

---

# 3000788 Intro to Comp Molec Biol

## Lecture 18: Biological networks

October 16, 2023



**Sira Sriswasdi, PhD**

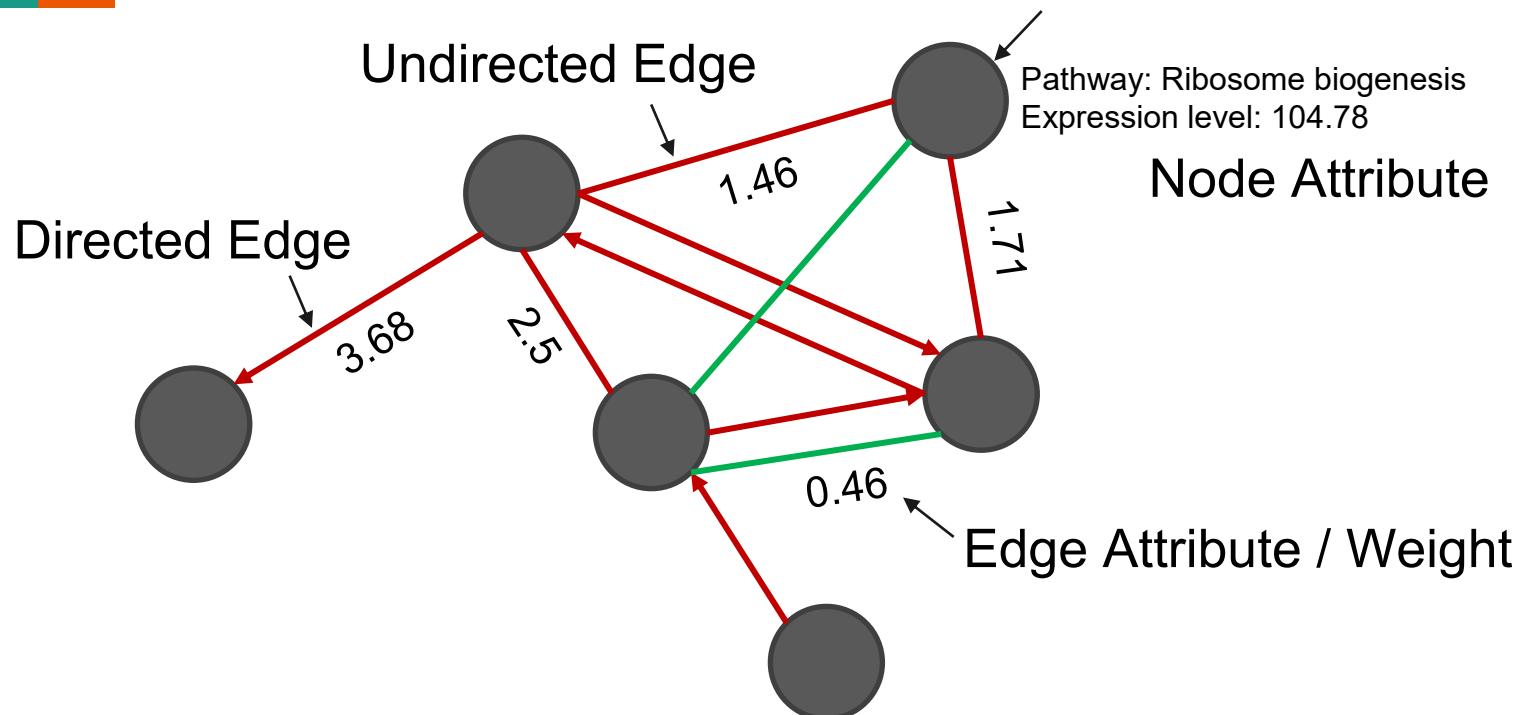
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)



# Network data

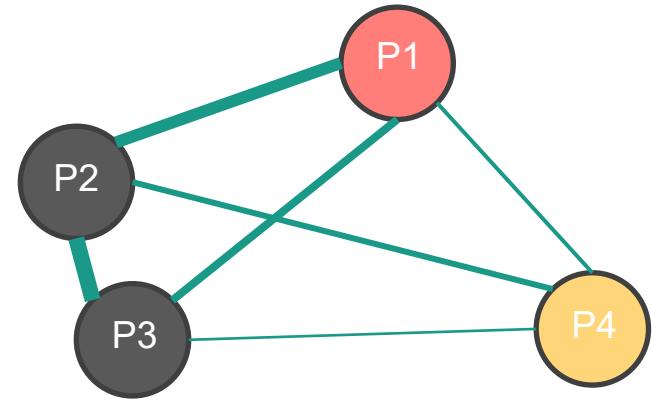
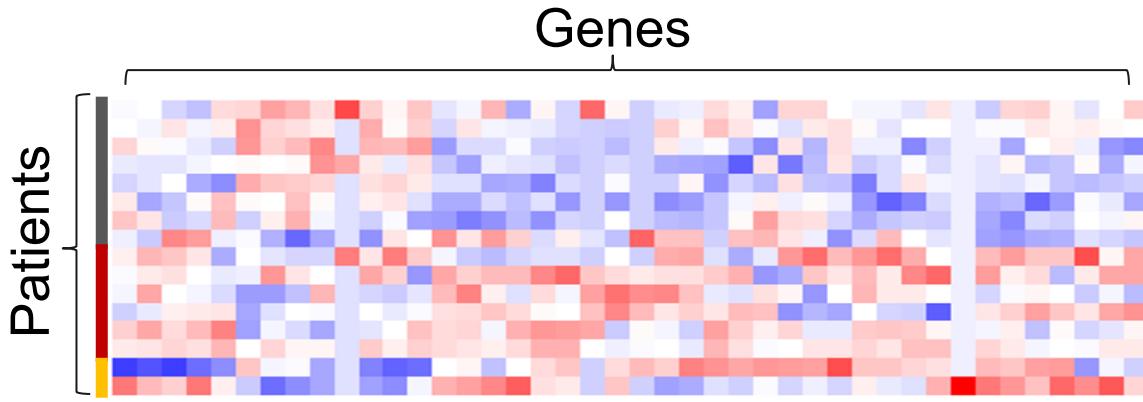
# Graph / Network

---



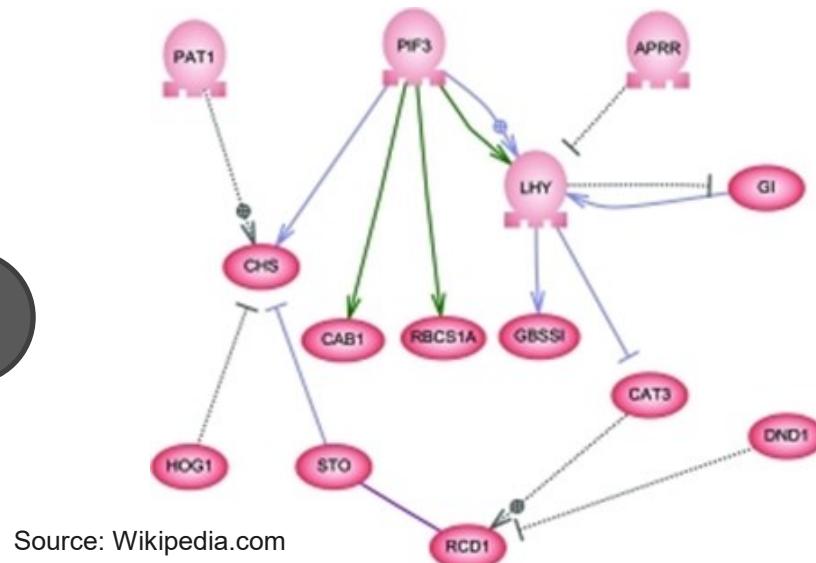
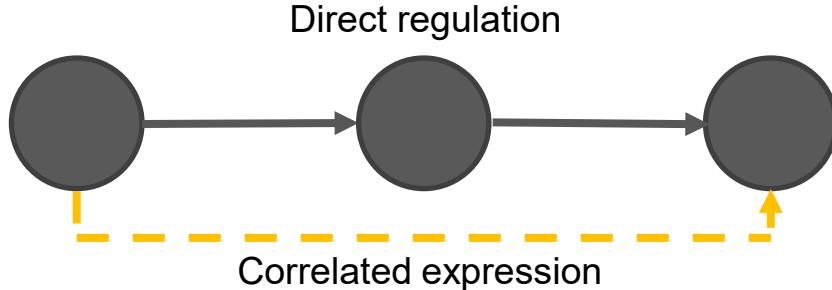
- Connection & relationship between entities

# Distance network



- Node = sample
- Edge weight = similarity / distance between samples
- Node attribute = sample metadata

# Causal inference from network



- Causation = direct interaction: protein binding, TF-DNA binding, etc.
- Type of interaction: activation, repression

# Real-world networks

---

- Computer network
- City-street
- Internet webpages
- Co-authorship
- Friendship
- River & sewage



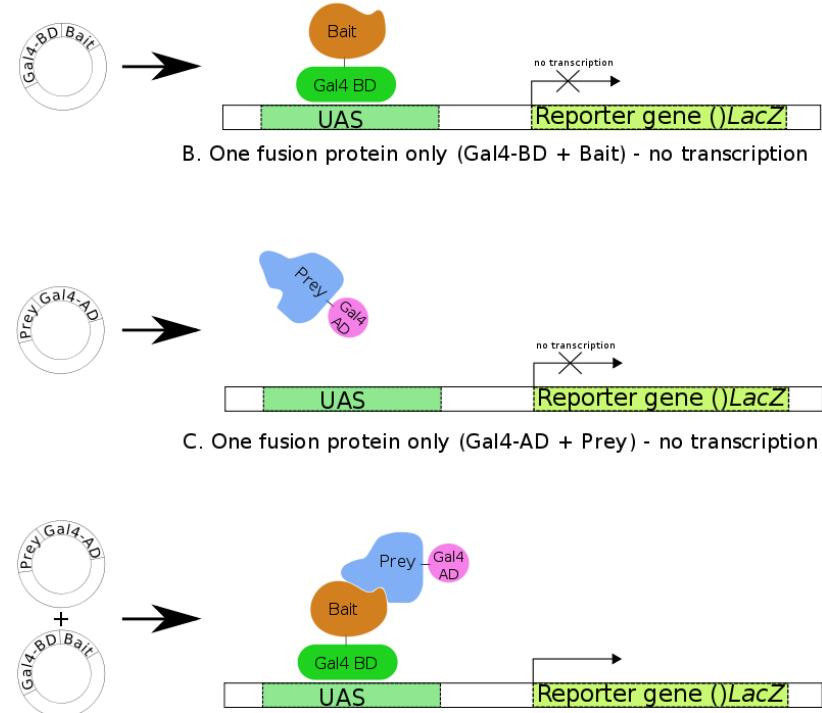
Image from <https://www.flickr.com/photos/caseorganic/4935751455>



# Biological networks

# Networks from omics data

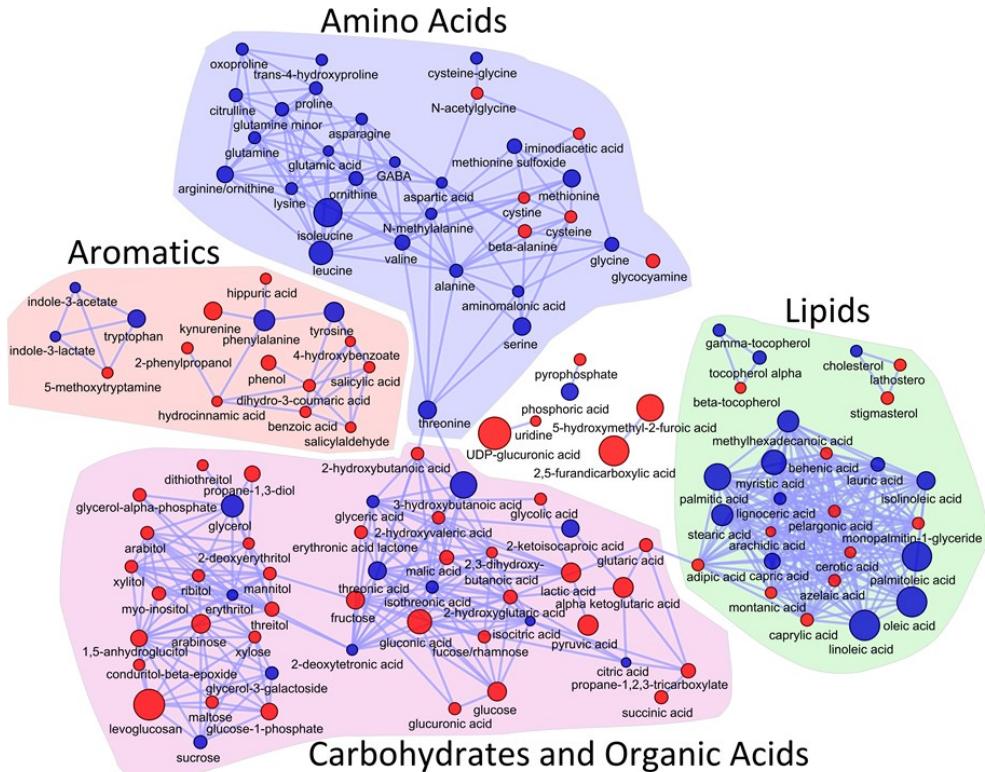
- Yeast-2-hybrid → protein-protein interaction networks
- Immunoprecipitation-MS → protein complex
- ChIP-seq → TF-gene regulatory network
- RNA-seq → gene co-expression network



