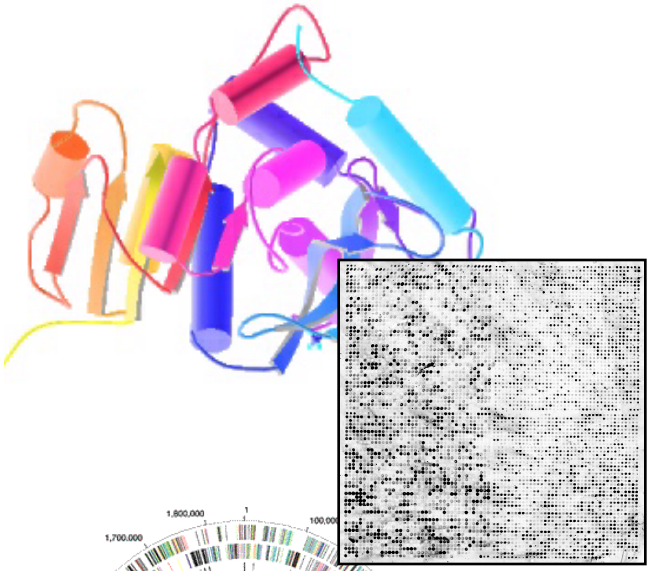
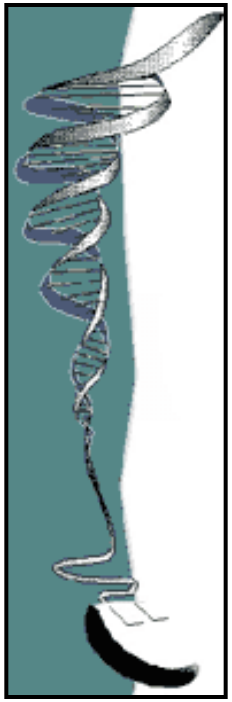
Biomedical Data Science: Analysis of Network Topology





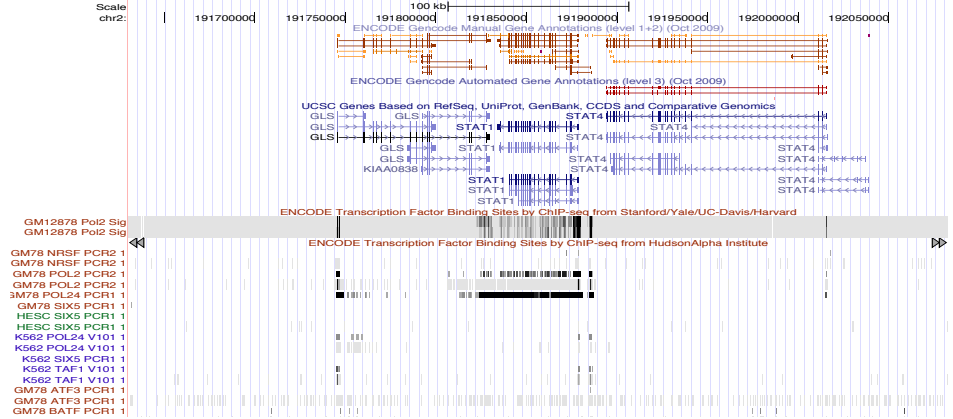
Mark Gerstein, Yale University

gersteinlab.org/courses/452

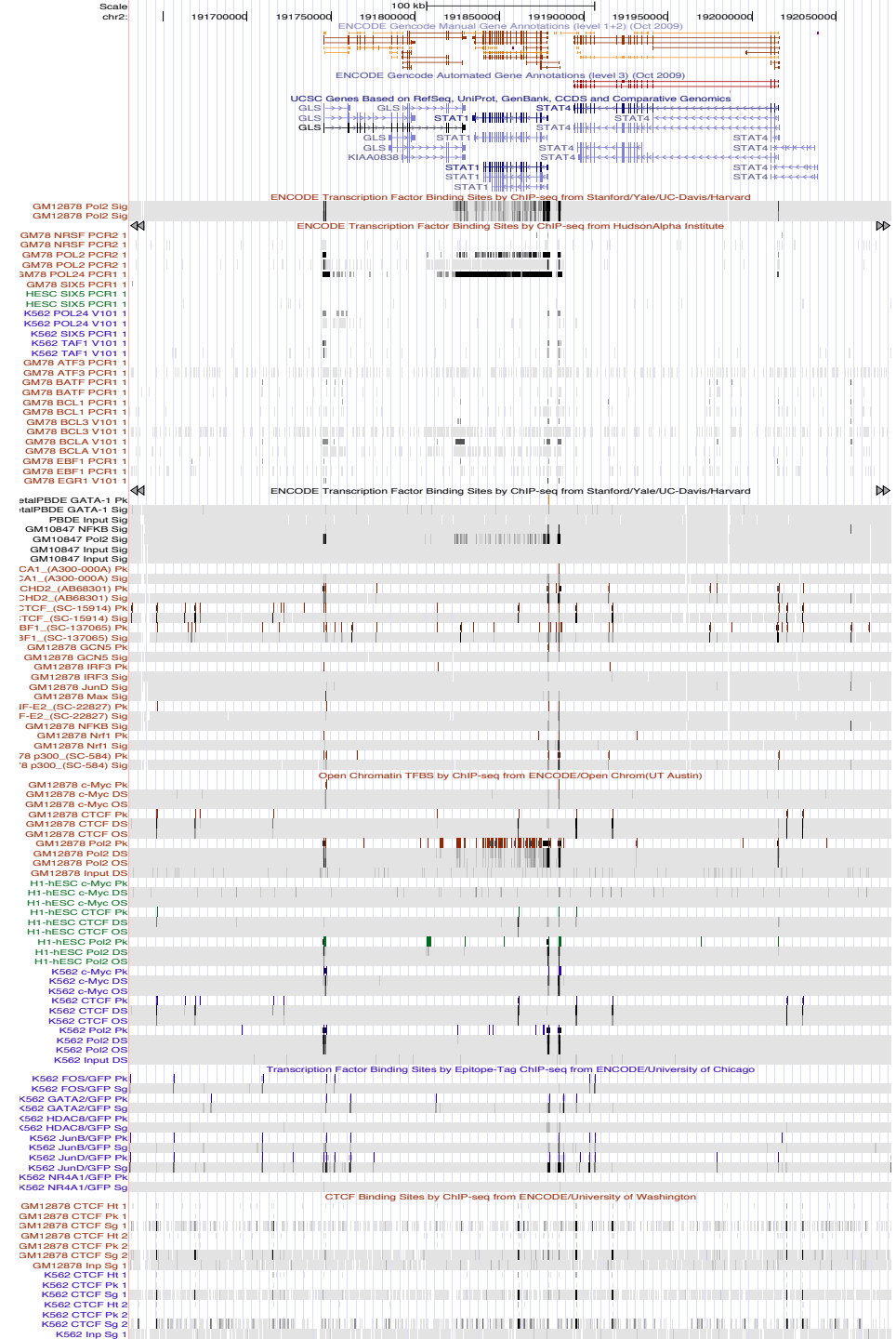
(last edit in spring '20)

Network Topology

Reasons for Networks: Overcome shortcomings of linear genome annotation



Current Annotation: 1D Browser Tracks

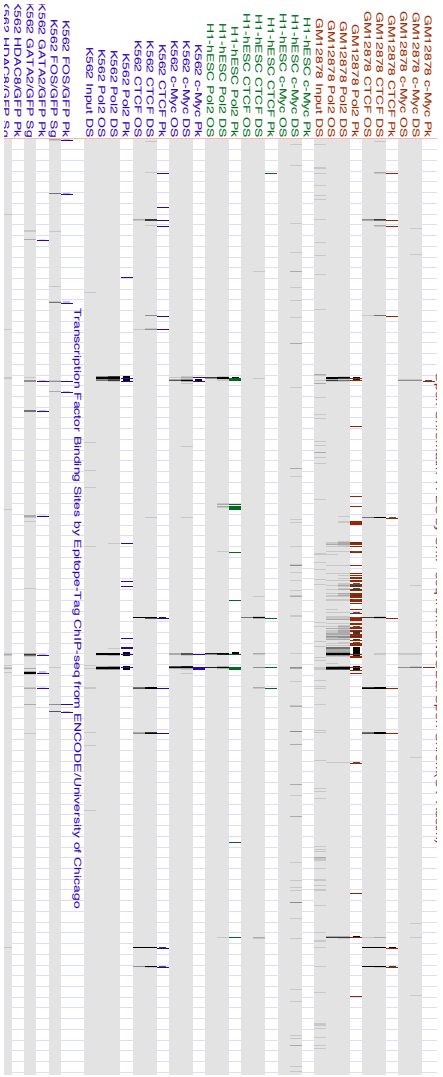


Current Annotation: 1D Browser Tracks

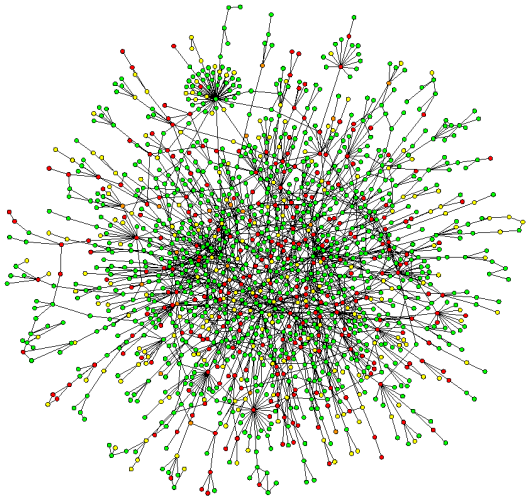
Will this scale to 1000+ tracks? What will next-gen annotation look like?...

Systems from Parts

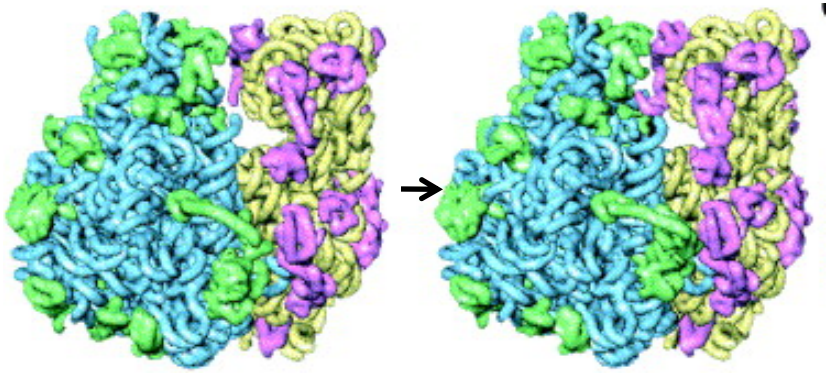
Networks occupy a midway point in terms of level of understanding



1D: Complete Partslist ("Elements" in genomic tracks)



~2D: Network Wiring Diagram of a Molecular System



3D & 4D:
Detailed structural understanding of cellular machinery
(e.g. ribosome in different functional states)

[UCSC genome browser]

[Jeong et al. Nature, 41:411]

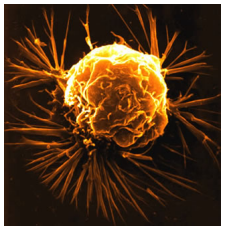
[Chiu et al. Trends in Cell Biol, 16:144]

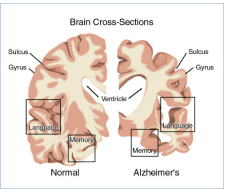
(c) M Gerstein, gerstein.info/talks

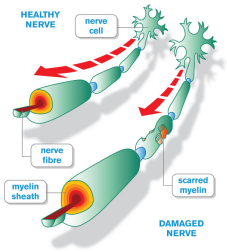
Network Topology

Reasons for Networks: Useful way of thinking about disease

Network pathology & pharmacology





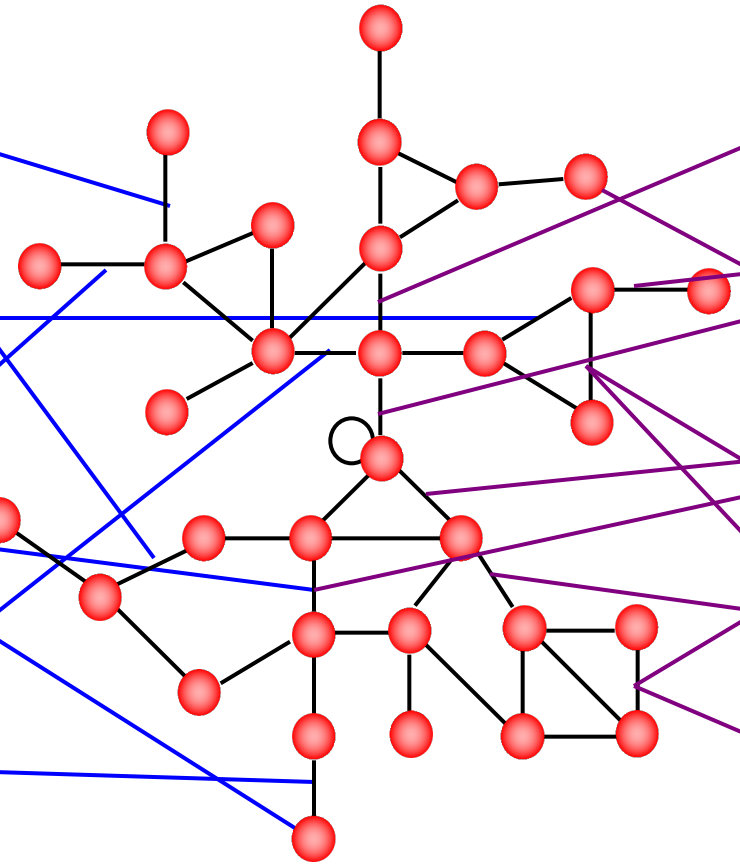


Breast Cancer

Alzheimer's Disease

Parkinson's Disease

Multiple Sclerosis



Interactome Networks and Human Disease

Vol 455|23 October 2008|doi:10.1038/nature07385 nature

ARTICLES

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

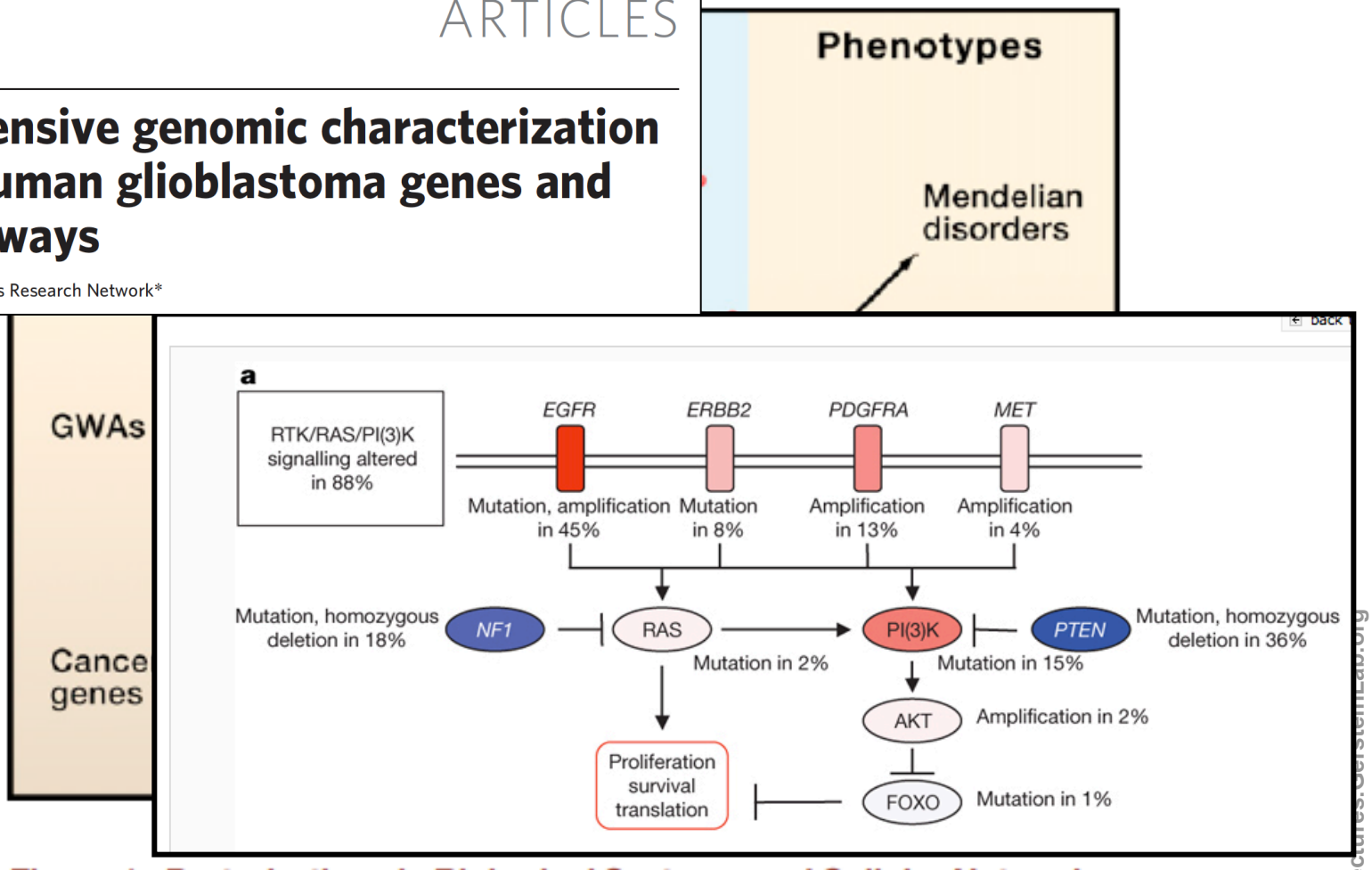
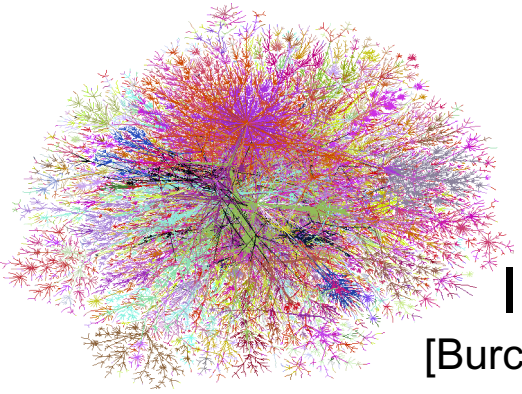


Figure 1. Perturbations in Biological Systems and Cellular Networks May Underlie Genotype-Phenotype Relationships

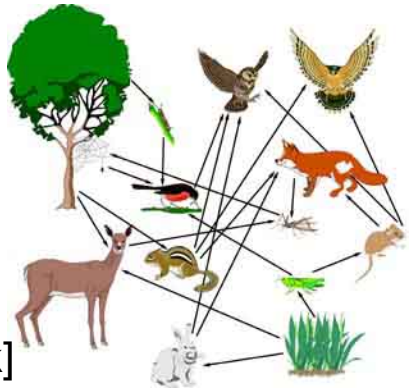
Network Topology

Reasons for Networks: Comprehensive representation, capable of representing many types of biological & non-biological data & bridging between disciplines

