Network theory

Part I

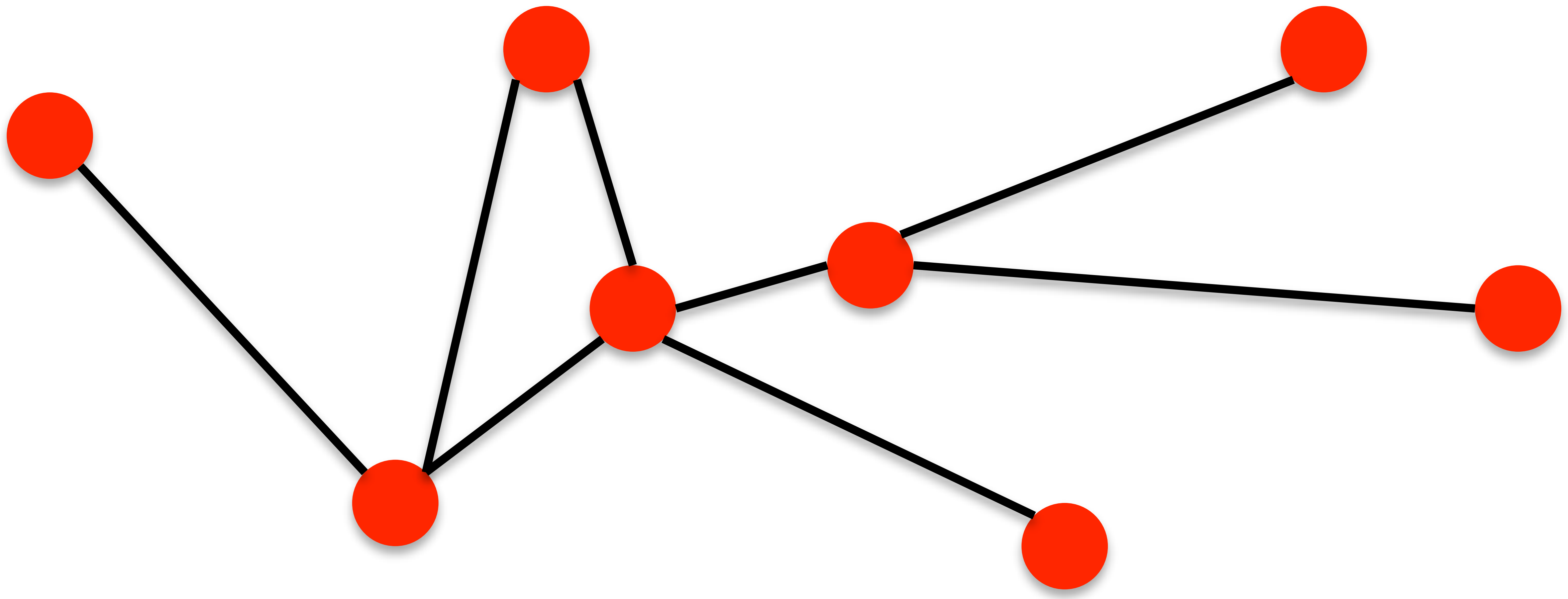
CENTAI

ISI Foundation

Complexity in Social Systems
AA 2023/2024

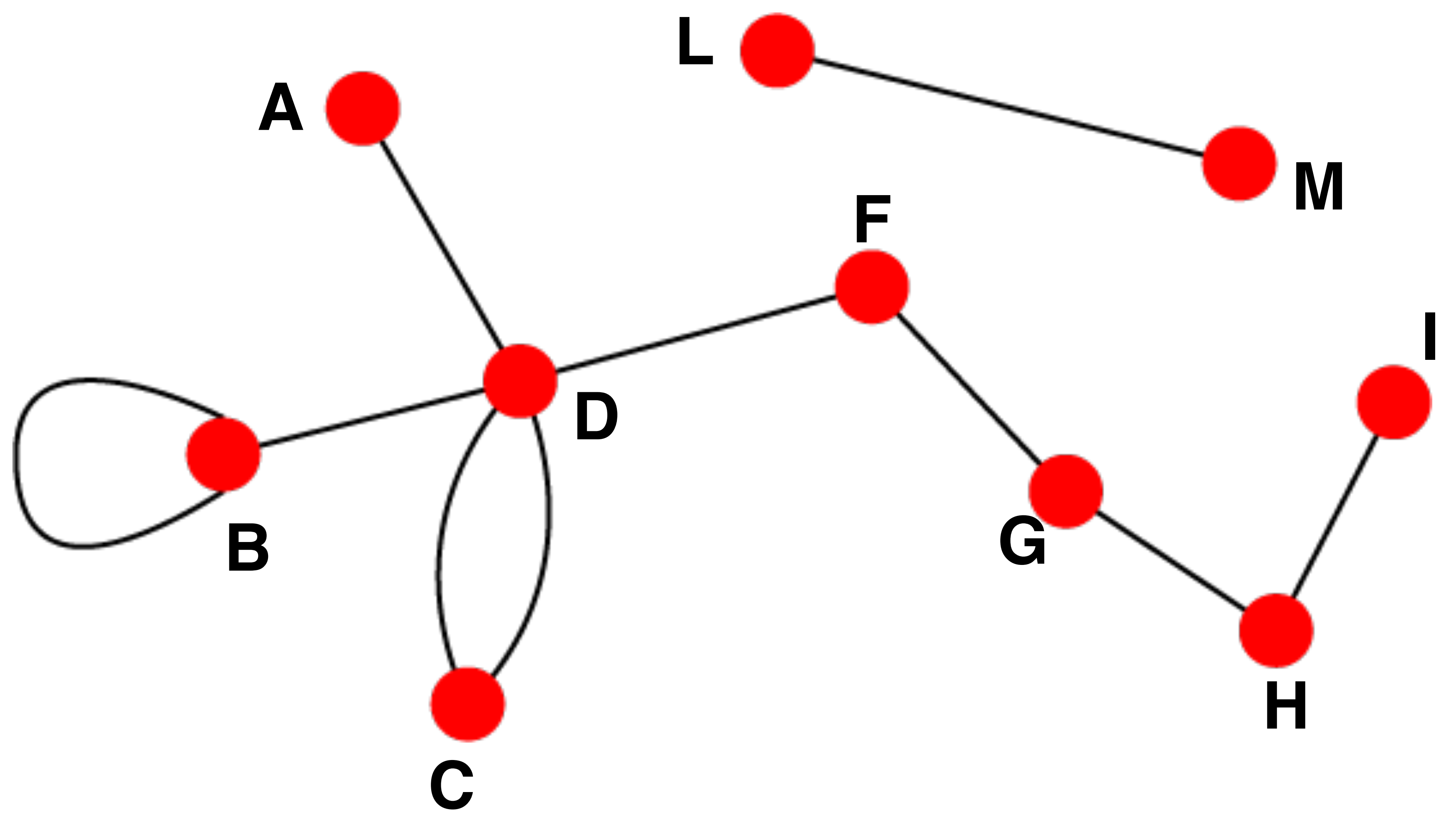
Maxime Lucas