Source: Wikipedia.com

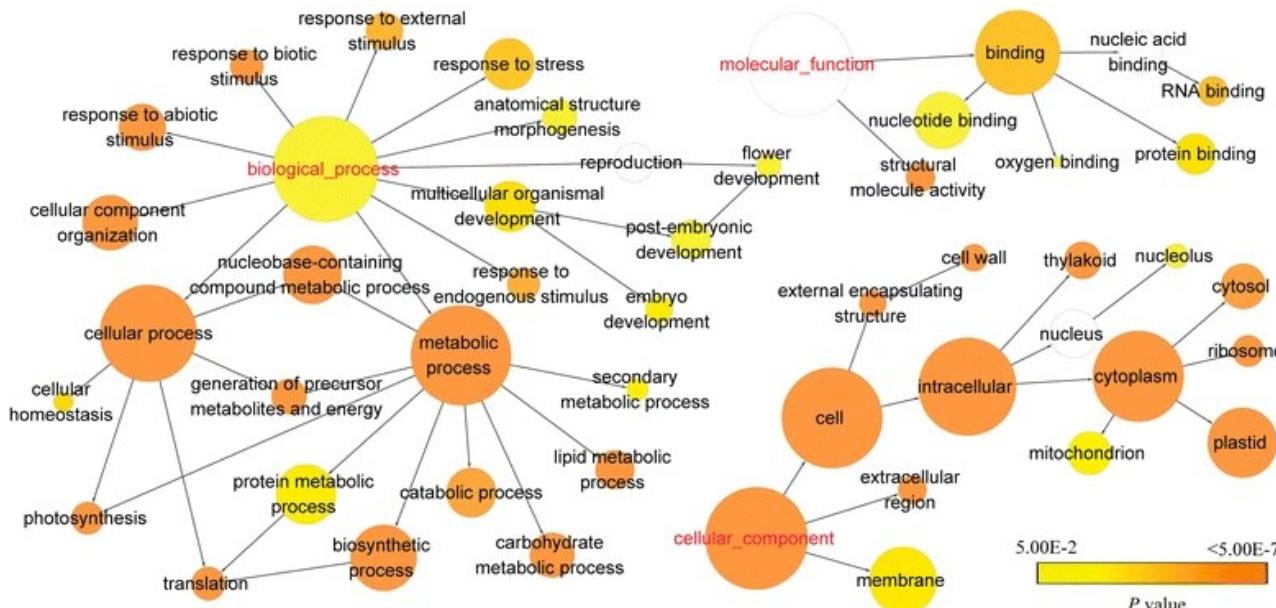
# More examples

- Structural / physicochemical similarity network
  - Cell-cell similarity network
  - Multi-omics networks
    - Drug-gene-disease
    - RNA + ATAC + Bisulfite



Source: imdevsoftware.wordpress.com

# Gene ontology network



Gao, B. et al. BMC Genomics 16:416 (2015)

- Generate a concise summary based on connected terms



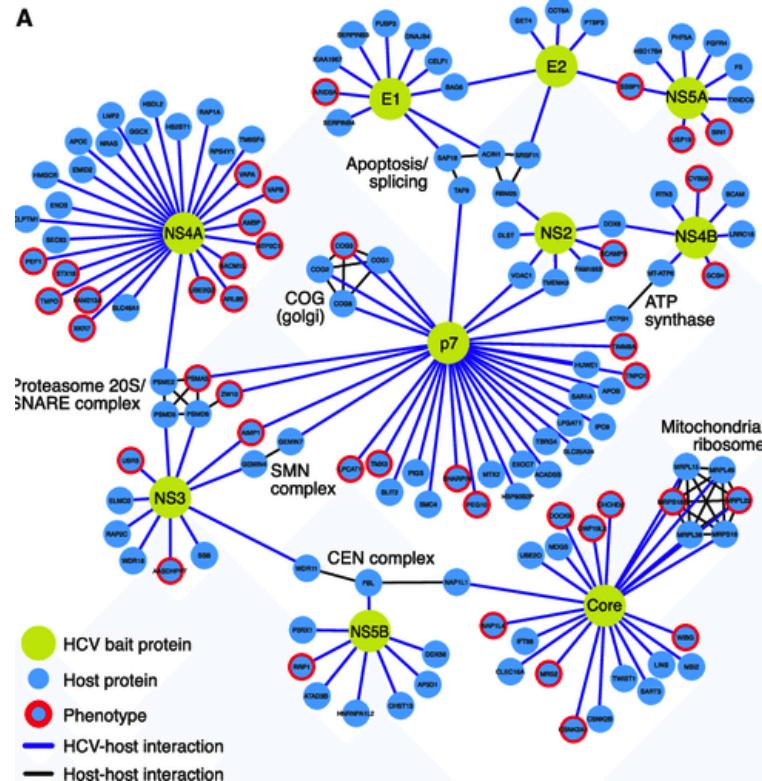
# The Connectivity Map

The Connectivity Map (CMap) at the Broad Institute is creating a genome-scale library of cellular signatures that catalogs transcriptional responses to chemical, genetic, and disease perturbation. To date, the library contains more than 1 Million profiles resulting from perturbations of multiple cell types.

- More than 1M gene expression profiles of cell lines with various disease states and treated with various small molecules
- Network of transcriptome similarity → characterize effect of new drugs

# Host-viral protein interaction

- Two node types: human, HIV
  - Two edge types
  - Node attribute: affected by infection
  - Propose mechanisms underlying the effect of infection
  - Prioritize targets for antibody design

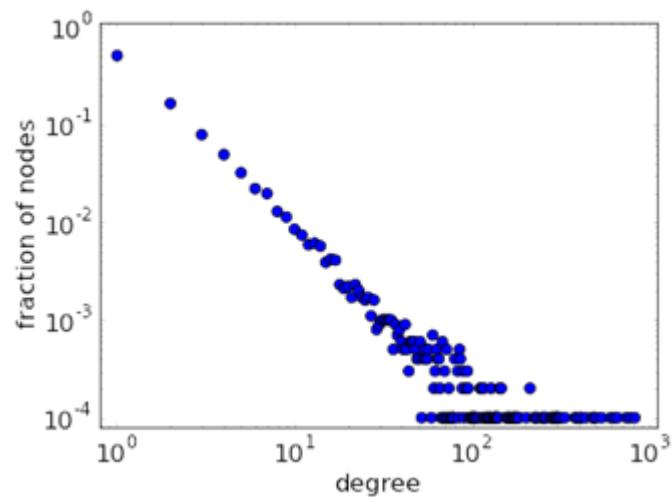


Source: Ramage et al. Mol Cell (2015)

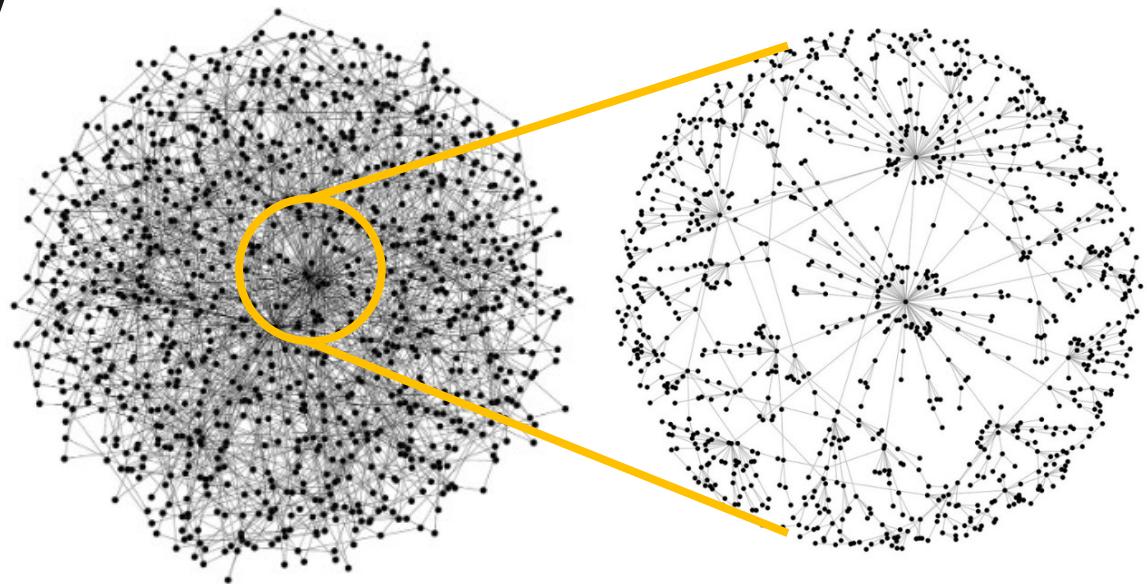


# Properties of real-world networks

# Scale-free property



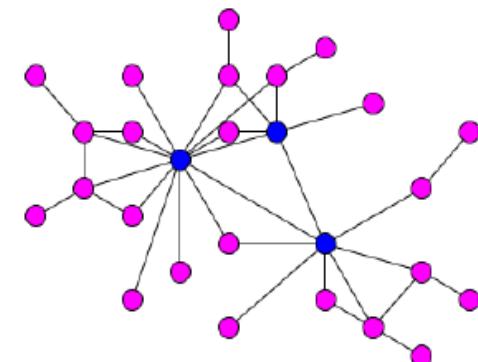
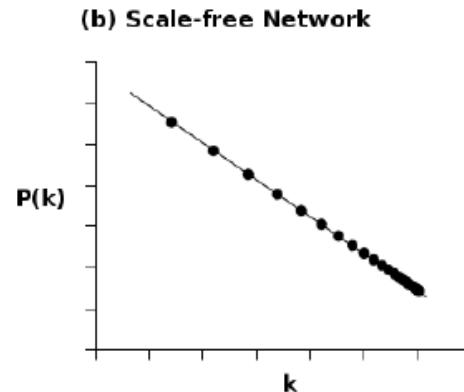
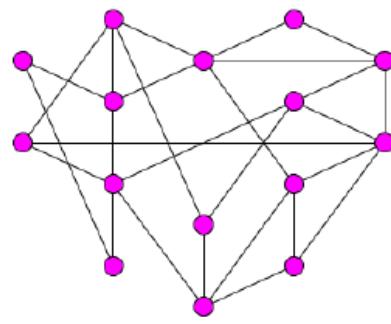
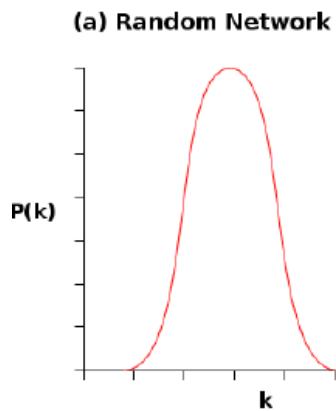
Source: [mathinsight.org](http://mathinsight.org)



Source: [flickr.com & pacojariego.me](http://flickr.com & pacojariego.me)

- Power law:  $P(\text{node connected to } k \text{ edges}) \sim 1/k^n$
- Same local structure as global structure
  - Node-edge distribution

