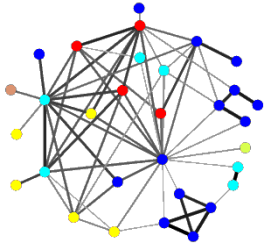
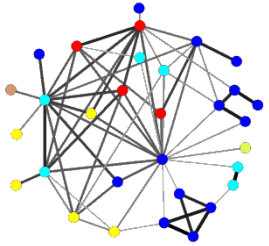


Graph Representation and Topological Analysis

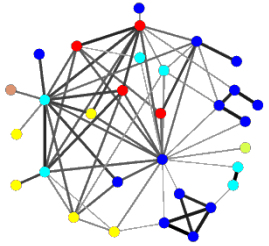
Yazdan Asgari

2020



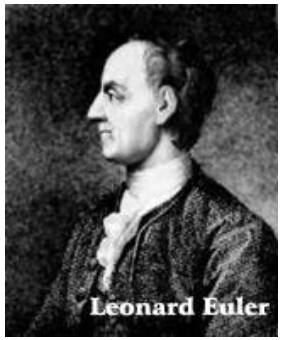
Previous Sessions

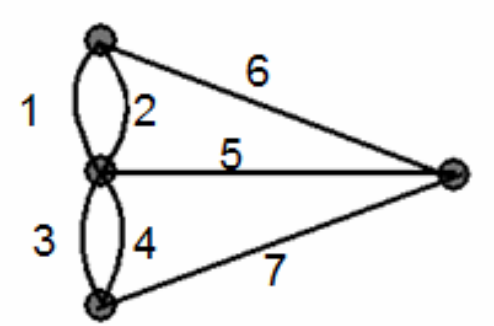
- ✓ Systems
- ✓ Working with Cytoscape
- ✓ Biological Networks

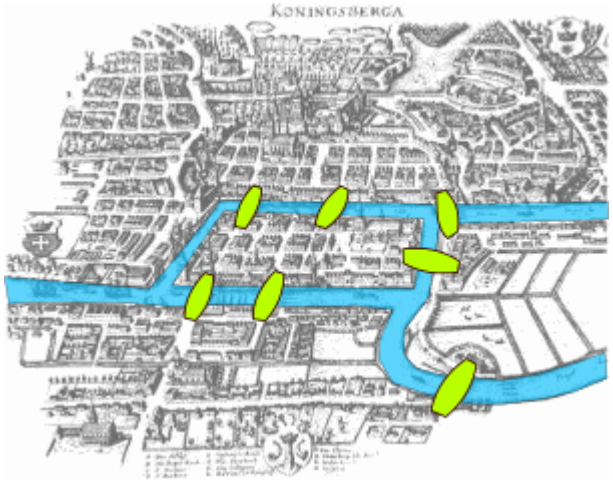


Introduction to Graph Theory

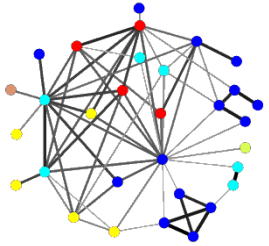
- (1736) Bridges of Konigsberg (today's Kaliningrad): walk all 7 bridges without crossing a bridge twice.







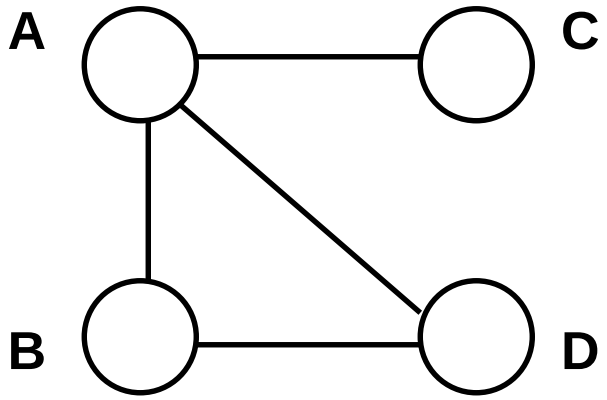
Biggs, N.; Lloyd, E. and Wilson, R. (1986), Graph Theory, 1736-1936, Oxford University Press

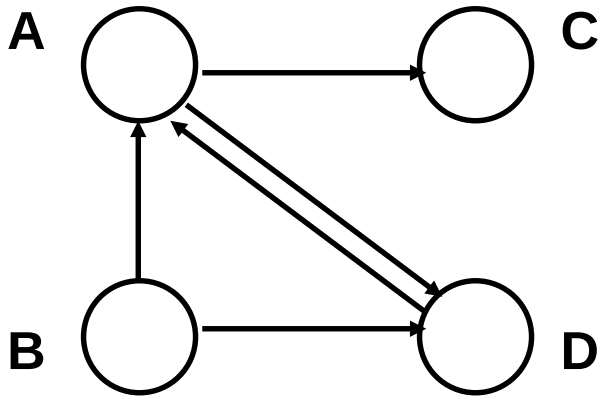


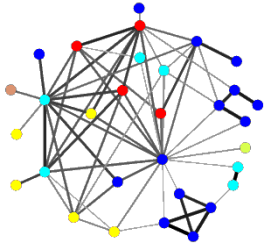
Introduction to Graph Theory

- **Graph theory** is the study of *graphs*, mathematical structures used to model pair wise relations between objects from a certain collection.

- A "graph" in this context refers to a collection of vertices or 'nodes' and a collection of *edges* that connect pairs of vertices.

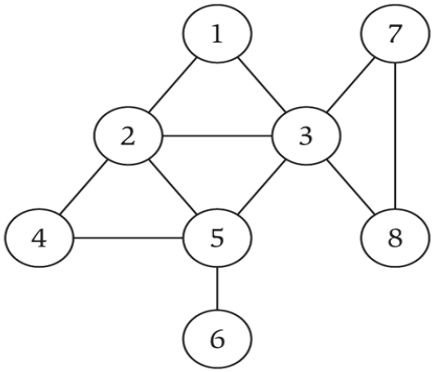






Introduction to Graph Theory

- **Graph** – mathematical object consisting of a set of:
  - *V* = **nodes** (vertices, points).
  - *E* = **edges** (links, arcs) between pairs of nodes.
  - **Graph size** parameters: *n = |V|, m = |E|*



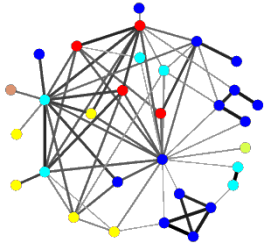
V = { 1, 2, 3, 4, 5, 6, 7, 8 }

E = { {1,2}, {1,3}, {2,3}, {2,4}, {2,5}, {3,5}, {3,7}, {3,8}, {4,5}, {5,6} }

n = 8

m = 11

West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall



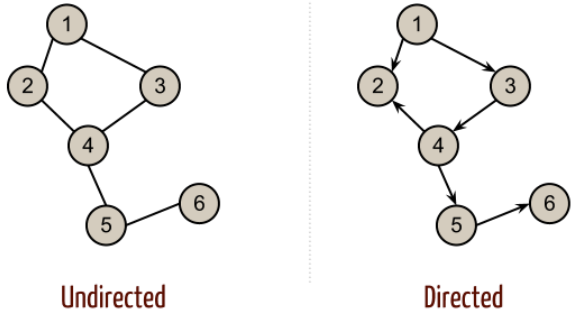
Introduction to Graph Theory

- A **subgraph** of a graph *H* is a graph whose vertex set is a subset of that of *G*, and whose adjacency relation is a subset of that of *G* restricted to this subset.

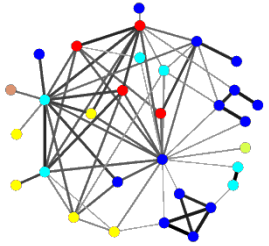
- **Direction:**

  digraph, or directed graph, or oriented graph

  undirected

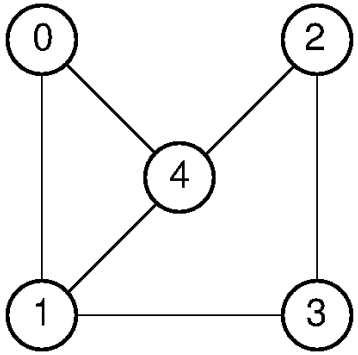


West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall



Introduction to Graph Theory

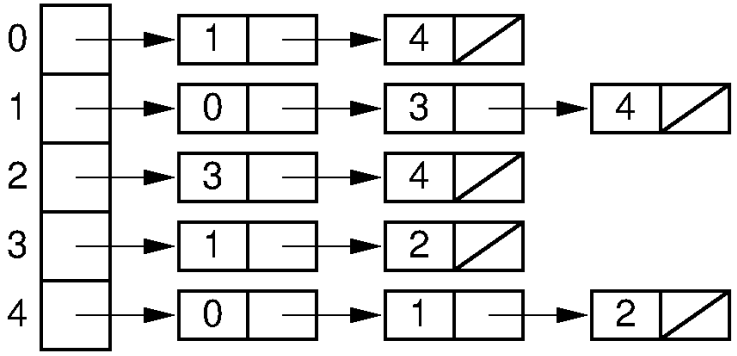
- Adjacency Matrix
- Adjacency list



(a)

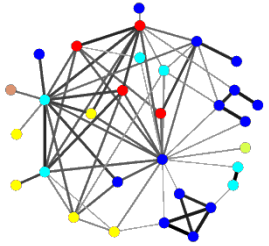
	0	1	2	3	4
0		1			1
1	1			1	1
2				1	1
3		1	1		
4	1	1	1		

(b)



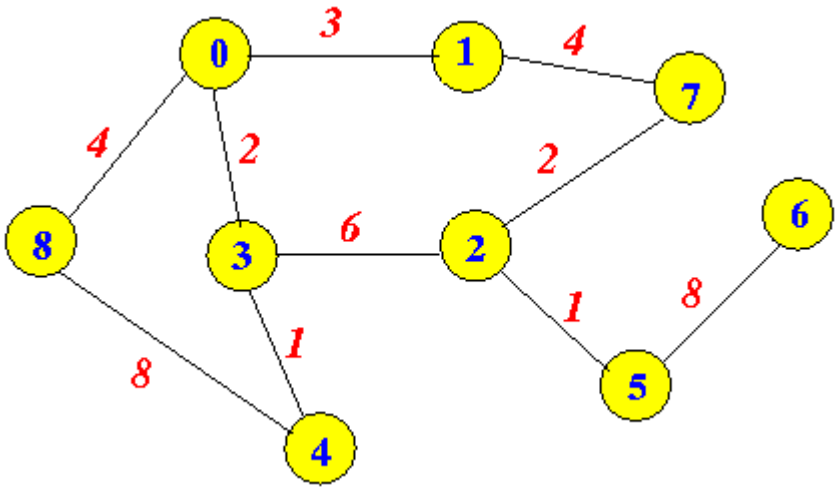
(c)

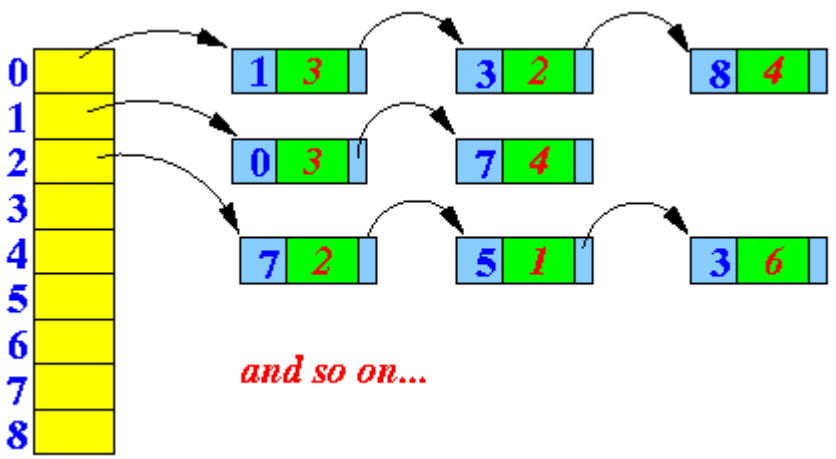
West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall



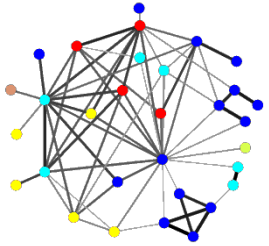
Introduction to Graph Theory

- Weighted Graph



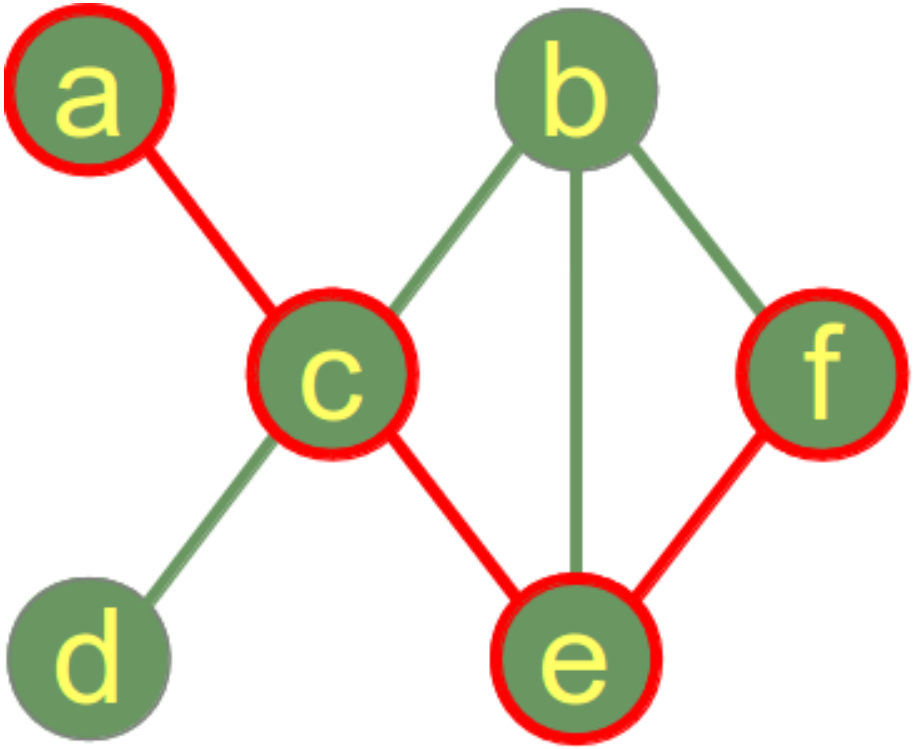


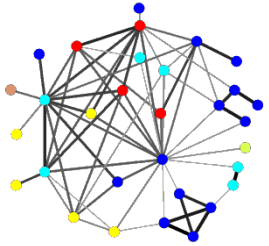
West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall



