evolve to allow flow of energy from top to bottom (?!**!???), whereas transcriptional networks evolve to process information

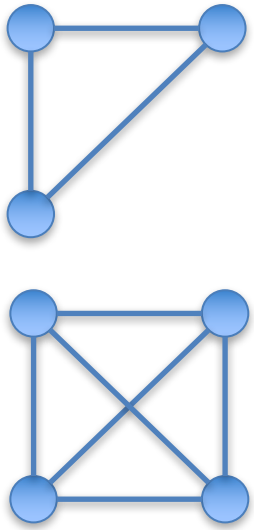
Ontologies

- Define a hierarchy among objects
- Gene Ontology (GO)
  - Curated from scientific literature
- Other examples
  - Disease – symptom relationships
  - Organ – tissue – cell-type relationships

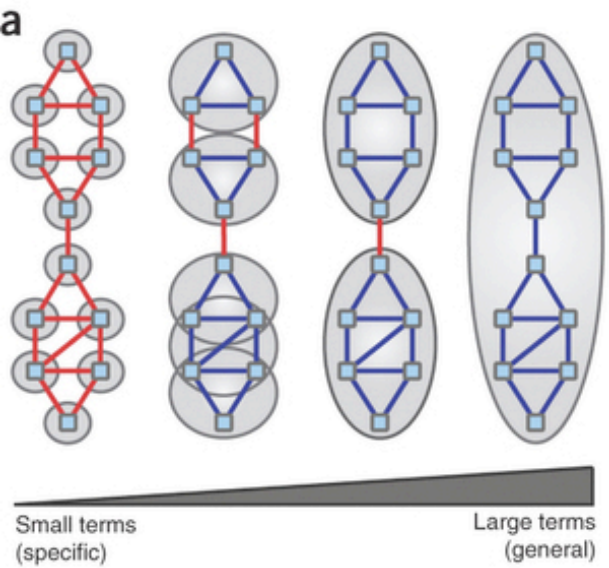


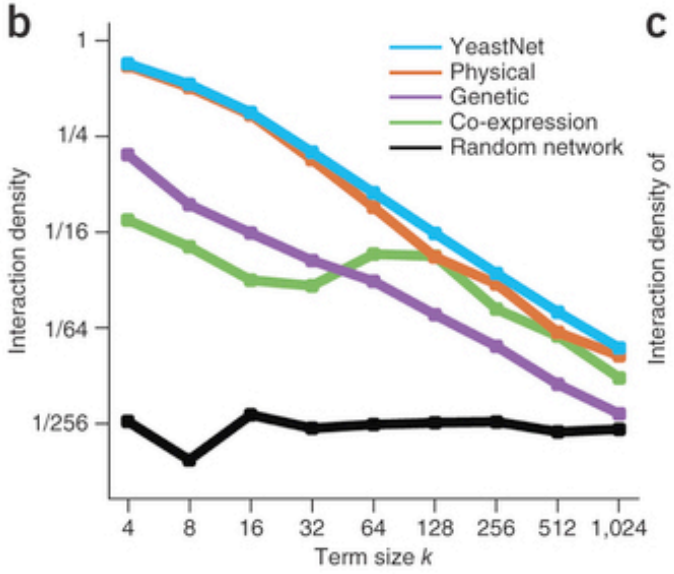
Modularity of Graphs

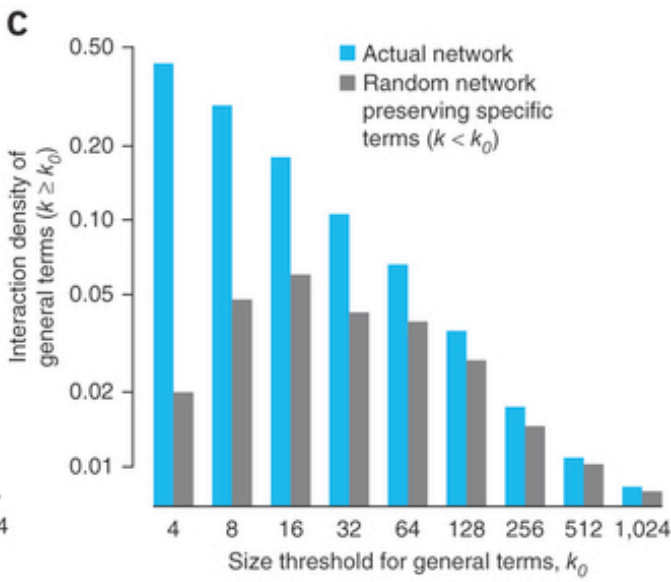
- Cliques
  - A set of nodes that are all directly connected to one another
  - “Complete” subgraph of G

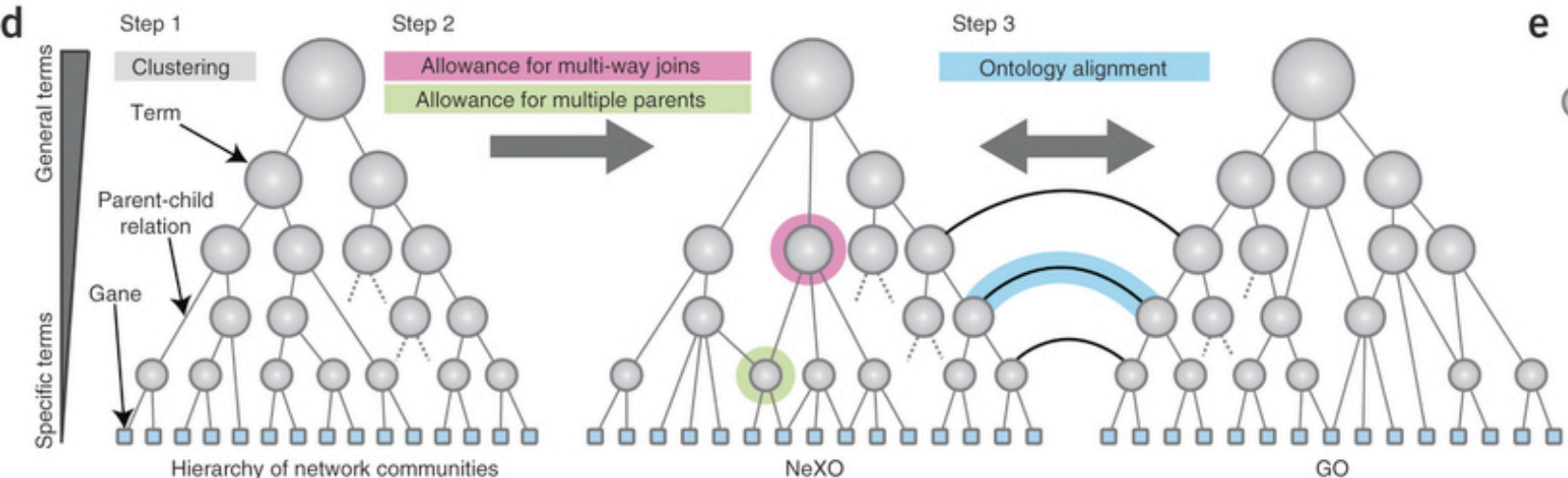


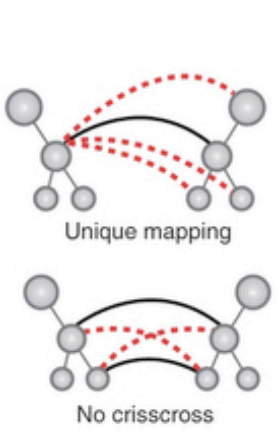
Constructing Ontologies from High-throughput Data