# Hub and small-world property



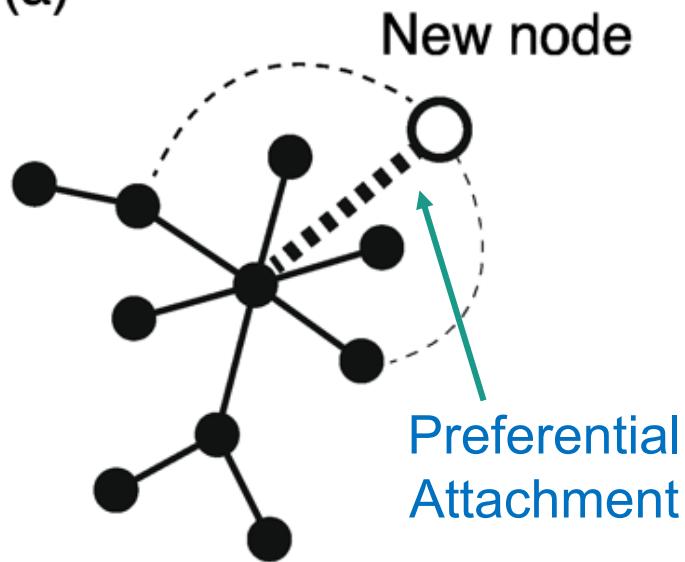
Source: Segura-Cembrera *et al.* Analysis of Protein Interaction Networks to Prioritize Drug Targets of Neglected-Diseases Pathogens

- Few nodes connected with many edges act as short-cut for traffic through the network
  - Transcription factors in biological networks
  - Social influencers on internet

# Emergence of real-world networks

---

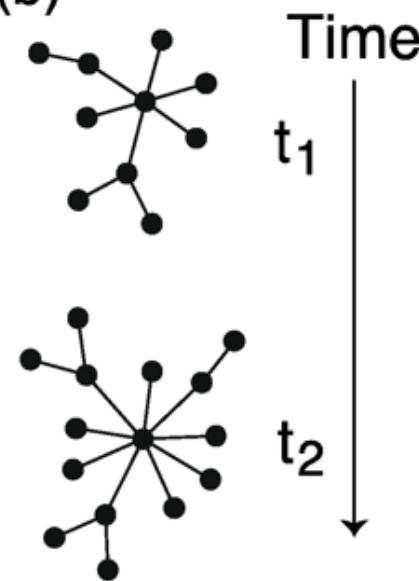
(a)



New node

Preferential  
Attachment

(b)



Time

$t_1$

$t_2$

Takemoto. Metabolites, 2:429-457 (2012)

- New edges prefer to attach to existing nodes with already many edges

# Implication of preferential attachment

---

- **Assumption:** Node connected with high number of edges will more easily attract more edges
- **Consequence 1:** There will be nodes connected to extremely high numbers of edges
- **Consequence 2:** A lot of nodes will not be connected to many edges
- Qualitatively agree with power law:  $P(\text{connected to } k \text{ edges}) \sim 1/k^n$

# The rise of power law

---

- Number of edges,  $E(t) = rt$ , grow linearly with time with factor  $r$
- Node  $N_i$  is connected to  $k_i(t)$  edges at time  $t$
- Rate of increase of  $k_i(t)$  is a competition with other nodes
  - $\frac{dk_i(t)}{dt} = \frac{k_i(t)}{2E(t)}$
  - $\frac{1}{k_i(t)} dk_i(t) = \frac{1}{2E(t)} dt = \frac{1}{rt} dt$
- Integrating both sides
  - $\ln(k_i(t)) = \frac{1}{r} \ln(t) + C_i$
  - $k_i(t) = C_i t^{1/r}$

# The rise of power law

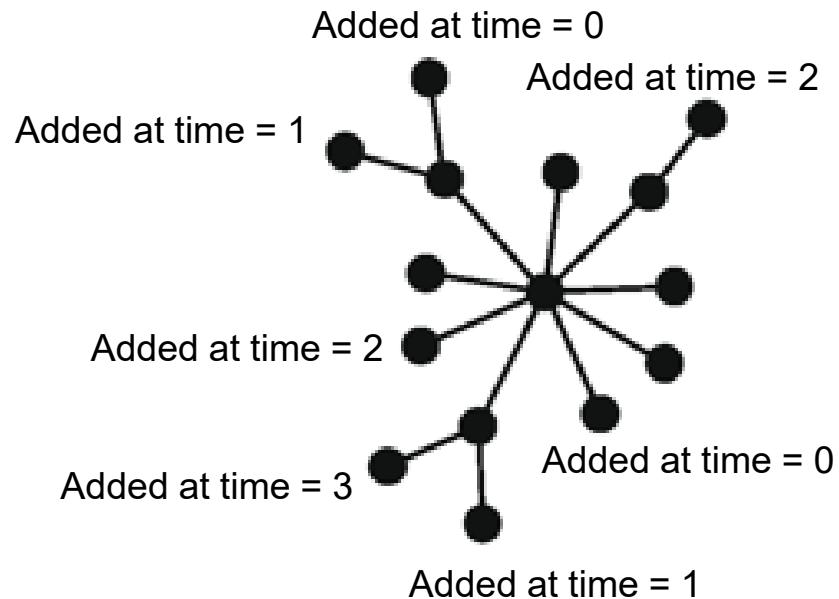
---

- Node  $N_i$  emerge in the network at time  $t_i$  with initial edges =  $m_0$ 
  - $m_0 = k_i(t_i) = C_i t_i^{1/\textcolor{teal}{r}}$
  - $C_i = \frac{m_0}{t_i^{1/\textcolor{teal}{r}}}$
- Let's look at the cumulative density of  $k_i(t)$ 
  - $P(k_i(t) < k) = P\left(\frac{m_0}{t_i^{1/\textcolor{teal}{r}}} t^{1/\textcolor{teal}{r}} < k\right) = P\left(t_i > \frac{m_0^{\textcolor{teal}{r}} t}{k^{\textcolor{teal}{r}}}\right) = 1 - P\left(t_i \leq \frac{m_0^{\textcolor{teal}{r}} t}{k^{\textcolor{teal}{r}}}\right)$
  - $P\left(t_i \leq \frac{m_0^{\textcolor{teal}{r}} t}{k^{\textcolor{teal}{r}}}\right)$  is equal to the probability of picking a node  $N_i$  that emerged in the network before time =  $\frac{m_0^{\textcolor{teal}{r}} t}{k^{\textcolor{teal}{r}}}$

# The rise of power law

---

- Number of nodes,  $N(t)$ , also grow linearly
- The probability of picking a node  $N_i$  that emerged in the network before time  $\frac{m_0^r t}{k^r}$  among  $N(t)$  nodes is **directly proportional** to  $\frac{m_0^r t}{N(t)k^r}$
- $P\left(t_i \leq \frac{m_0^r t}{k^r}\right) = \frac{m_0^r t}{p N(t) k^r}$



# The rise of power law

---

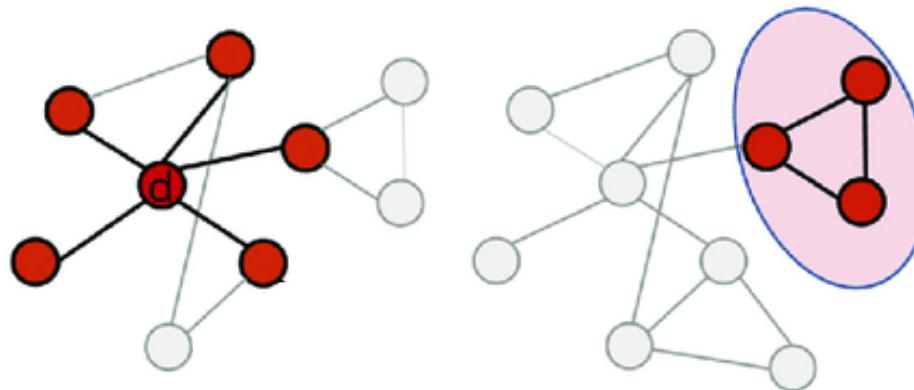
- We now know the cumulative density
  - $P(k_i(t) < k) = 1 - P\left(t_i \leq \frac{m_0^r t}{k^r}\right) = 1 - \frac{m_0^r t}{p^{N(t)} k^r}$
- Take derivative to get the probability density
  - $P(k_i(t) = k) = \frac{dP(k_i(t) < k)}{dk} = \frac{rm_0^r t}{p^{N(t)} k^{r+1}} \propto \frac{1}{k^{r+1}}$
  - The order of the power law is linked to the growth rate of edges



# Topological properties

# Degree and complete subgraph

---

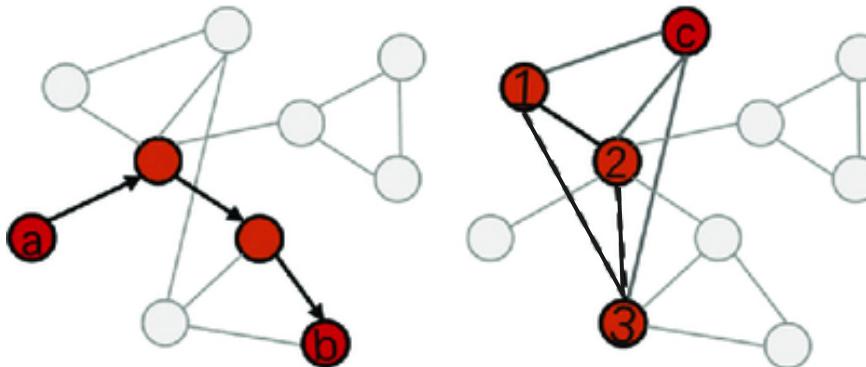


Cai and Niu. Dev Cog Neuroscience (2018)

- **Degree** = number of edges connected to a certain node
- **Clique, complete subgraph** = region of a network whose nodes are fully connected with  $\frac{n(n-1)}{2}$  edges

# Path and clustering coefficient

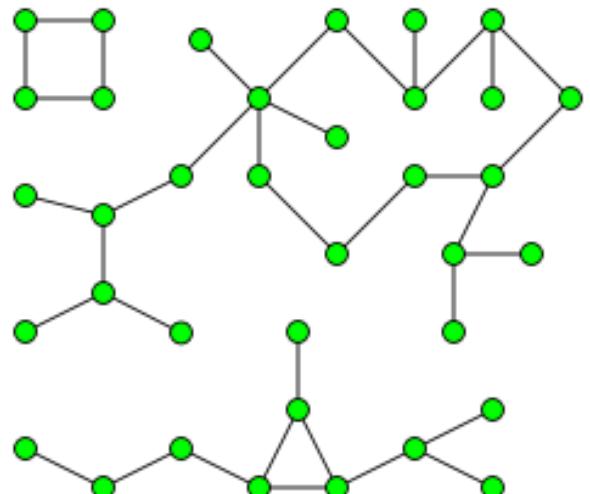
---



Cai and Niu. Dev Cog Neuroscience (2018)

- **Path** = connection from one node to another through several edges
  - Serve as distance between nodes on the network
- **Clustering coefficient** = proportion of neighbors that are also connected
  - Indicate the extent of local connectivity / redundancy of the network

# Connected components



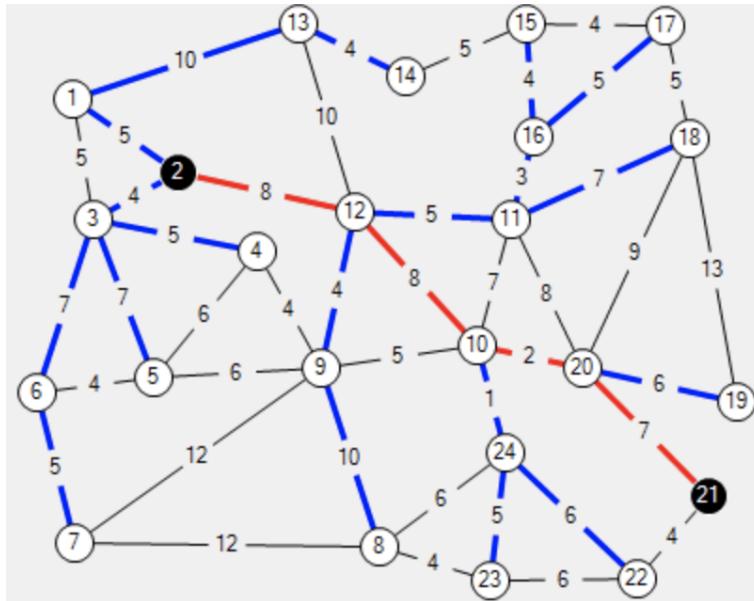
Source: Wikipedia.com

- Some network might consist of small disconnected subnetworks
  - Number of connected components can indicate the complexity and organization of the whole network
  - Can be caused by incomplete data
    - Missing edges in biological networks