Introduction to Graph Theory

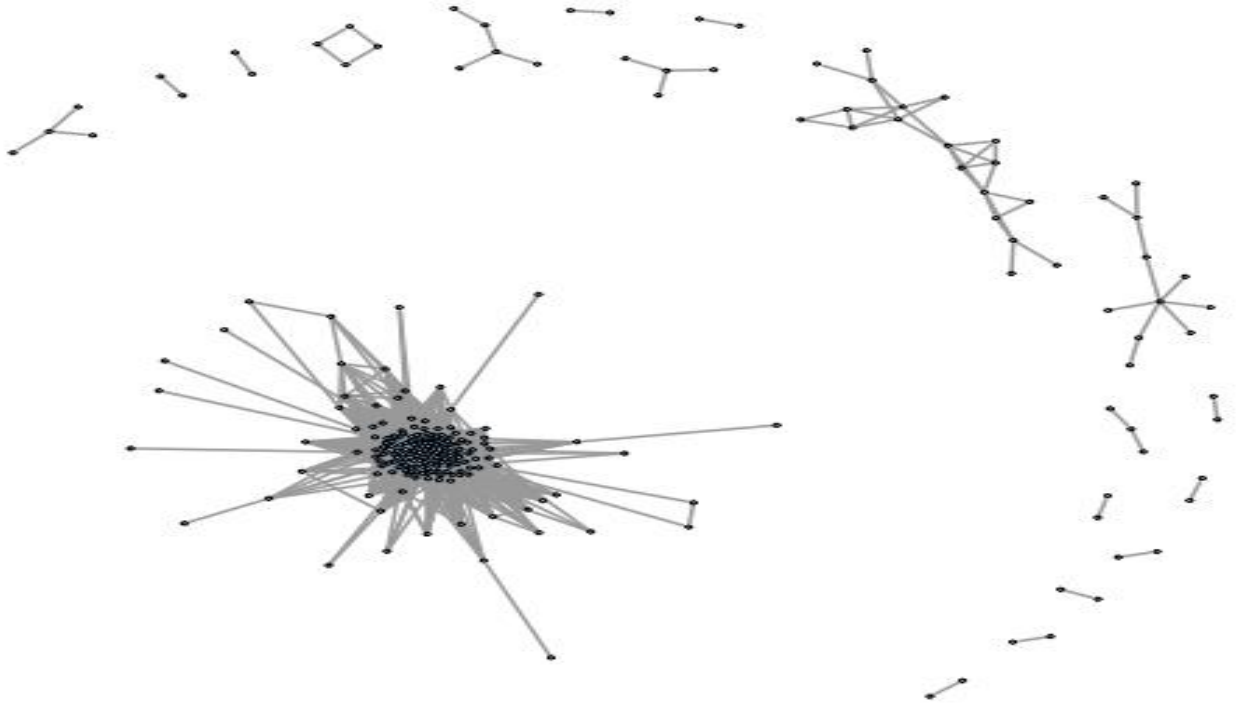
- Walk
  - Path
  - Simple Path
  - Length
  - Cycle



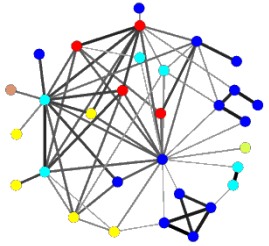


Introduction to Graph Theory

- Connected Components

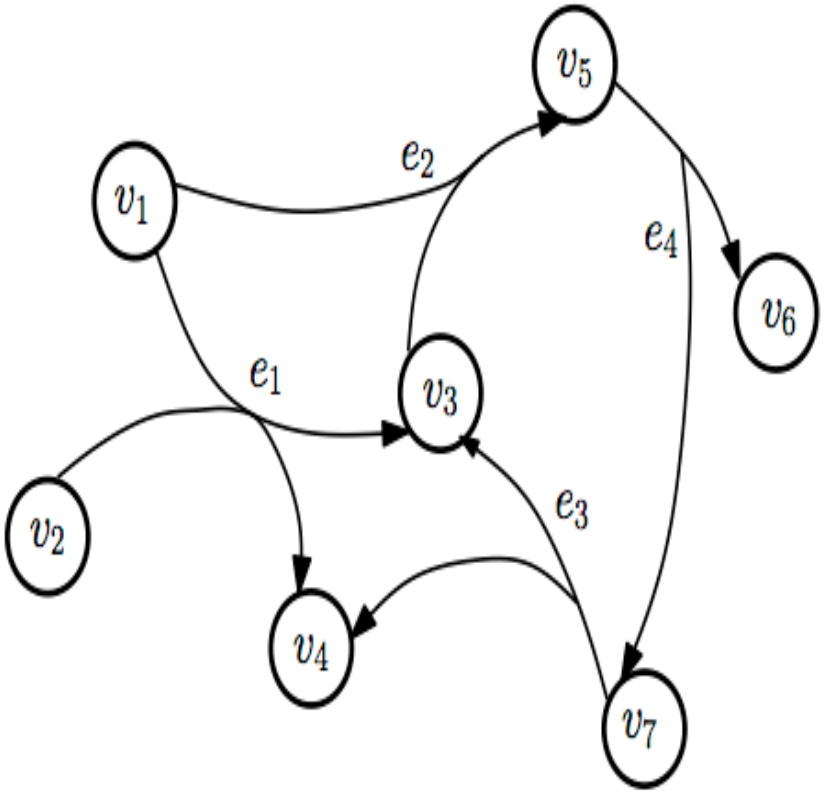


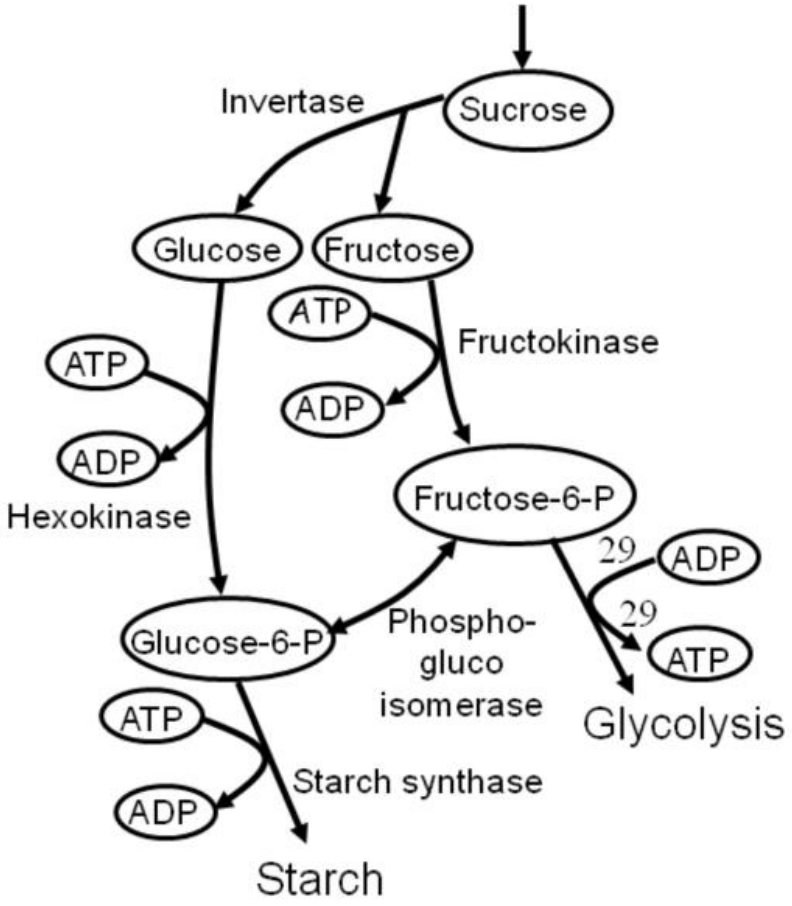
West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall



Introduction to Graph Theory

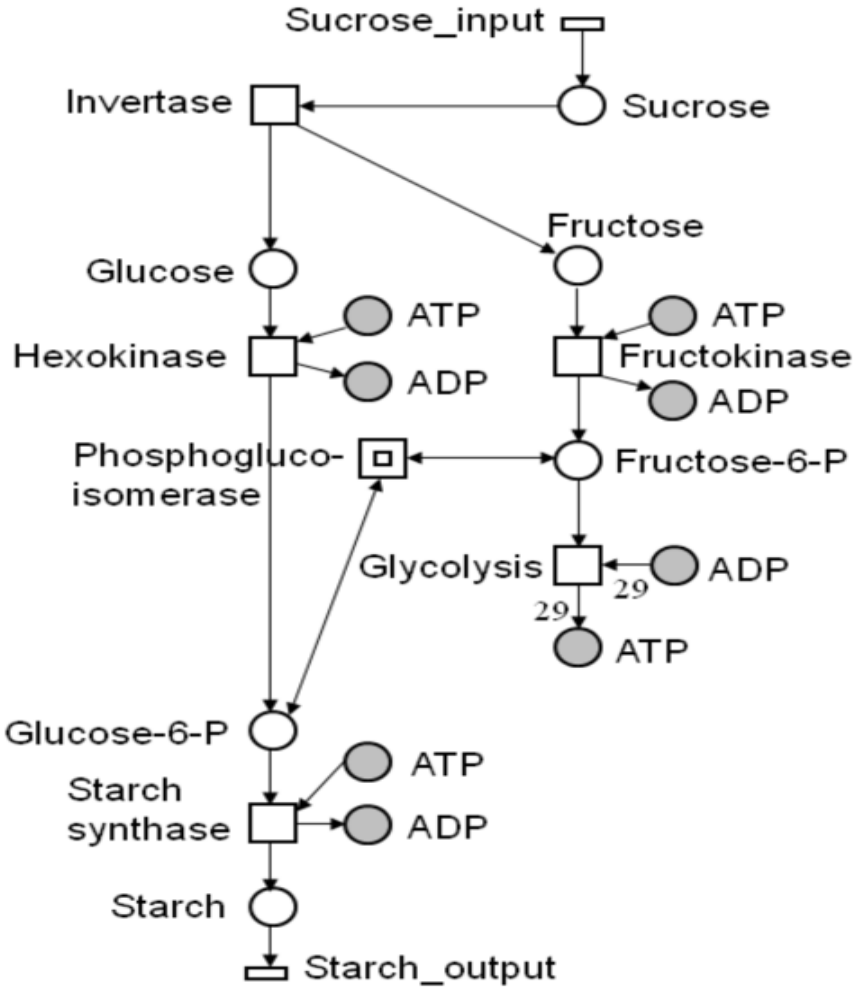
- Hypergraph



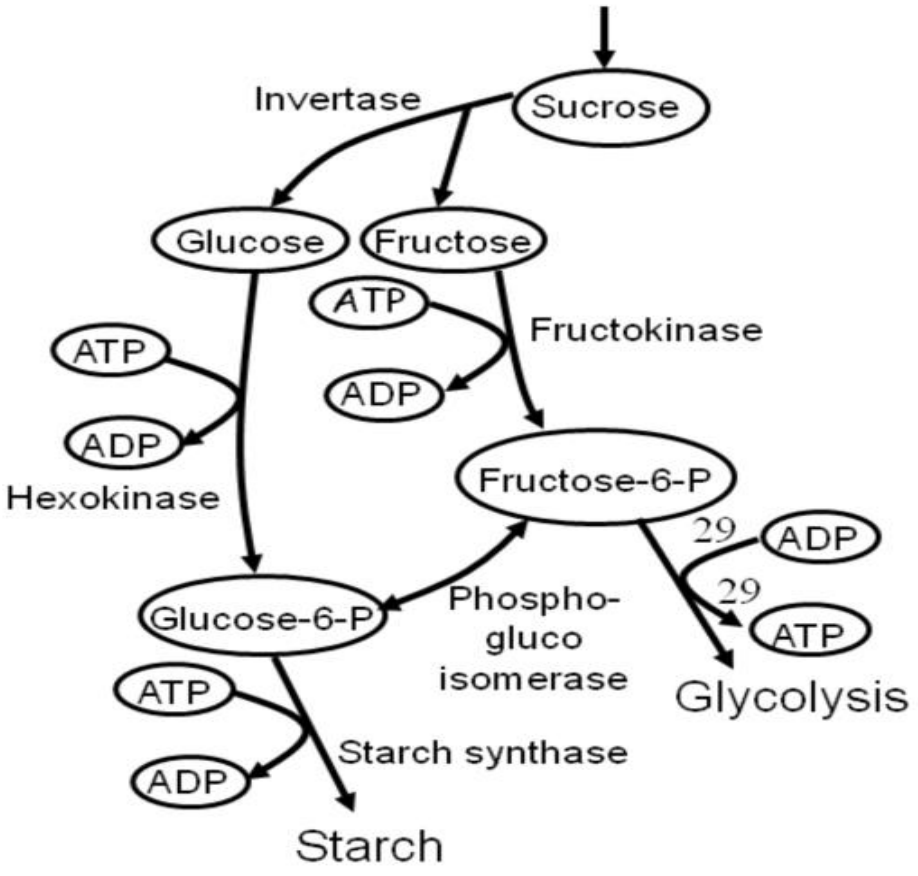


West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall

Hypergraph to Bipartite Graph

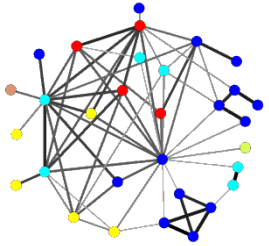


b) The bipartite graph



a) The hypergraph

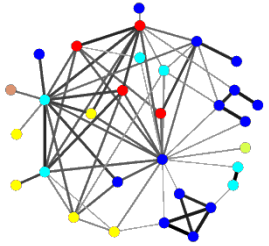
West, Douglas B., 2001, Introduction to Graph Theory, 2nd ed., Prentice Hall



Properties of Large Networks

- **Global network properties**
  1) Degree distribution
  2) Clustering coefficient
  3) Clustering spectrum
  4) Average Diameter
  5) Shortest path lengths
  6) Centralities
  7) Scale-free
  8) Small World

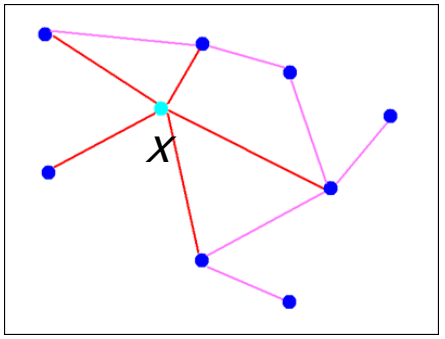
- **Local network properties**
  1) Network motifs



Degree Distribution (Undirected)

Degree distribution (undirected)

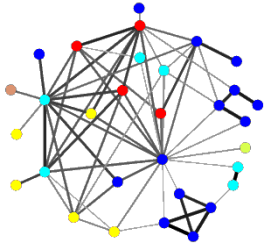
- The **degree** of a node in a network is the number of connections it has to other nodes.