Networks as a universal language

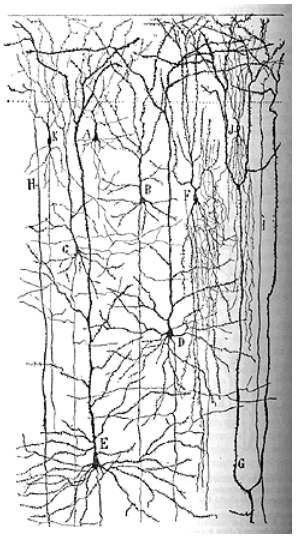


Internet
[Burch & Cheswick]

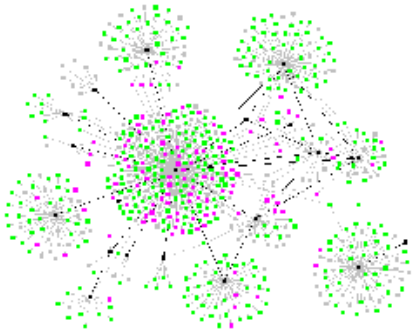


Food Web

Electronic Circuit

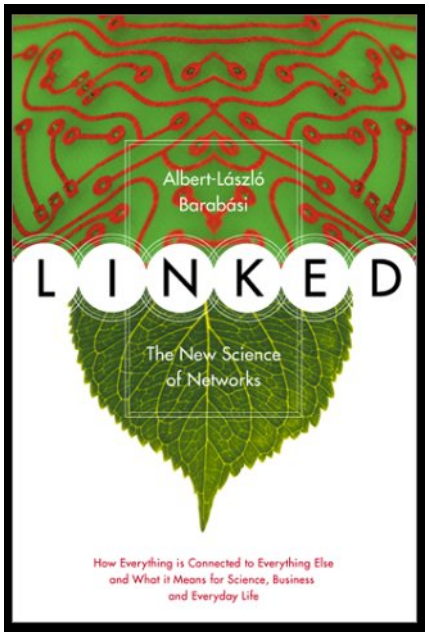


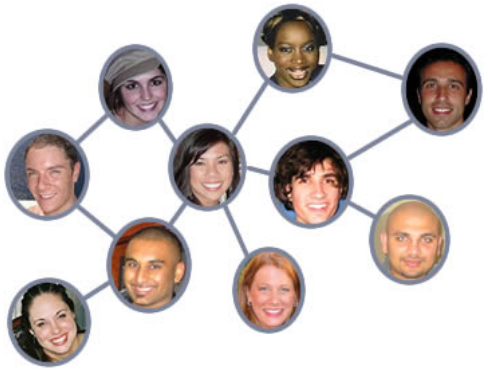
Neural Network
[Cajal]



Disease Spread
[Krebs]

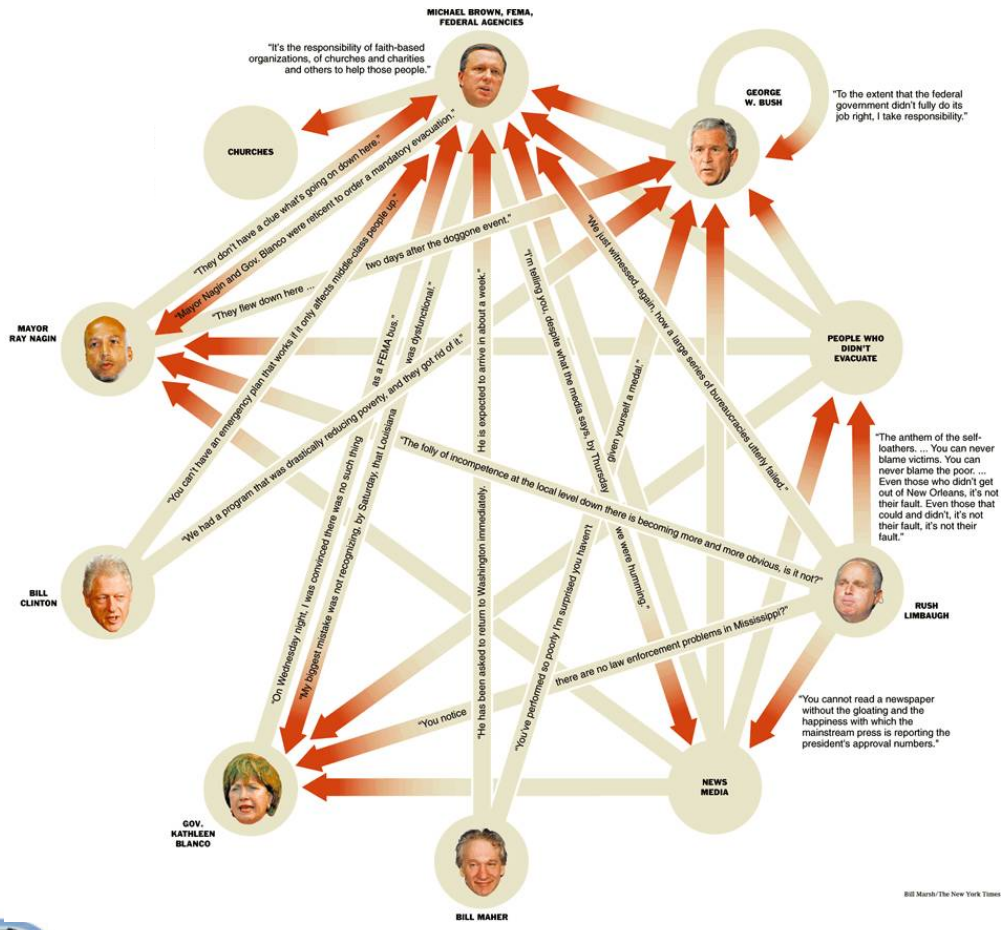
Protein Interactions
[Barabasi]



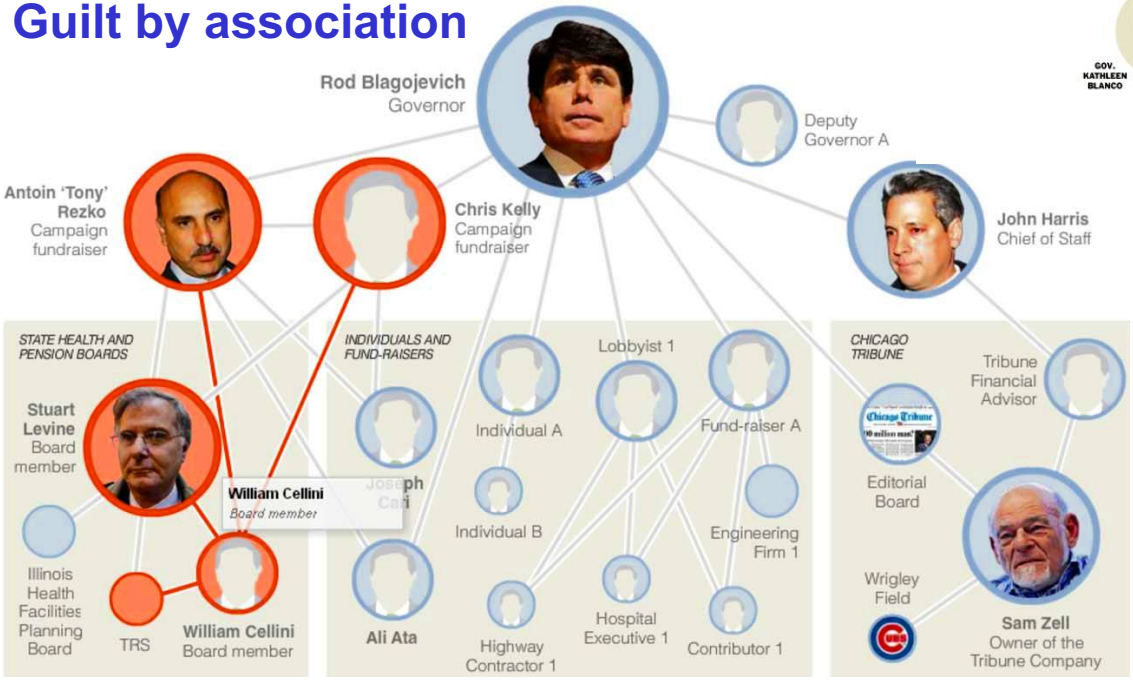


Social Network

Using the position in networks to describe function



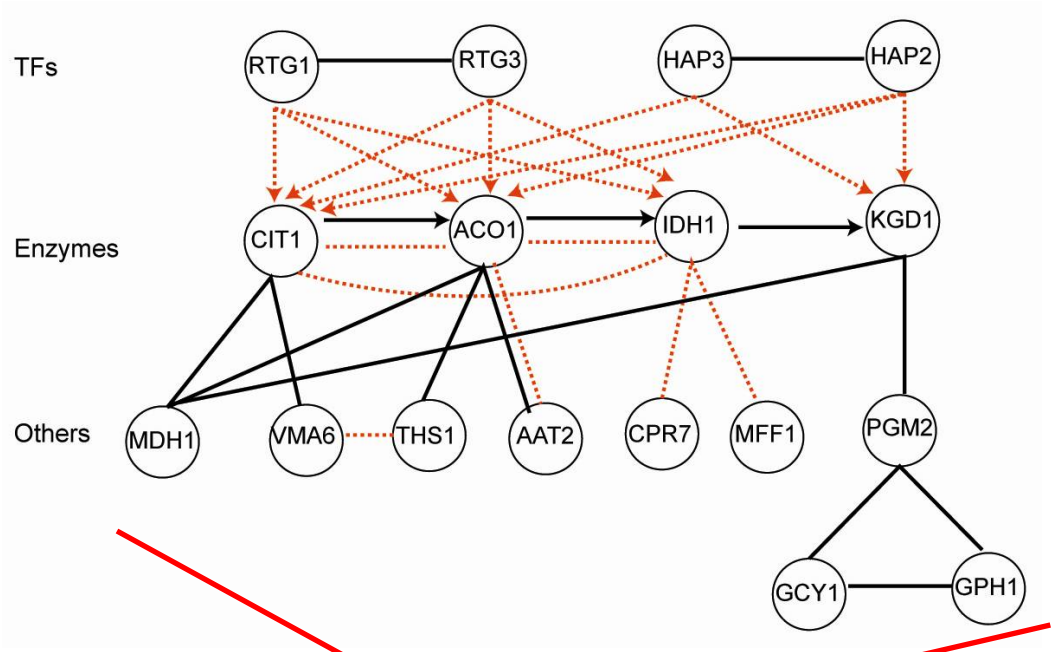
Guilt by association

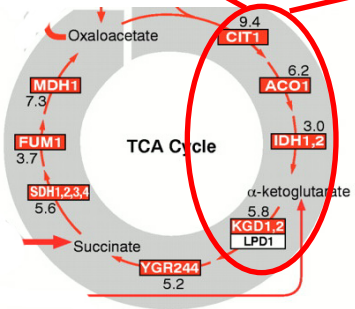


Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]

Combining networks forms an ideal way of integrating diverse information





Part of the TCA cycle

Metabolic pathway

Transcriptional regulatory network

Physical protein-protein Interaction

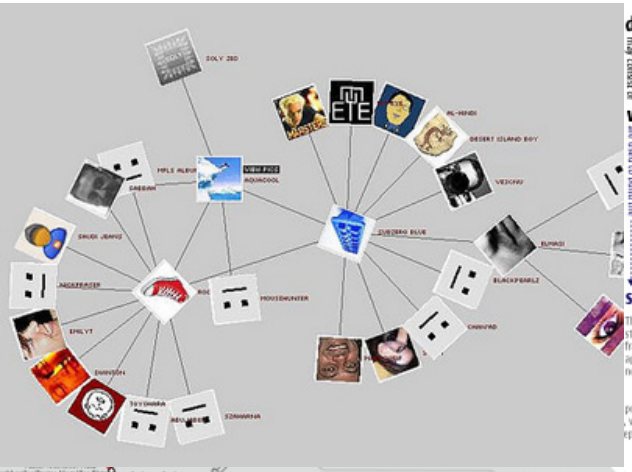
Co-expression Relationship

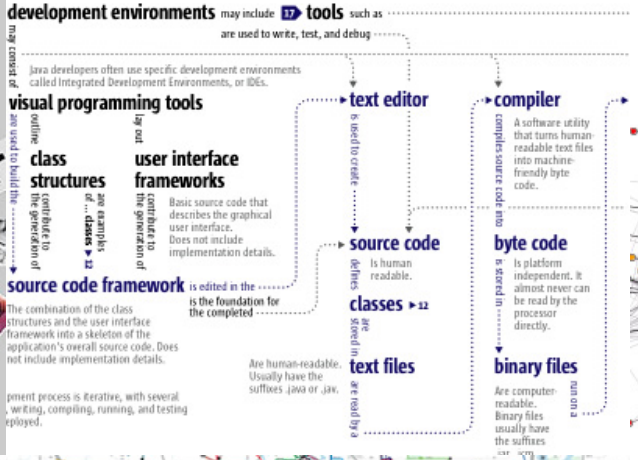
Genetic interaction (synthetic lethal)
Signaling pathways

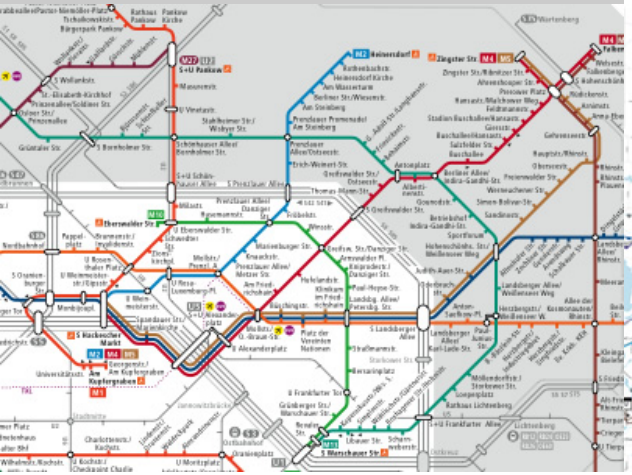
Richness of the Visual Representation of Networks

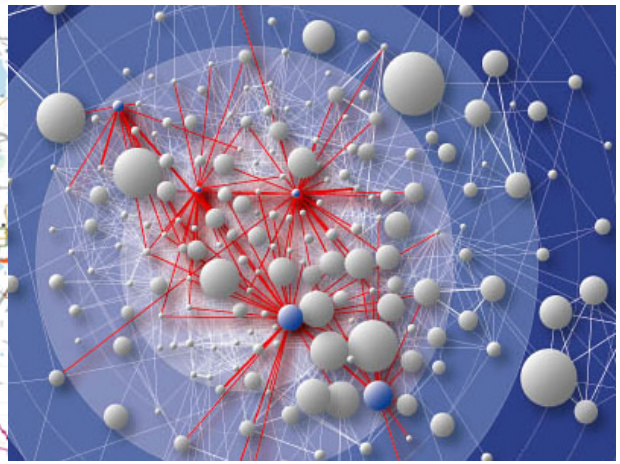
VisualComplexity.com

Some structure (connectivity) but some flexibility (e.g. edge colors, node positions and shapes) that can used to encode additional information









(c) M Gerstein, gerstein.info/talks

Network Topology

Building Networks in Genomics

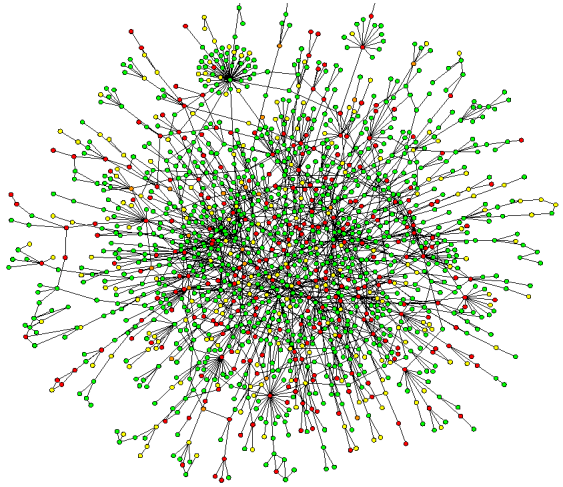
Origin of Networks

- Protein-protein interactions
  - ◊ Phosphorylation networks
- Metabolic Networks
- Regulatory networks
  - ◊ from Chip-Seq (see next slide)

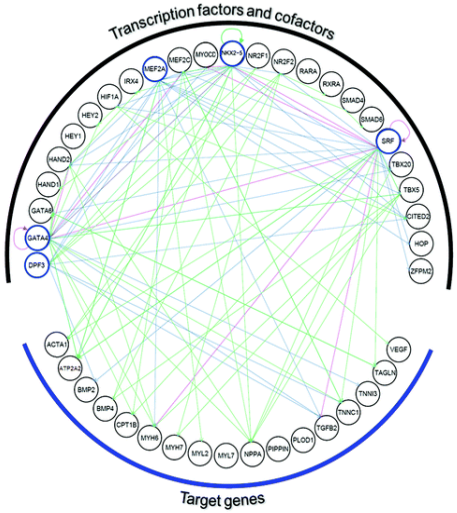
- "Squared" scale
  - ◊ 6K genes in yeast but ~18M potential interactions (6000 chose 2 pairs of interactions)

(c) M Gerstein, gerstein.info/talks

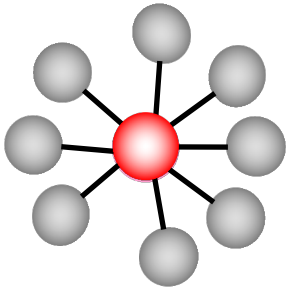
Different Types of Molecular Networks



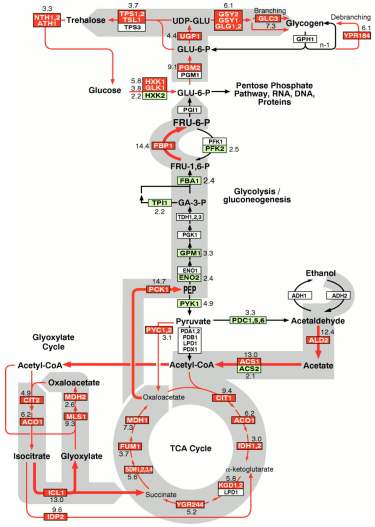
Protein-protein Interaction networks



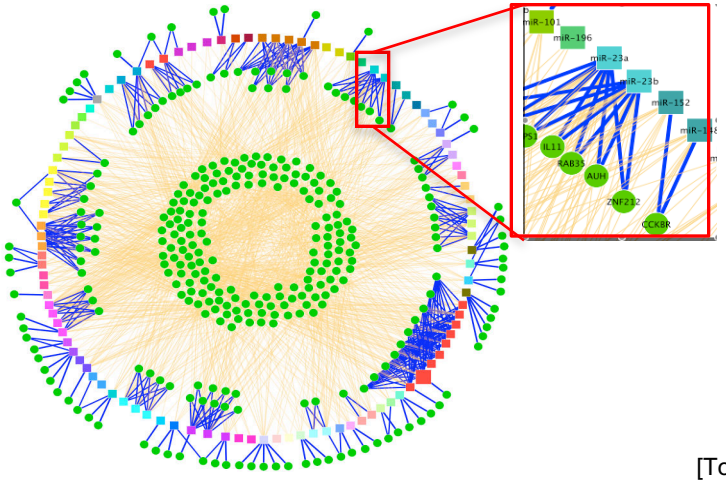
TF-target-gene Regulatory networks



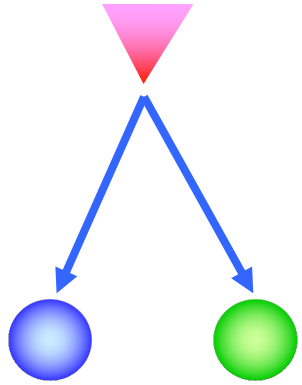
Undirected



Metabolic pathway networks



miRNA-target networks



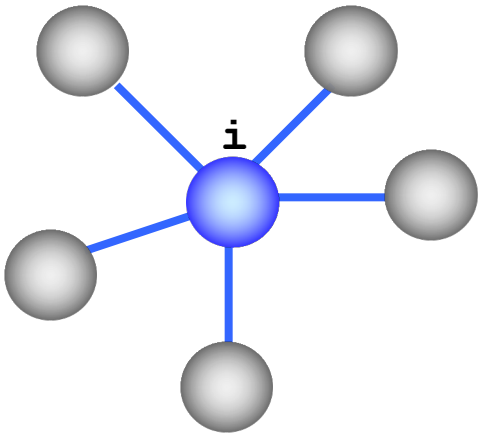
Directed

[Toenjes, *et al*, *Mol. BioSyst.* (2008); Jeong *et al*, *Nature* (2001); [Horak, et al, Genes & Development, 16:3017-3033; DeRisi, Iyer, and Brown, Science, 278:680-686]

Network Topology

What are the Main Quantities that Can be Calculated from Networks?