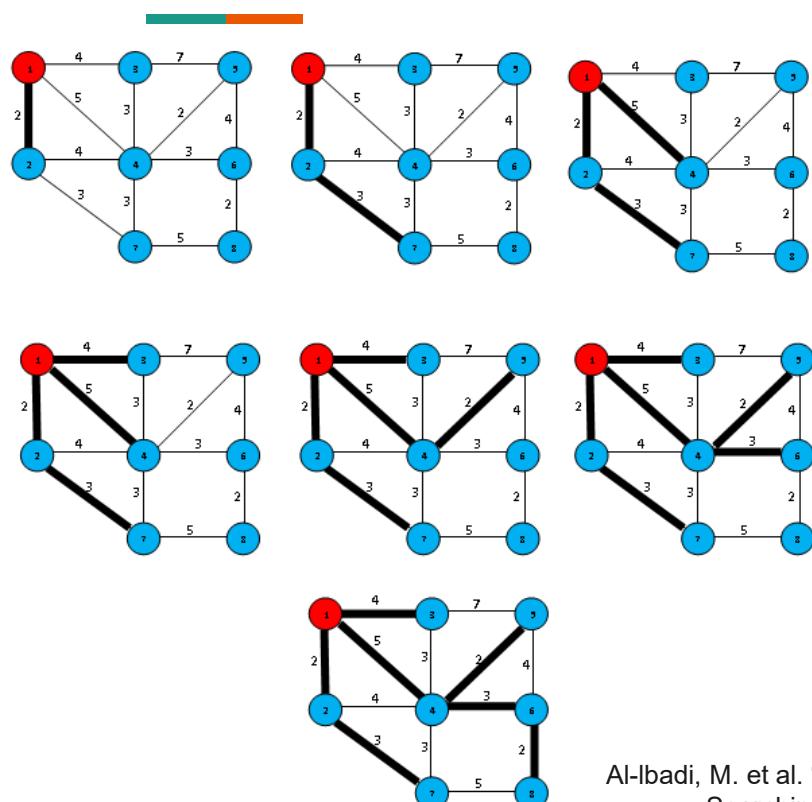
# **Connectivity measures and interpretation**

# Importance of path and path length



- Network operates by transmitting signals from one node to another
- Nodes frequently involved in the transmission are important
- **Model 1:** Shortest paths only
- **Model 2:** Weighted across all paths

# Dijkstra algorithm



- Dynamic programming
- From a start node, traverse the edges and record current distances for visited nodes
- Update distance  $d(i, j)$  if a shorter path is found

# Network as flows

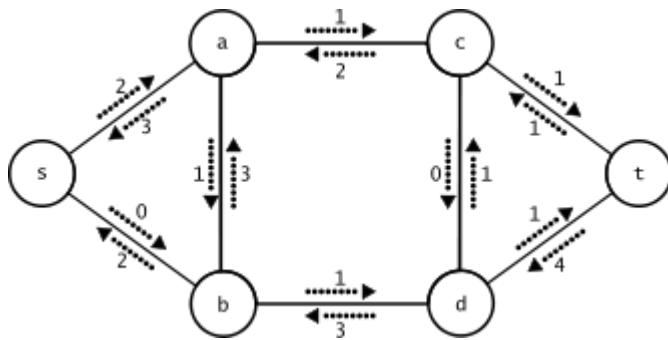


Image from [https://en.wikipedia.org/wiki/Flow\\_network](https://en.wikipedia.org/wiki/Flow_network)

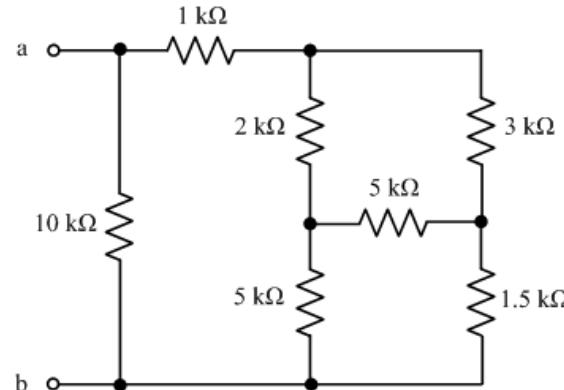
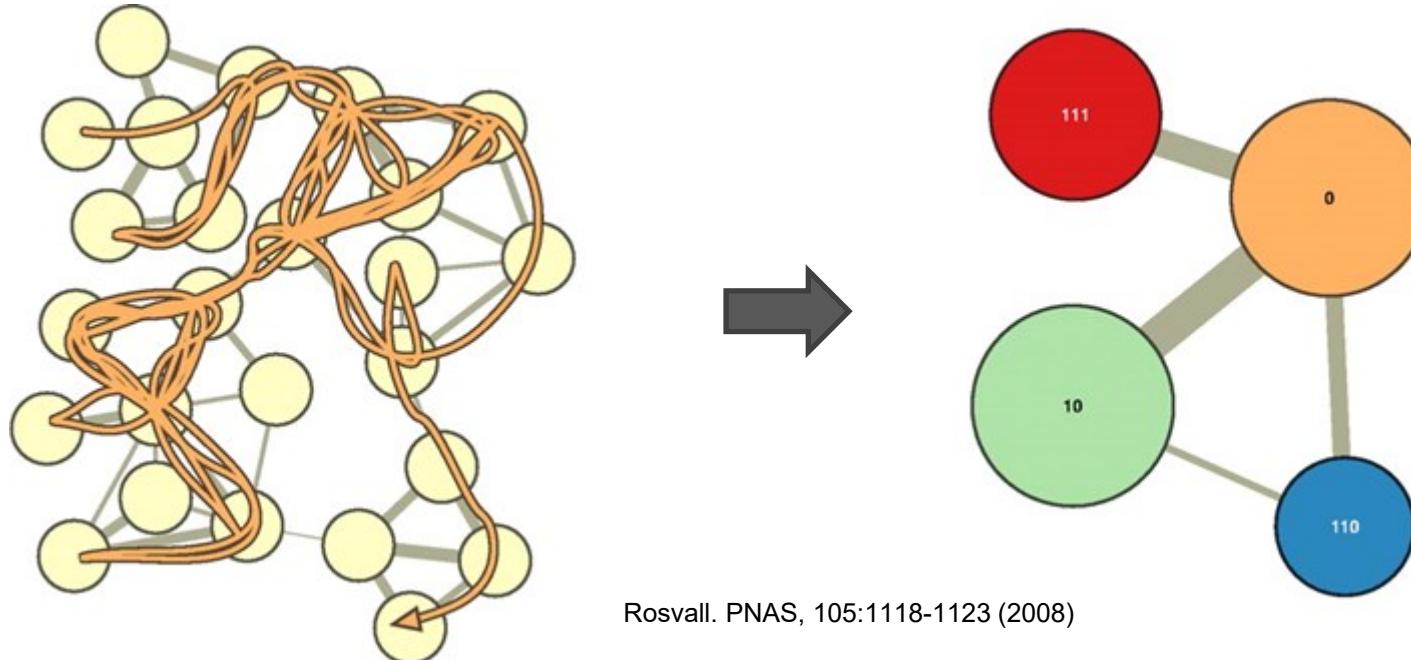


Image from <http://www.roose-hulman.edu/CLEO/browse/?path=1/2/79/91/92/19>

- Signal as fluid flowing through pipes with various diameters
- Signal as electric current flowing through circuit with various resistances
- Study the dynamics and stationary states with simulations

# Network as random walks

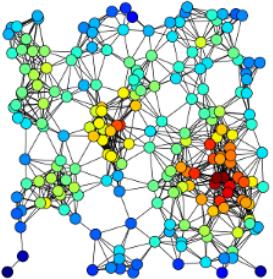


Rosvall. PNAS, 105:1118-1123 (2008)

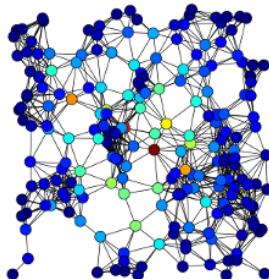
- Discrete particles travel from node to node with probability (edge weight)

# Centrality scores

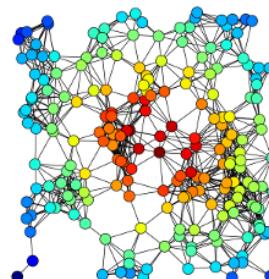
Degree



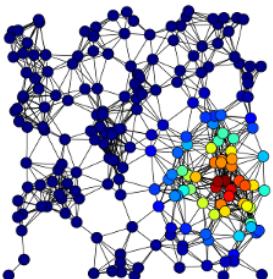
Betweenness



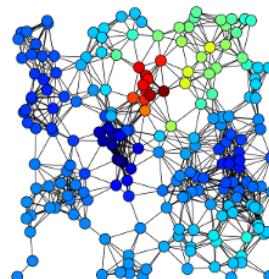
Closeness



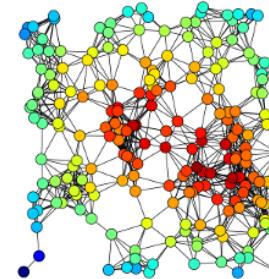
Eigenvector



Katz



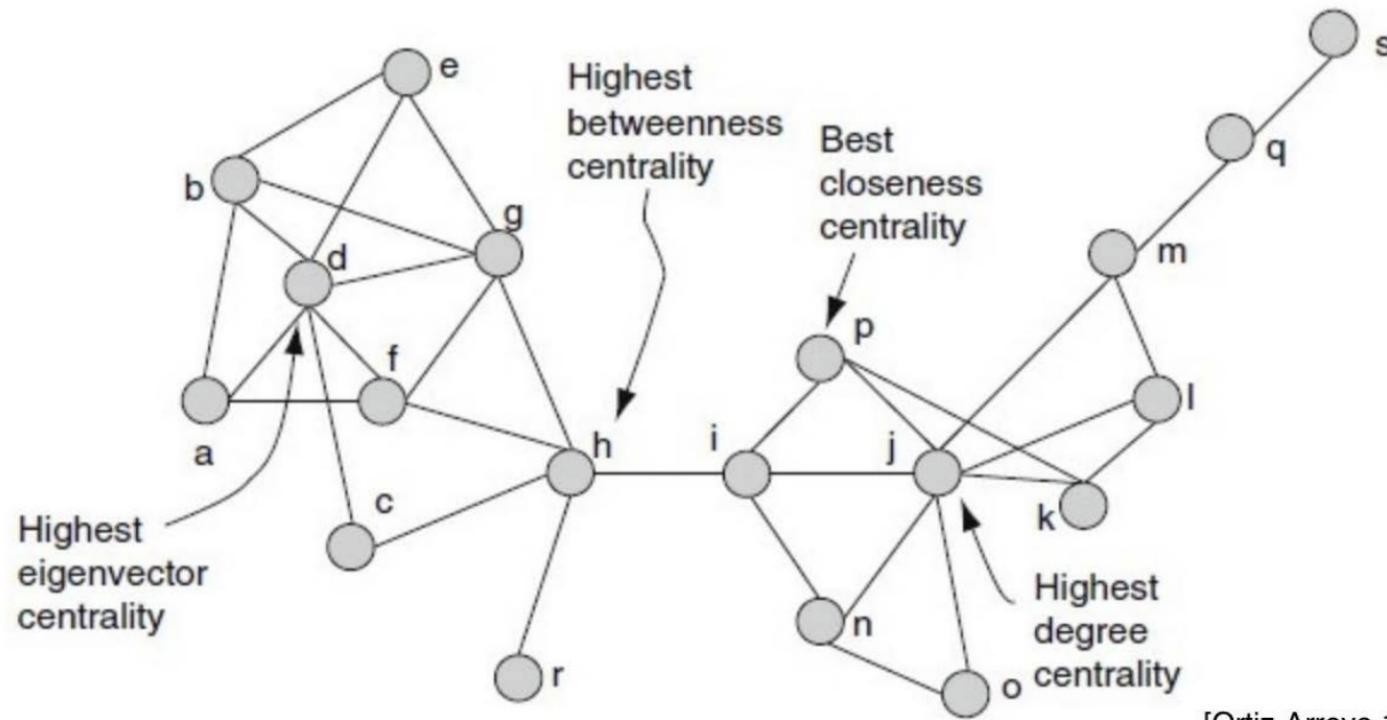
Harmonic



- Indicate the importance of a node in the context of the connectivity of the network
- **Degree** = local connectivity
- **Betweenness** = fraction of shortest paths
- **Closeness** = inverse distance to other nodes