- N(x)=5

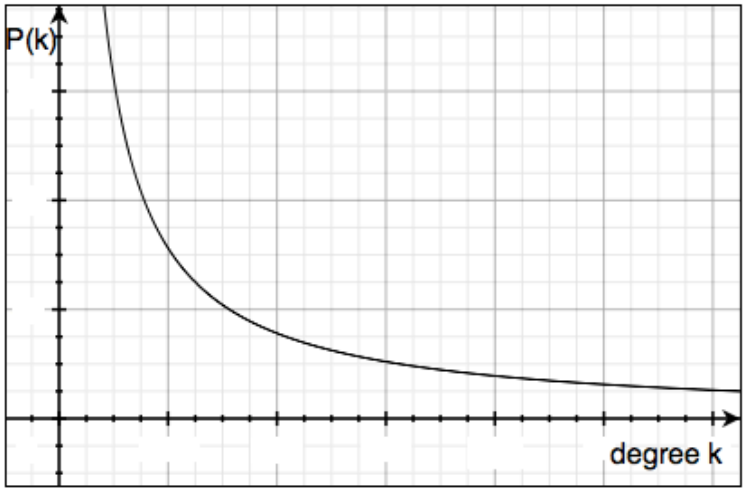
- The **degree distribution** $P(k)$ gives the probability that a selected node has exactly k links

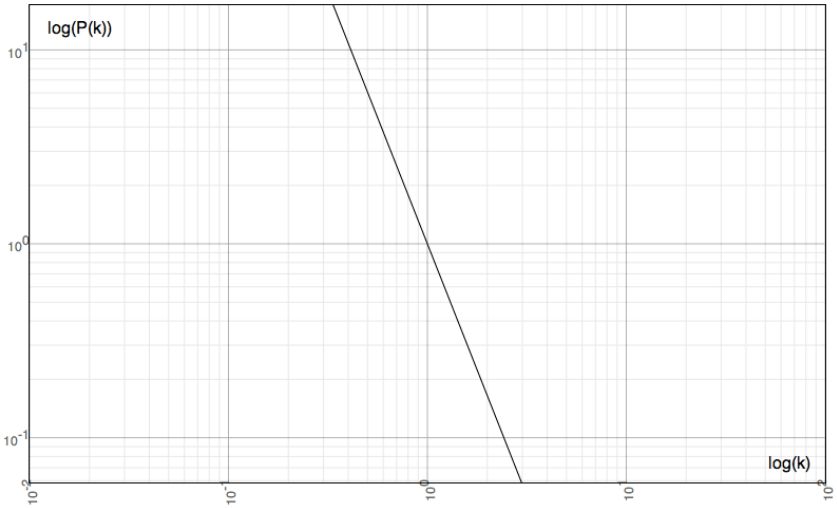
- $P(k)$ is obtained by counting the number of nodes N (k) with $k = 1, 2,\ldots$ links and dividing by the total number of nodes N.



Degree Distribution (Undirected)

- Example:



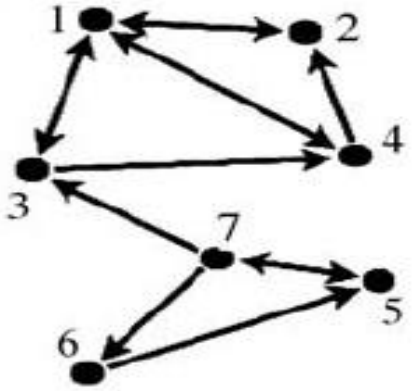


(log-log plot)

- Here $P(k) \sim k^{-\gamma}$, where often $2 \leq \gamma < 3$. This is a *power-law* distribution.
- Networks with power-law degree distributions are called ***scale-free*** networks.
- Most of the nodes are of low degree, but there is a small number of highly-linked nodes.

Degree Distribution (Directed)

(a)



(b) **Adjacency Matrix**

	1	2	3	4	5	6	7
1	0	1	1	1	0	0	0
2	1	0	0	0	0	0	0
3	1	0	0	1	0	0	0
4	1	1	0	0	0	0	0
5	0	0	0	0	0	0	1
6	0	0	0	0	1	0	0
7	0	0	1	0	1	1	0

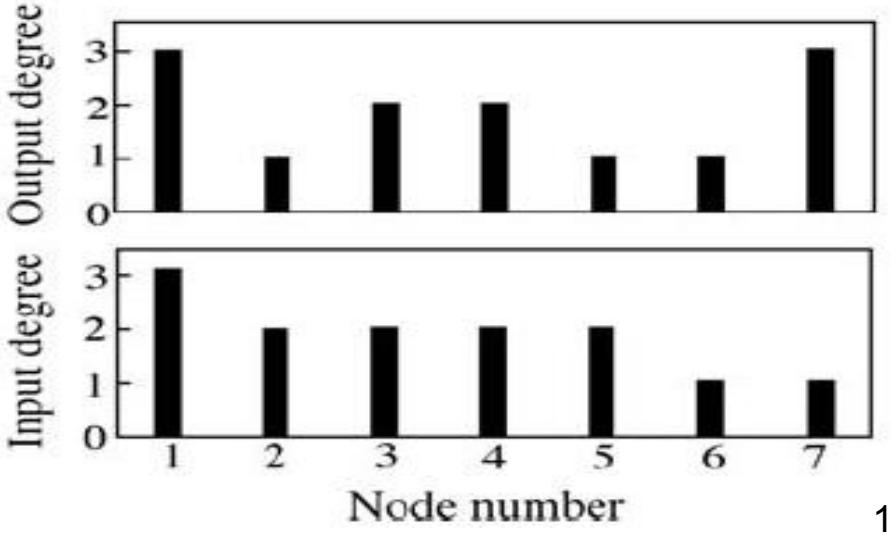
(c) **Adjacency list**

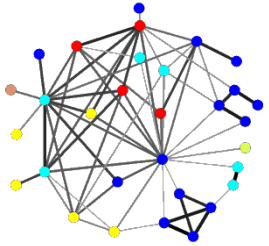
1	2, 3, 4
2	1
3	1, 4
4	1, 2
5	7
6	5
7	3, 5, 6

(d) **Path**

	1	2	3	4	5	6	7
1	2	1	1	1	∞	∞	∞
2	1	2	2	2	∞	∞	∞
3	1	2	2	1	∞	∞	∞
4	1	1	2	2	∞	∞	∞
5	3	4	2	3	2	2	1
6	4	5	3	4	1	3	2
7	2	3	1	2	1	1	2

(e)



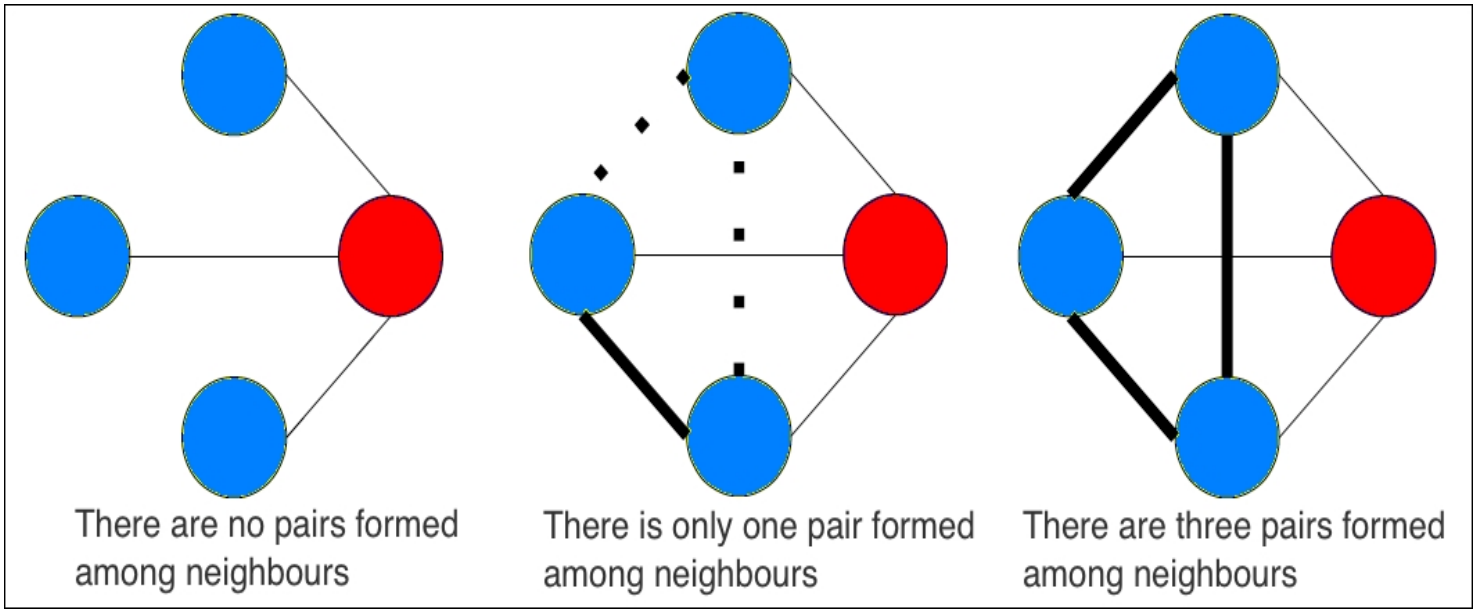


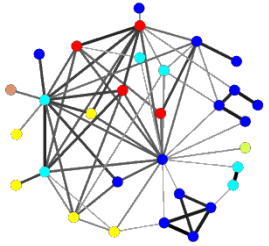
Clustering Coefficient (Local)

clustering coefficient C_v of a node v

$$C_v = 2E_{N(v)} / k_v(k_v - 1)$$

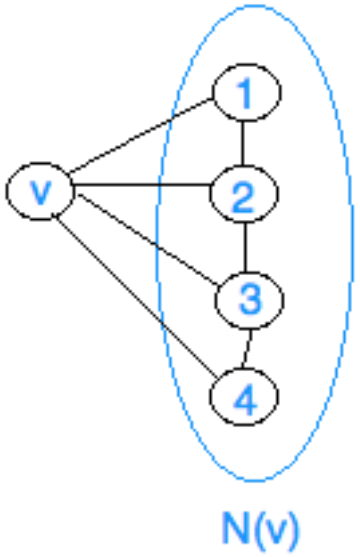
C_v can be viewed as the probability that two neighbors of v are connected. Thus $0 \leq C_v \leq 1$.





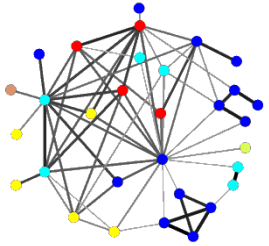
Clustering Coefficient (Local)

- Example:



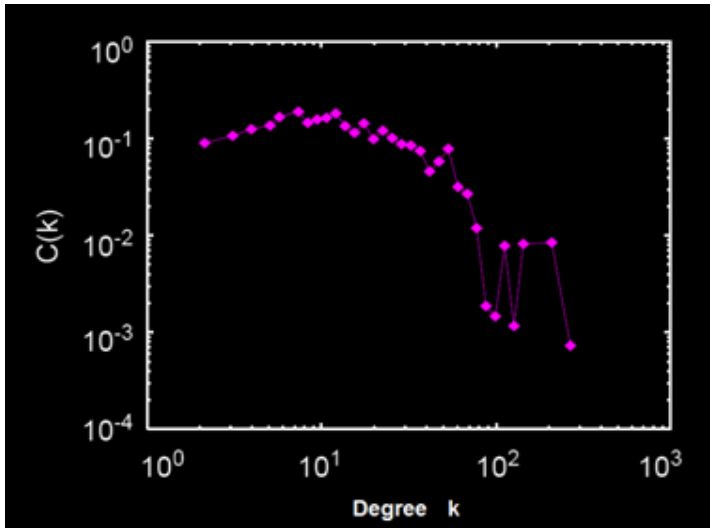
$$C_v = 2E_{N(v)} / k_v(k_v - 1)$$

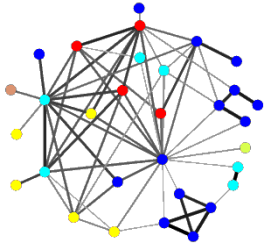
- $|E_{N(v)}| = 3$ (edges in *N(v)*)
- $k_v(k_v-1)/2 = 4(4-1)/2 = 6$ (Maximum Possible Edges in N(v))
- Therefore $C_v = 3/6 = 1/2$



Clustering Coefficient (Local)

- ***Average clustering coefficient***, C, of a network is the average C_v over all the nodes $v \in V$.
- ***Clustering spectrum***, $C(k)$ is the distribution of the average clustering coefficients of all nodes of degree k in the network, over all k.