- Degree of a node: the number of edges incident on the node

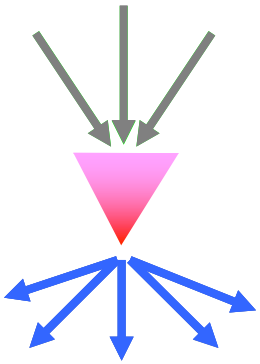


Degree of node i = 5

(c) M Gerstein, gerstein.info/talks

Network parameters

Number of incoming and outgoing connections



Connectivity

Incoming connections = 2.2

→each gene is regulated by ~2 TFs

In-degree

Outgoing connections = 20.2

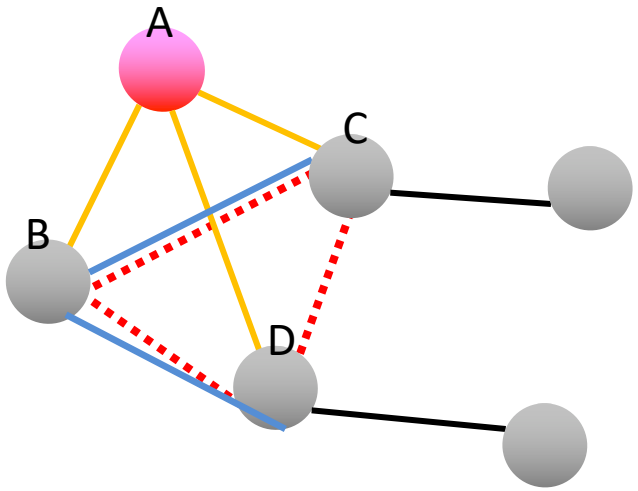
→each TF regulates ~20 genes

Out-degree

(c) M Gerstein, gerstein.info/talks

Clustering coefficient

- Clustering Coefficient:
  - Ratio of existing links to maximum number of links for neighbouring nodes
  - Example:
    - For A:
      - 3 neighbours
      - 2 existing link
      - 3 possible links
    - Clustering coefficient
      - $C_A=2/3$



Example from: http://www.learner.org/courses/mathilluminated/units/11/textbook/04.php

Clustering coefficient

- Average Coefficient:
  - Average of clustering coefficients of all nodes n

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

  - Measure of inter-connectedness of the network
  - Global property

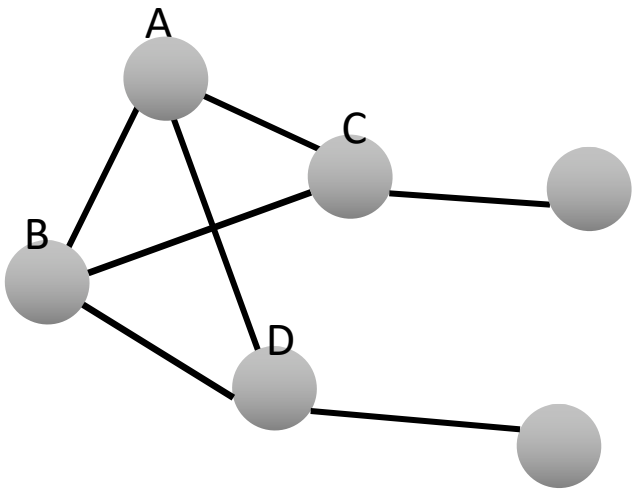
- Example:
  - Clustering coefficient:

C_A=2/3 C_B=2/3

C_C=1/3 C_D=1/3

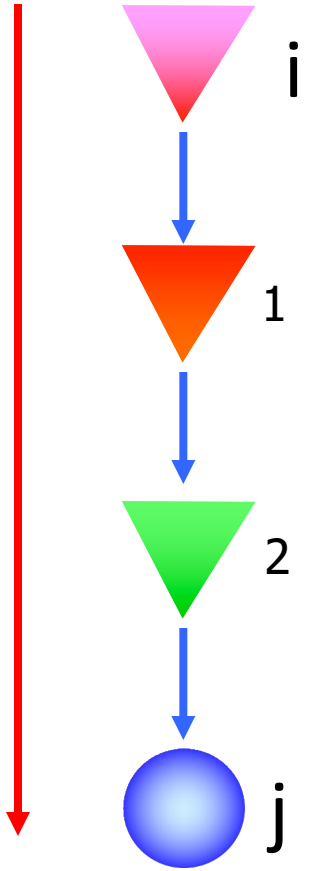
  - Average coefficient =

1/4(2/3+2/3+1/3+1/3)= 0.5



Example from: http://www.learner.org/courses/mathilluminated/units/11/textbook/04.php

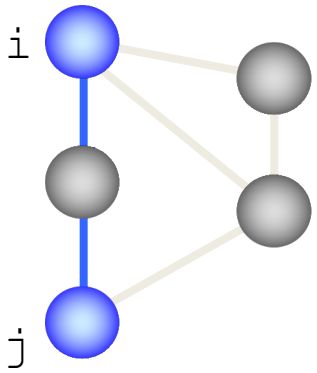
Path length



- Number of edges along a path
- Path length = 3
- Meaning:
  - Number of intermediate TFs to reach final target
  - Indication of how immediate a response is

Path length

- Shortest path length:
  - $L_{(i,j)}$ is the minimum number of edges that must be traversed to travel from a vertex i to another vertex j of a graph G



$$L_{(i,j)} = 2$$

Graph Theory Terminology (Batten, pp. 92-105)

- Characteristic path length (Average path length)
  - The characteristic path length *L* of a graph is the average of the $L_{(i,j)}$ for every possible pair (i,j)

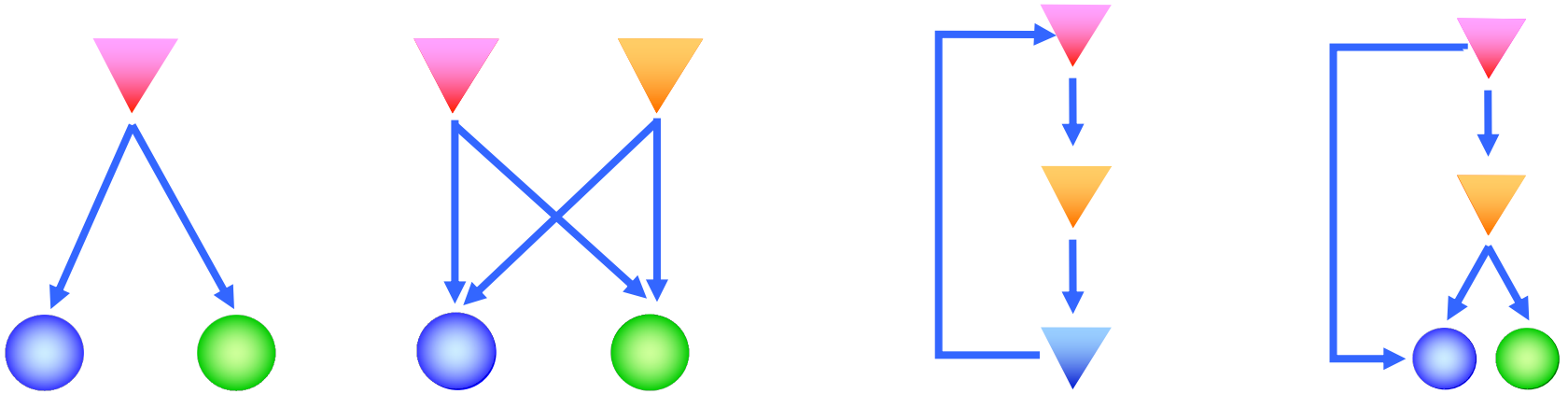
$$L = \frac{1}{n(n-1)} \sum_{i,j} L_{(i,j)}$$

  - Networks with small values of L are said to have the "small world property"

http://en.wikipedia.org/wiki/Average_path_length

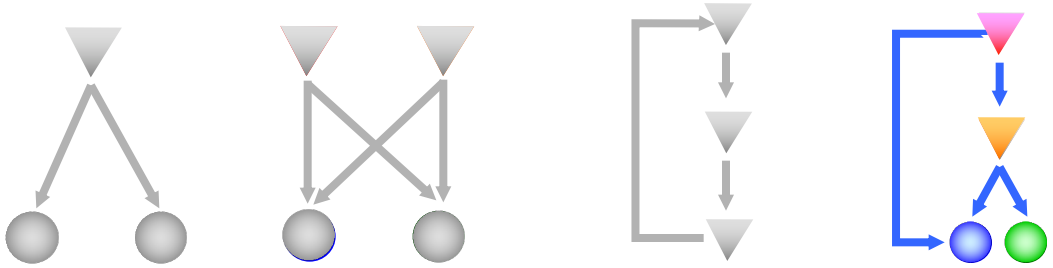
Network motifs

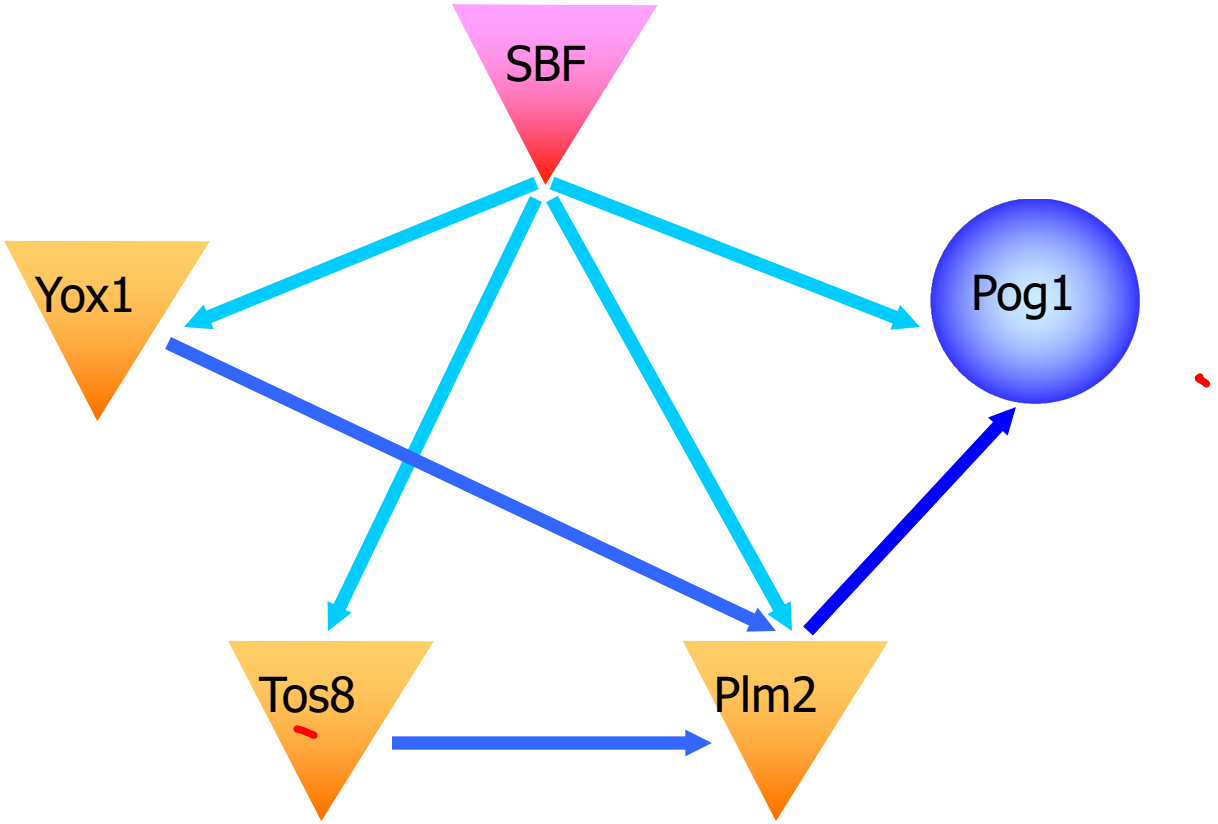
Regulatory modules within the network



[Alon U, 2007, Nature Reviews Genetics, Network motifs: theory and experimental approaches]

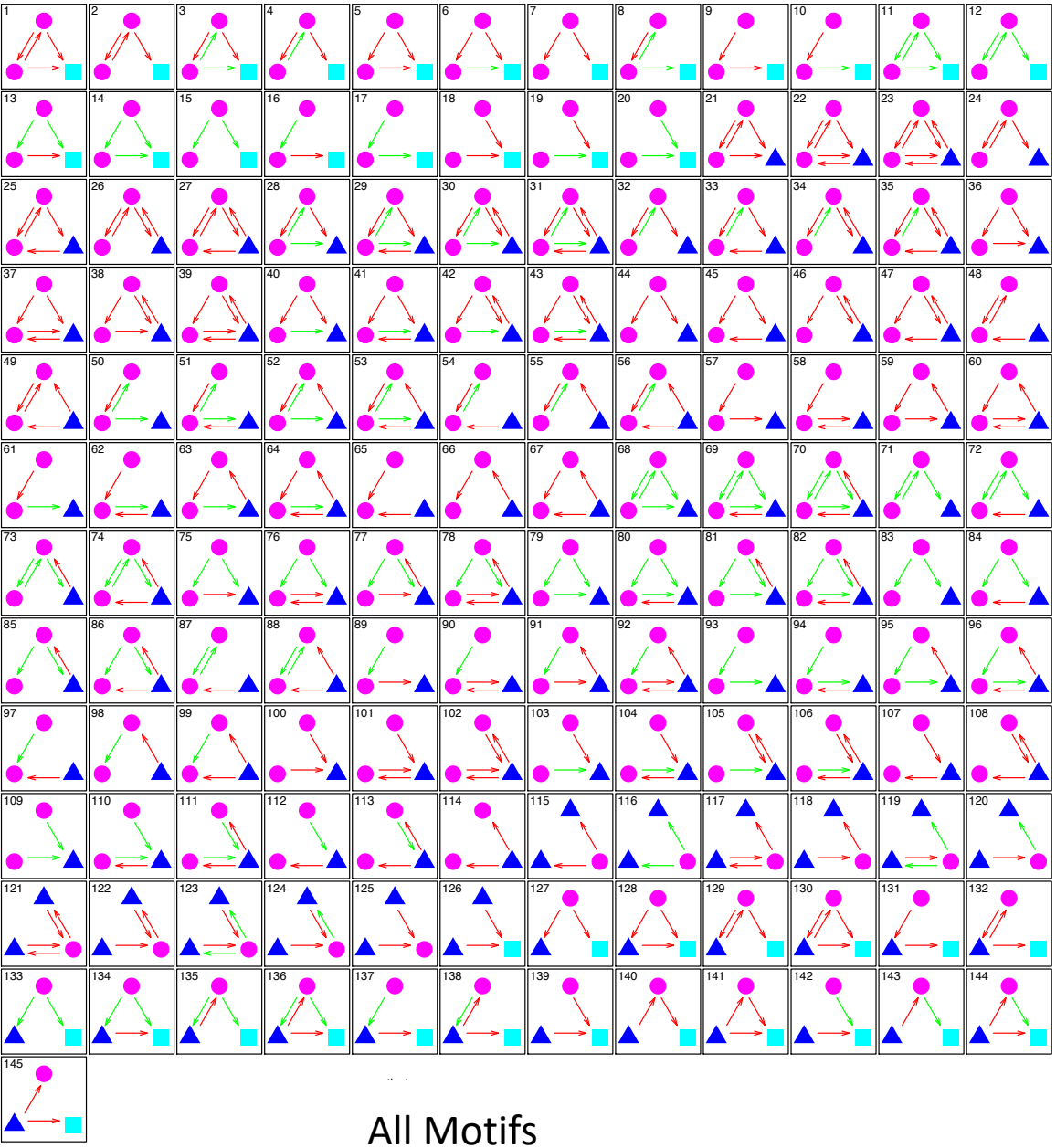
FFL = Feed-forward loops





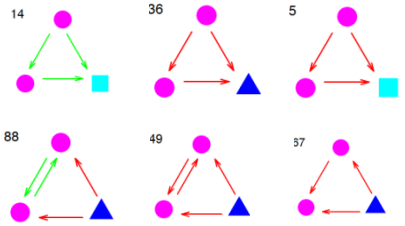
[Alon U, 2007, Nature Reviews Genetics]

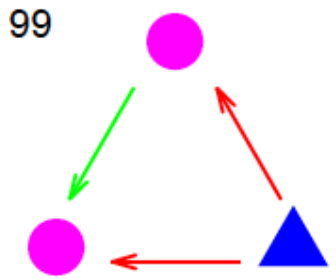
Network Motifs



All Motifs

7 Motifs
Over-represented





FFL involving
miRNA & 2 TFs

TF MIR GENE

activate
repress

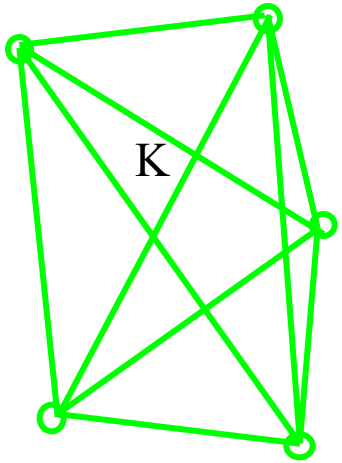
[*Science* 330:6012; Cheng et al. *PLOS CB* '11]