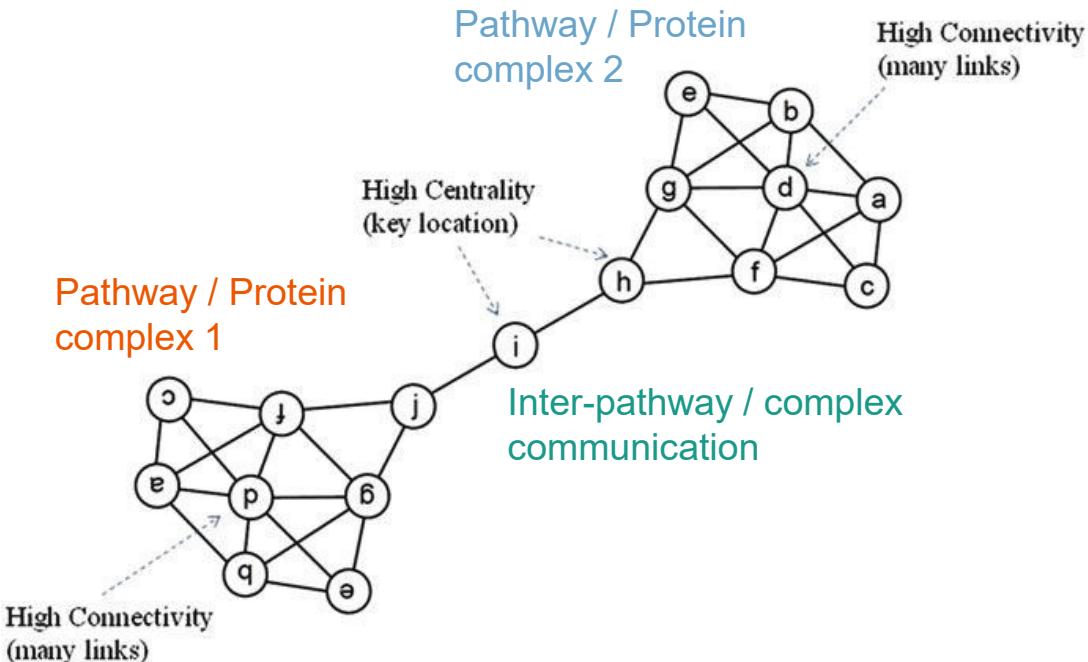
# Different scores, different meaning

---



[Ortiz-Arroyo 10]

# Biological interpretation



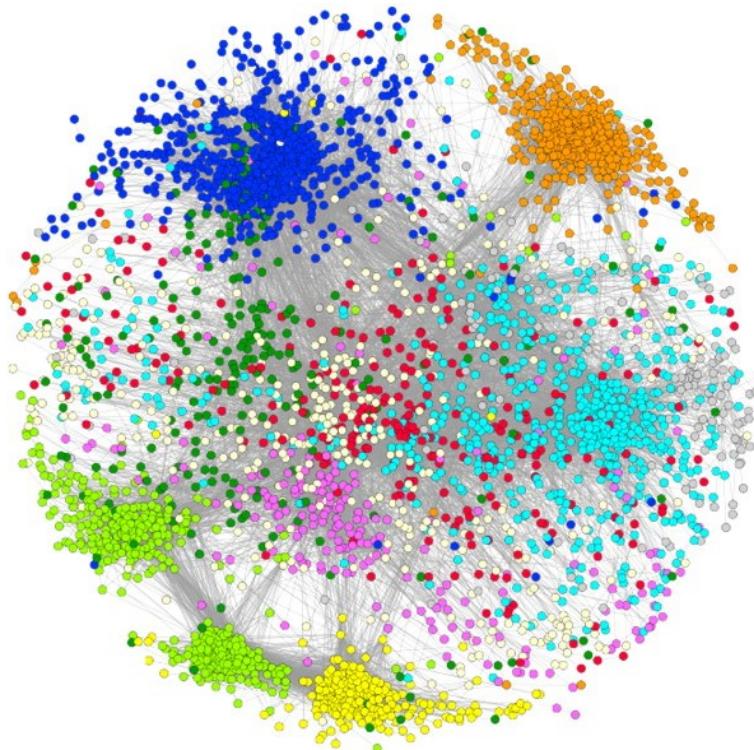
- **Low degree, High betweenness** = connect multiple functional pathways
- **High degree, Low betweenness** = core protein of a complex, transcription factor with multiple downstream targets



# Network clustering

# Dissect local characteristics

---

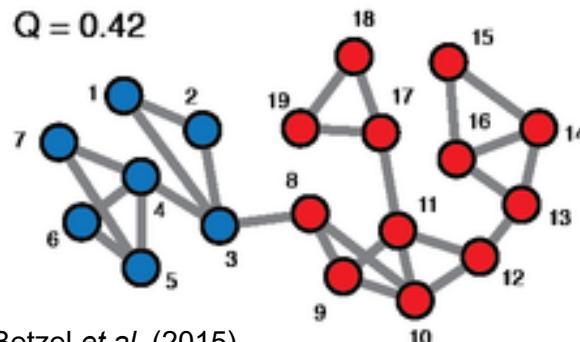


- GO: RNA Metabolism
- GO: Primary Metabolism
- GO: Protein Biosynthesis
- GO: DNA Metabolism
- GO: Transcription
- GO: Cell Cycle
- GO: Other
- GO: Biological Process
- GO: Signal Transduction
- GO: Transport

# Modularity score

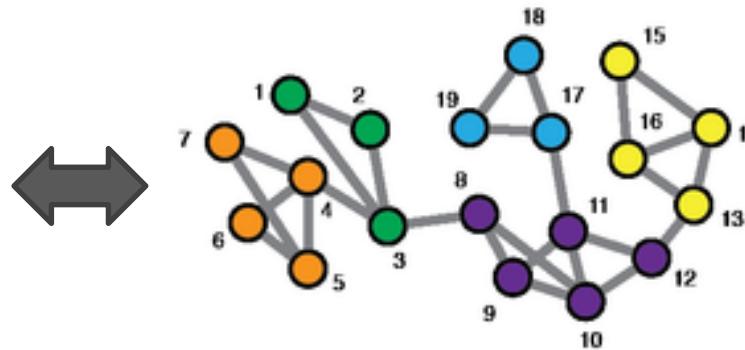
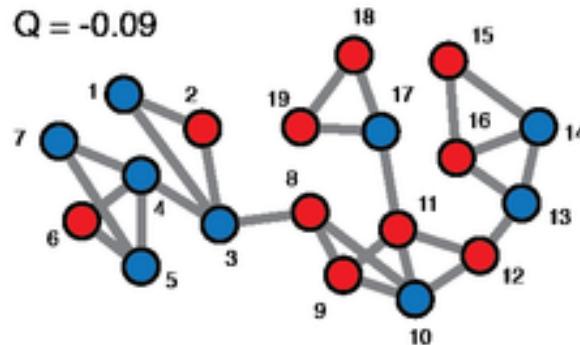


$Q = 0.42$



Betzel et al. (2015)

$Q = -0.09$



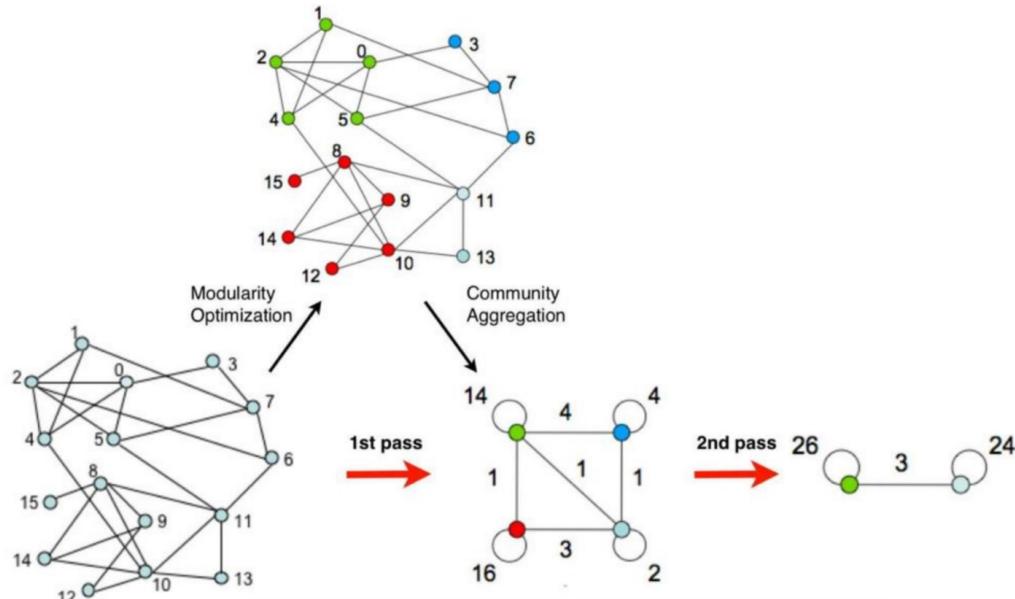
- Number of within-cluster edges compared to expectation (based on number of nodes and global number of edges)
- Multiple resolutions

## A simple modularity score

---

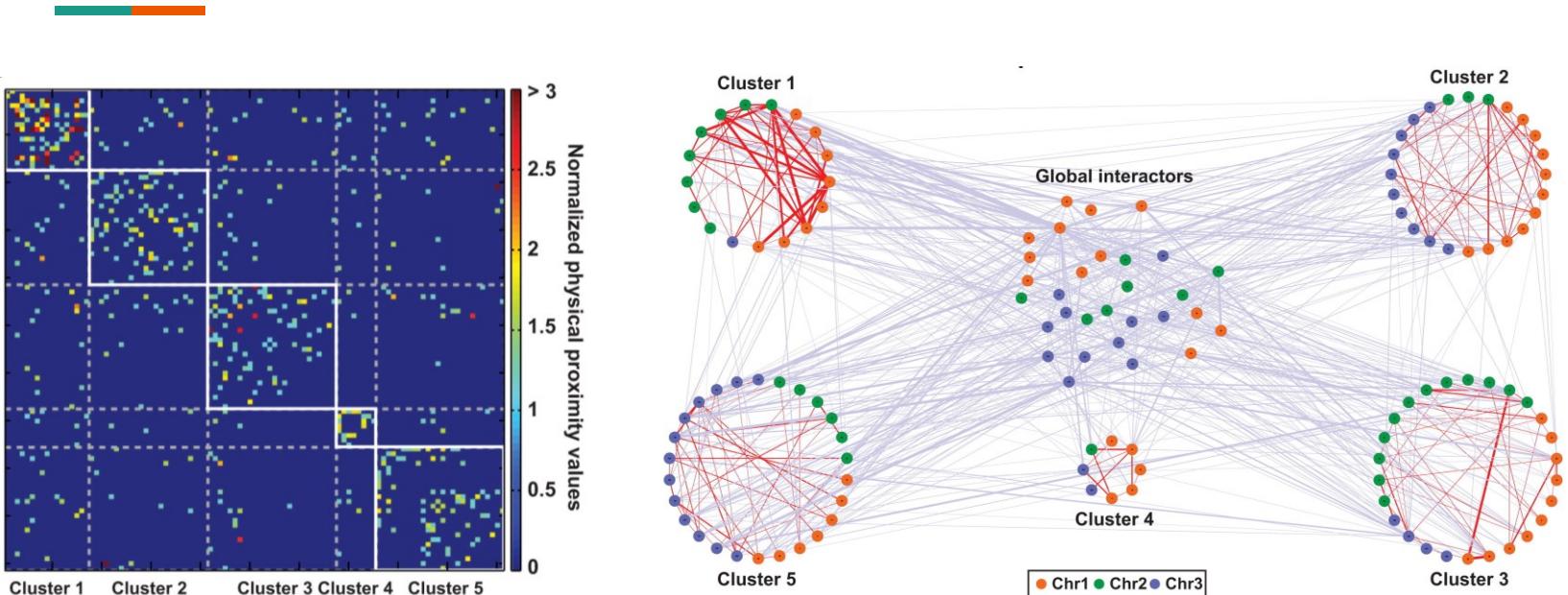
- Node  $N_i$  with degree  $d_i$
- Node  $N_j$  with degree  $d_j$
- Edge weights are all 1
- $P(\text{edge between } N_i \text{ and } N_j \text{ by chance}) \sim d_i d_j / 2 \times \# \text{ edges}$
- Modularity score of a cluster of nodes  $(N_1, N_2, \dots, N_n)$  is then:
  - $Q = \# \text{ within-cluster edges} - \sum_{i,j} \frac{d_i d_j}{2e}$