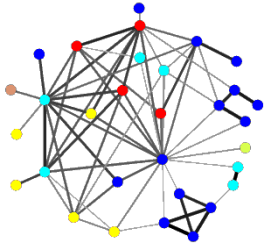


Clustering Coefficient (Global)

- The global gives an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes.
- The global clustering coefficient is based on triplets of nodes.
- A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties.
- The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed).

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triplets of vertices}}.$$

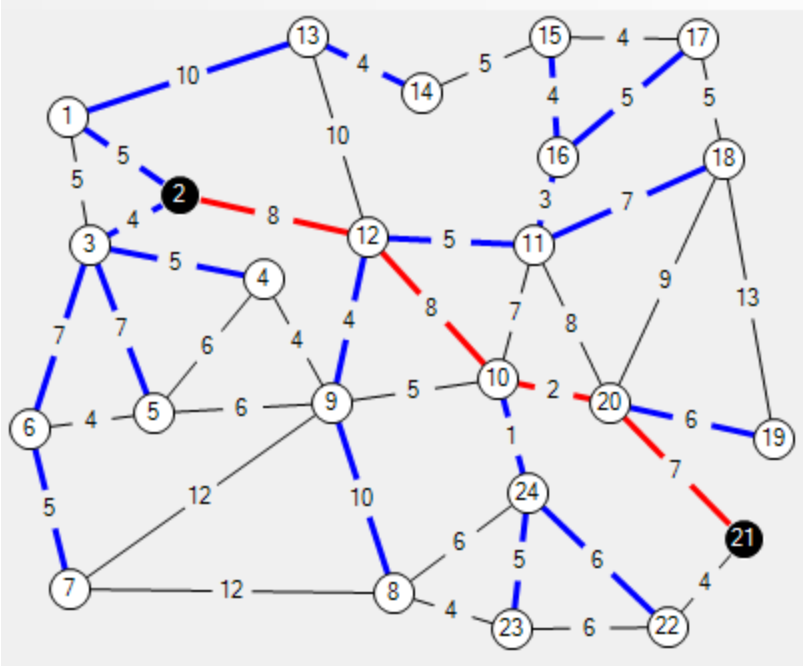
- A triangle consists of three closed triplets, one centered on each of the nodes.

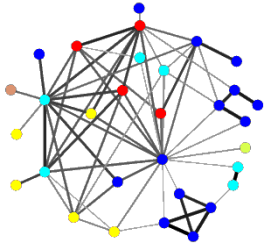


Distance, Shortest Path, Network Diameter

- *Distance between two nodes* is the smallest number of links that have to be traversed to get from one node to the other.

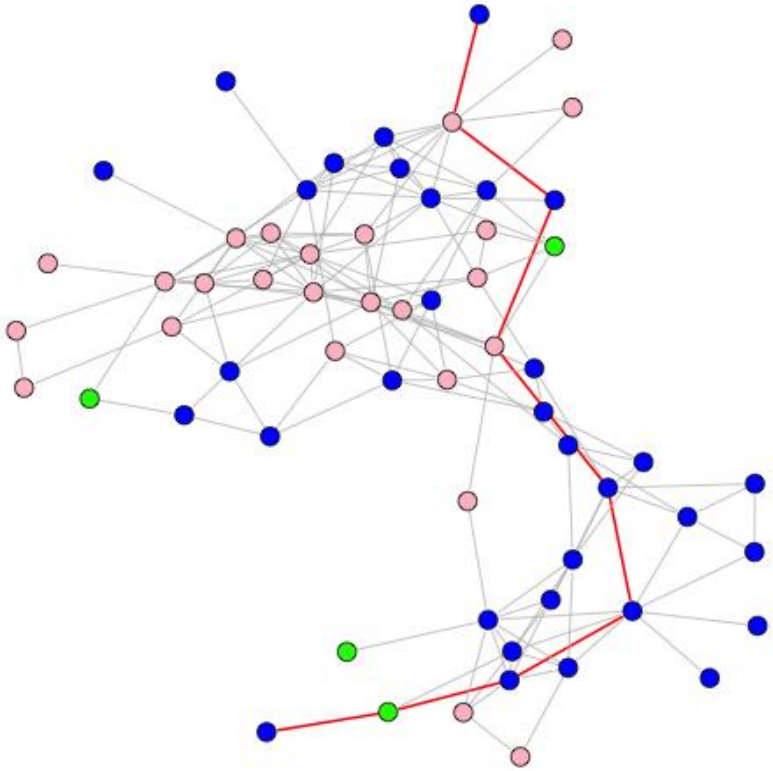
- *Shortest Path* is the path that achieves that distance.

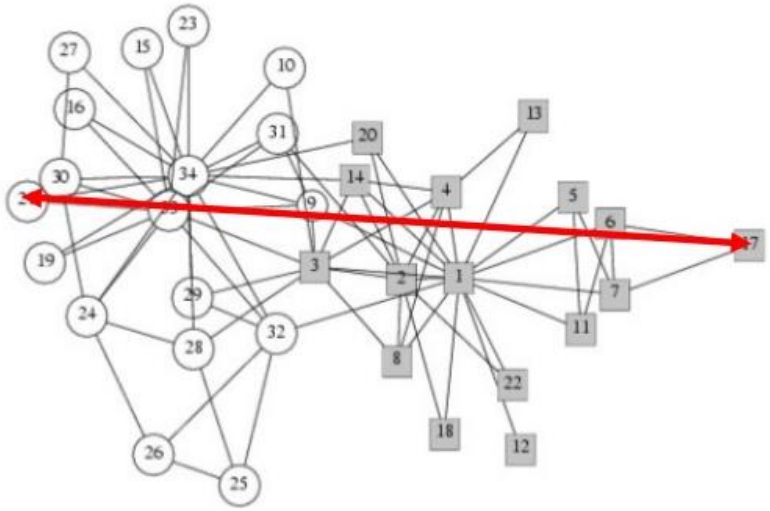




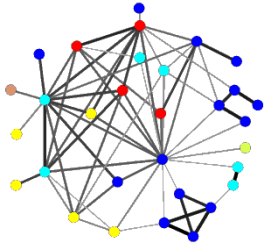
Distance, Shortest Path, Network Diameter

- *Network Diameter* is the average of shortest path lengths over all pairs of nodes in a network.





Tijana Milenkovic, Graph-theoretical approaches for studying biological networks, 2008. The university of California

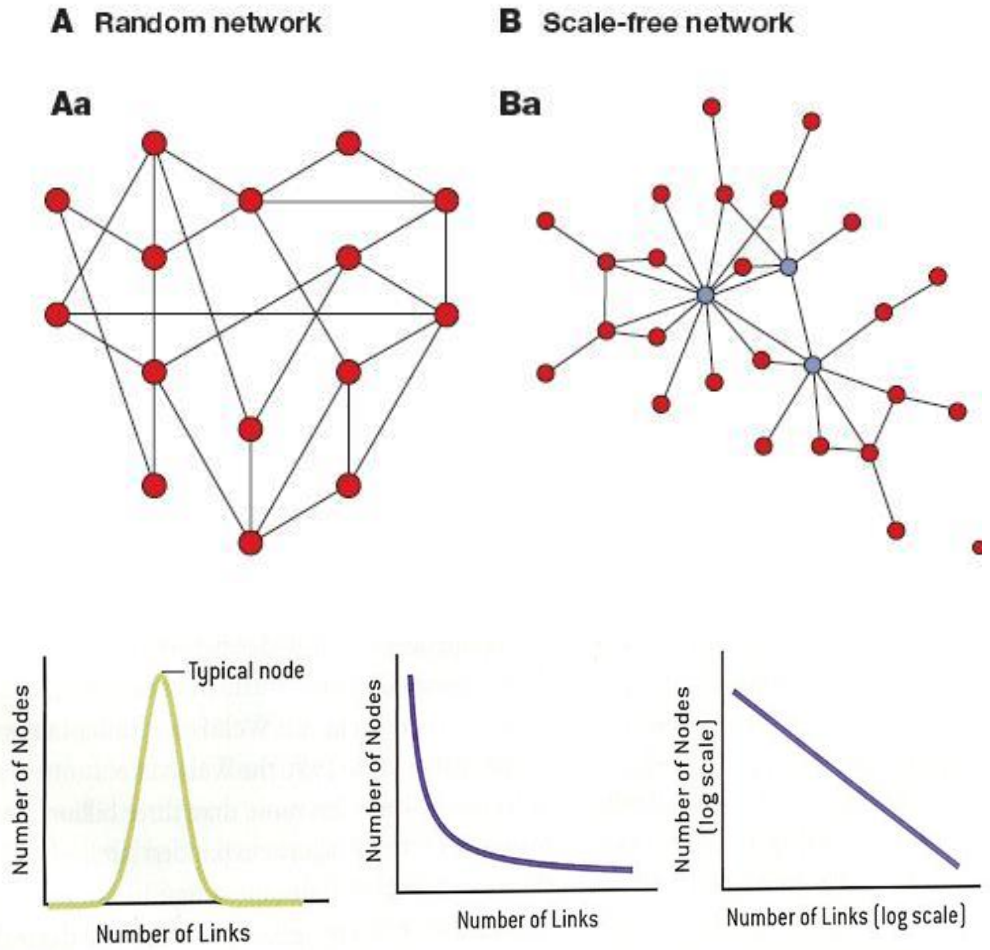


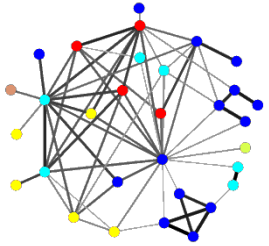
Scale-free vs Random

A scale-free network is a network whose degree distribution follows a power law $P(k) \sim k^{-\gamma}$

Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree.

Network	n	z	C measured	C for random graph
Internet	6,374	3.8	0.24	0.00060
World Wide Web (sites)	153,127	35.2	0.11	0.00023
power grid	4,941	2.7	0.080	0.00054
biology collaborations	1,520,251	15.5	0.081	0.000010
mathematics collaborations	253,339	3.9	0.15	0.000015
film actor collaborations	449,913	113.4	0.20	0.00025
company directors	7,673	14.4	0.59	0.0019
word co-occurrence	460,902	70.1	0.44	0.00015
neural network	282	14.0	0.28	0.049
metabolic network	315	28.3	0.59	0.090
food web	134	8.7	0.22	0.065

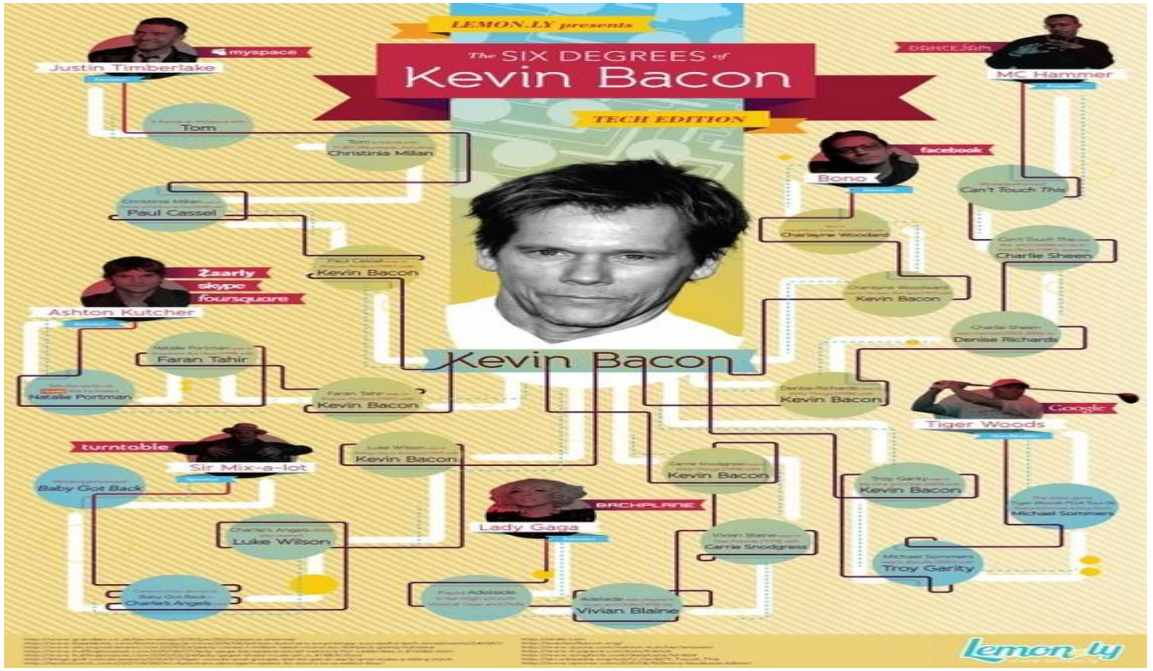


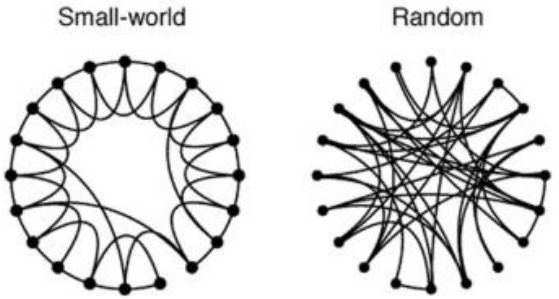


Small World

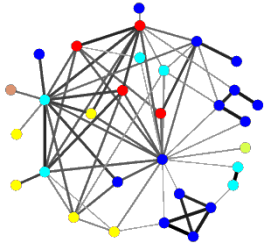
a small-world network is defined to be a network where the typical distance L between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes N in the network.