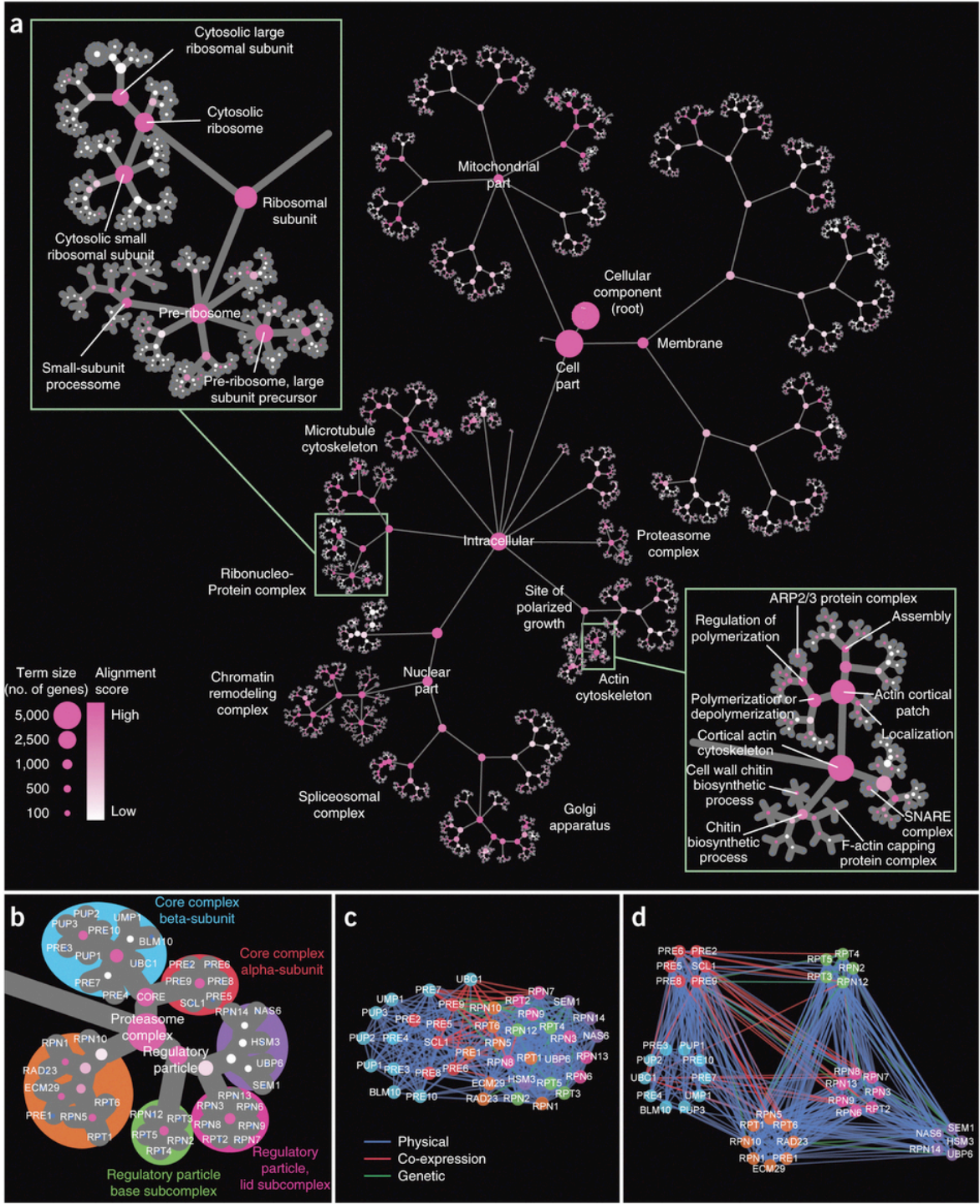








NEXO for Yeast



Nature Biotechnology (2013)

Properties of Biological Networks

- Scale Free
  - Generally see a small number of highly connected hubs and a much larger number of less connected nodes
  - Fault tolerance – essential nodes less likely to be “attacked” by random error
- Small world
  - Most nodes in biological networks are not that distant from one another
  - Evolutionary advantage – deleting a node generally doesn’t greatly increase the mean shortest path length
- Modular
  - Guilt-by-association: Neighboring nodes in biological networks are usually involved in the same biological function