# Louvain / Leiden algorithm



- Iteratively find clusters of node that maximize modularity score
- Collapse nodes of a cluster into a representative node to speed up refinement

# Application on chromatin interaction



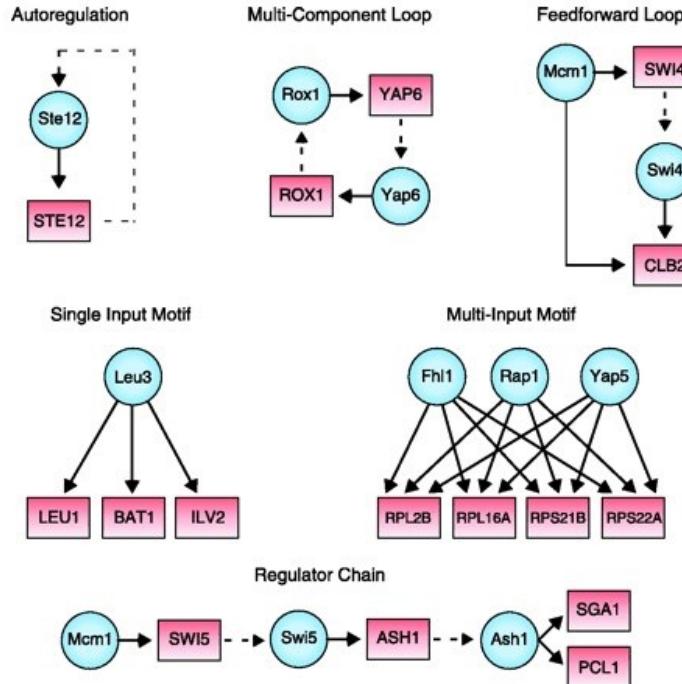
Tanaka et al. Molecular Cell, 48:532-46 (2012)

- Identify clusters of genomic loci that are nearby in 3D + global interactors

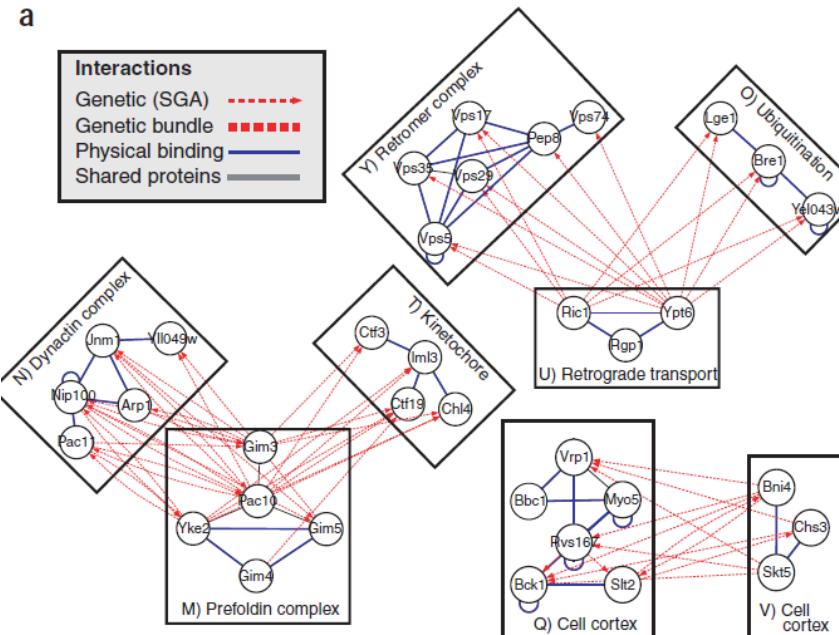


# Network motifs and homology

# Motif = recurring patterns

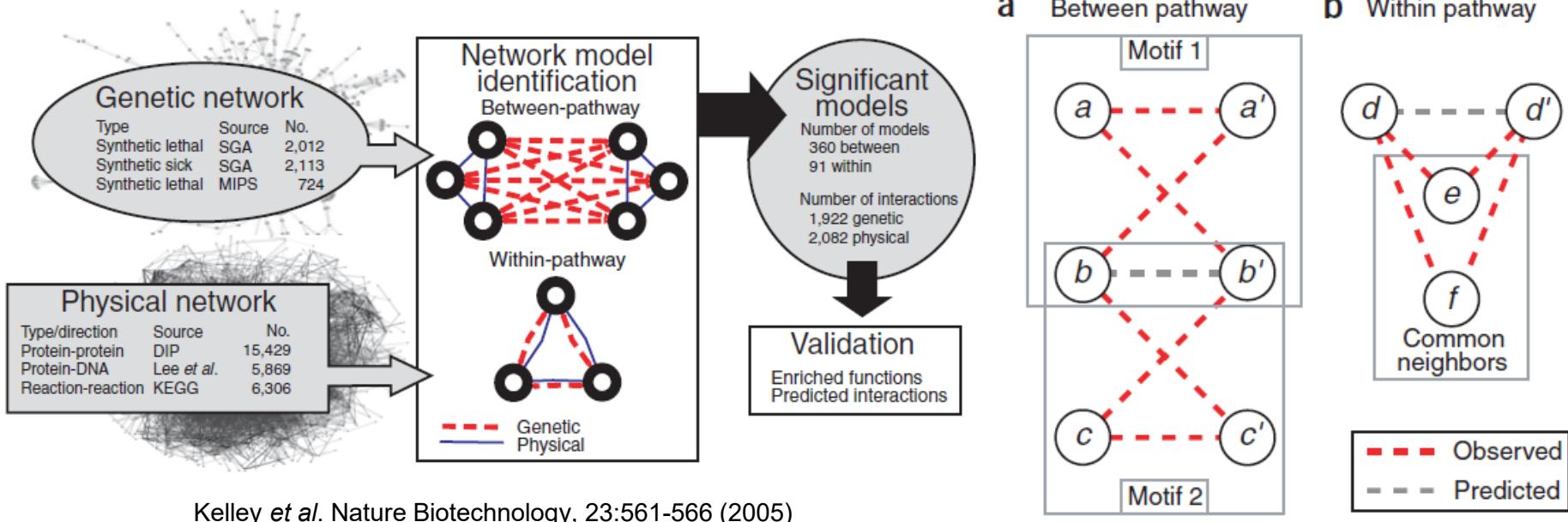


Lee et al. Science. 298:799-804 (2002)



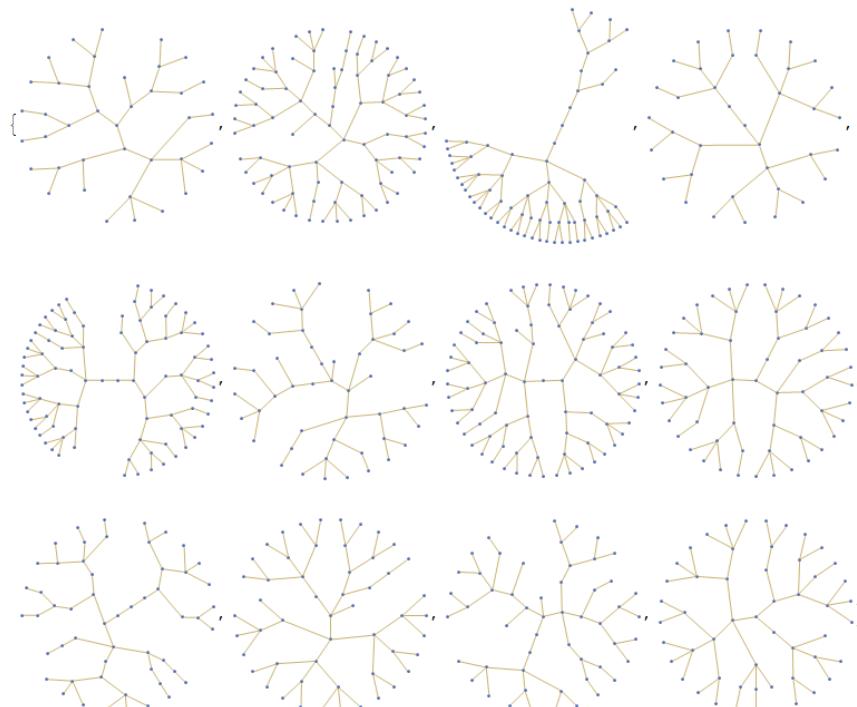
Kelley et al. Nature Biotechnology, 23:561-566 (2005)

# Incomplete motif indicates missing data

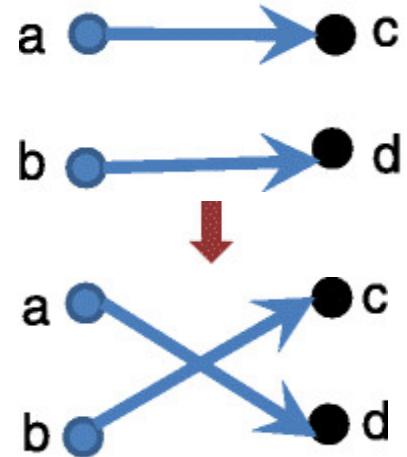


- Predict missing data and prioritize for validation

# Finding motif with permutation test



Edge switching



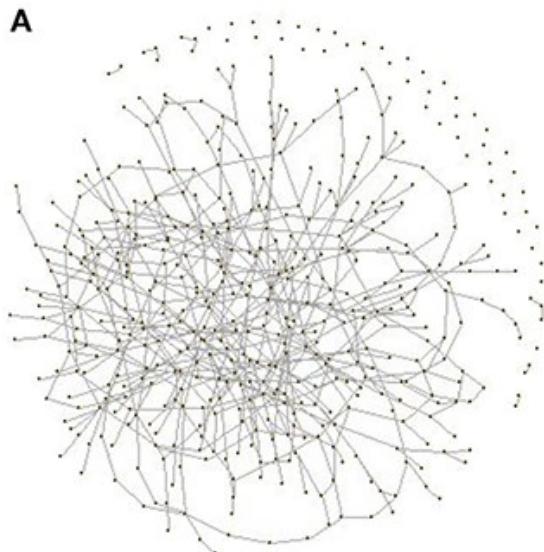
Temate-Tiageru et al. BMC Genomics, 17:542 (2016)

Preserve degree distribution!