$$L \propto \log N$$



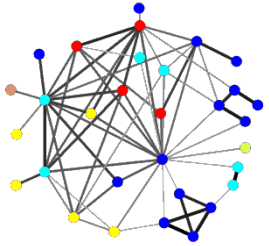


Tijana Milenkovic, Graph-theoretical approaches for studying biological networks, 2008. The university of California

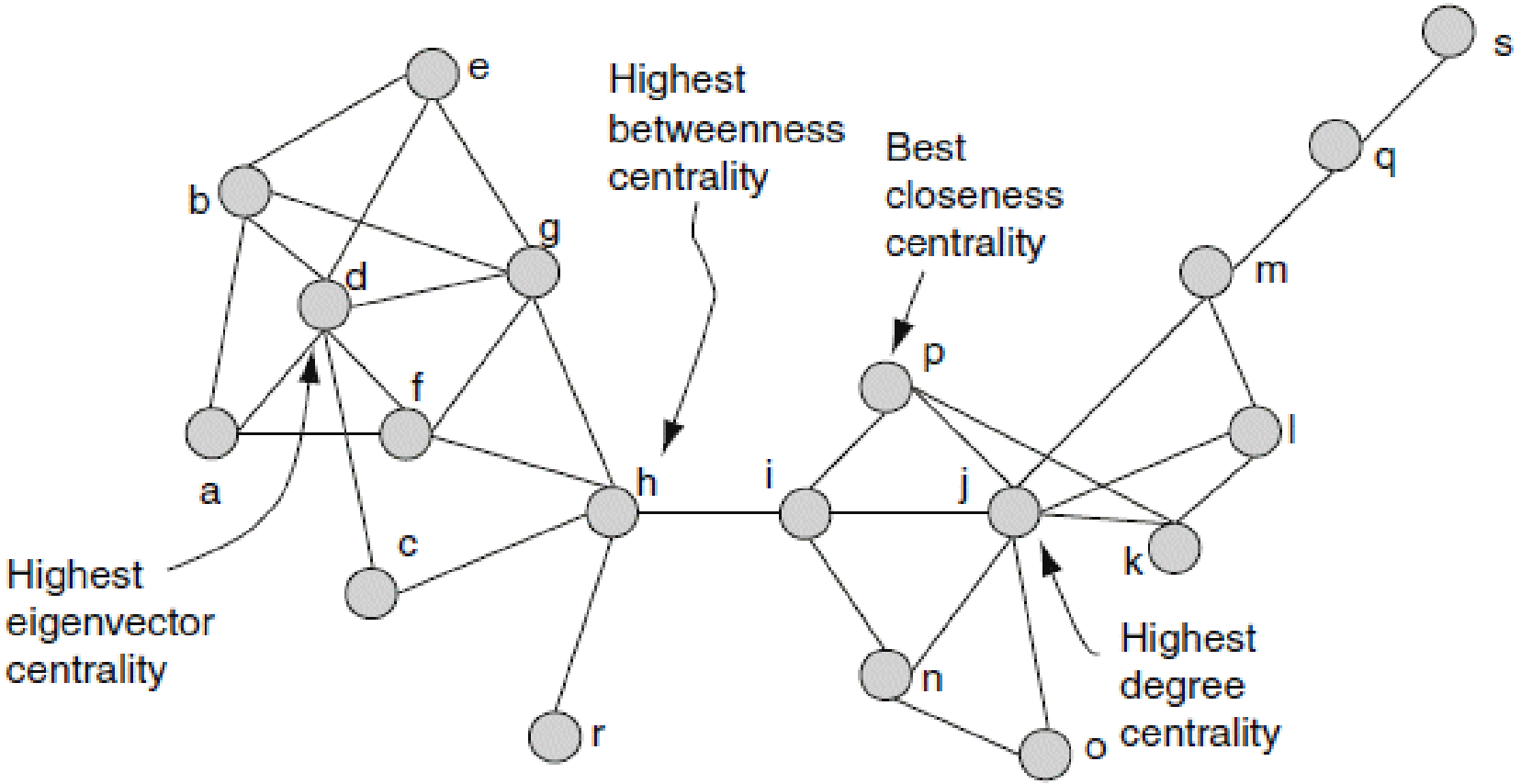


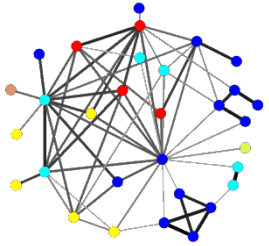
Centrality

- Rank nodes according to their “topological importance”
  - Degree
  - Closeness
  - Betweenness
  - Eigenvector
  - BottleNeck
  - Clustering coefficient
  - Stress
  - Radiality
  - EcCentricity
  - EPC
  - MNC
  - DMNC
  - MCC
  - …



Centrality

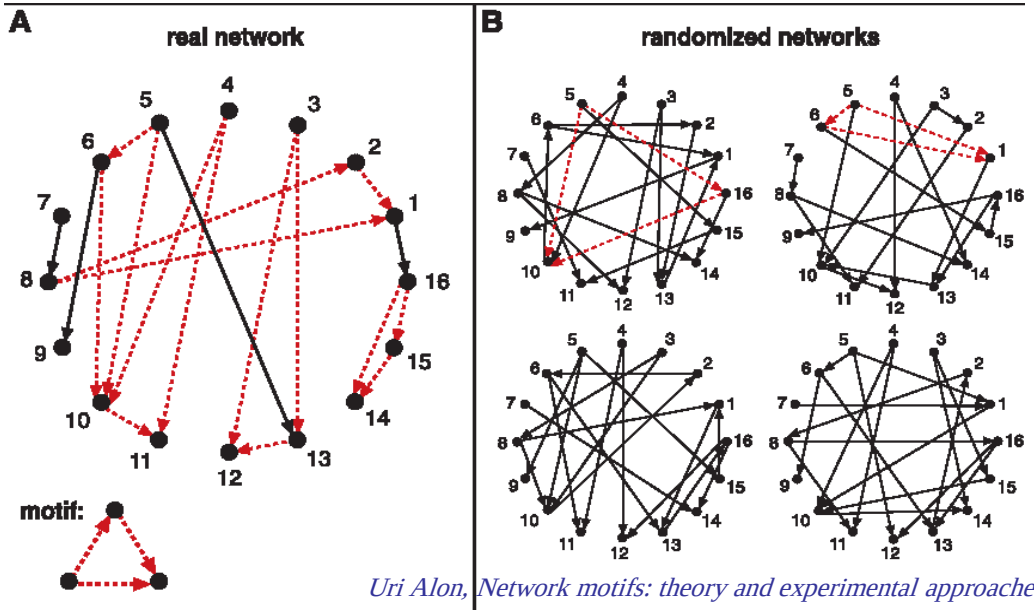

e
b
g
d
f
a
c
h
r
i
j
p
n
o
k
m
q
s
l
Highest
betweenness
centrality
Best
closeness
centrality
Highest
eigenvector
centrality
Highest
degree
centrality

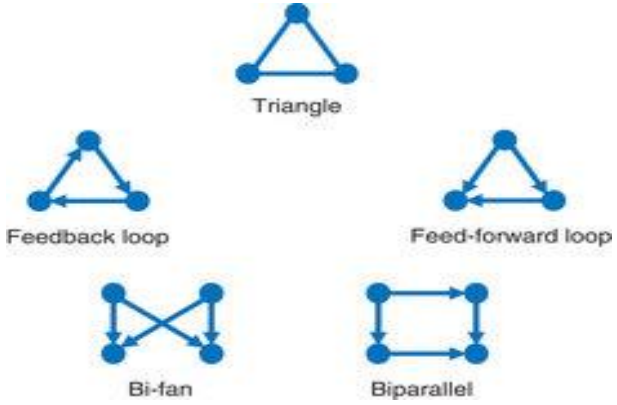


Network Motifs

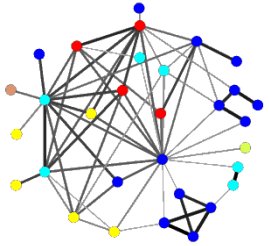
Network motifs are connectivity-patterns (sub-graphs) that occur much more often than they do in random networks.

These small circuits can be considered as simple building blocks from which the network is composed.
This idea was first presented by Uri Alon and his group when network motifs were discovered in the gene regulation network of the bacteria E. coli.



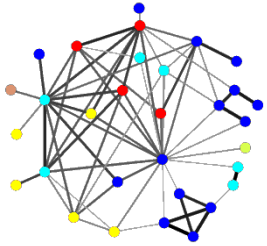


Uri Alon, Network motifs: theory and experimental approaches, Nature Reviews Genetics, 8, 451, 2007

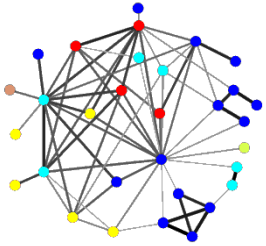


Network Clustering

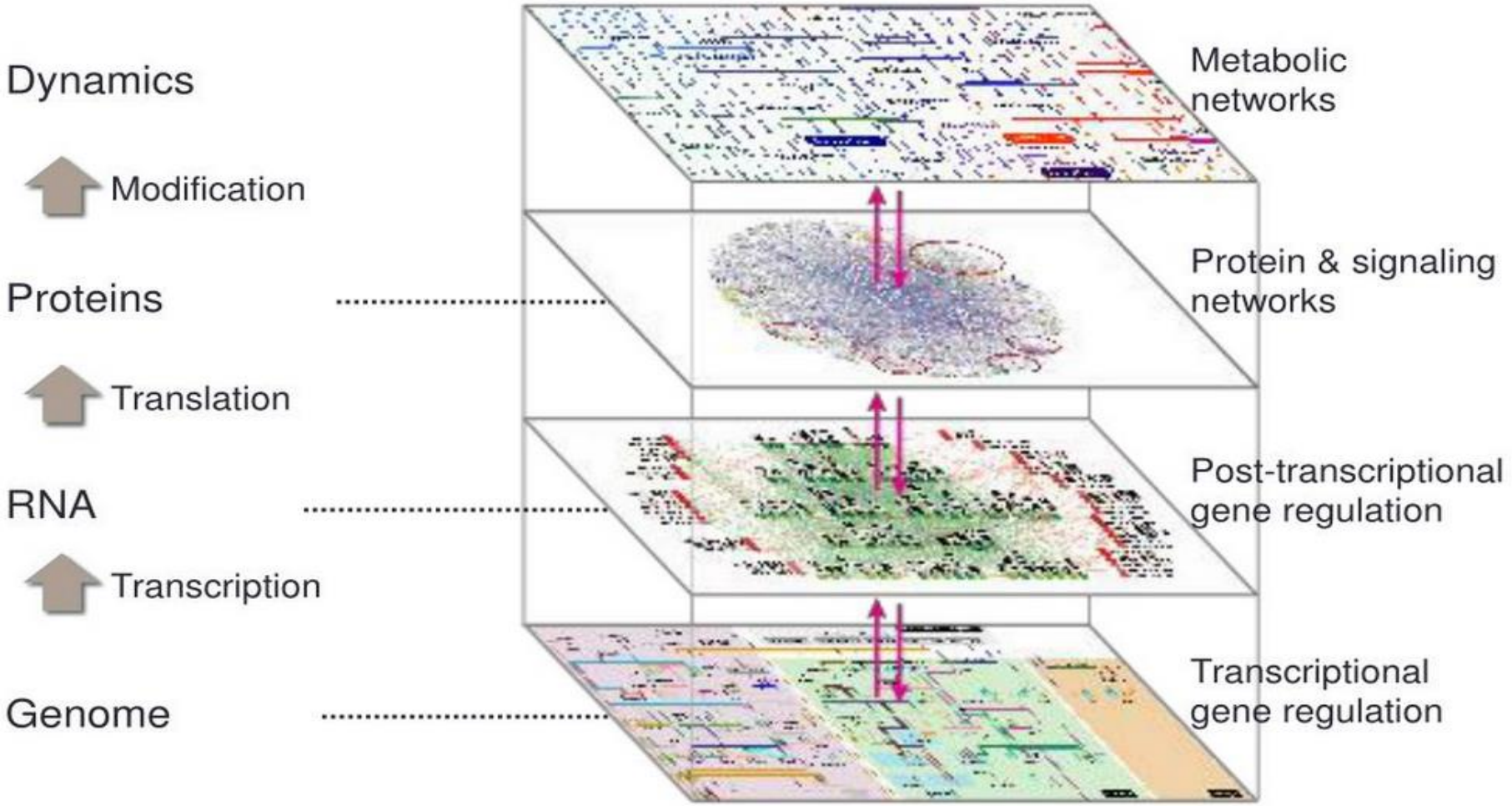
Tijana Milenkovic, Graph-theoretical approaches for studying biological networks, 2008. The university of California