Lorenzo Dall'Amico

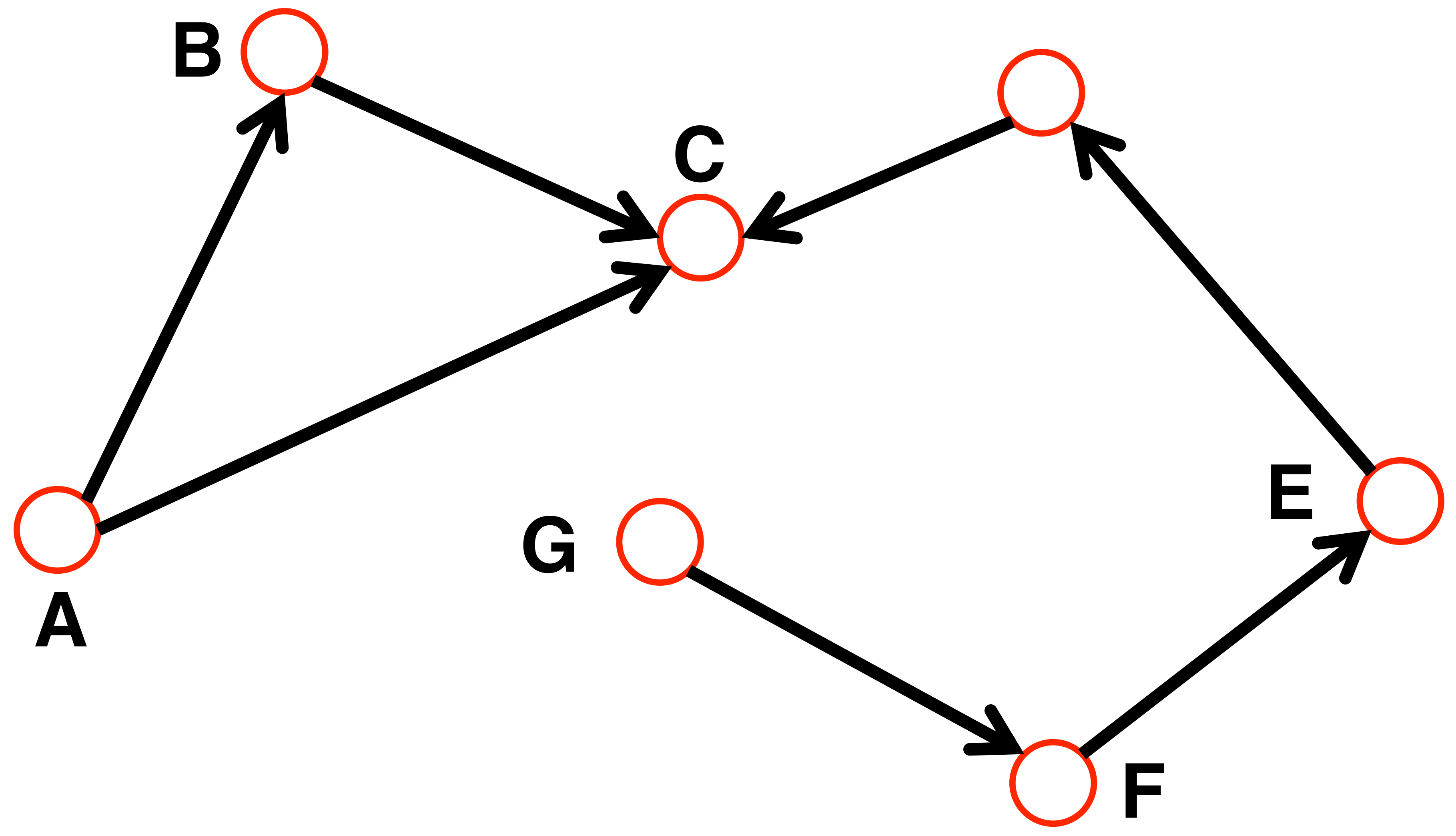
components



- **components**: nodes, vertices — N
- **interactions**: links, edges — L
- **system**: network, graph — (N,L)