Cliques

- Fully connected sub-components
- Related measures

  k-cores : For all vertices in a graph G have degree at least k

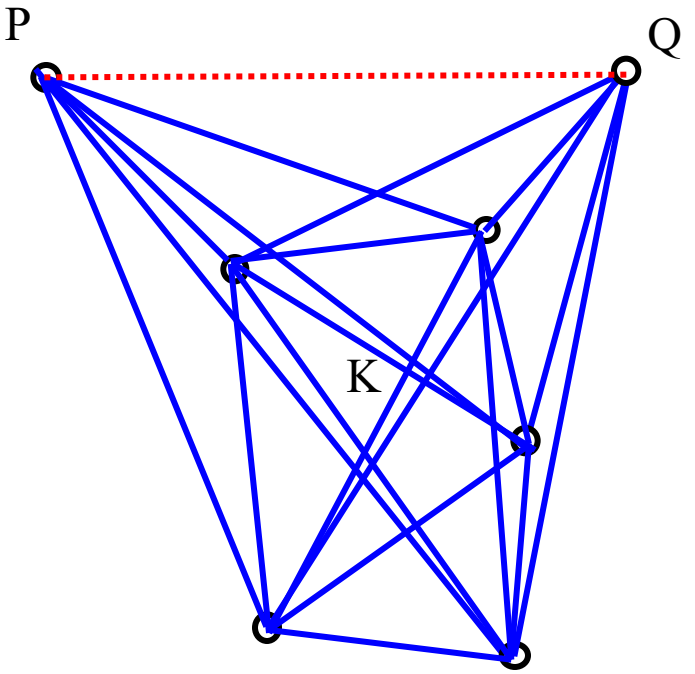
[Hogue et al, BMC BIOINFORMATICS, 2003]



Problem: High-throughput experiments are prone to missing interactions

One solution—defective cliques

- If proteins P and Q interact with a clique K of proteins which all interact with each other, then P and Q are more likely to interact with each other
- P, Q, and K form a defective clique



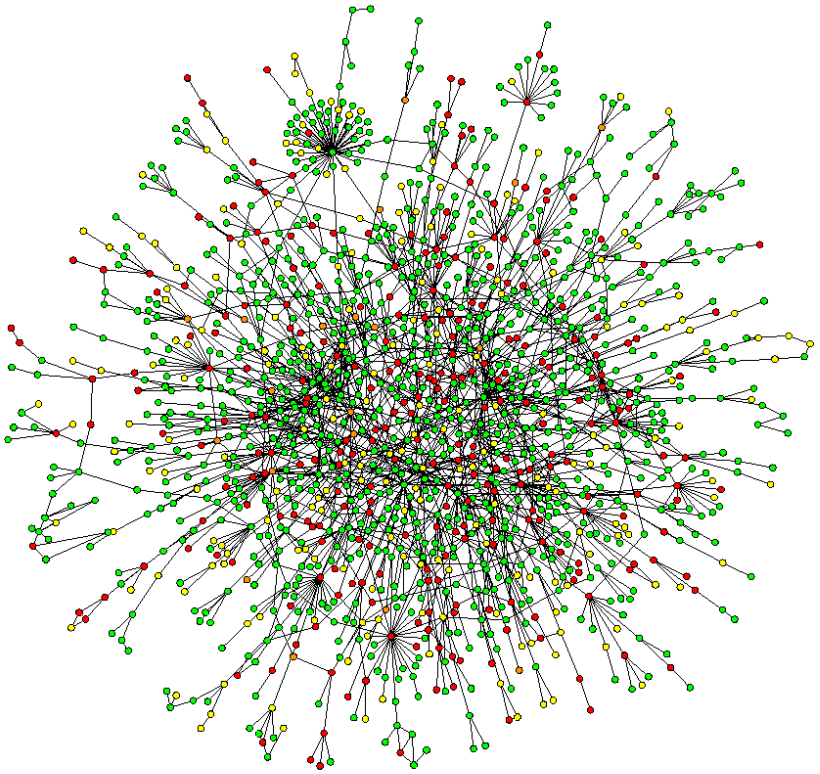
Predicting protein interactions by completing defective cliques

[Yu et al. Bioinformatics (2006)]

Network Topology

Simple Mathematical Models for Interpreting Complex Topology: Scale Free Networks & Hubs

Models for networks of complex topology



- Erdos-Renyi (1960)
- Watts-Strogatz (1998)
- Barabasi-Albert (1999)

A Barabási & R Albert "Emergence of scaling in random networks," *Science* 286, 509-512 (1999).

(c) M Gerstein, gerstein.info/talks

The Erdős-Rényi [ER] model (1960)

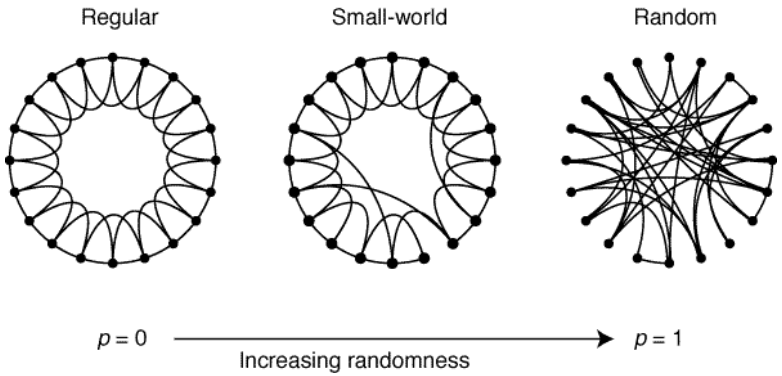
- Start with N vertices and no edges
- Connect each pair of vertices with probability P_{ER}

Important result: many properties in these graphs appear quite suddenly, at a threshold value of $P_{ER}(N)$

-If $P_{ER} \sim c/N$ with $c<1$, then almost all vertices belong to isolated trees
-Cycles of all orders appear at $P_{ER} \sim 1/N$

(c) M Gerstein, gerstein.info/talks

The Watts-Strogatz [WS] model (1998)



- Start with a regular network with N vertices
- Rewire each edge with probability p

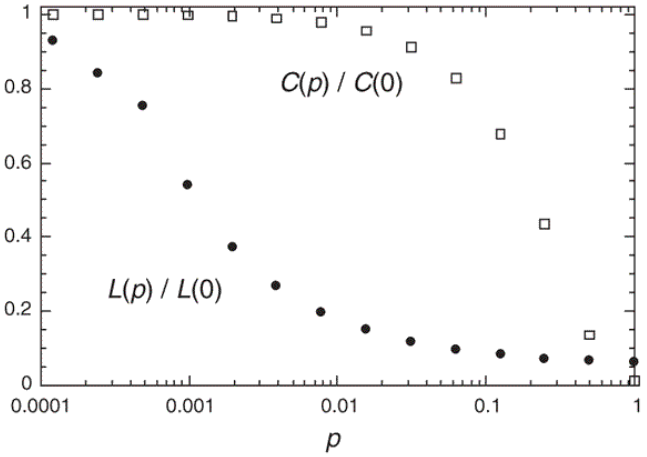
For p=0 (Regular Networks):
- high clustering coefficient
- high characteristic path length

For p=1 (Random Networks):
- low clustering coefficient
- low characteristic path length

QUESTION: What happens for intermediate values of p?

(c) M Gerstein, gerstein.info/talks

1) There is a broad interval of p for which L is small but C remains large



2) Small world networks are common :

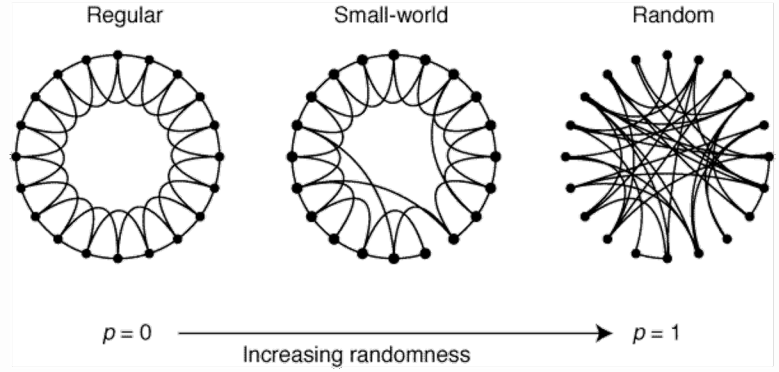
Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

(c) M Gerstein, gerstein.info/talks

Small world network

- A simple connected graph G exhibiting two properties:
  - **Large Clustering Coefficient:** Each vertex of G is linked to a relatively well-connected set of neighboring vertices, resulting in a large value for the clustering coefficient C(G);
  - **Small Characteristic Path Length**: The presence of short-cut connections between some vertices results in a small characteristic path length L(G).

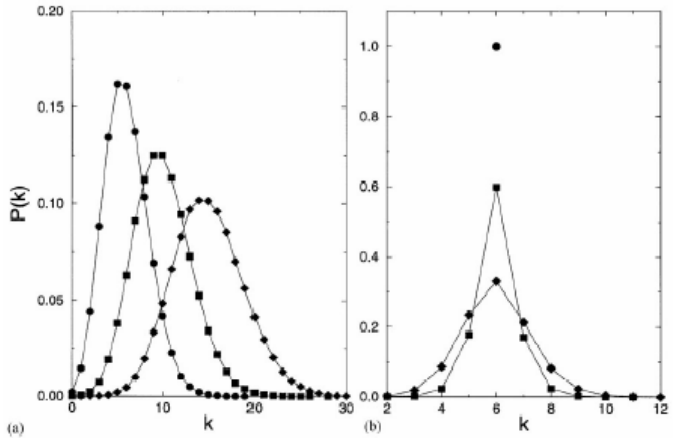


- local connectivity and global reach

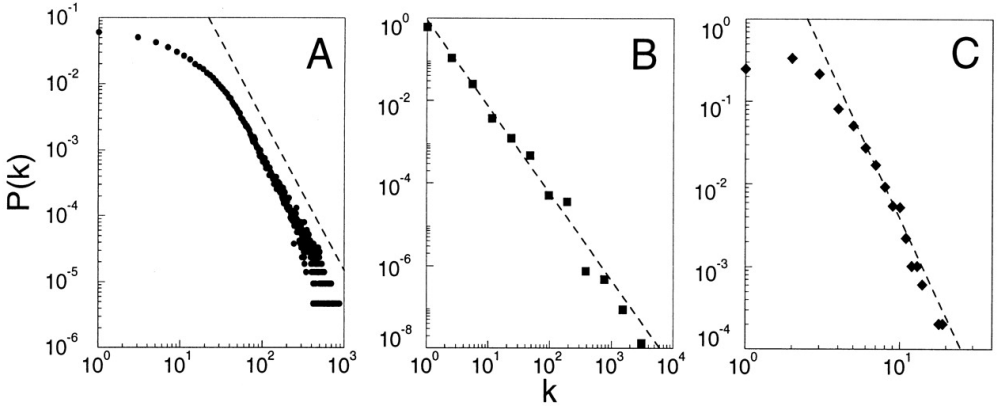
Watts and Strogatz (1998), Nature, Collective dynamics of 'small-world' networks

The Barabási-Albert [BA] model (1999)

Look at the distribution of degrees



ER Model WS Model



actors power grid www

The probability of finding a highly connected node decreases exponentially with k

$$P(K) \sim K^{-\gamma}$$

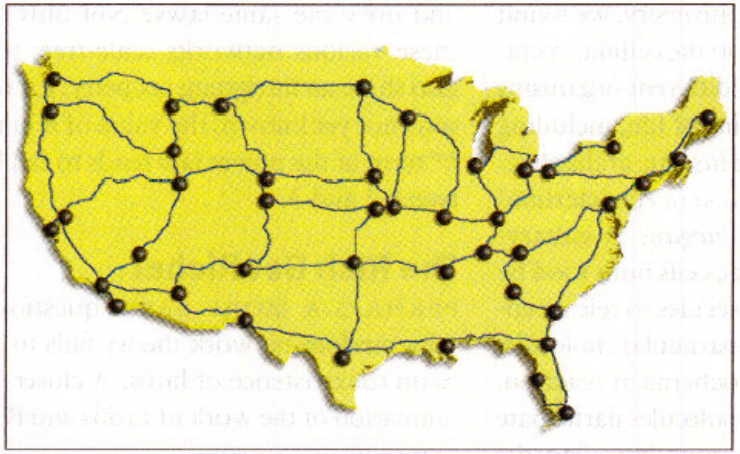
(c) M Gerstein, gerstein.info/talks

Random v Scale-free Networks

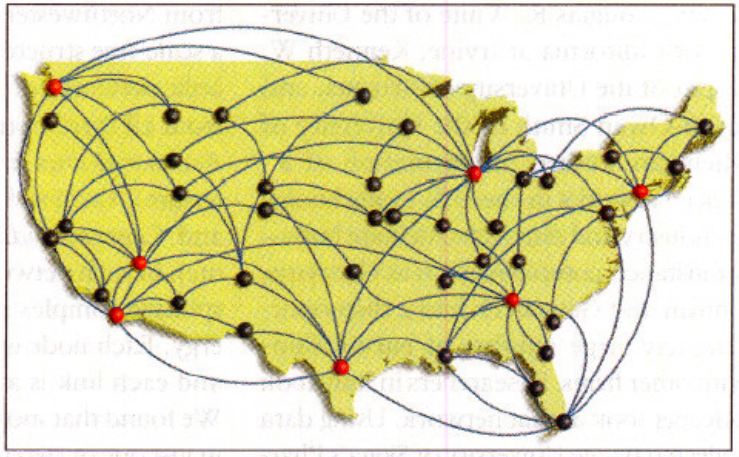
RANDOM NETWORKS, which resemble the U.S. highway system (*simplified in left map*), consist of nodes with randomly placed connections. In such systems, a plot of the distribution of node linkages will follow a bell-shaped curve (*left graph*), with most nodes having approximately the same number of links.

In contrast, scale-free networks, which resemble the U.S. airline system (*simplified in right map*), contain hubs (*red*)—nodes with a very high number of links. In such networks, the distribution of node linkages follows a power law (*center graph*) in that most nodes have just a few connections and some have a tremendous number of links. In that sense, the system has no "scale." The defining characteristic of such networks is that the distribution of links, if plotted on a double-logarithmic scale (*right graph*), results in a straight line.

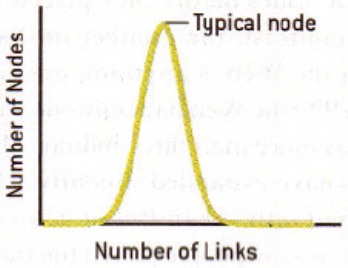
Random Network



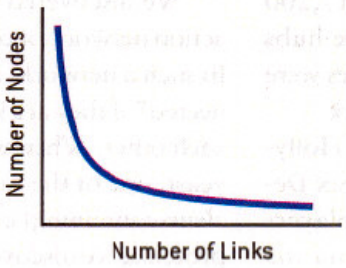
Scale-Free Network

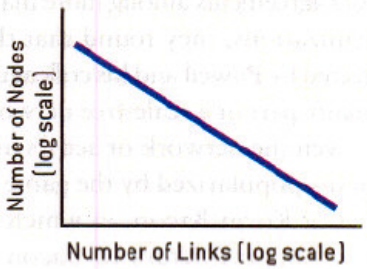


Bell Curve Distribution of Node Linkages



Power Law Distribution of Node Linkages



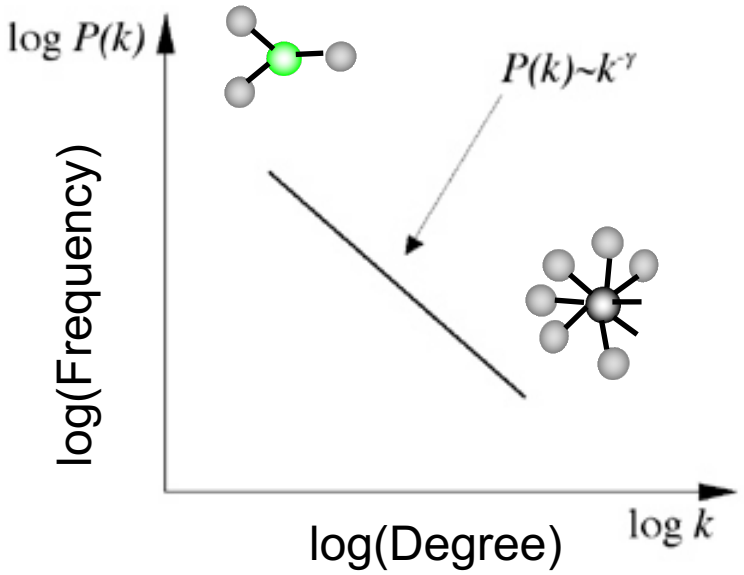


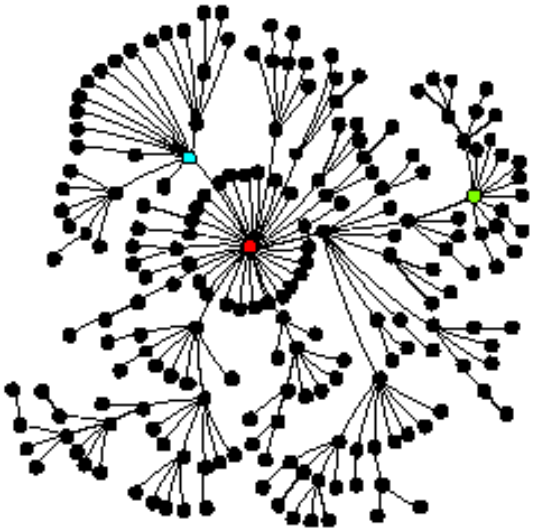
[From Barabasi & Bonabeau, *Sci. Am.*, May '03]

(c) M Gerstein, gerstein.info/talks

Scale-free networks in Biology

Power-law distribution





Hubs dictate the structure of the network

[Barabasi]

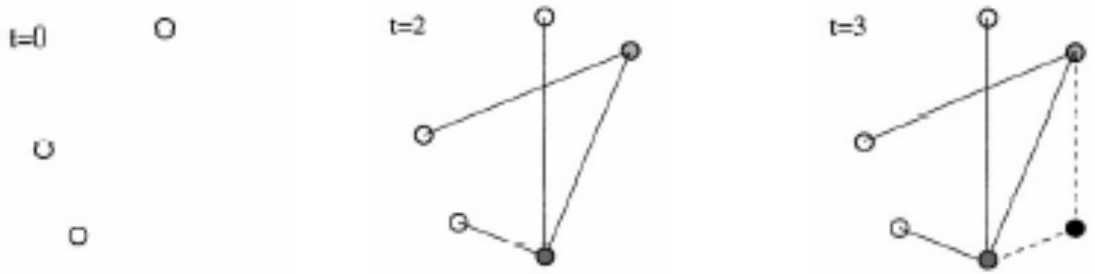
(c) M Gerstein, gerstein.info/talks

- two problems with the previous models:
    1. N does not vary
    2. the probability that two vertices are connected is uniform

• GROWTH: starting with a small number of vertices m_0 at every timestep add a new vertex with $m \leq m_0$