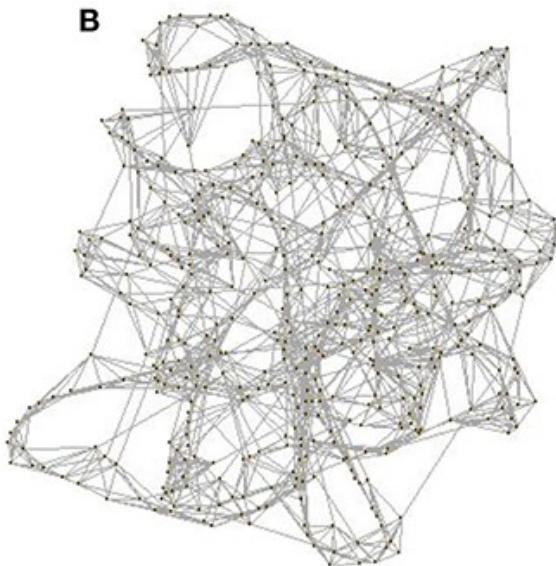
# Random network models

---

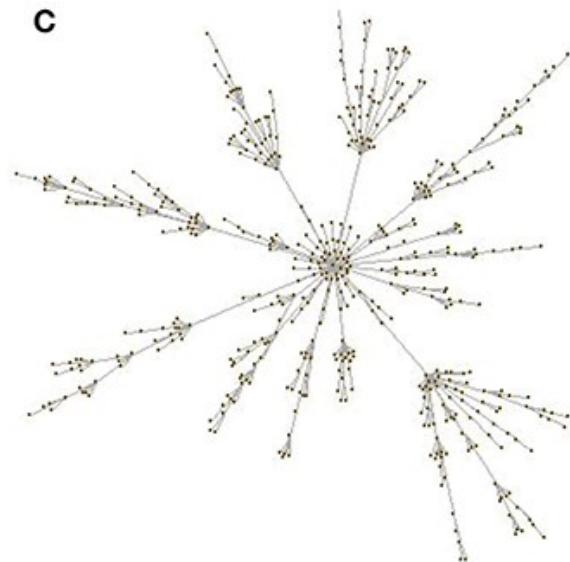
A



B



C

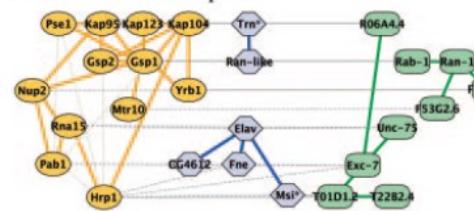


Koutrouli, M. et al. Front. Bioeng. Biotechnol. 31 (2020)

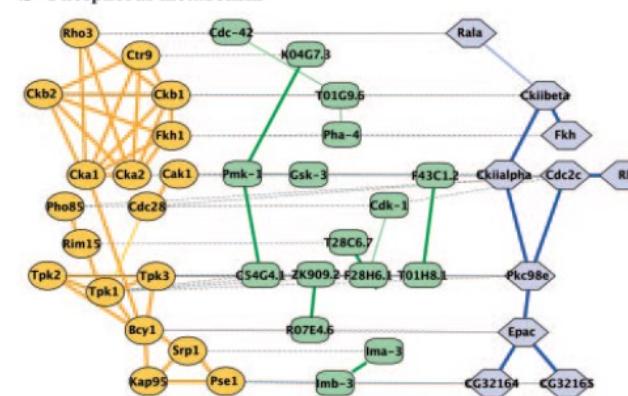
- A) Erdos-Renyi, B) Watts-Strogatz, C) Barabasi-Albert

# Homology between nodes and network structures

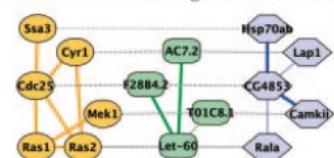
a Intracellular transport



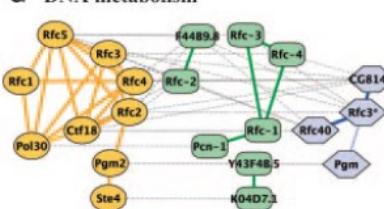
b Phosphorus metabolism



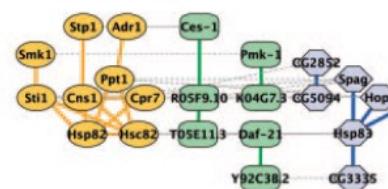
c Ras-mediated regulation of cell cycle



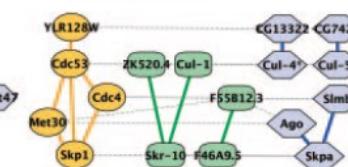
d DNA metabolism



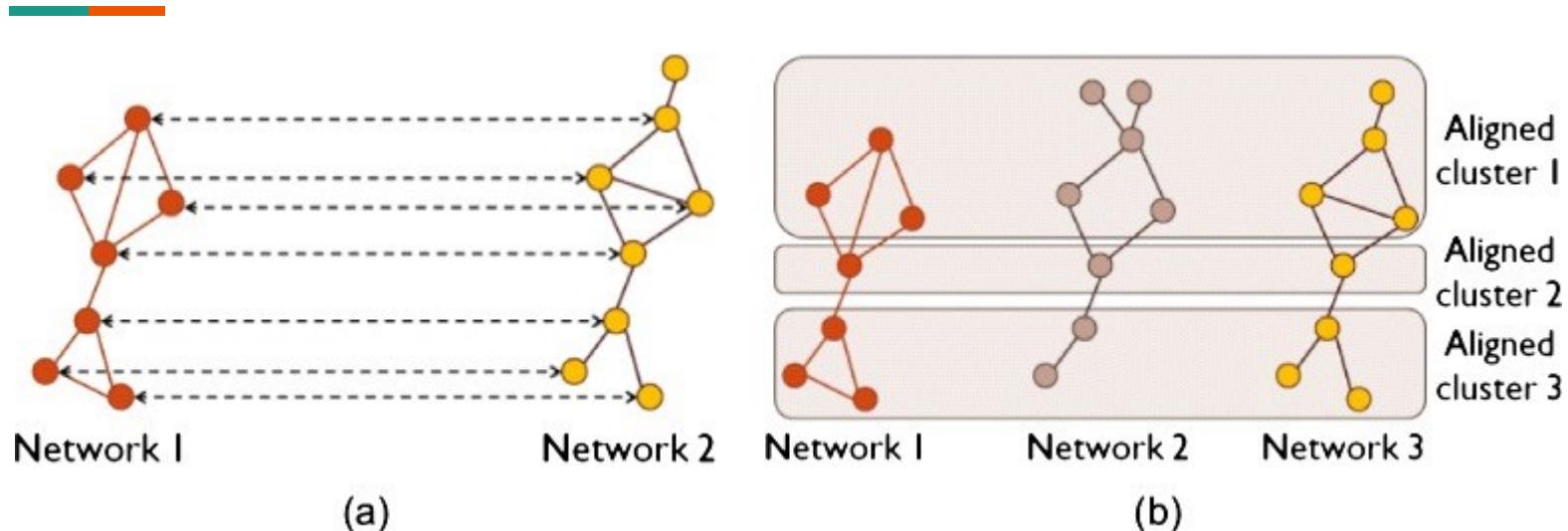
e Protein folding



f Cell proliferation



# Evolution as change in networks



Faisal et al. EURASIP J Bioinform Syst Biol (2015)

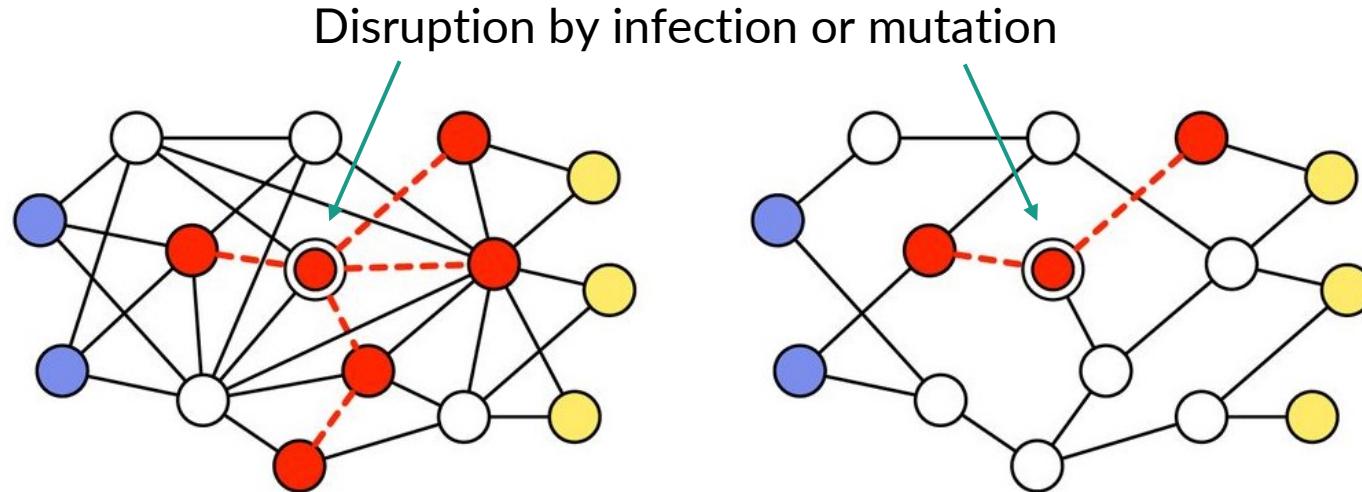
- Identify gene expansion and emergence of new interactions
- Predict missing interactions for validation



# Applications of network analyses

# What if nodes/edges were removed?

---

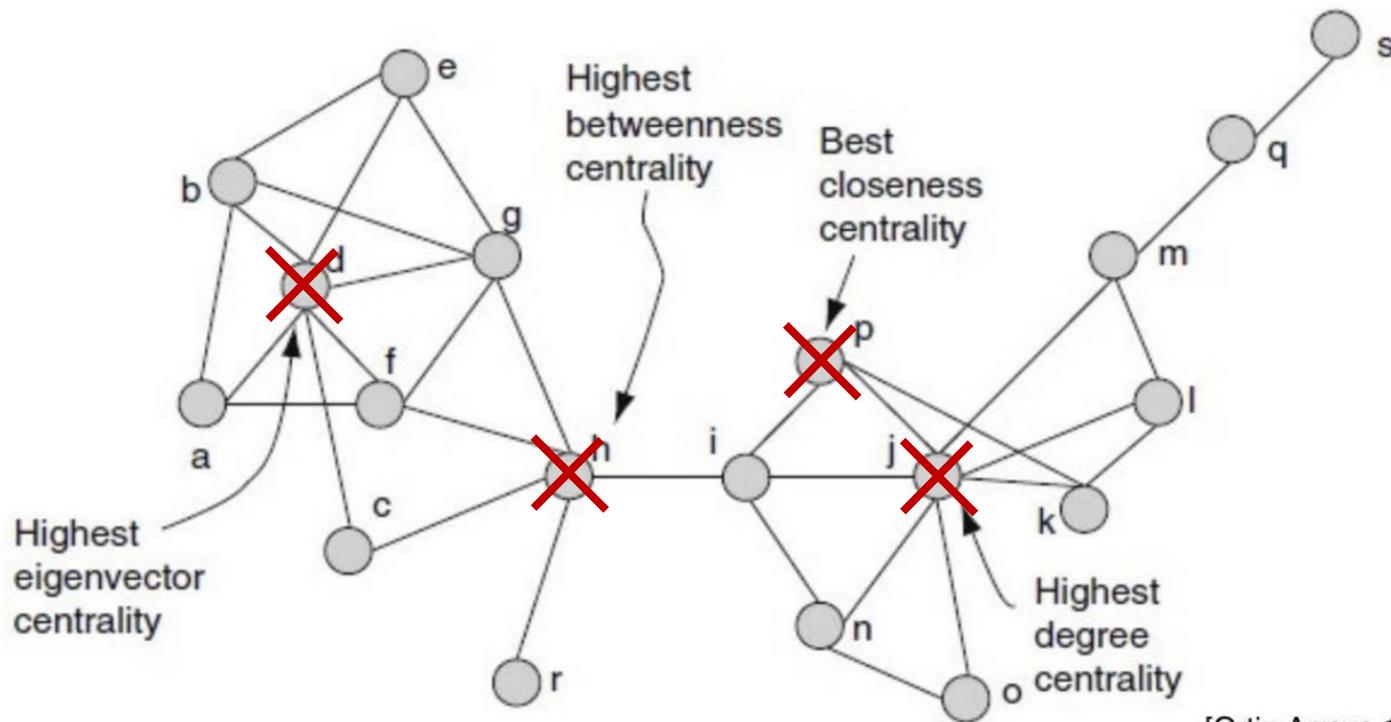


Navlakha *et al.* J of the Royal Society Interface, 11 (2014)