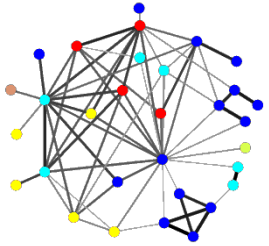
Graph Theory in Biological Networks



Graph Theory in Biological Networks

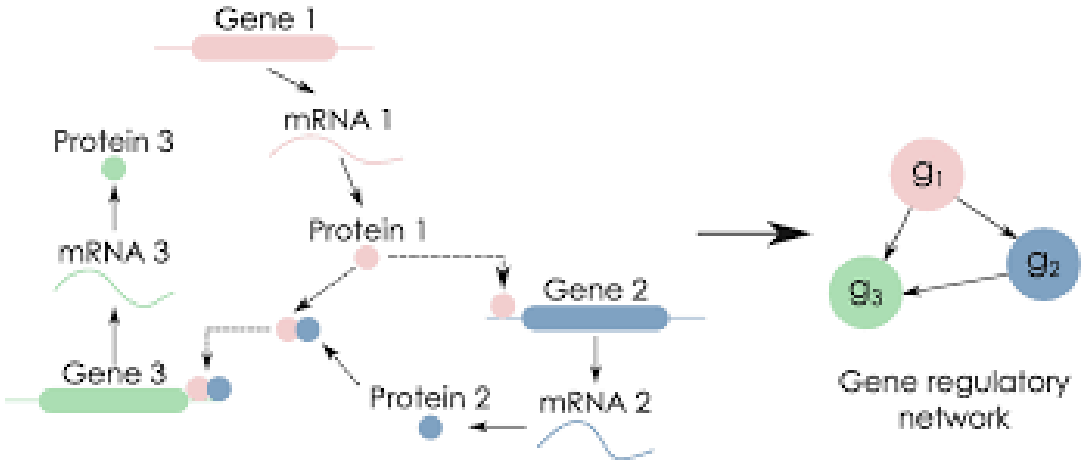


http://www.slideserve.com/ashling/by-soheil-feizi-final-project-presentation-18-338-mit

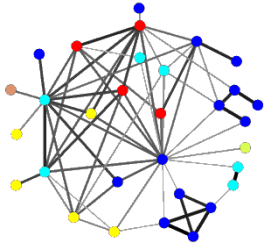


Gene Regulatory Network (GRN)

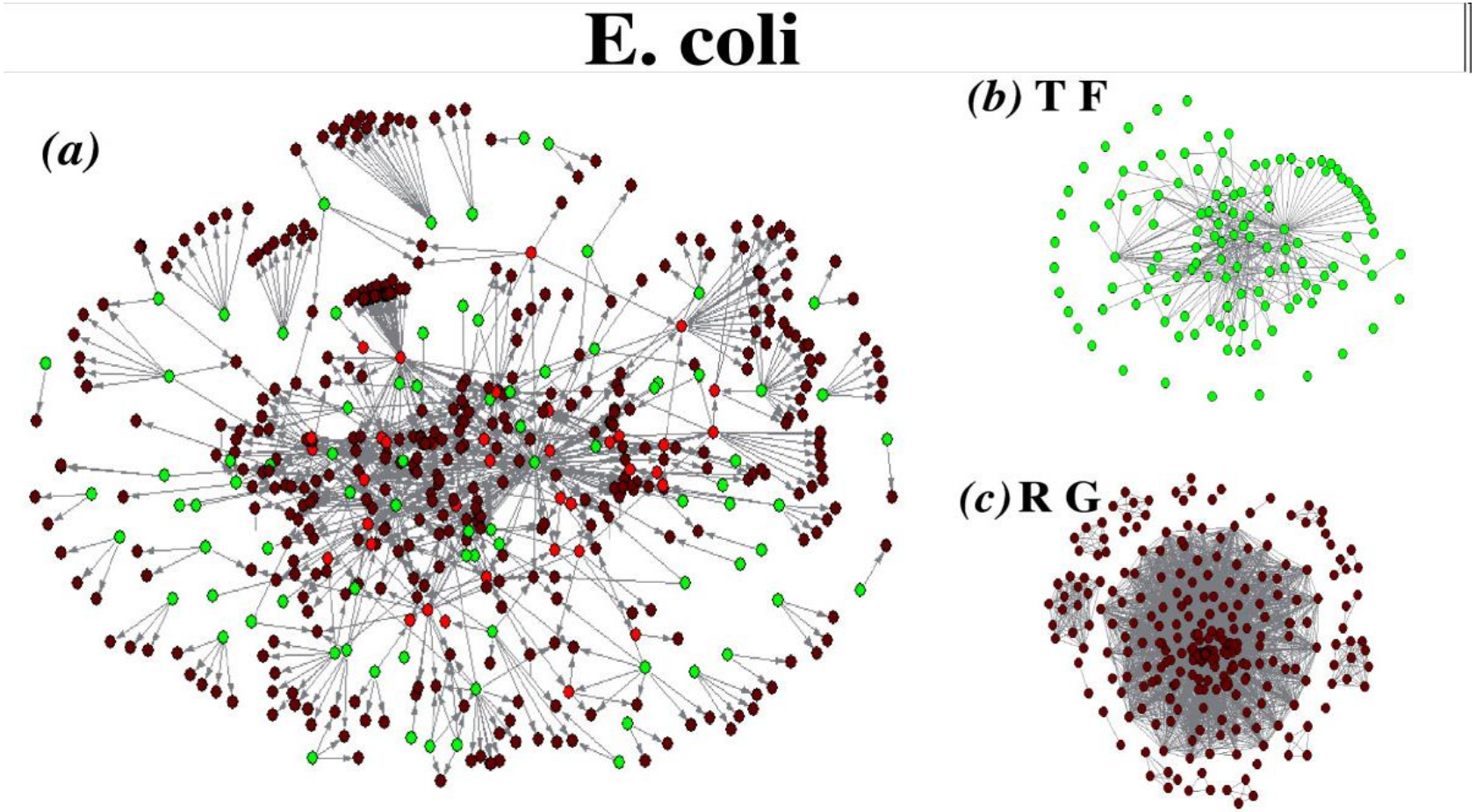
- Nodes correspond to genes
- Directed edges correspond to interactions through which the products of one gene affect those of another
  - Protein-protein, protein-DNA and protein-mRNA interactions



- *Transcription factor* X (protein product of gene X) binds regulatory DNA regions of gene Y to regulate the production rate (i.e., stimulate or repress transcription) of protein Y

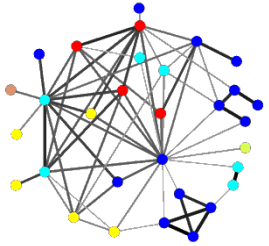


Gene Regulatory Network (GRN)



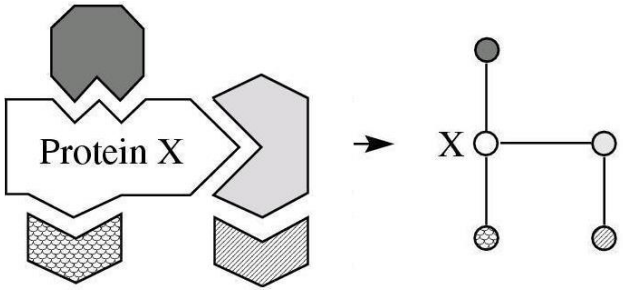
Representation of the *E. coli* transcriptional regulatory network. a) Representation of the transcription-factor gene regulatory network of *E. coli*. Green circles represent transcription factors, brown circles denote regulated genes, and those with both functions are coloured in red. Projections of the network onto b) transcription factor and onto c) regulated gene nodes are also shown.

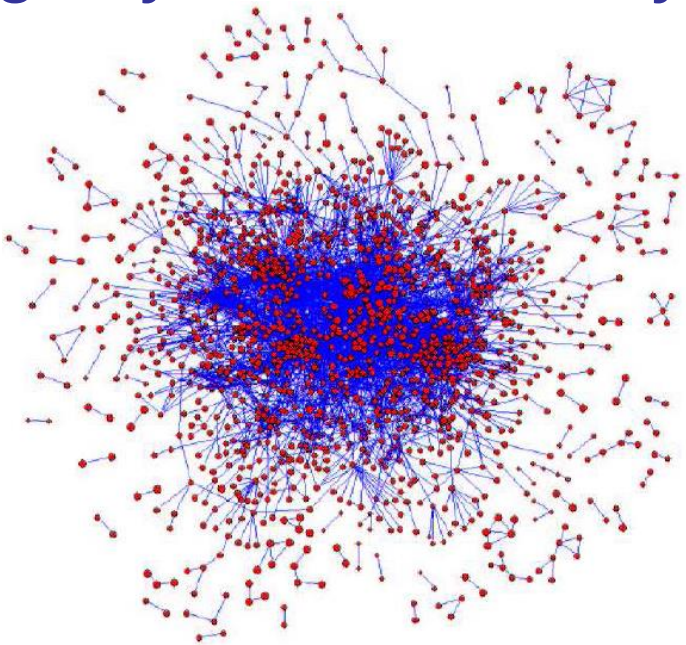
Guzmán-Vargas and Santillán *BMC Systems Biology* 2008 **2**:13 doi:10.1186/1752-0509-2-13



Protein-Protein Interaction (PPI) Network

- A *protein-protein interaction (PPI)* usually refers to a physical interaction, i.e., binding between proteins
- Can be other associations of proteins such as functional interactions – e.g., synthetic lethality



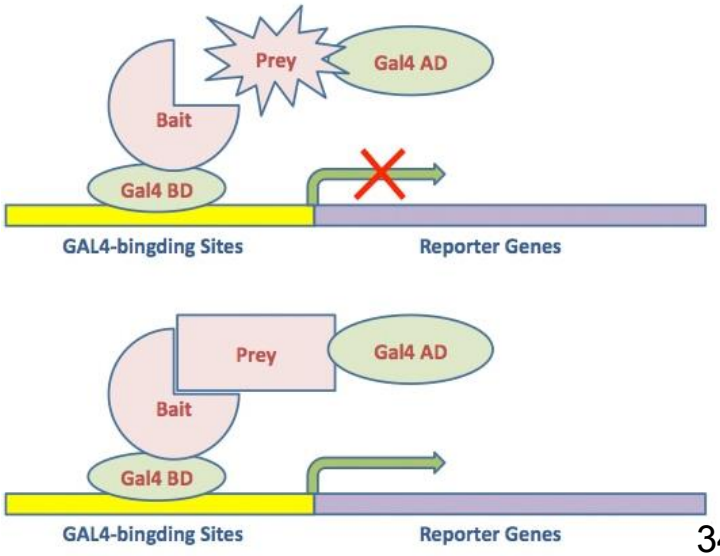


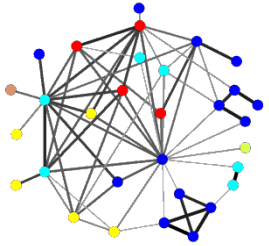
Protein-Protein Interaction (PPI) Network

- The premise behind the test is the activation of downstream reporter gene by the binding of a transcription factor onto an upstream activating sequence (UAS).
- For two-hybrid screening, the transcription factor is split into two separate fragments, called the DNA-binding domain (DBD or often also abbreviated as BD) and activating domain (AD).

The BD is the domain responsible for binding to the UAS and the AD is the domain responsible for the activation of transcription.

The Y2H is thus a protein-fragment complementation assay.