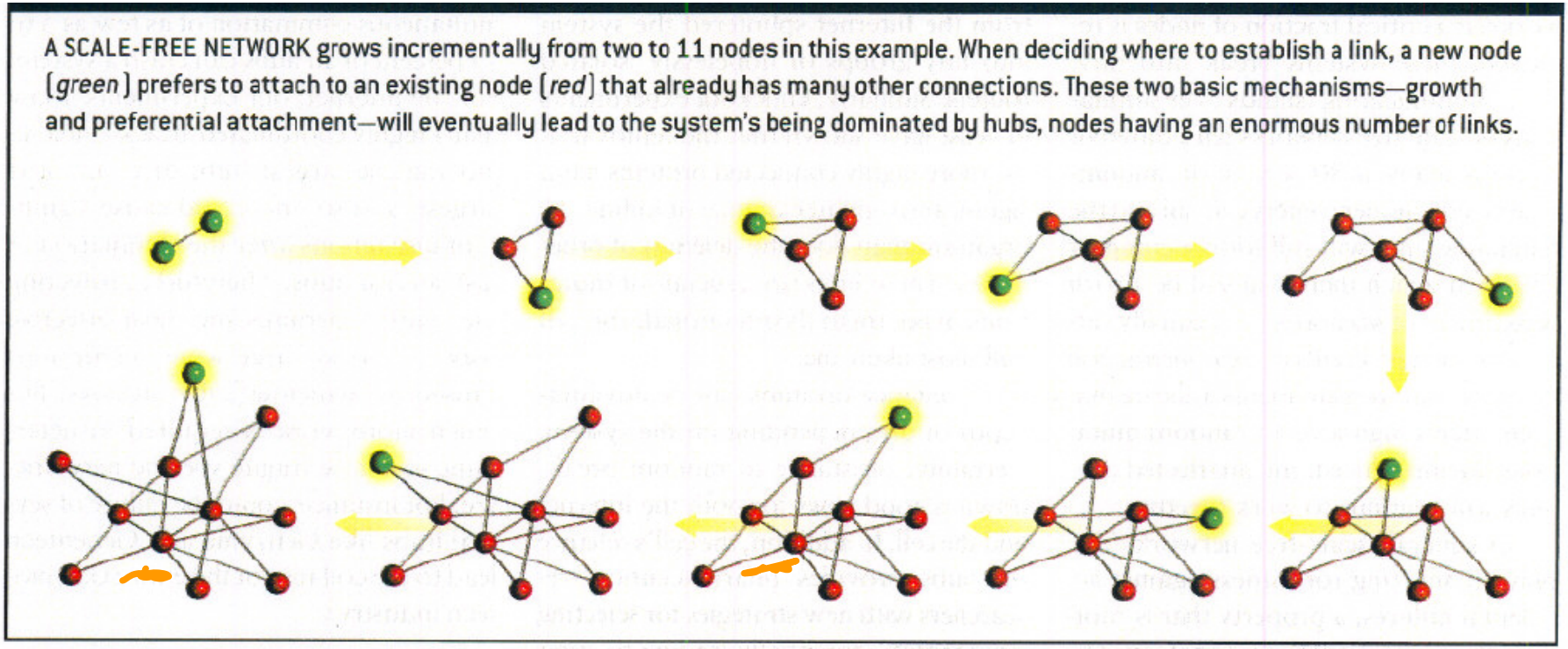
• PREFERENTIAL ATTACHMENT: the probability Π that a new vertex will be connected to vertex i depends on the connectivity of that vertex:

$$\prod(k_i) = \frac{k_i}{\sum_j k_j}$$



(c) M Gerstein, gerstein.info/talks

Birth of Scale-Free Network

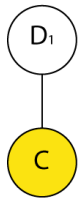


[From Barabasi & Bonabeau, *Sci. Am.,* May '03]

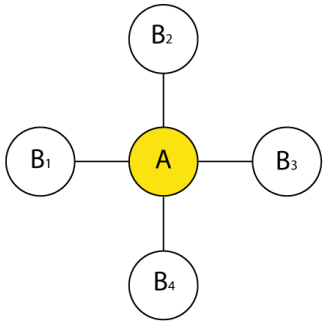
(c) M Gerstein, gerstein.info/talks

SCALE FREENESS GENERALLY EVOLVES THROUGH PREFERENTIAL ATTACHMENT (THE RICH GET RICHER)

The Duplication Mutation Model



Gene duplication



The interaction partners of A are more likely to be duplicated

Description

- Theoretical work shows that a mechanism of preferential attachment leads to a scale-free topology

  ("The rich get richer")

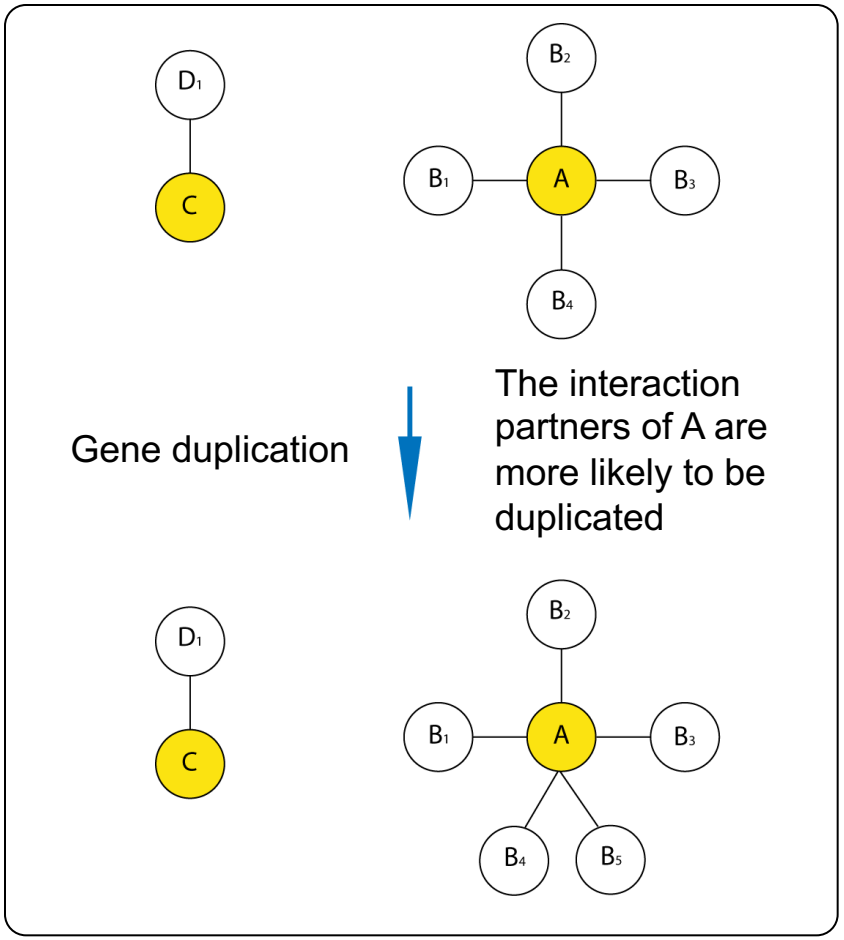
- In interaction network, gene duplication followed by mutation of the duplicated gene is generally thought to lead to preferential attachment

- Simple reasoning: The partners of a hub are more likely to be duplicated than the partners of a non-hub

Source: Albert et al. *Rev. Mod. Phys.* (2002) and Middendorf et al. *PNAS* (2005)

SCALE FREENESS GENERALLY EVOLVES THROUGH PREFERENTIAL ATTACHMENT (THE RICH GET RICHER)

The Duplication Mutation Model



Description

- Theoretical work shows that a mechanism of preferential attachment leads to a scale-free topology

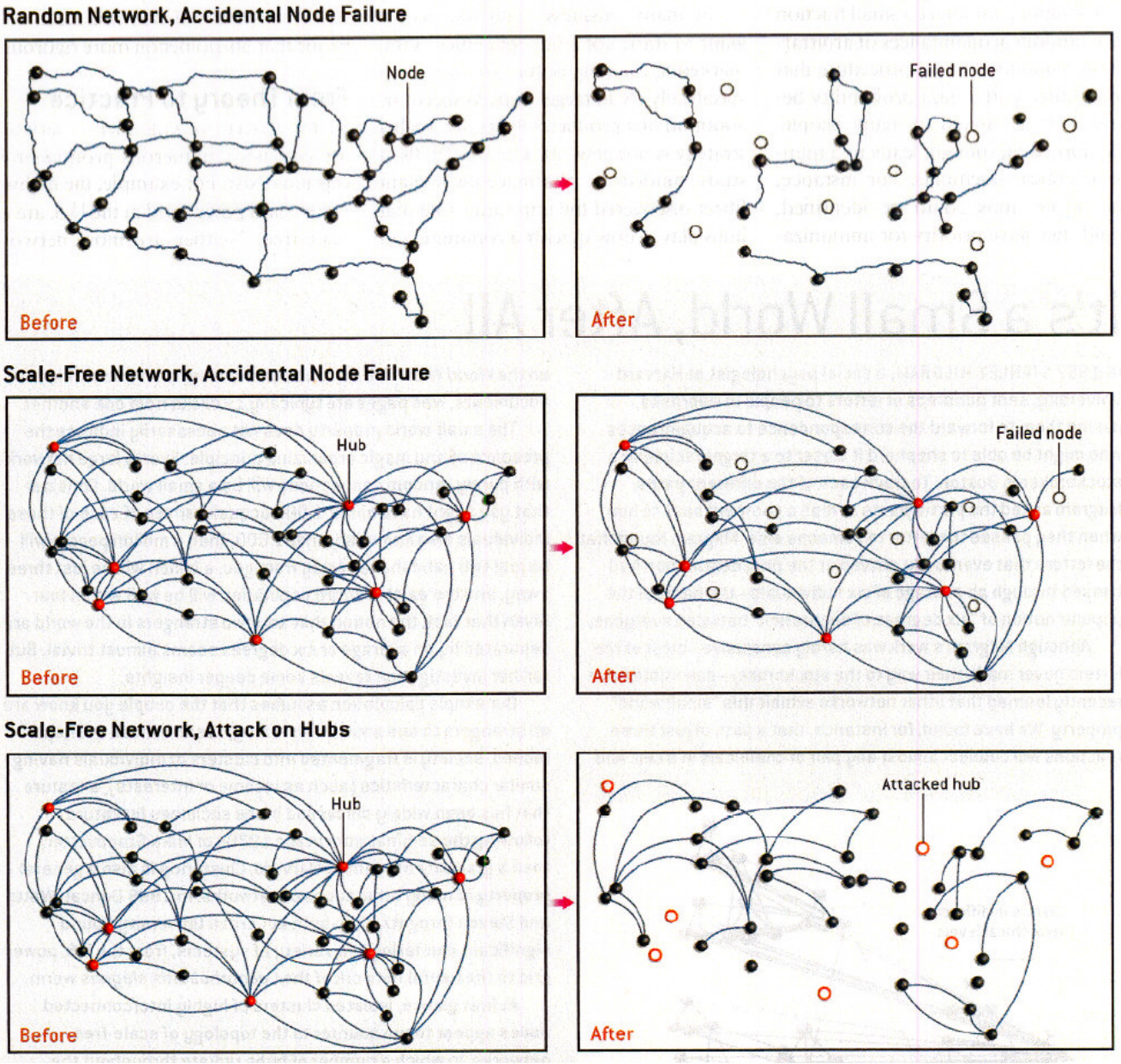
  ("The rich get richer")

- In interaction network, gene duplication followed by mutation of the duplicated gene is generally thought to lead to preferential attachment

- Simple reasoning: The partners of a hub are more likely to be duplicated than the partners of a non-hub

Source: Albert et al. *Rev. Mod. Phys.* (2002) and Middendorf et al. *PNAS* (2005)

Knocking Out Nodes in Scale-free and Random Networks



[From Barabasi & Bonabeau, *Sci. Am.,* May '03]

(c) M Gerstein, gerstein.info/talks

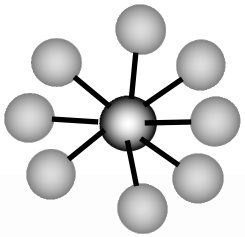
Network Topology

Relating Hubs to Biological Variation & Essentiality

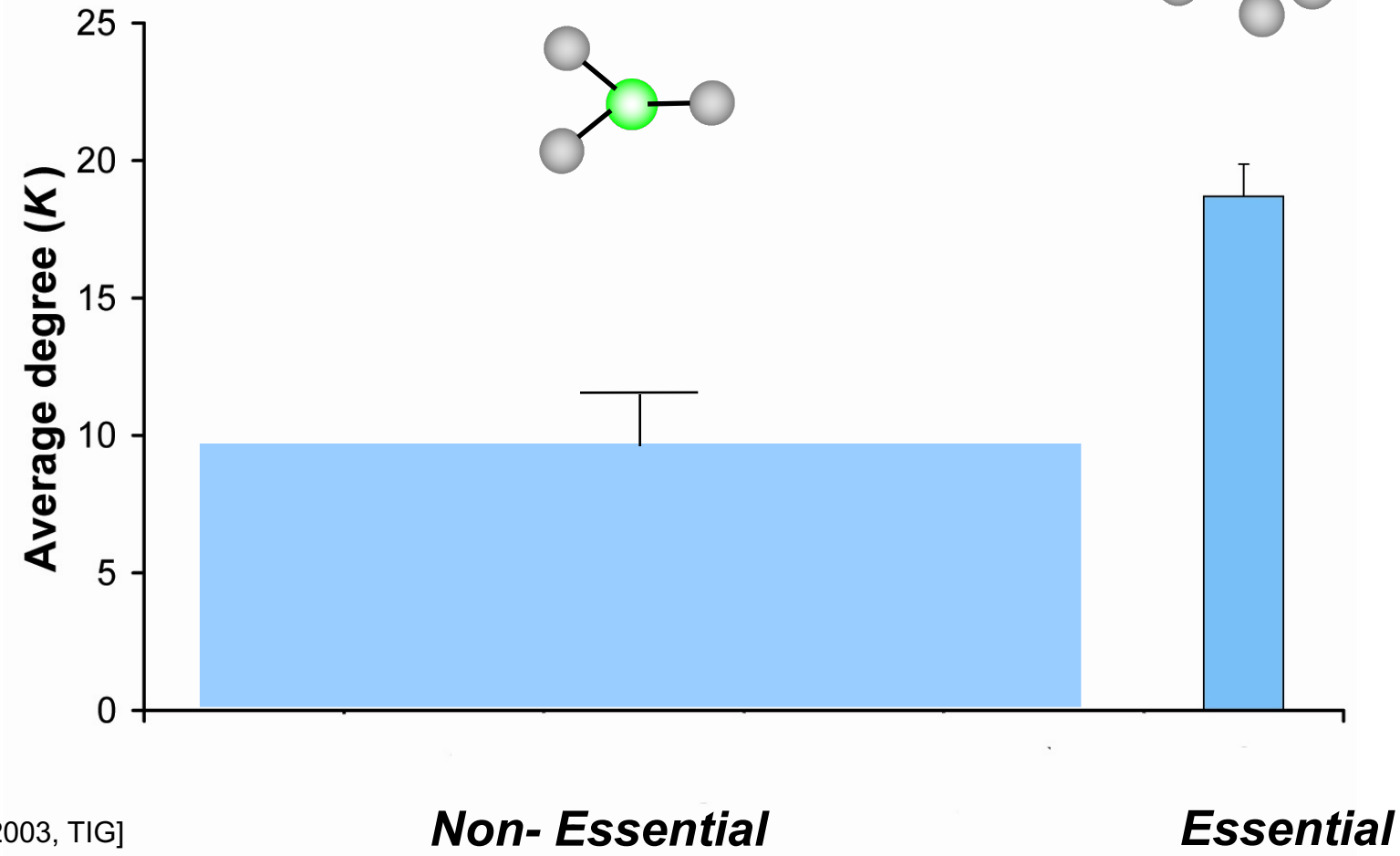
Hubs tend to be Essential

Integrate gene essentiality data with protein interaction network. Perhaps hubs represent vulnerable points?

[Lauffenburger, Barabasi]



"hubbiness"



[Yu et al., 2003, TIG]

(c) M Gerstein, gerstein.info/talks

- **Notably, the most highly conserved proteins were highly connected in *E.coli* protein interaction network. (*E.coli*)**

  - Butland et al. 2004, Nature

- **Connectivity of well-conserved proteins in network is negatively correlated with their rate of evolution. (*Saccharomyces cerevisiae*)**

  - Fraser et al. 2002, Science
  - Fraser et al. BMC Evol. Biol. (2003)

- **Proteins that have a more central position evolve more slowly and are more likely to be essential for survival. (*Saccharomyces cerevisiae; C. elegans; Drosophila*)**

  - Hahn et al. 2005, Mol Biol Evol

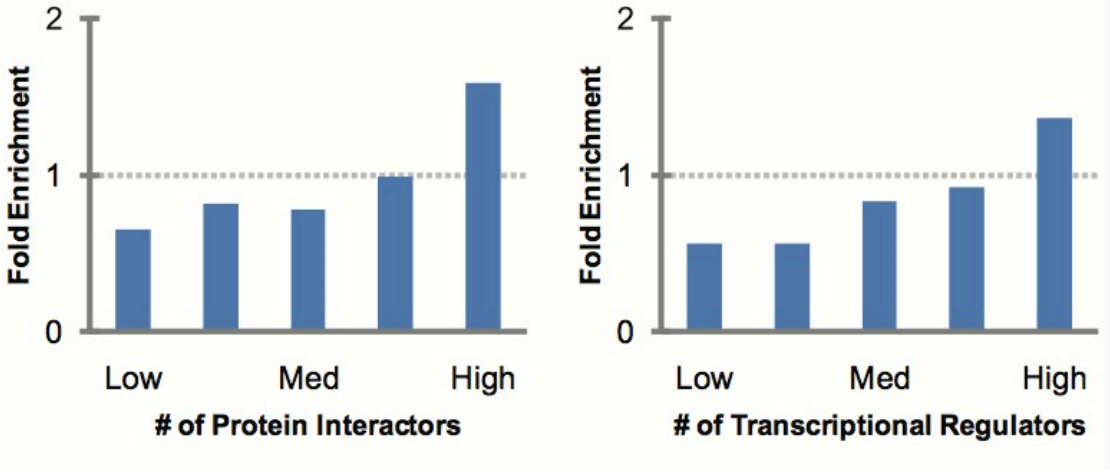
- **More miRNA regulation of a target gene associated with lower dN/dS (r = -.21) *(M. musculus)***

  - Cheng et al. 2009, BMC Genomics

- **Slowly evolving proteins tend to have more interaction partners. (*Saccharomyces cerevisiae*)**