undirected vs directed





co-authorship
actor networks
co-occurrence

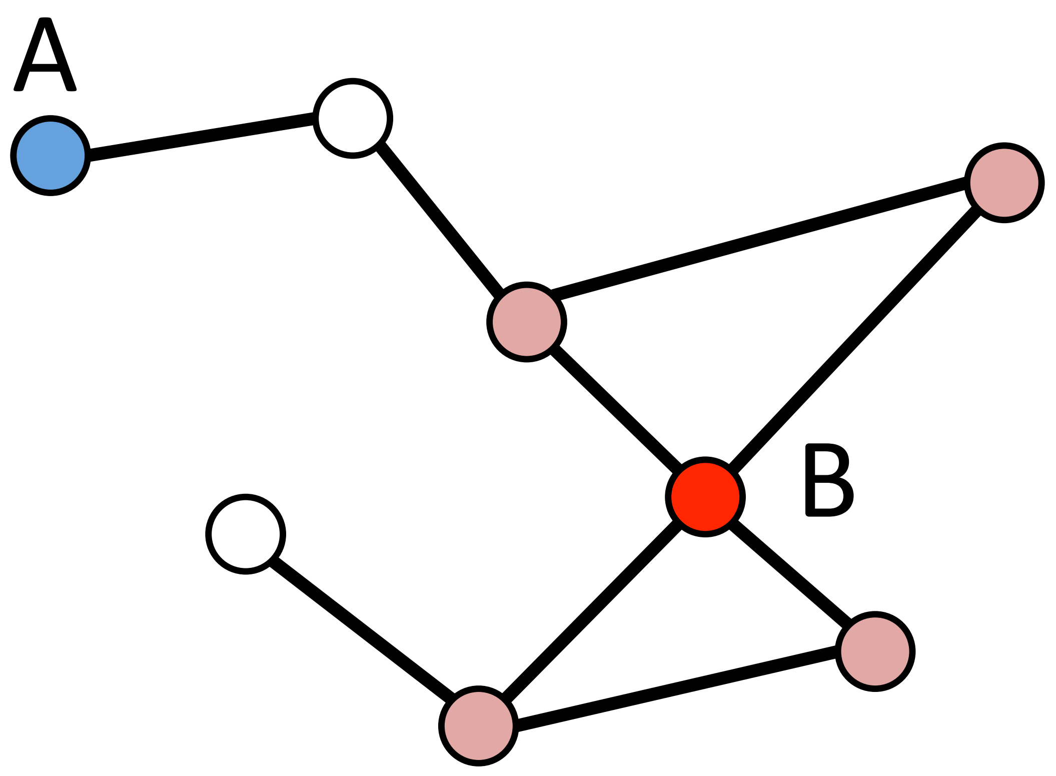
phone calls
hyperlinks
scientific citations

reference networks

Network	Nodes	Links	Directed / Undirected	N	L	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

degree and degree distribution

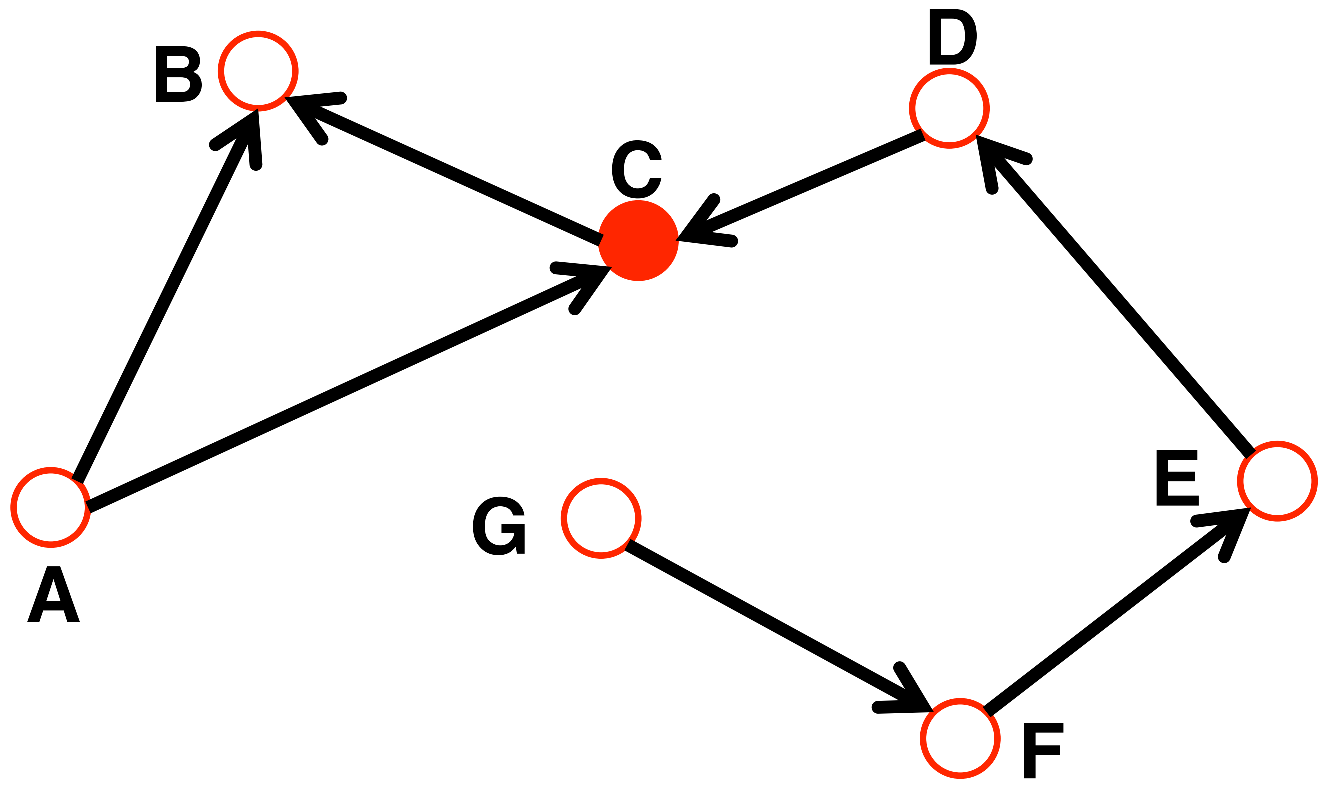
Undirected



$k_A = 1 \qquad k_B = 4$

***node degree**: the number of links connected to the node*

Directed



$k_C^{in} = 2 \quad k_C^{out} = 1 \quad k_C = 3$

***Source:** degree in = 0*

***Sink:** degree out = 0*

BRIEF STATISTICS REVIEW

Four key quantities characterize a sample of N values $x_1, \ldots, x_N$:

Average (mean):

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The n^{th} moment:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

Standard deviation:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

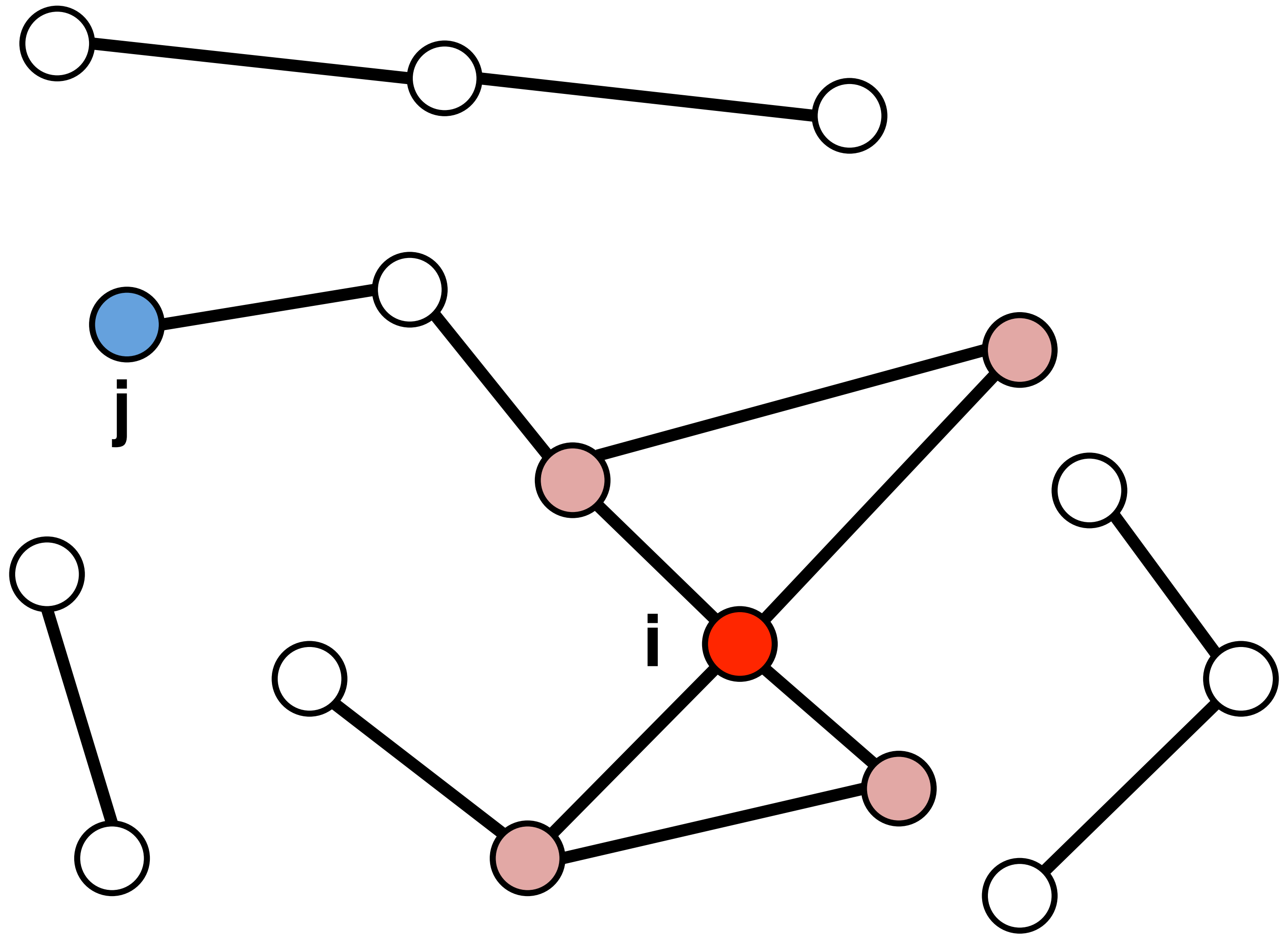
Distribution of x:

$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

where p_x follows

$$\sum_i p_x = 1 \left(\int p_x \, dx = 1\right)$$

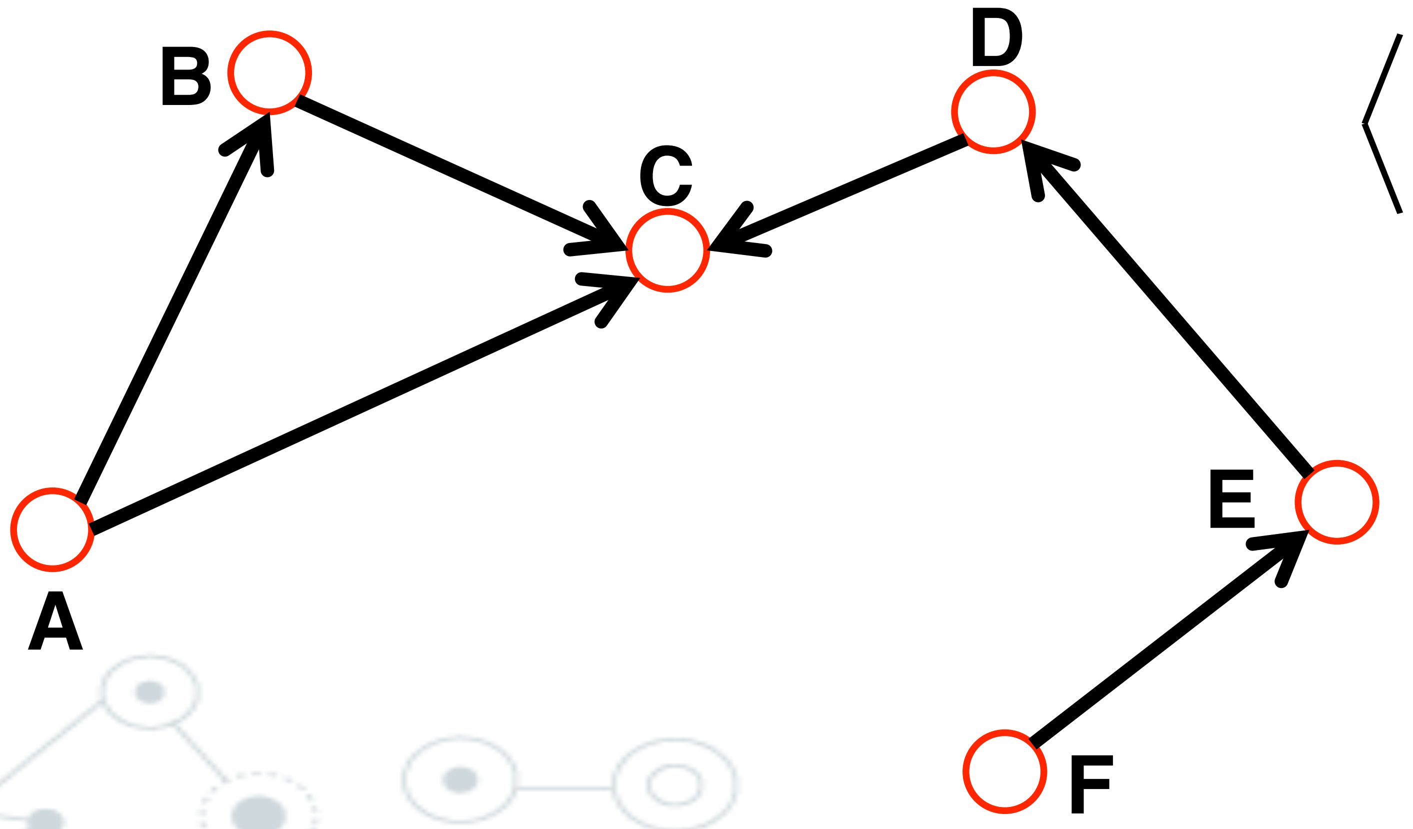
Undirected



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – the number of nodes in the graph

Directed



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

$$\langle k \rangle \equiv \frac{L}{N}$$

degree distribution

$$P(k) = \frac{N_k}{N}$$

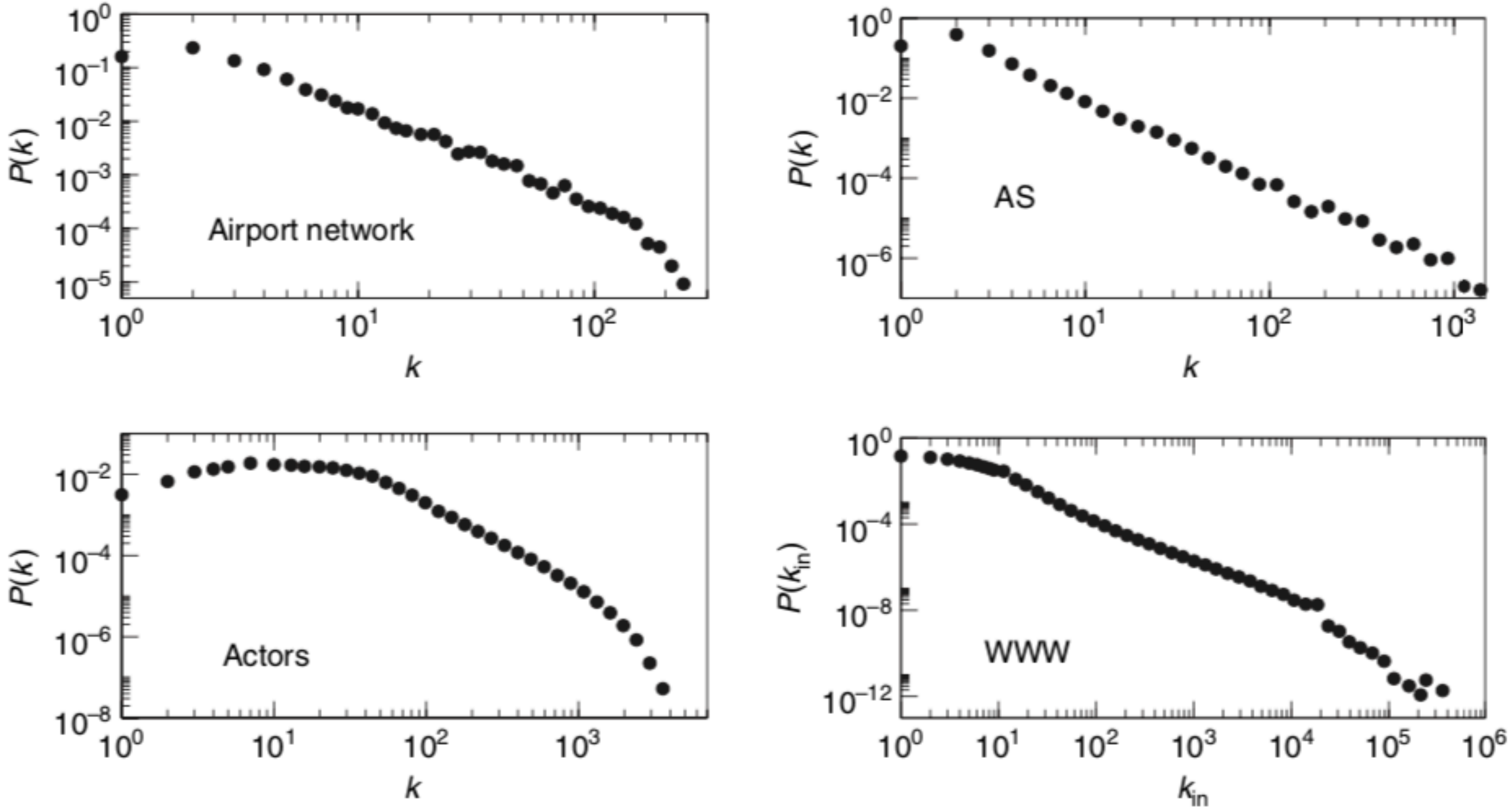
probability that a random chosen node has degree k