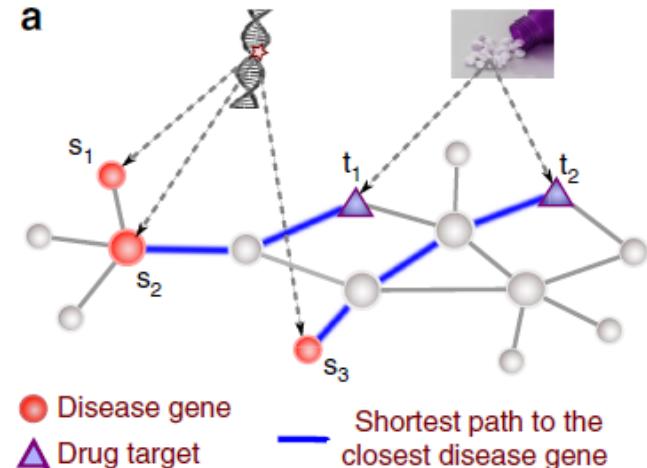
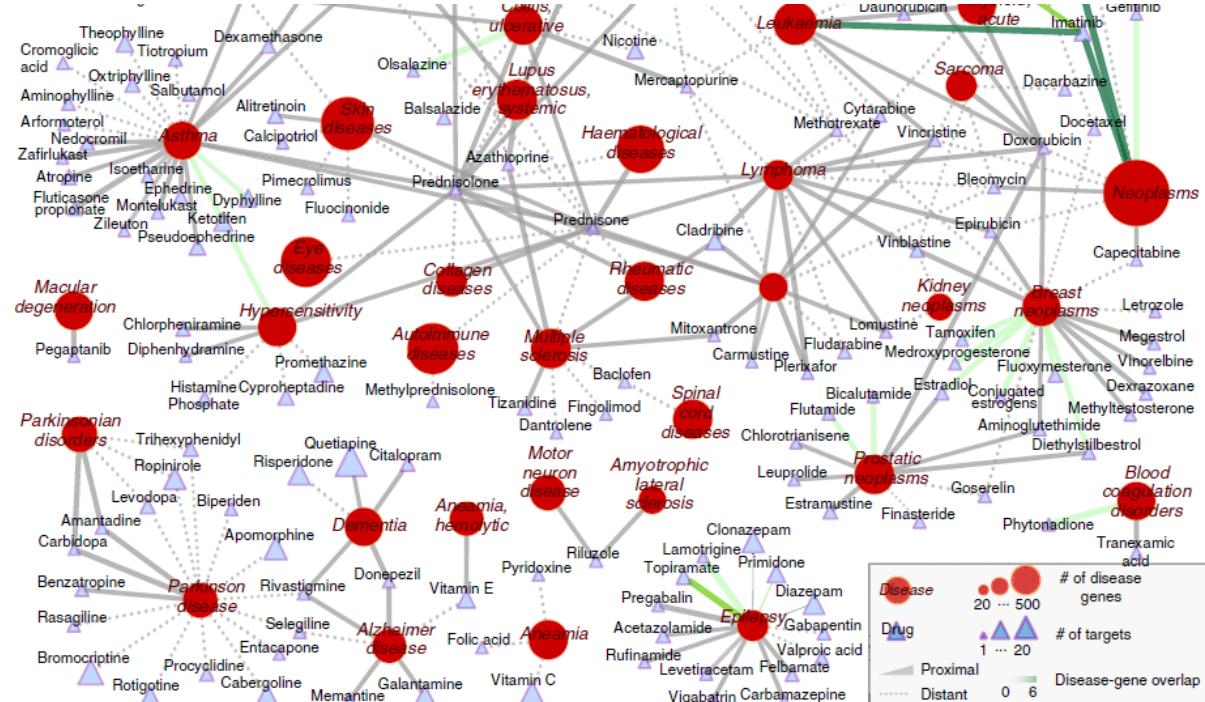
- Analysis of network-level changes induced by node/edge-level changes
- Complement centrality scores

# Variety of expected impacts

---



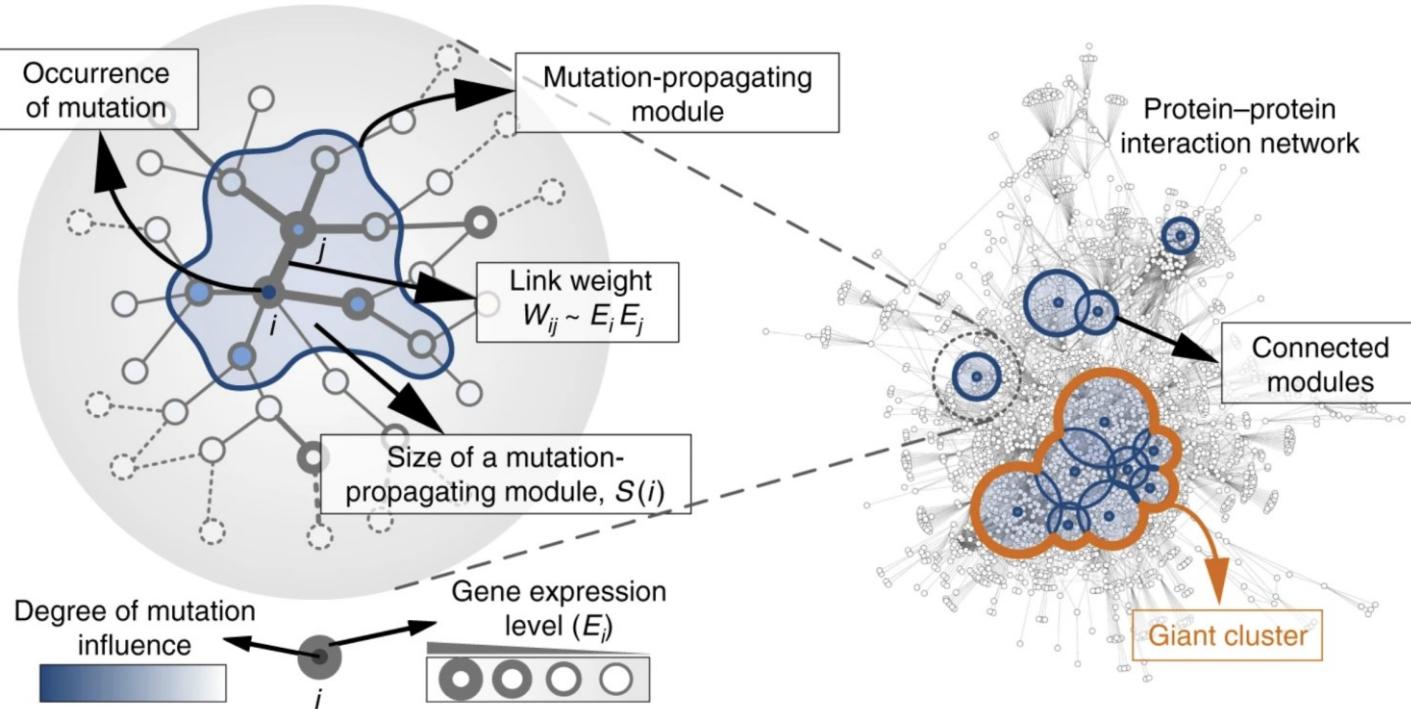
# Linking drug to disease via gene network



Guneys et al. Nature Comm, 7:10331 (2016)

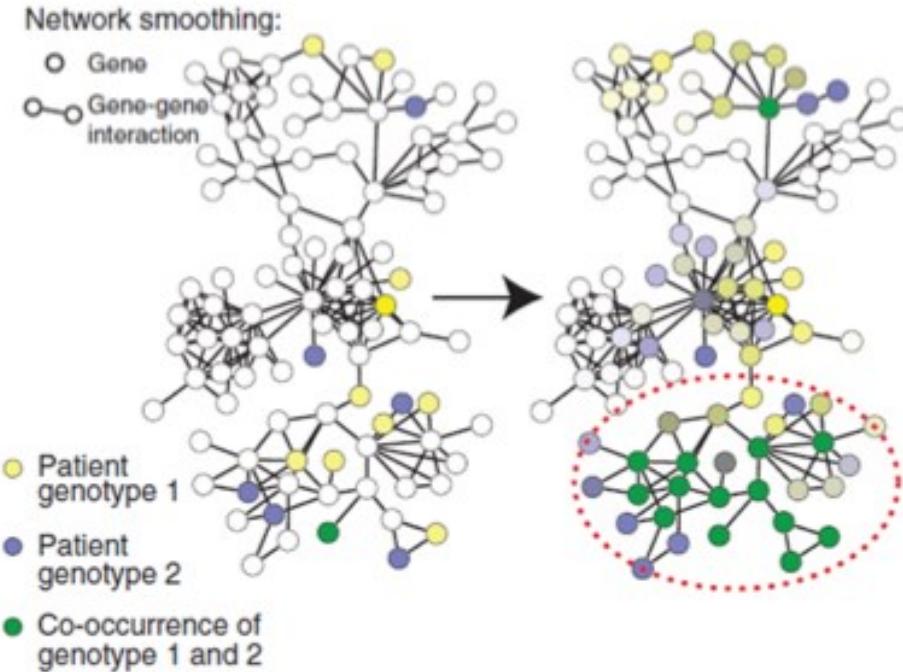
# Propagate effect of mutations through network

a



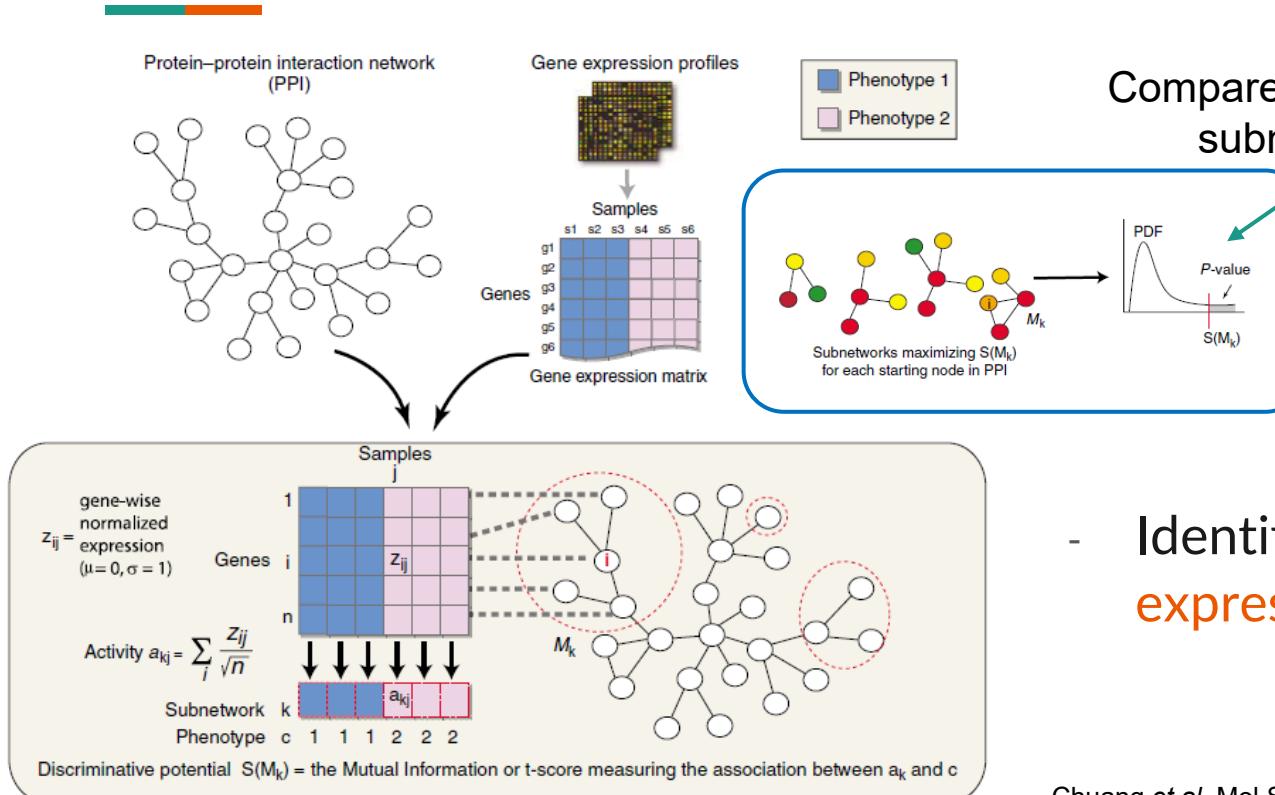
# Network-based patient stratification

---



- Different patients have different mutation profiles
- Different mutation profiles may have similar impacts on gene-gene network
- Identify commonly affected gene subnetworks

# Network-based differential expression



Chuang et al. Mol Syst Biol, 3:140 (2007)

# Summary

---

- Network data capture interaction information
  - Provide a bird-eye view of the biological systems
  - Lead to mechanistic understanding
- Connectivity analysis can reveal important components and interactions
  - Path & flow techniques
  - Perturbation
- Motif and homology analysis can help us discover missing data

# Any question?

---

- See you on October 19