  - Xia et al. 2009, Plos Comput Biol

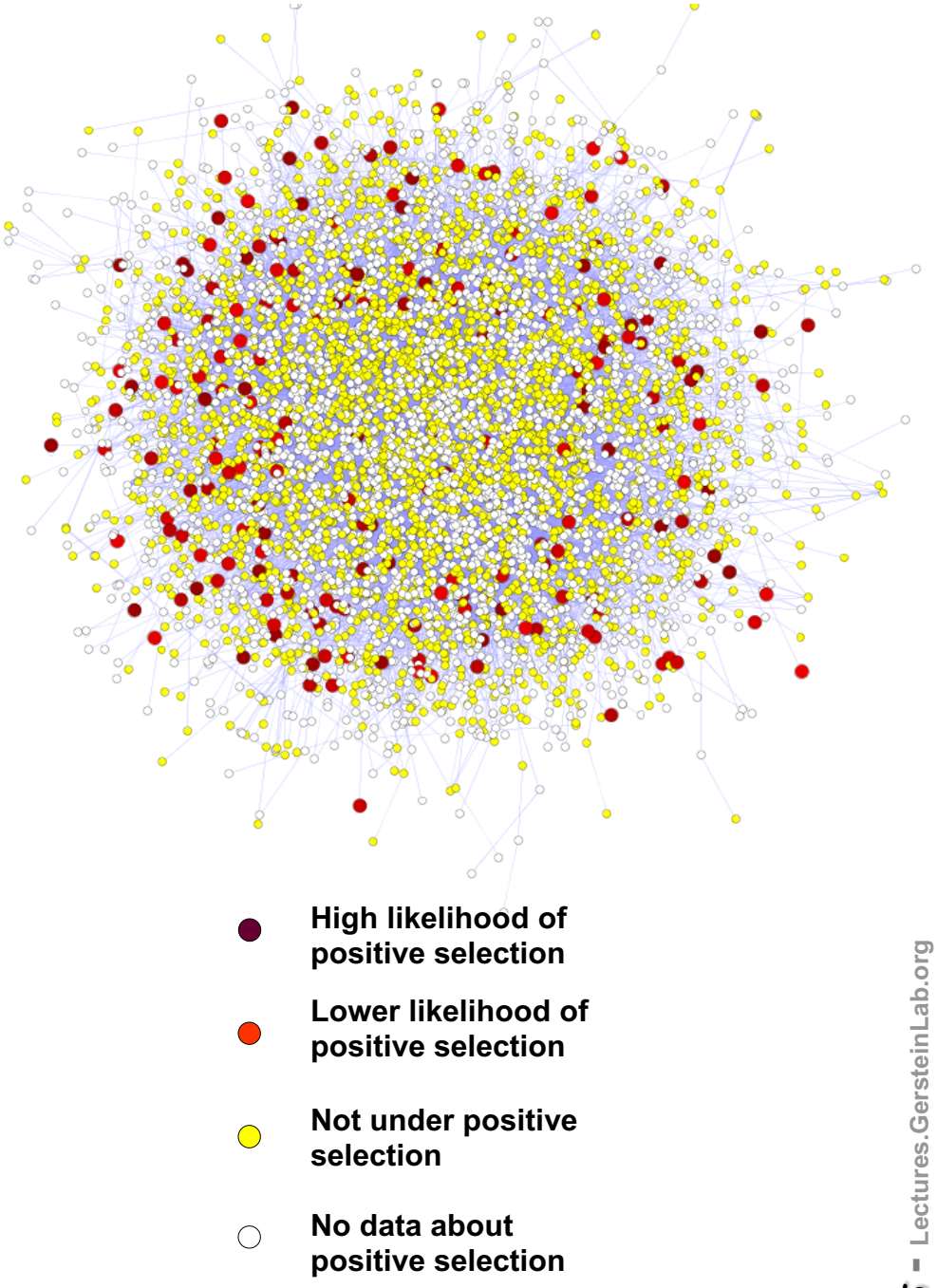
More Connectivity, More Constraint : A theme borne out in many studies



Fold Enrichment of slowly evolved proteins - *Xia et al.*

Rapid Evolution in the interaction network takes place at the periphery

- **Pos. sel. v. deg. centrality:**
  ρ = -.06, P < 1.2e-6
  - Effect is independent of any bias due to gene expression differences
  - Update w. 1000G Phase I
    SNP dens. v. centrality:
    ρ = -.1, p< 2.2e-16



Source: Nielsen et al. *PLoS Biol.* (2005), HPRD, and Kim et al. PNAS (2007)

Network Topology

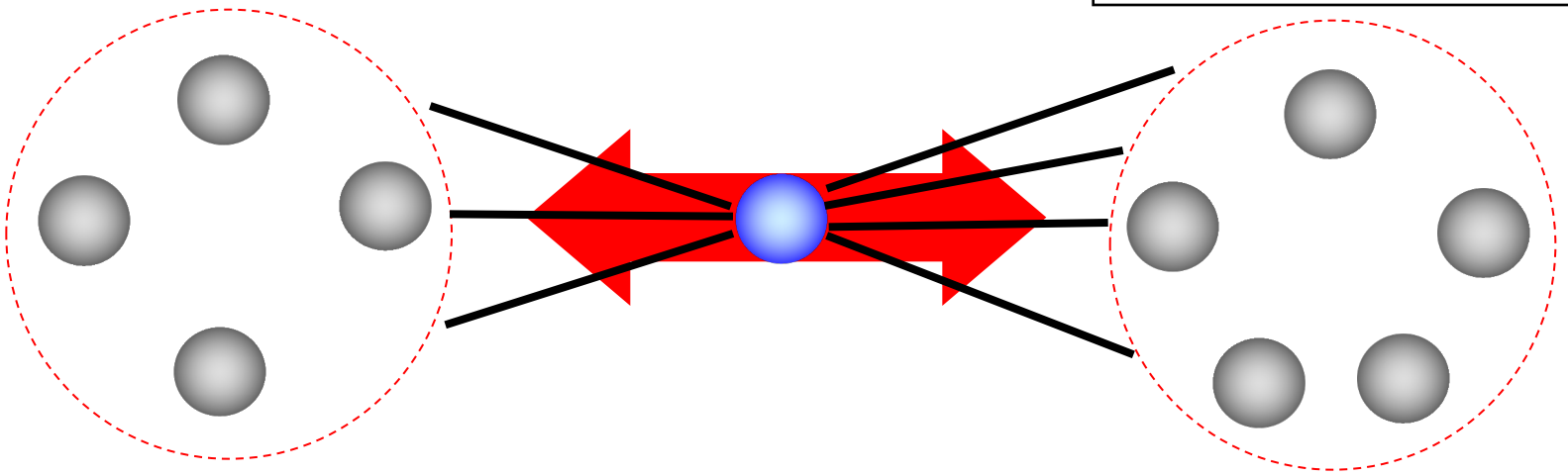
Other Measures of Centrality besides Hubs: Bottlenecks & their Calculation

Another measure of Centrality: Betweenness centrality

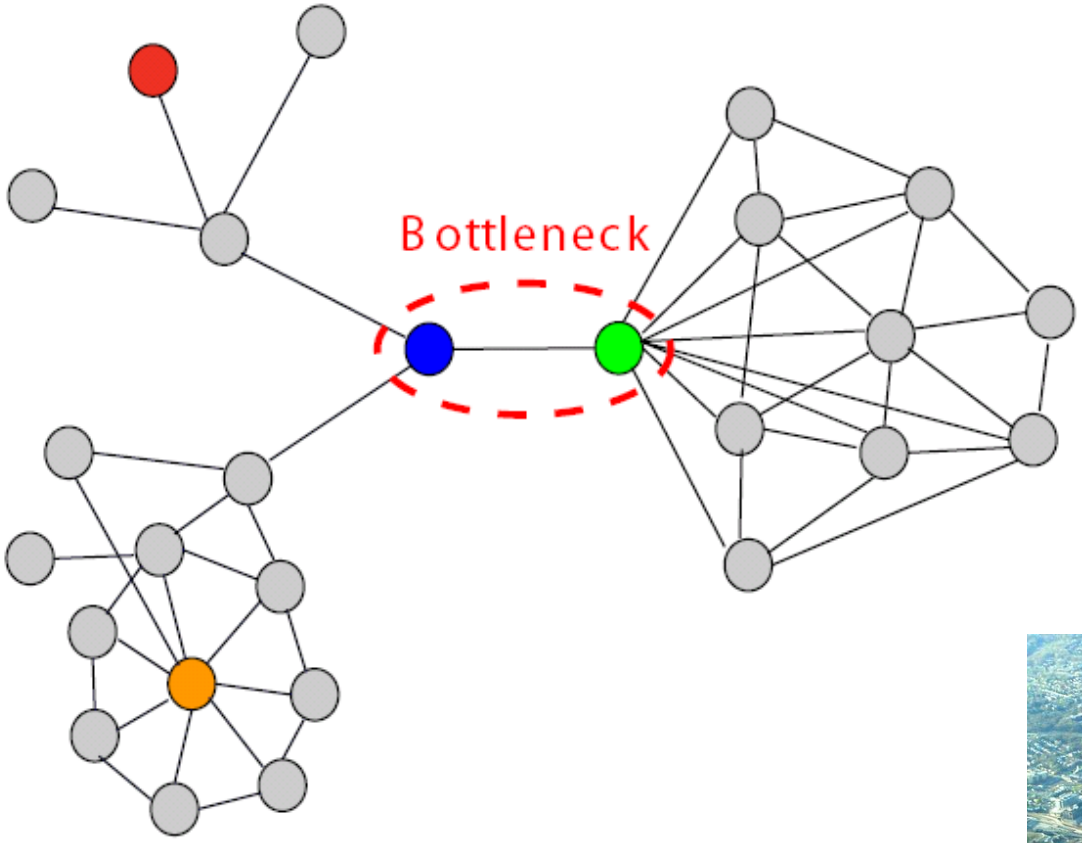
Betweenness of a node is the number of shortest paths of pairs of vertices that run through it -- a measure of information flow.

Freeman LC (1977) Set of measures of centrality based on betweenness. Sociometry 40: 35–41.

Girvan & Newman (2002) PNAS 99: 7821.



Bottlenecks & Hubs



Hub-bottleneck **node**

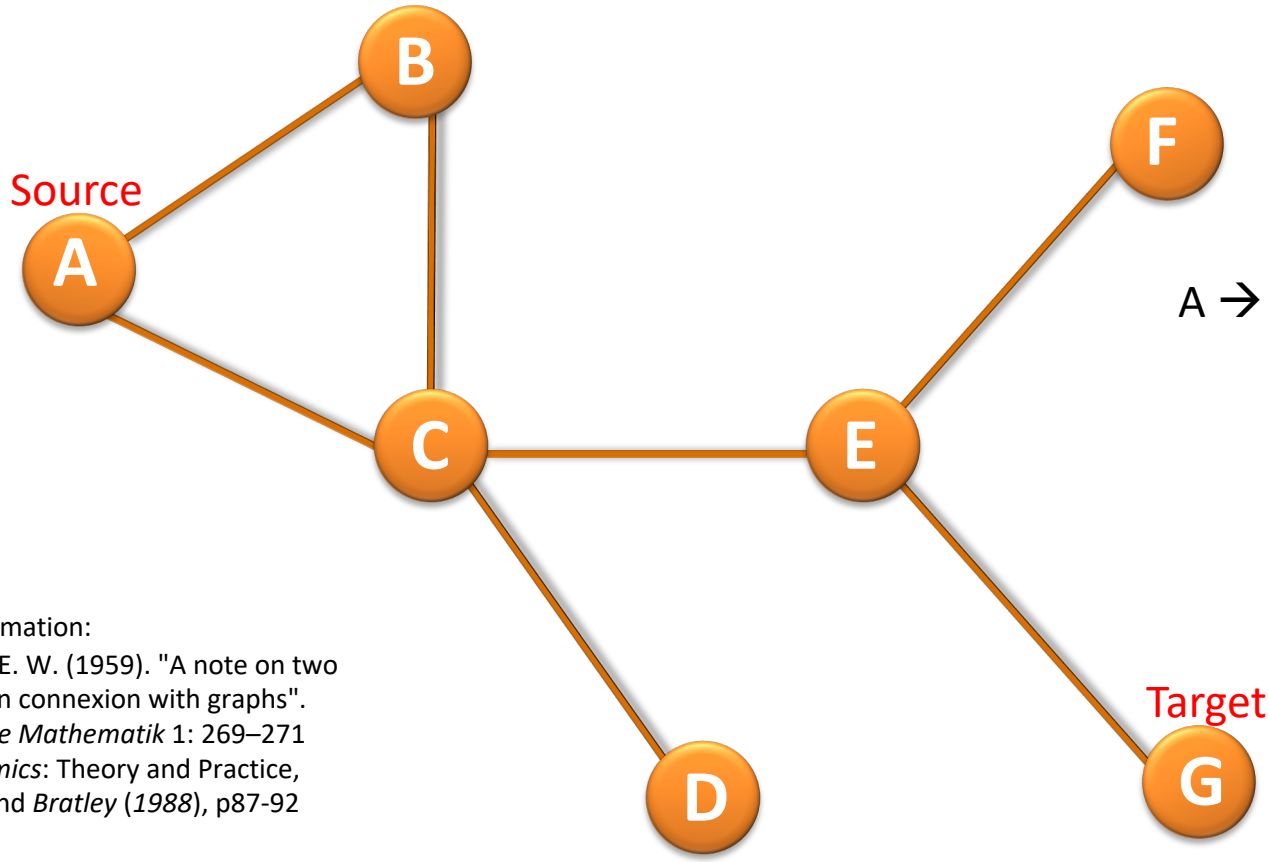
Non-hub-bottleneck **node**

Hub-non-bottleneck **node**

Non-hub-non-bottleneck **node**

[Yu et al., PLOS CB (2007)]

Shortest Paths: The Dijkstra Algorithm



More information:

1.Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". *Numerische Mathematik* 1: 269–271

2.Algorithmics: Theory and Practice, *Brassard* and *Bratley* (*1988*), p87-92

Betweenness of C

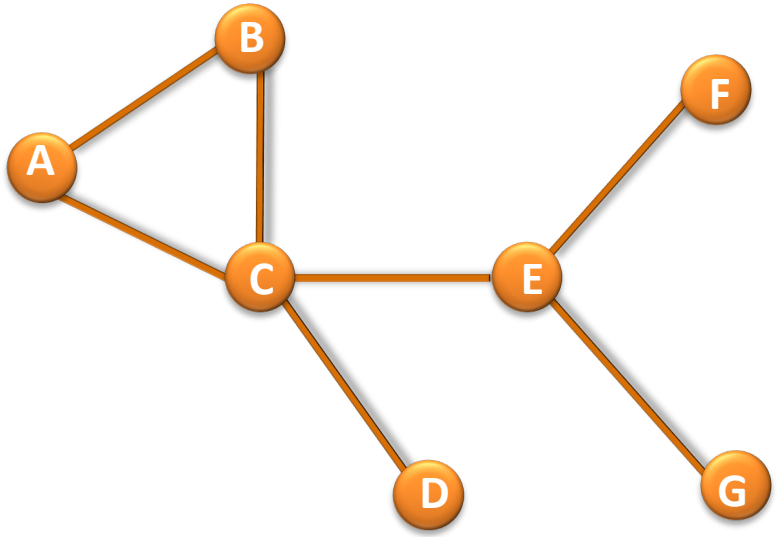
Pair of nodes

Source	Target
A	B
A	C
A	**D**
A	**E**
A	**F**
A	**G**
B	C
B	**D**
B	**E**
B	**F**
B	**G**
C	D
C	E
C	F
C	G
D	**E**
D	**F**
D	**G**
E	F
E	G
F	G

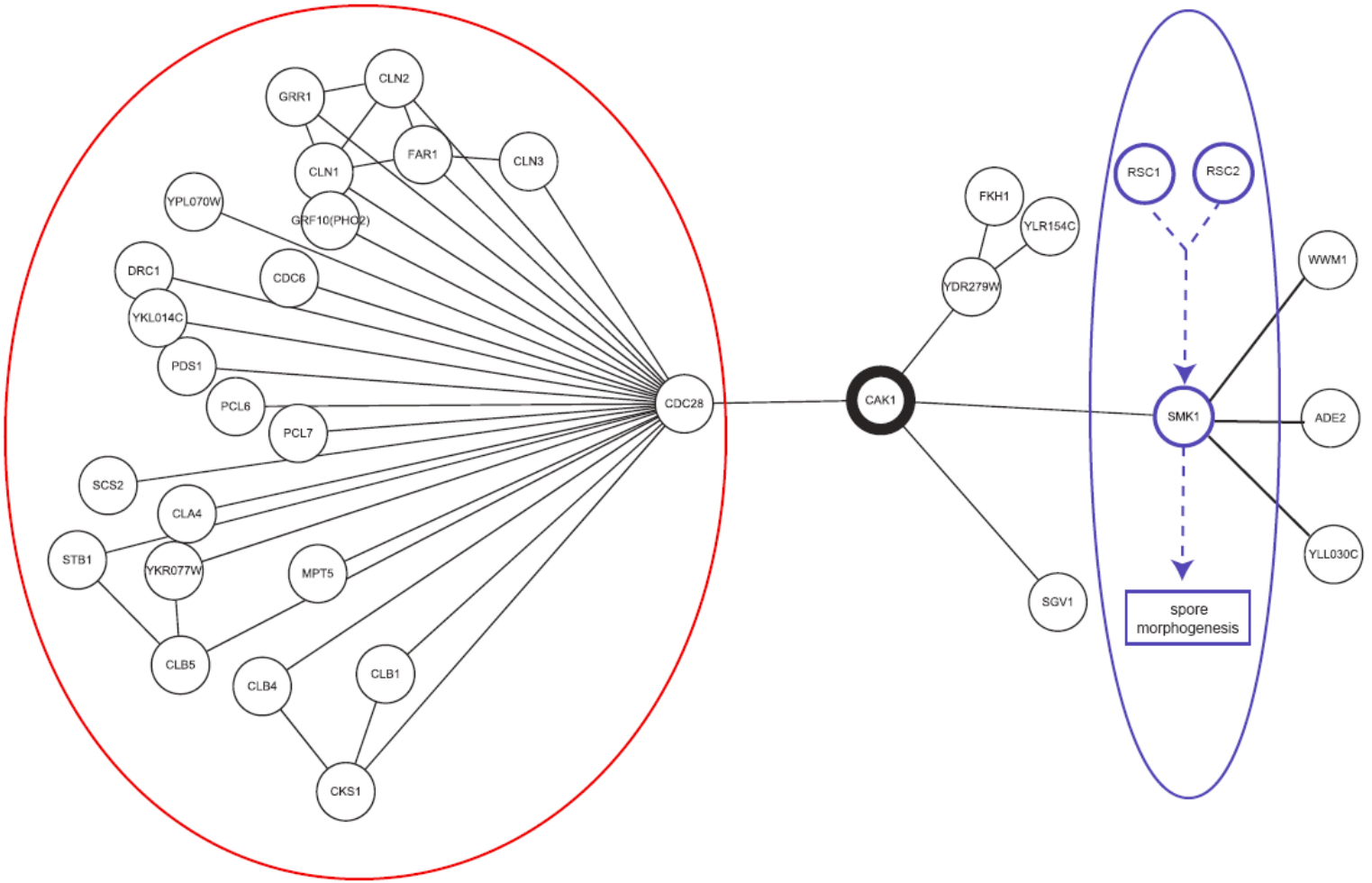
$$Betweenness(v)=\sum_{s\neq v\neq t}(\sigma_{st}\ (v)/\sigma_{st})$$

where σ_{st} is the total number of shortest paths from source s to target t and $\sigma_{st}(v)$ is the number of those paths that pass through *v*.