$$\langle k \rangle = \sum_k k P(k)$$

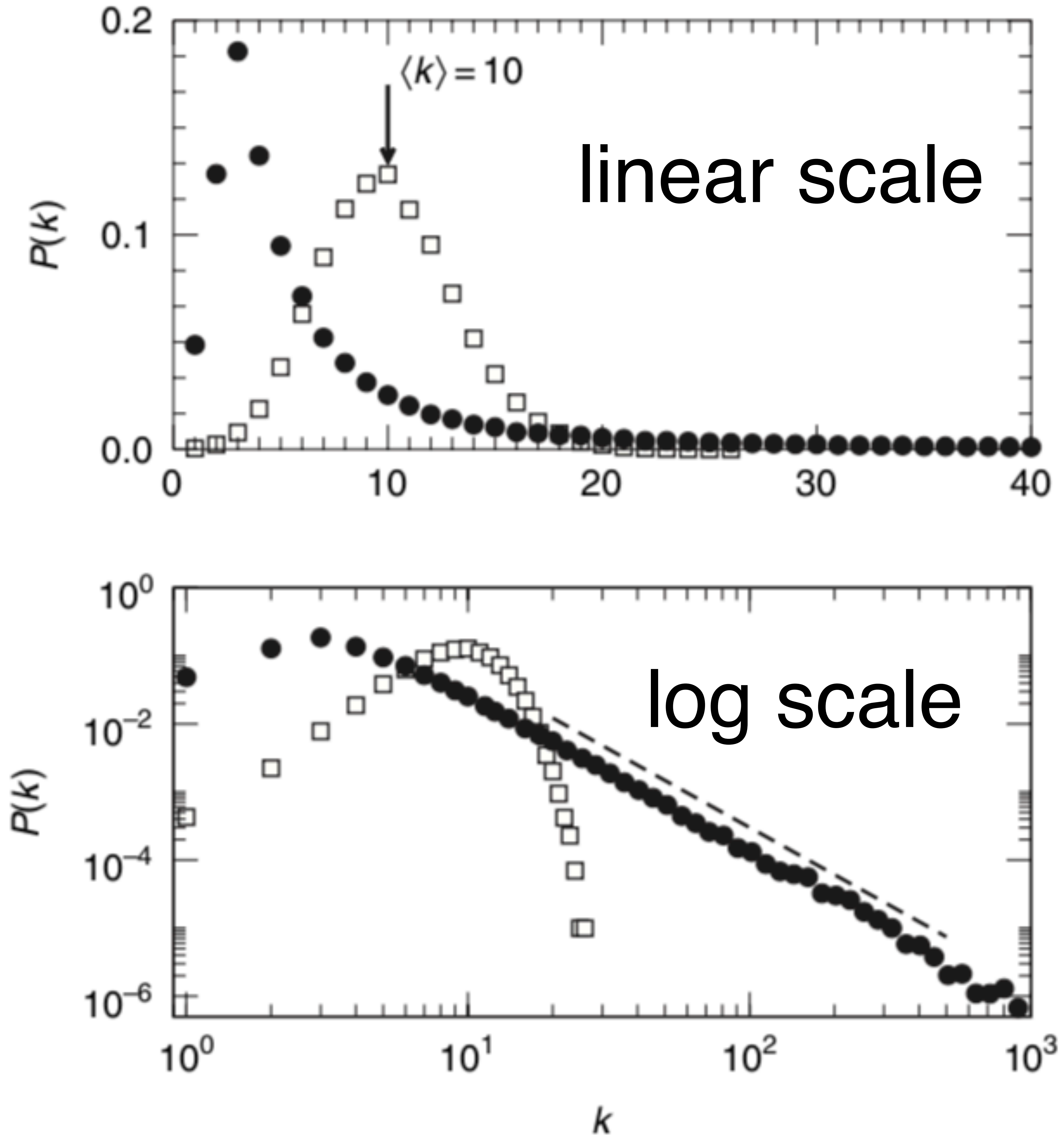
$$\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$$

$$\langle k^2 \rangle = \sum_k k^2 P(k)$$

real world networks



real world networks



Poisson
(homogeneous)

vs

power-law
(heterogeneous)