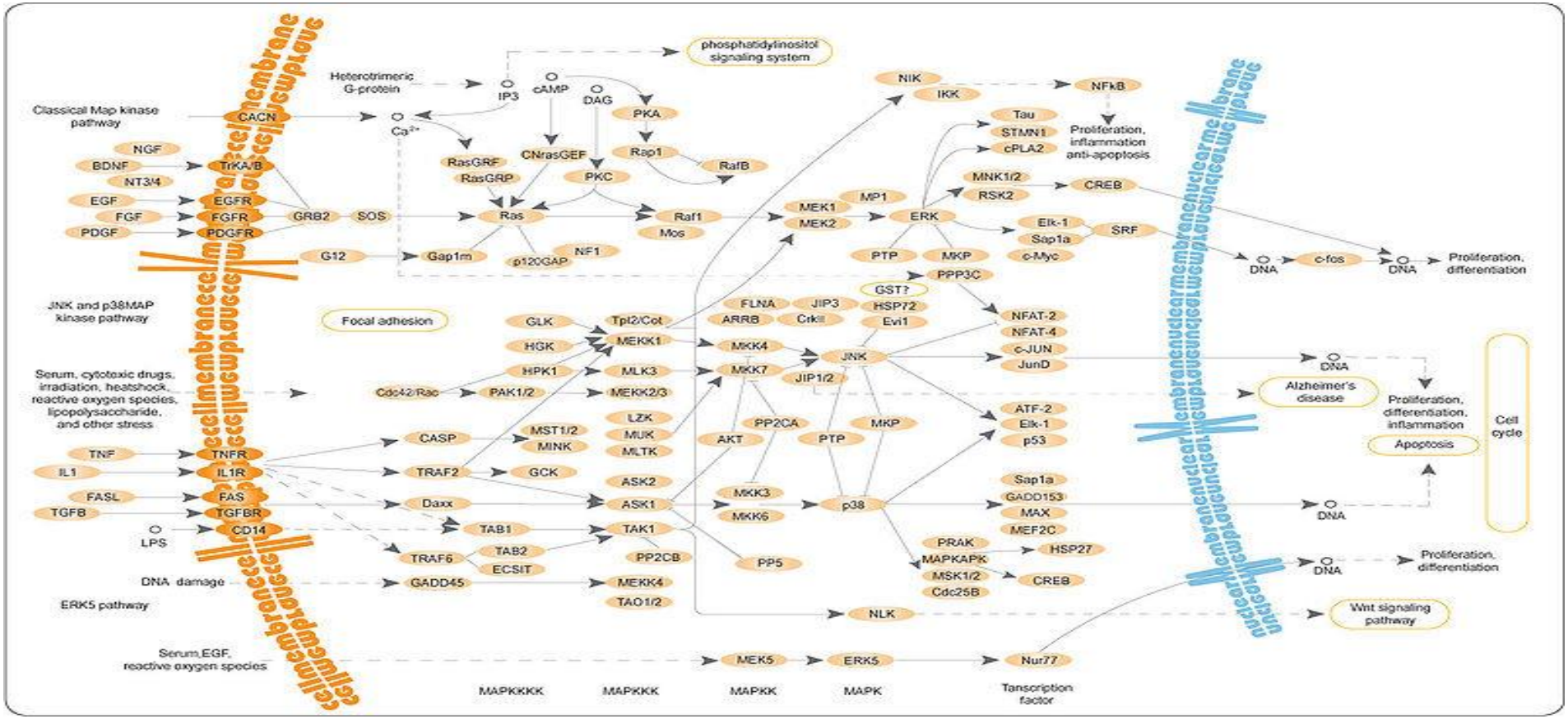




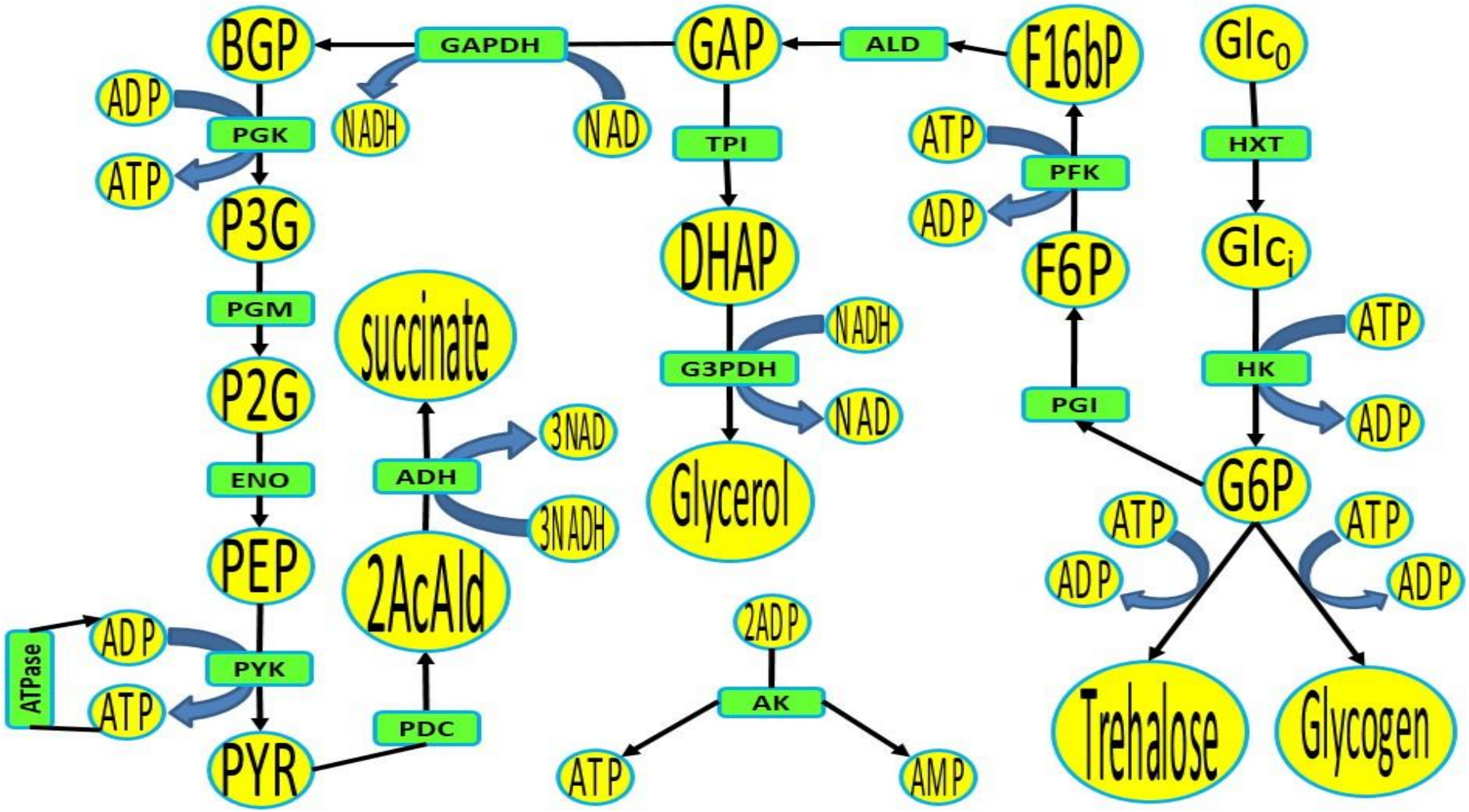
Signaling network

Nodes are proteins and edges are directed

example: MAPK/ERK pathway (Ras-Raf-MEK-ERK pathway)

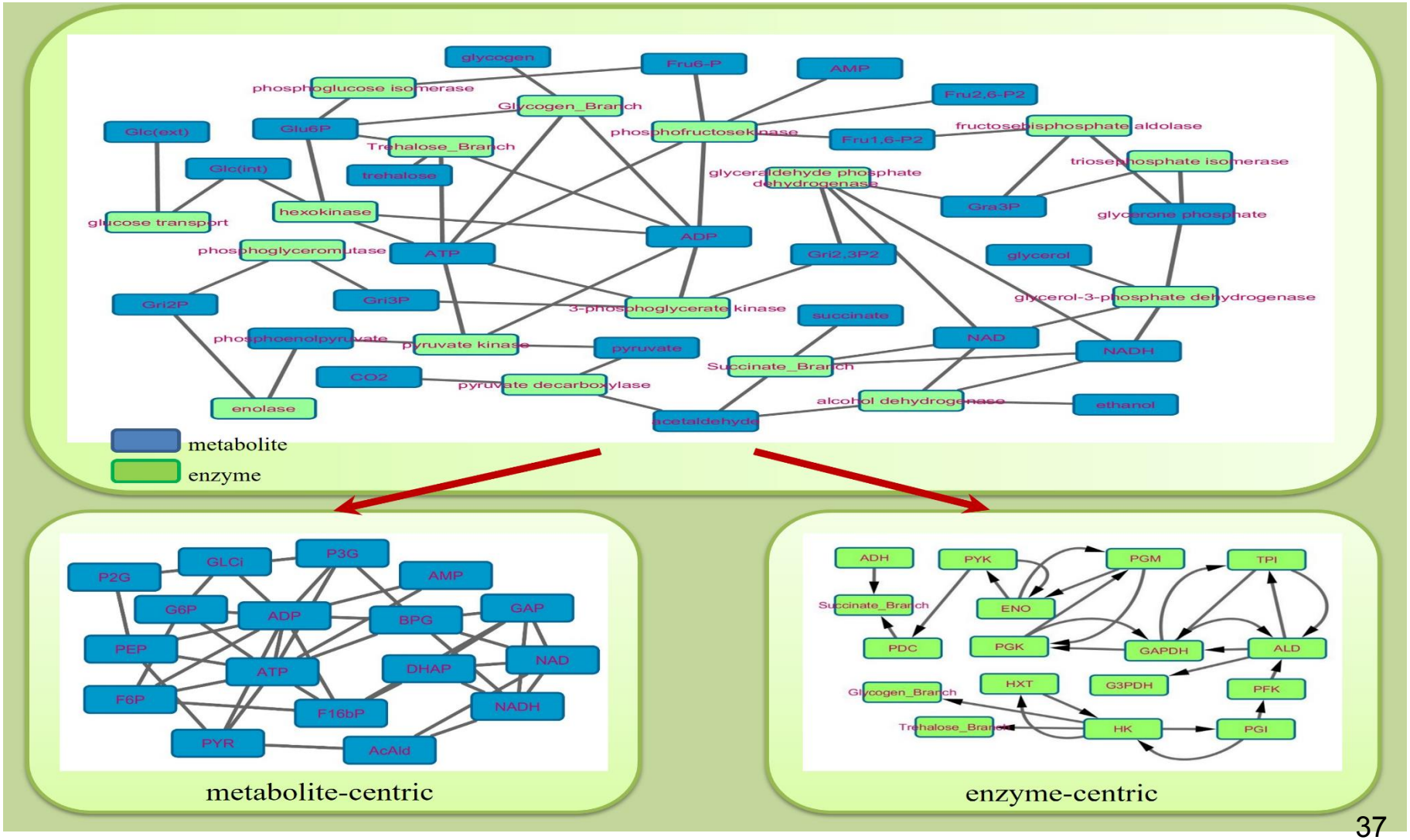


Metabolic Network-Example: Glycolysis

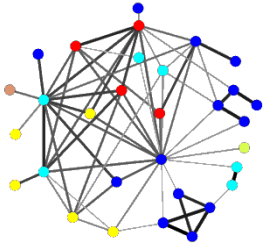


Asgari, et. al., Structural COBRA Add-oN (SCAN) for Analysing Large Metabolic Networks, 2018

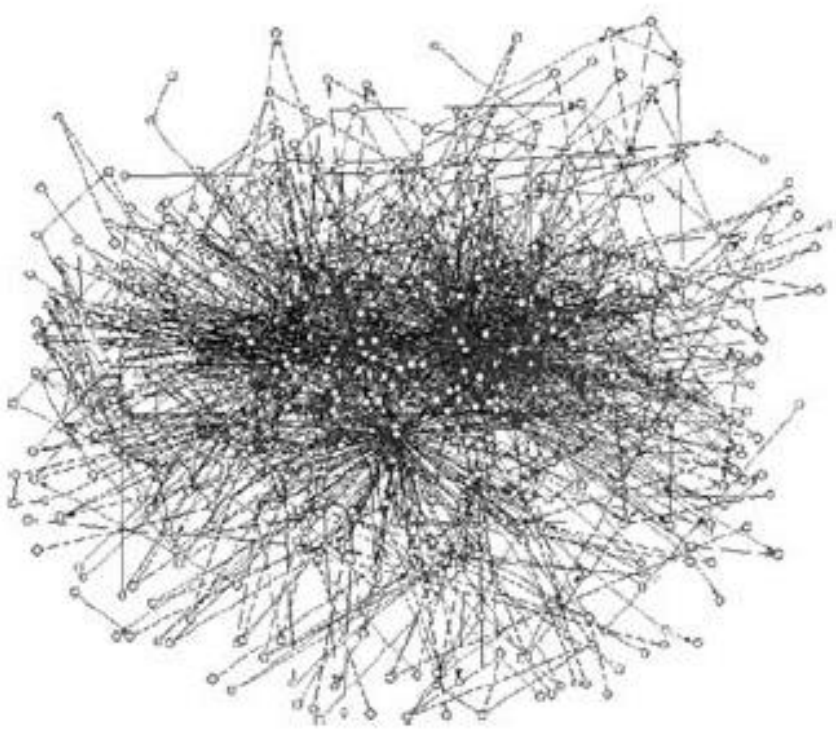
Metabolite-centric, Enzyme-centric

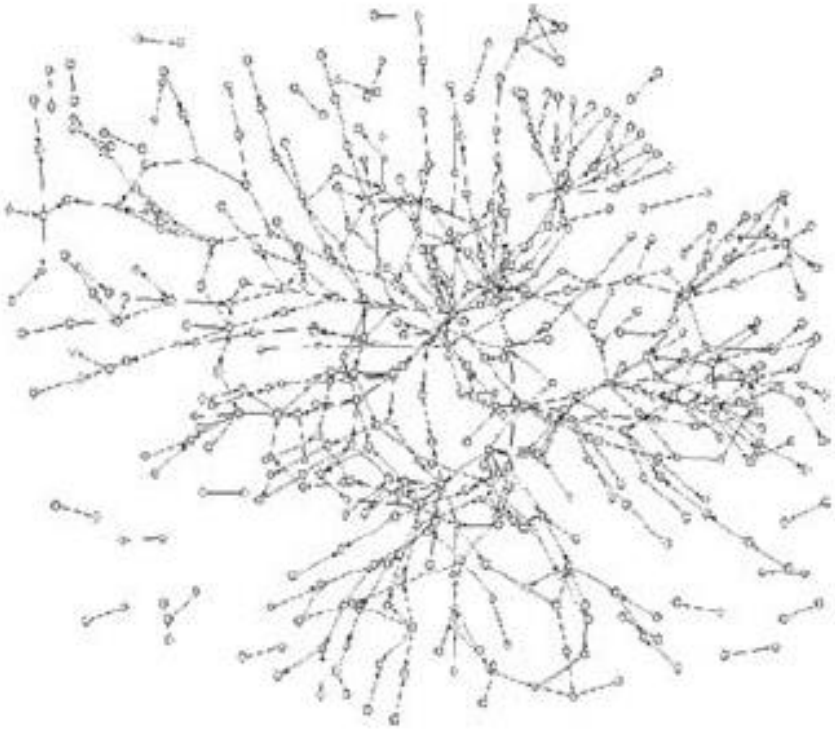


Asgari, et. al., Structural COBRA Add-oN (SCAN) for Analysing Large Metabolic Networks, 2018