11 shortest paths going through C



Bottleneck bridging between processes

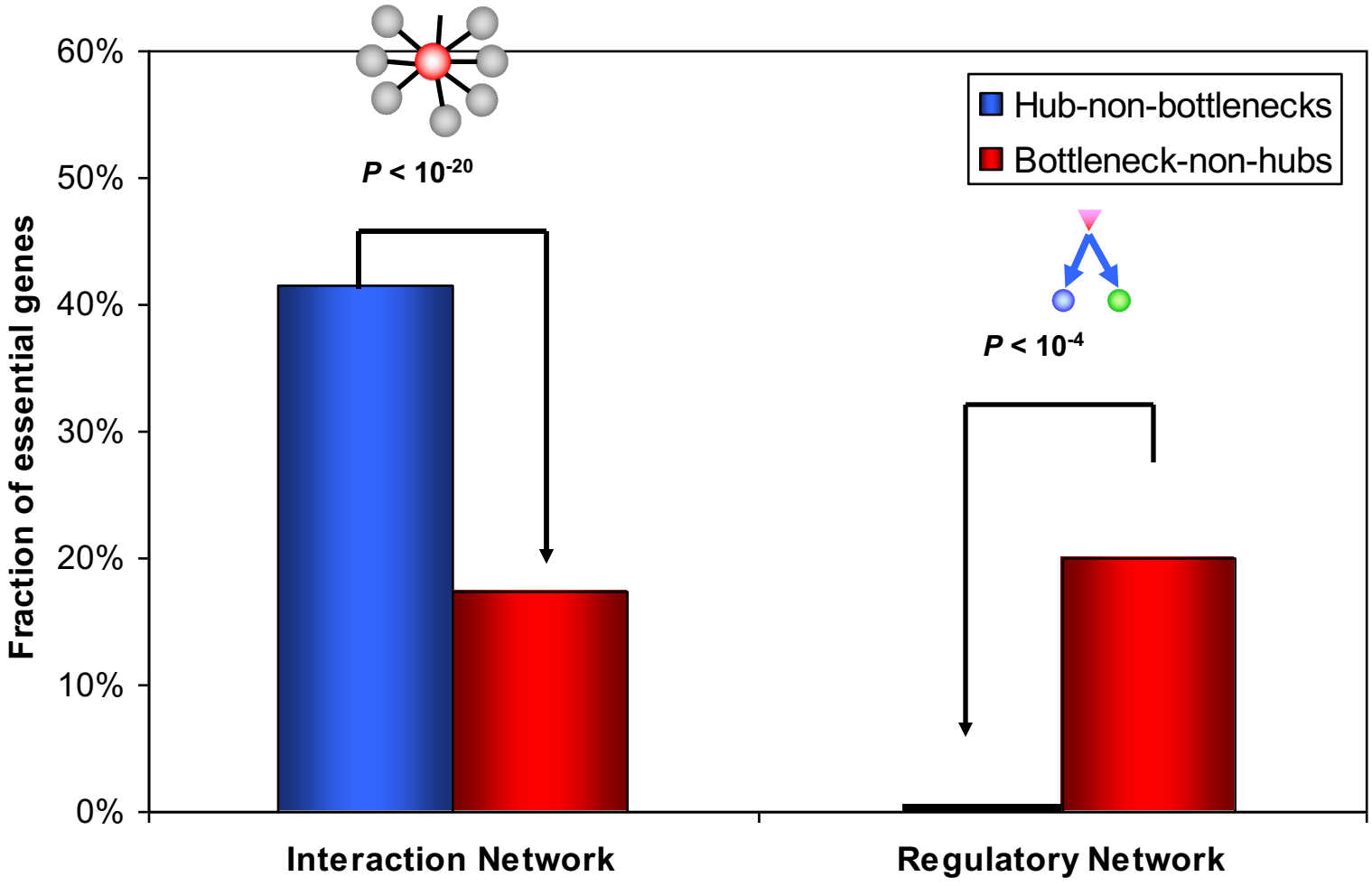


Regulation of mitotic cell cycle

MAP Kinase pathway regulating spore morphogenesis

[Yu et al., PLOS CB (2007)]

Bottlenecks are what matters in regulatory networks

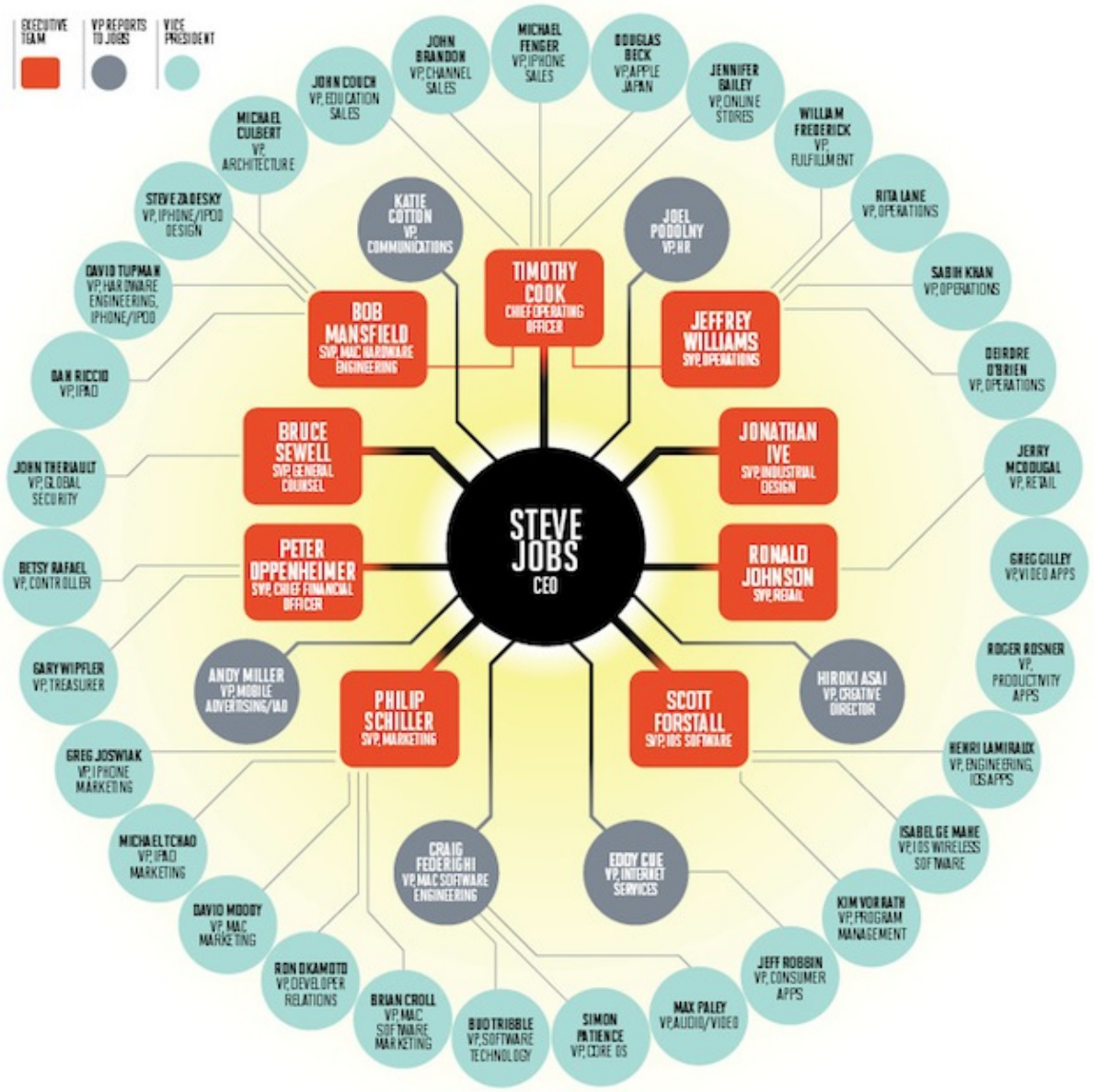


[Yu *et al.*, *PLoS Comput Biol* (2007)]

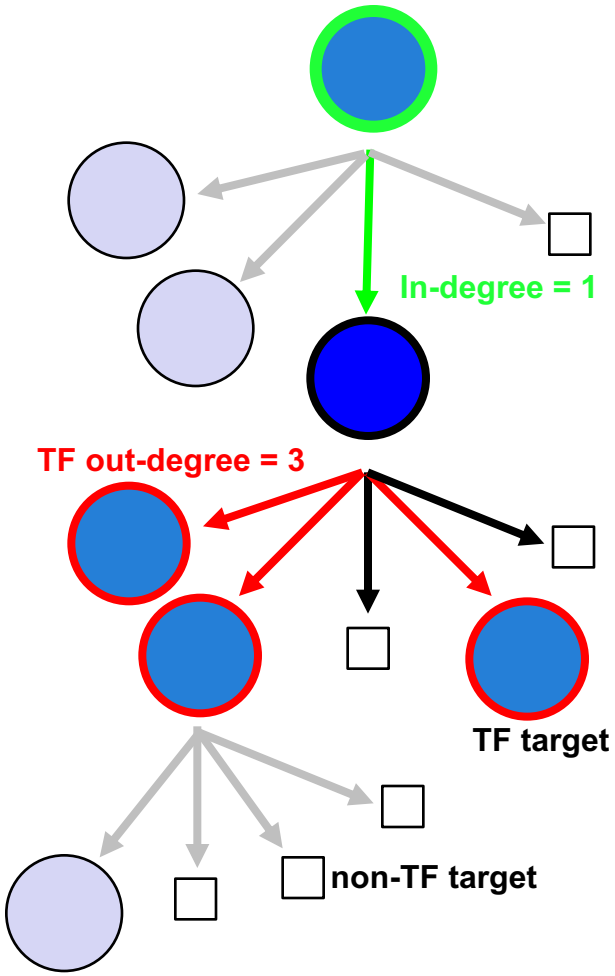
Network Topology

Other Measures of Centrality besides Hubs: Hierarchy

Network Hierarchy



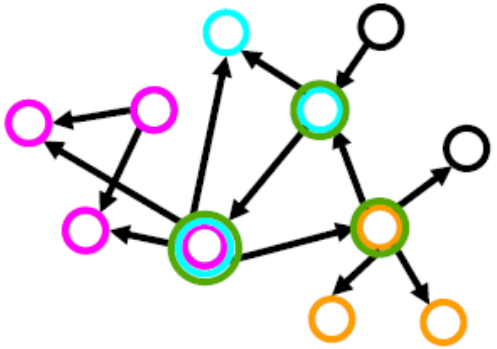
Network Stats to Identify Hierarchy



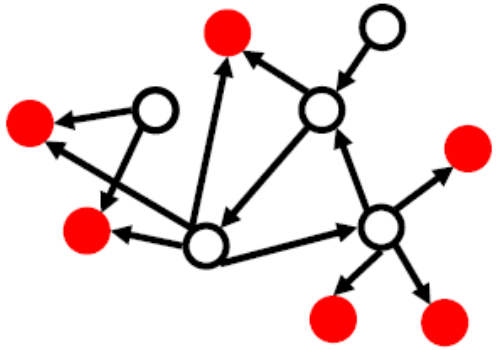
Hierarchy Height Statistic =
(normalized TF Out deg. – In deg.)

Determination of "Level" in Regulatory Network Hierarchy with Breadth-first Search

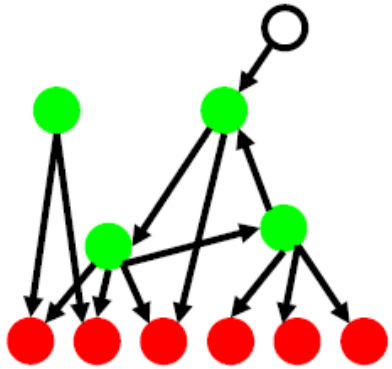
I. Example network with all 4 motifs



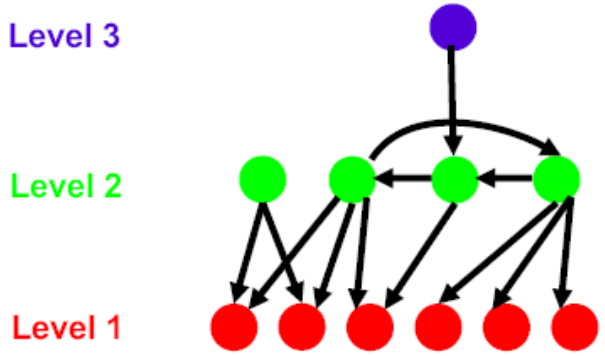
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

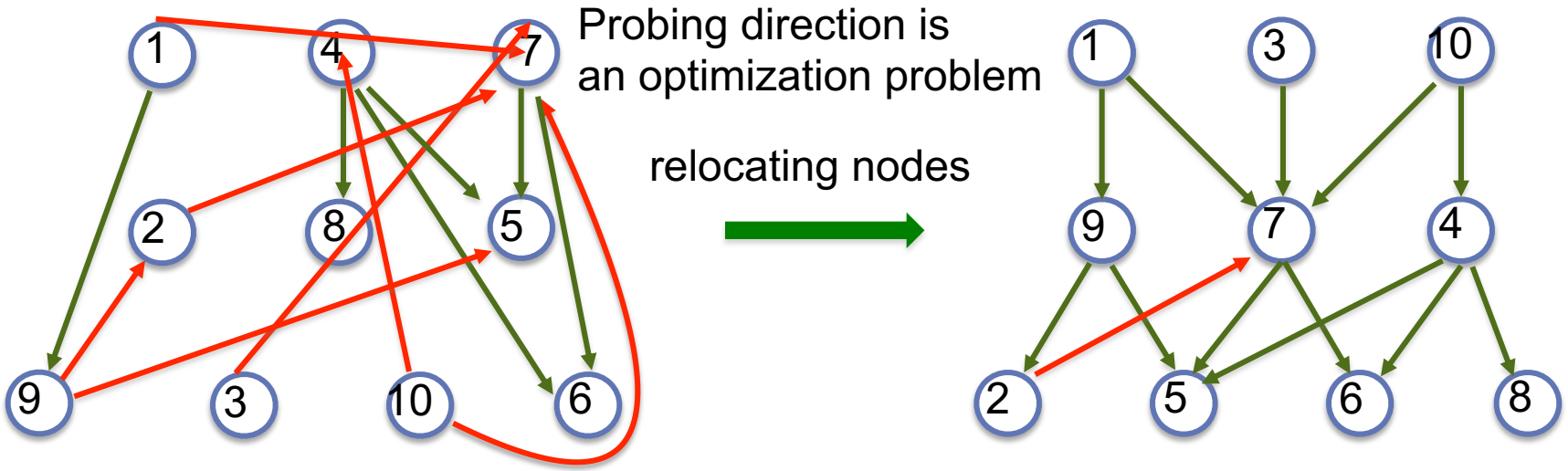


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Using Simulated Annealing to Globally Minimize the Number of Upward Pointing Edges



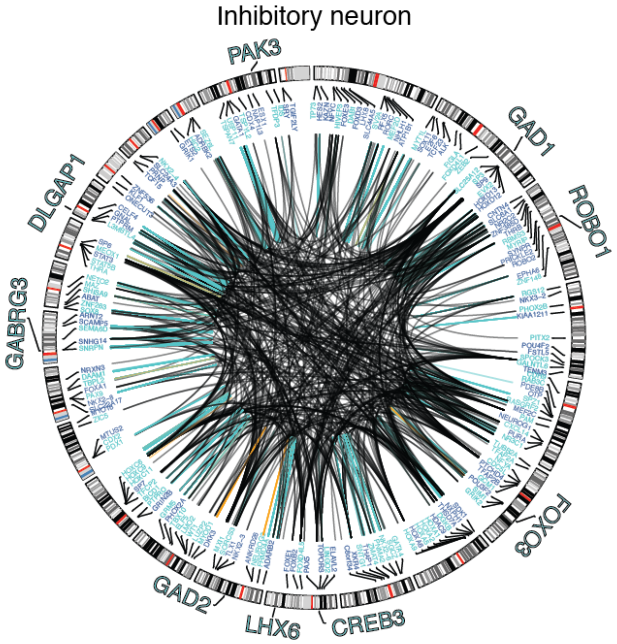
[Gerstein et al. Nature ('12)]

Network Topology

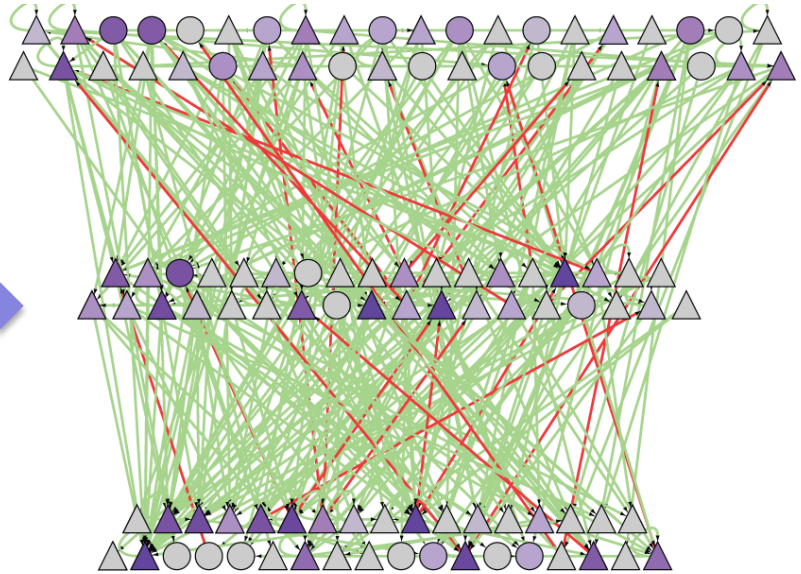
Hierarchy application

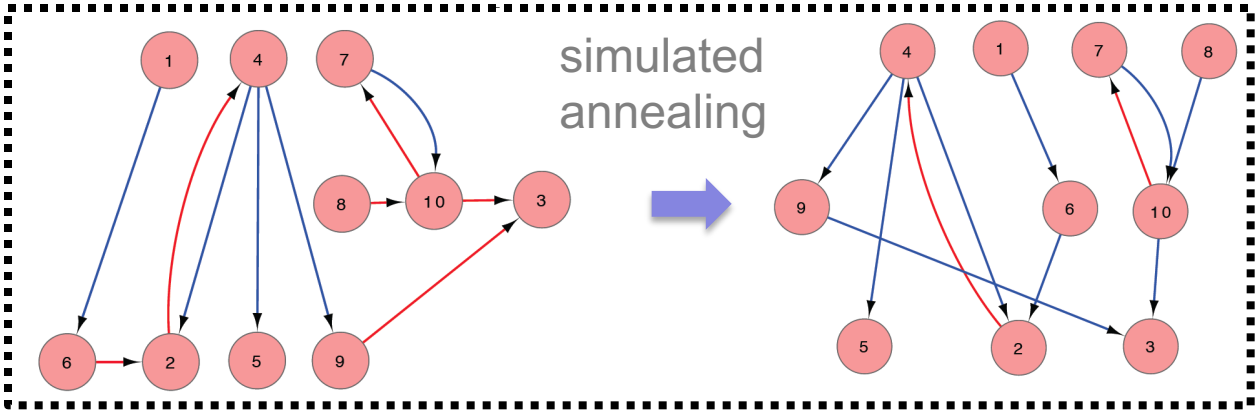
Transforming a Regulatory Network into a Hierachy