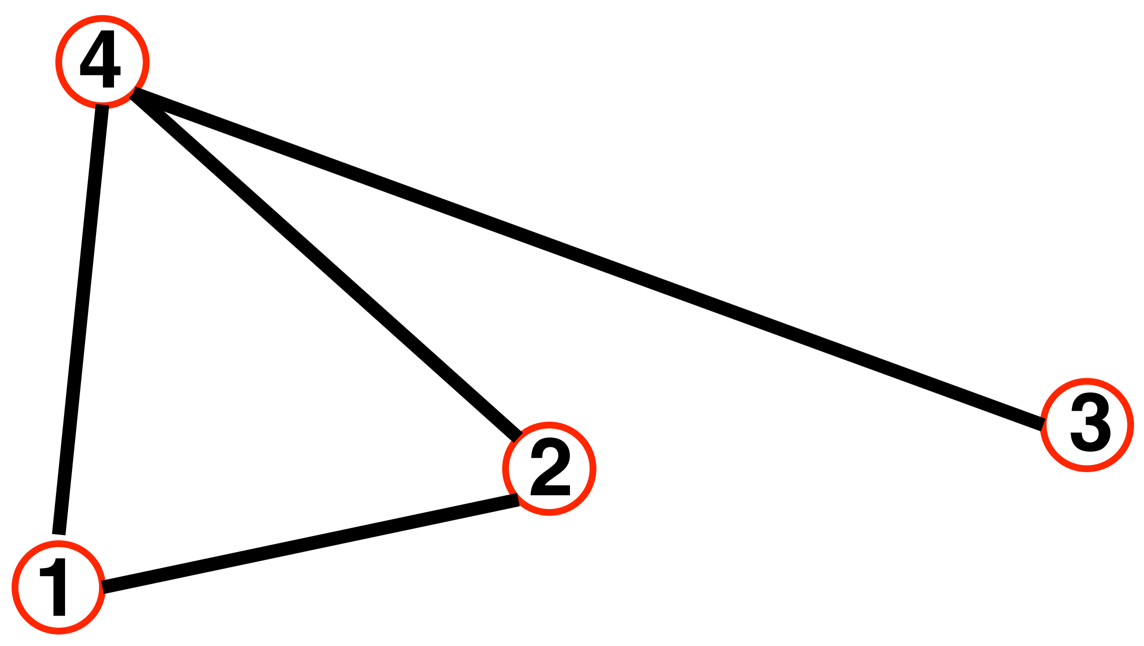
Broad degree distributions

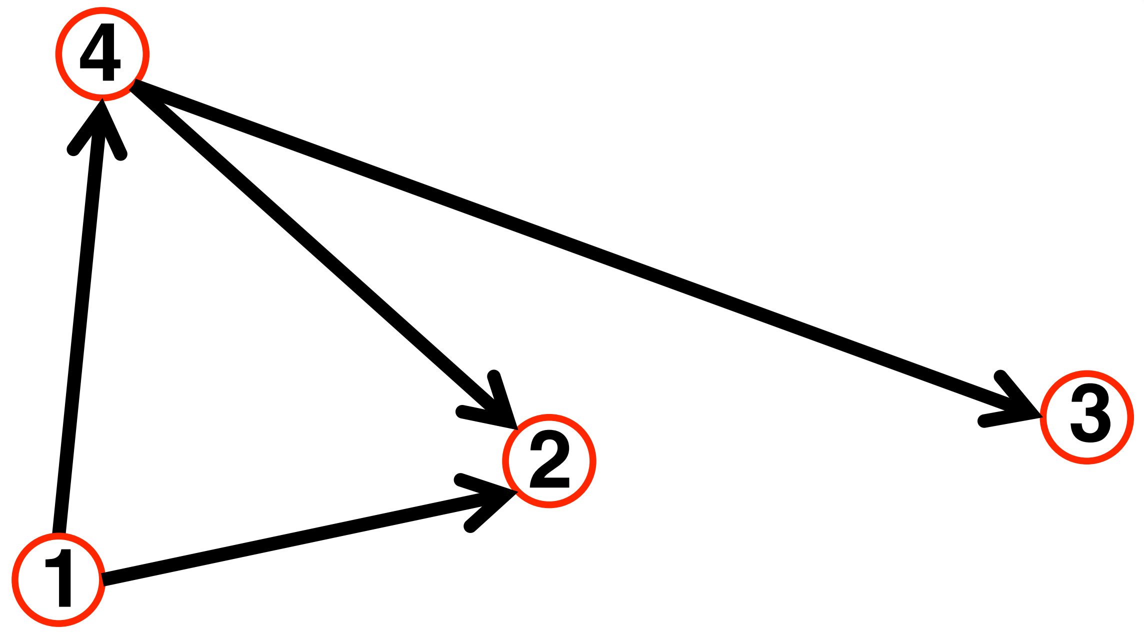
Power-law tails

$$P(k) \sim k^{-\gamma}, 2 < \gamma < 3$$

No characteristic scale

adjacency matrix





A_{ij}=1 if there is a link between node *i* and *j*

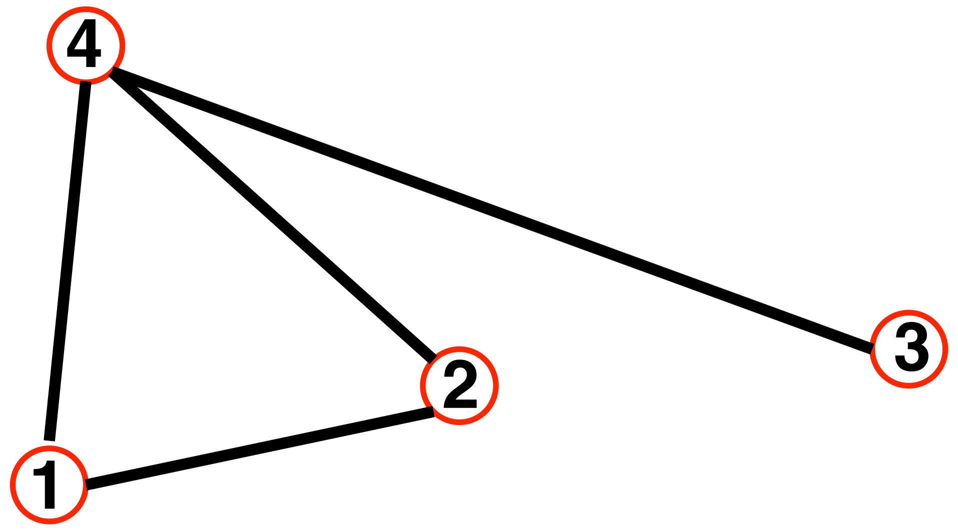
A_{ij}=0 if nodes *i* and *j* are not connected to each other.