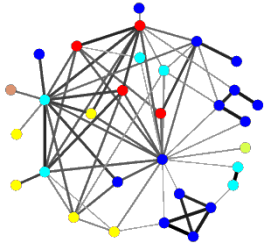


Currency Metabolites





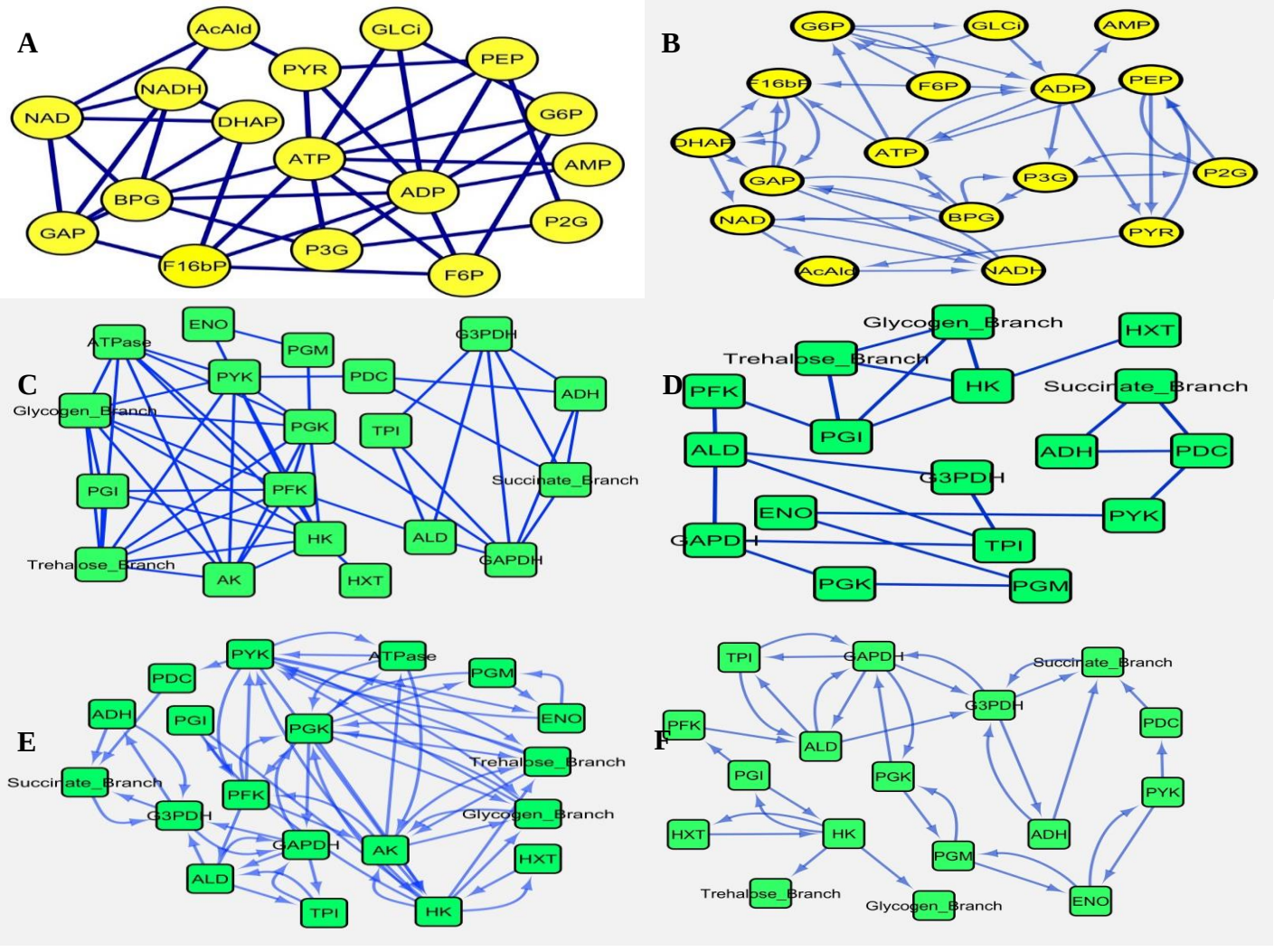
B. Junker, F. Schreiber, Analysis of Biological Network, 2008



Metabolic Network-Example: Glycolysis

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		HXT	HK	PGI	PFK	ALD	TPI	GAPDH	PGK	PGM	ENO	PYK	PDC	ADH	ATPase	AK	G3PDH	Glycogen_Branch	Trehalose_Branch
2	GLCi	1	-1																
3	ATP		-1		-1				1			1			-1	1		-1	-1
4	G6P		1	-1														-1	-1
5	ADP		1		1				-1			-1			1	-2		1	1
6	F6P			1	-1														
7	F16bP				1	-1													
8	AMP															1			
9	DHAP						1										-1		
10	GAP					1	-1	-1											
11	NAD							-1						3			1		
12	BPG							1	-1										
13	NADH							1						-3			-1		
14	P3G								1	-1									
15	P2G									1	-1								
16	PEP										1	-1							
17	PYR											1	-1						
18	AcAld												2	-2					

Glycolysis BioModel (BIOMD0000000172)



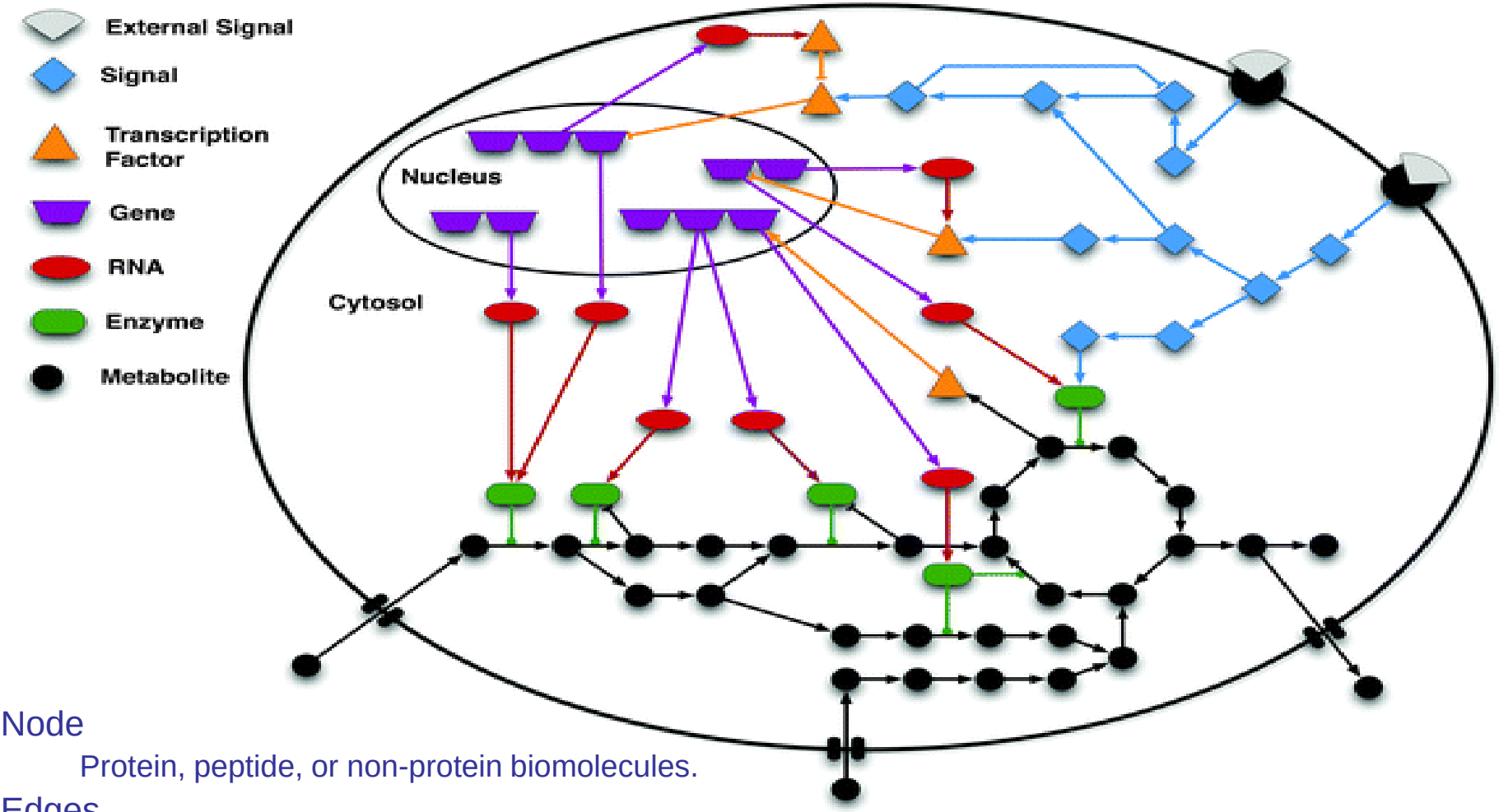
A) Undirected metabolite-centric network (Nodes=17, Edges=40).
B) Directed metabolite-centric network (Nodes=17, Edges=42).

C) Undirected enzyme-centric network without removing currency metabolites (Nodes=19, Edges=52).
D) Undirected enzyme-centric network after removing currency metabolites (Nodes=17, Edges=22).

E) Directed enzyme-centric network without removing currency metabolites (Nodes=19, Edges=62).
F) Directed enzyme-centric network after removing currency metabolites (Nodes=17, Edges=32).

Asgari, et. al., Structural COBRA Add-oN (SCAN) for Analysing Large Metabolic Networks, 2018

Integrated Networks



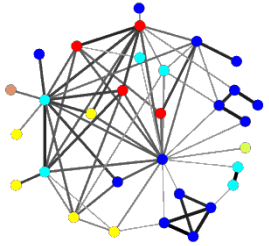
Node

Protein, peptide, or non-protein biomolecules.

Edges

Biological relationships, interactions ,regulations, reactions, transformations, activation, inhibitions

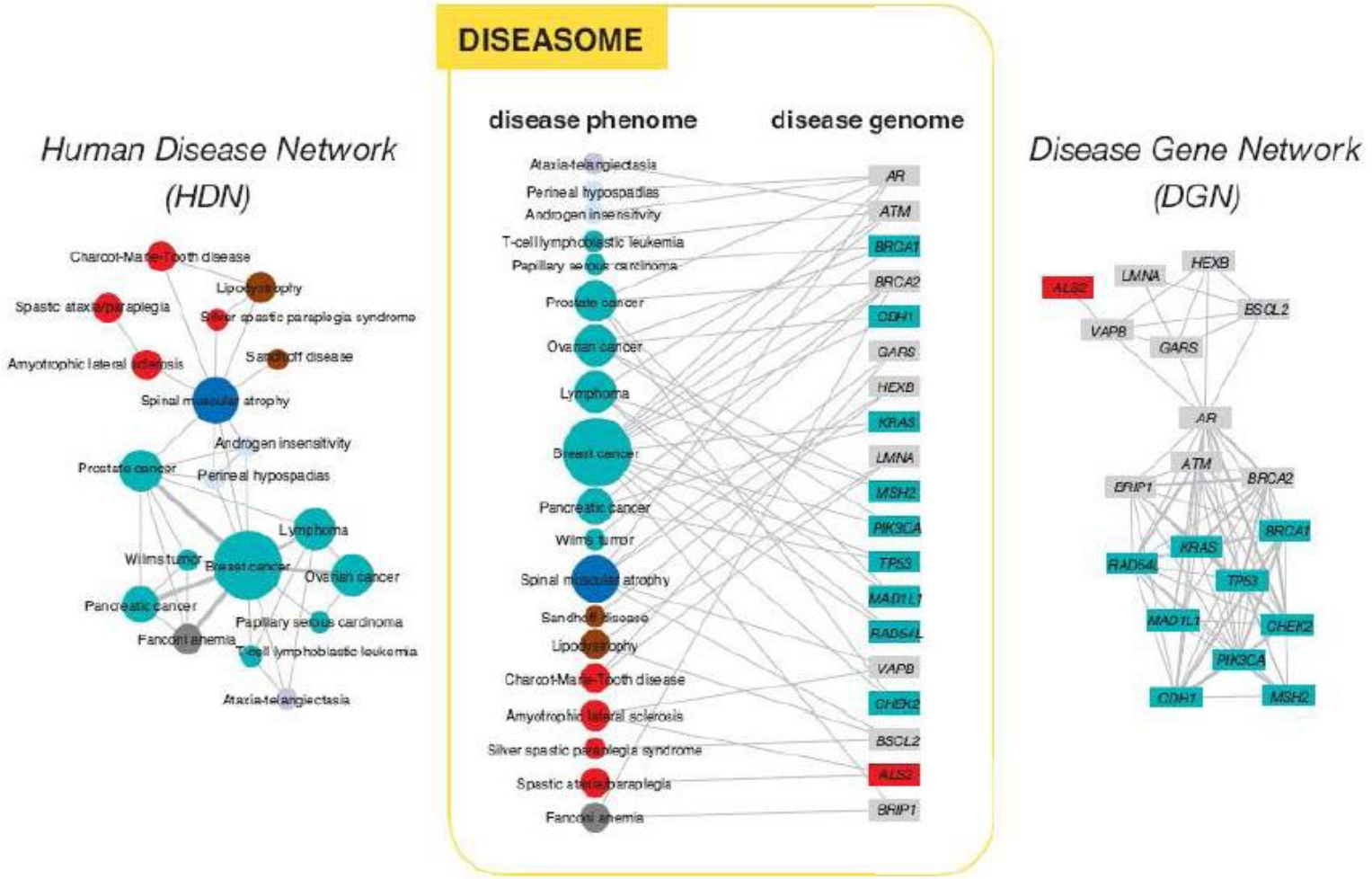
Mol. Biosyst. 2013, 9(7):1576-83



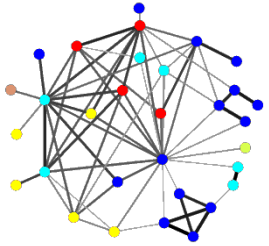
Other Biological Networks

- Disease – “disease gene” association networks
  - Link diseases that are caused by the same gene
  - Link genes if they cause the same disease
- Drug – “drug target” association networks
  - Link drugs if they target the same gene (protein)
  - Link genes (protiens) if they are targeted by the same drug

Other Biological Networks

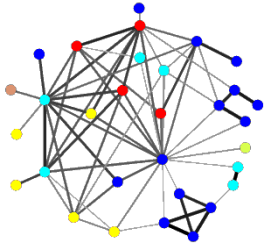


Goh K et al., The human disease network. *PNAS* 2007, 104(21):8685–8690.



Research debates…

- Check Scale-free degree distribution
- Clustering Coefficients
- Centrality Analysis
  - Essential Genes
- Do high-degree nodes interact with high-degree nodes?
- Structural robustness and attack tolerance:
  - Robust vs. Fragile
- Motif discovery
- Clustering Analysis



Summary

- Graph Theory
  - Global and local properties
    - Degree, clustering coefficients, motifs, …
- Graph representation of Biological Networks
  - Bipartite, Met-centric, Enz-centric

Questions?