Regulatory Network



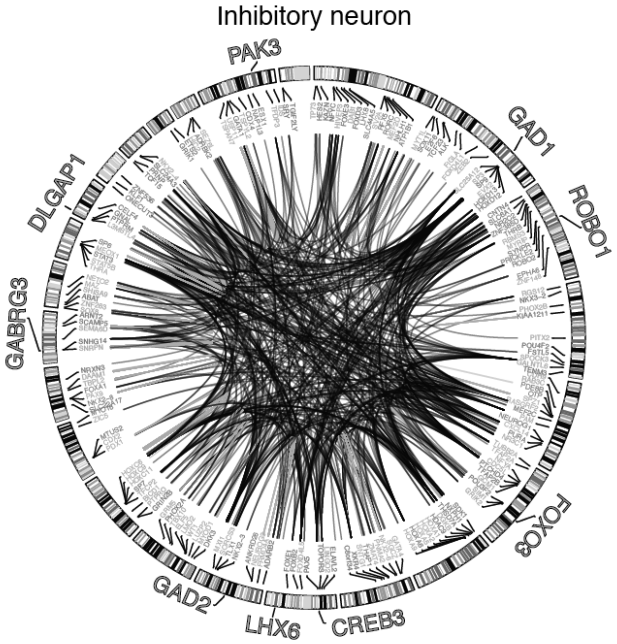
Network Hierarchy

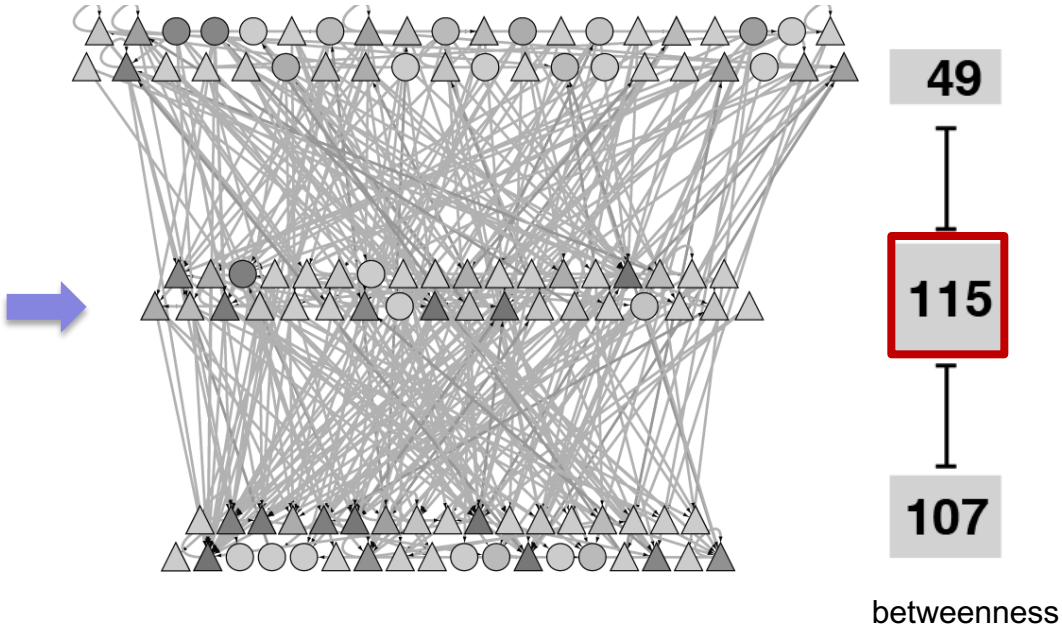


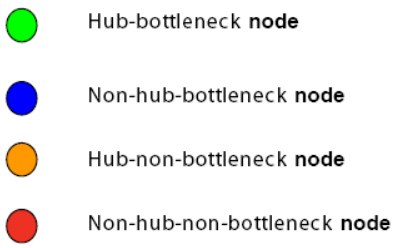


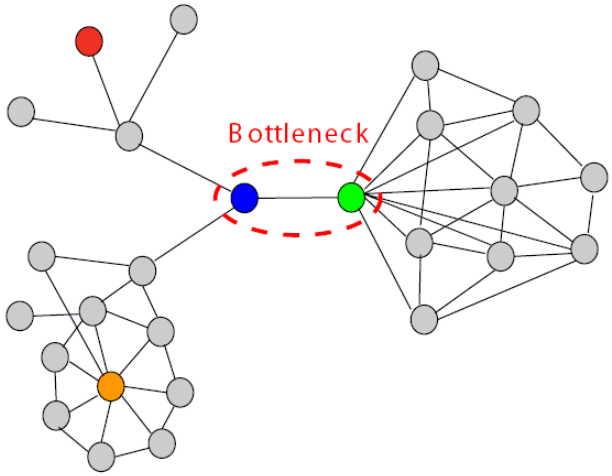
[Wang et al. ('18) Science, Gerstein et al. Nature ('12)]

Mid-level of the hierarchy has many high-connectivity bottlenecks



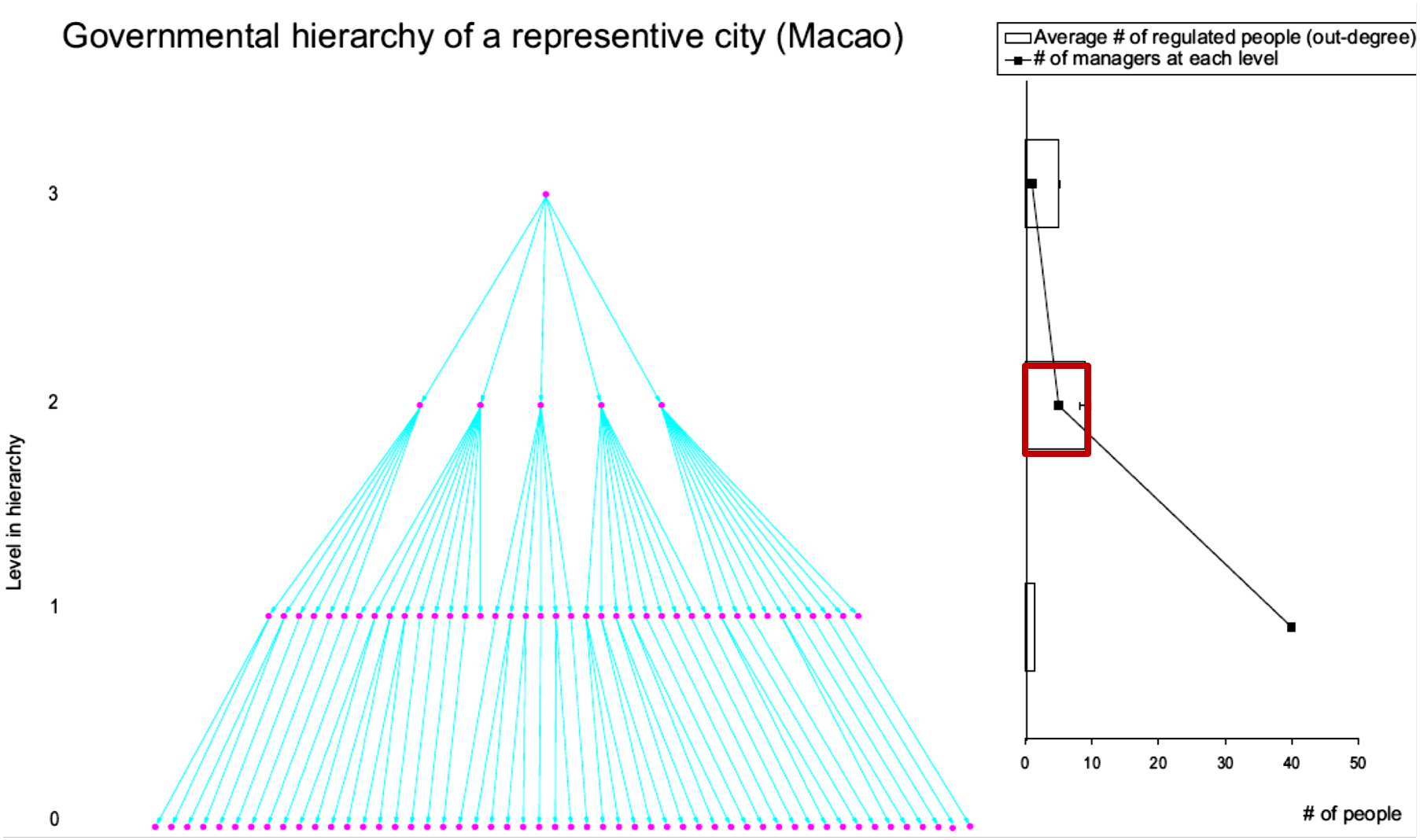






[Wang et al. ('18) Science, Gerstein et al. Nature ('12)]

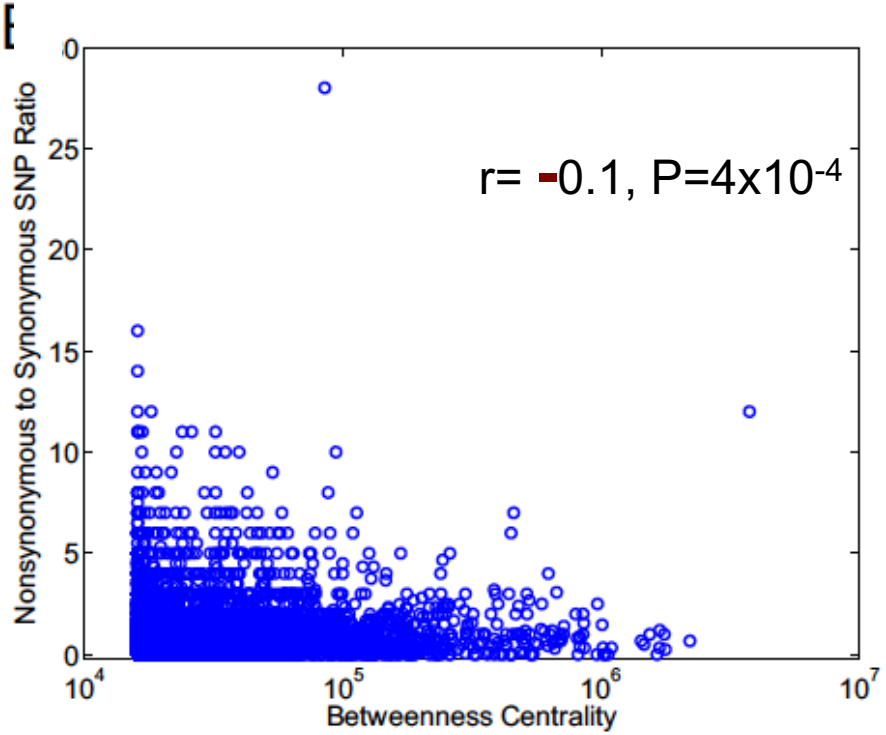
Governmental hierachies exhibit even more pronounced mid-level bottlenecks



[Yu et al., PNAS ('06)]

[Yu et al., *PNAS* '06]

Connectivity v. Constraint: in biological networks, less mutations tolerated at hubs & bottlenecks



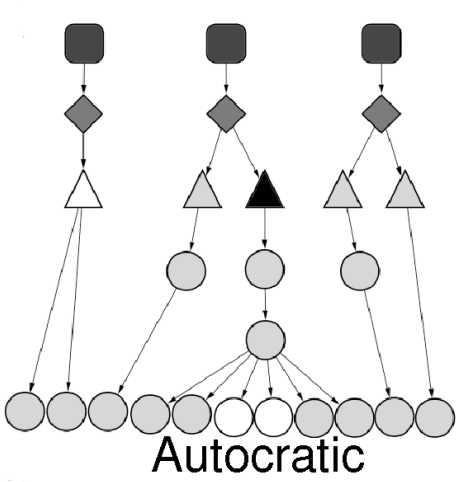
Kim et a. PNAS 2007

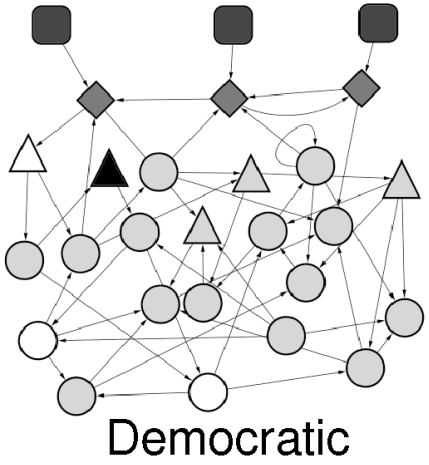
==Rate of Mutation==>

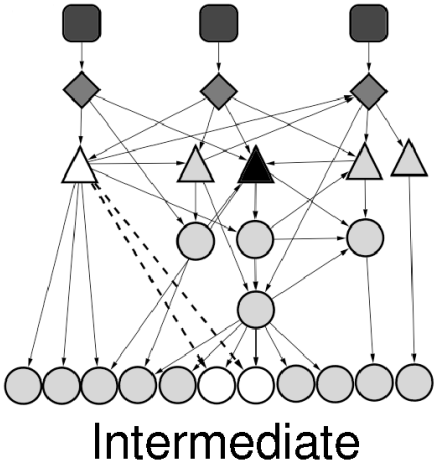
==Centrality=>

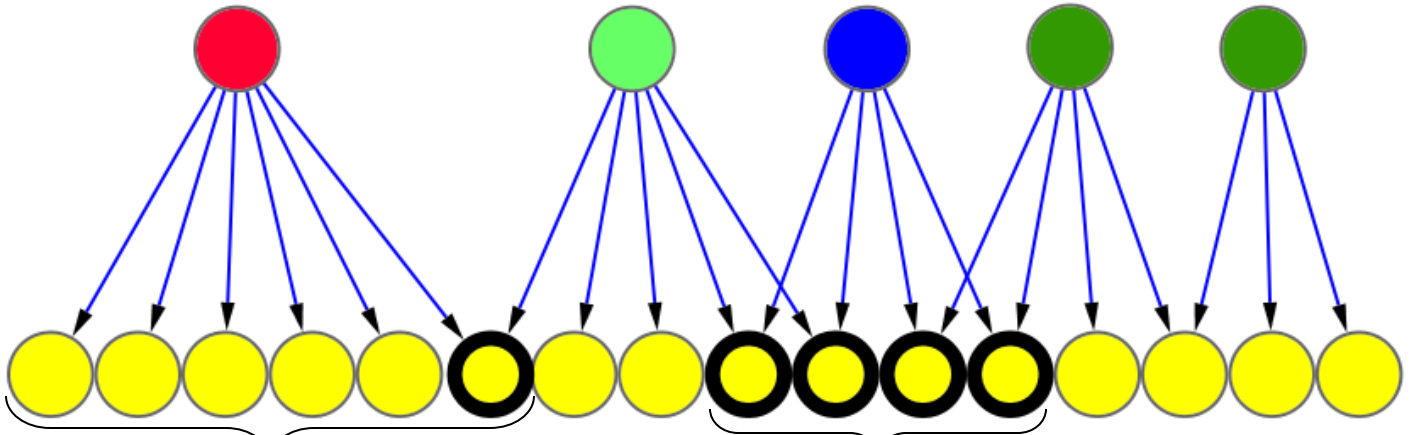
Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential. This phenomenon is observed in **many organisms & different kinds of networks** – e.g. Protein-Protein Interaction Networks (human, yeast, E coli, worm, fly), miRNA networks, regulatory networks, &c

Biological Networks tend to have more "democratic" hierarchies, easing bottlenecks, than many social ones









Degree of Collaboration

1/6=0.16

"Autocratic"

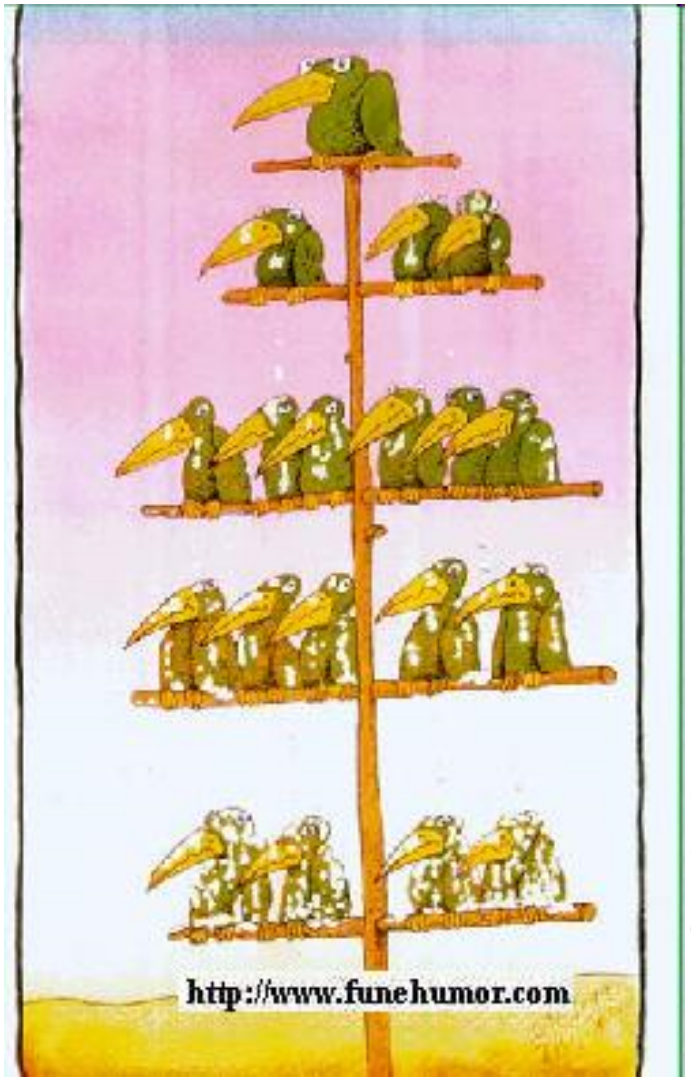
4/4=1

"Democratic"

[Bhardwaj et al., PNAS (2010)]

Middle Managers Interact the Most in Efficient Corporate Settings

- Floyd, S. W. et al (1992) **Middle management involvement in strategy and its association with strategic type** *Strategic Management Journal* 13, 153-167.
- Woodward, J. (1982) Industrial Organization: Theory and Practice (Oxford University Press, Oxford).
- Floyd, S. W. et al (1993) **Dinosaurs or Dynamos? Recognizing Middle Management's Strategic Role** The Academy of Management Executive 8, 47-57.
- Floyd, S. W. et al (1997) **Middle management's strategic influence and organizational performance** *Journal of Management Studies* 34, 465-485.



[Bhardwaj et al., PNAS (2010), in press]