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Note that for a directed graph (right) the matrix is not symmetric.

adjacency matrix

Undirected



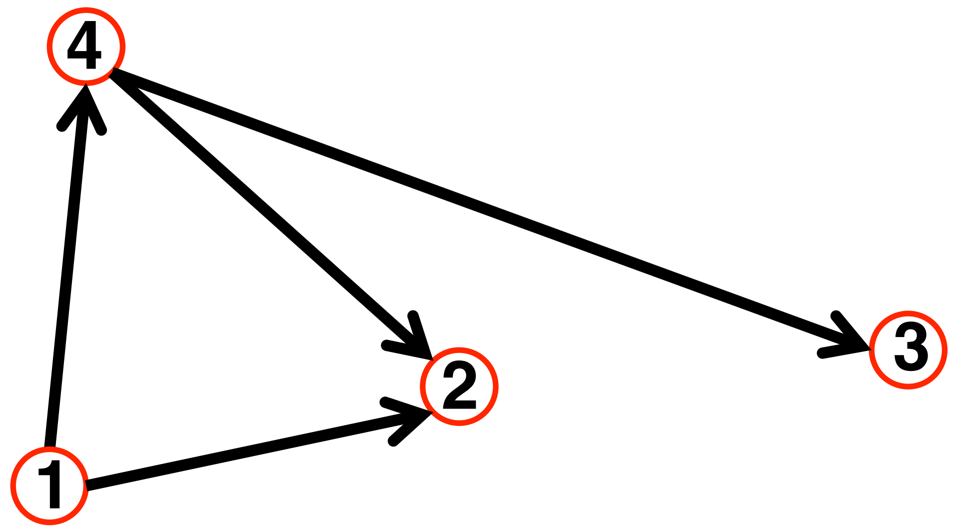
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

$$k_j = \sum_{i=1}^{N} A_{ij}$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij}^{N} A_{ij}$$

Directed



$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

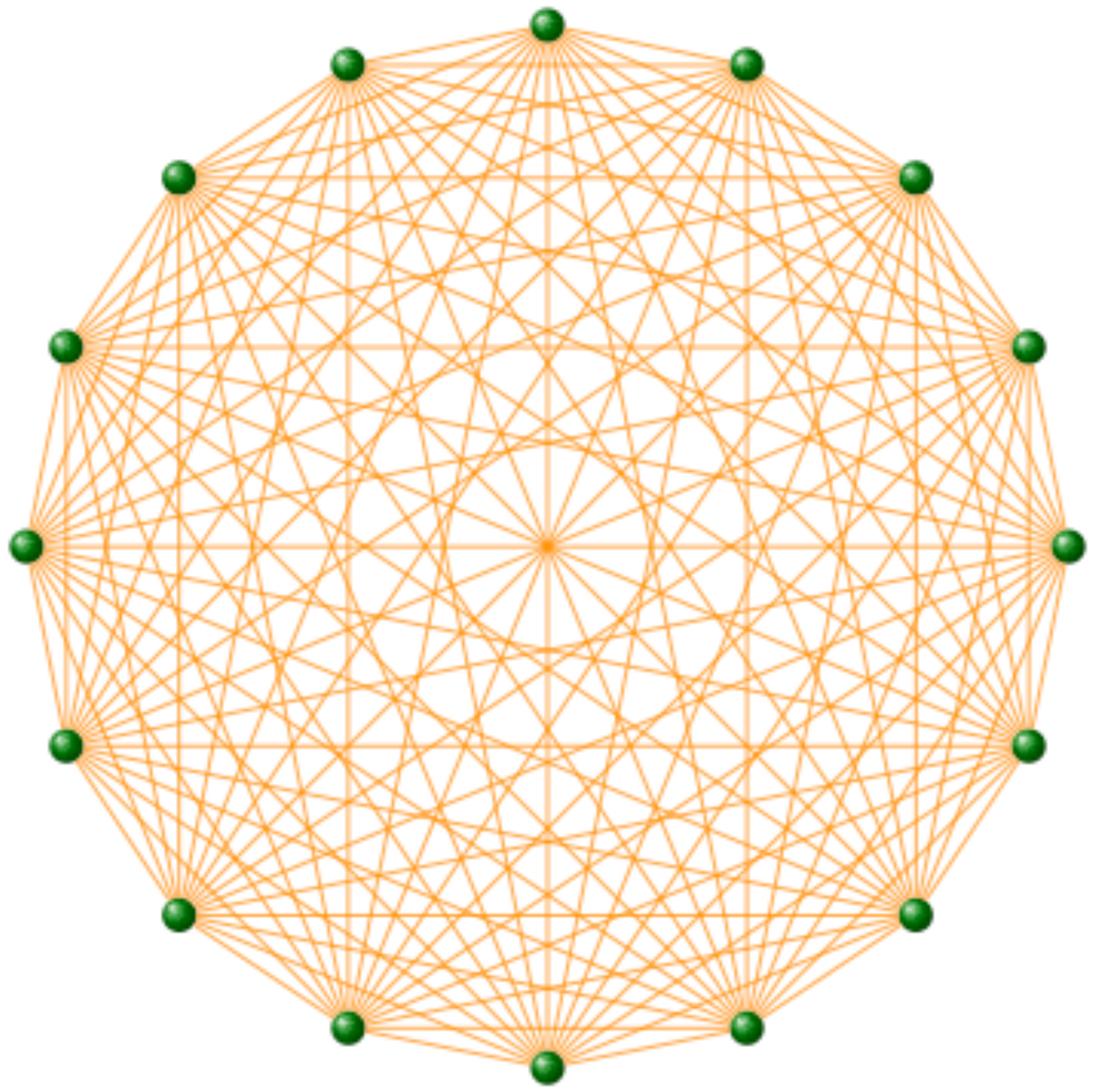
$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

Real networks are sparse!

The maximum number of links a network of N nodes can have is: $L_{max} = \binom{N}{2} = \frac{N(N-1)}{2}$



A graph with degree $L=L_{max}$ is called a complete graph, and its average degree is **<k>=N-1**

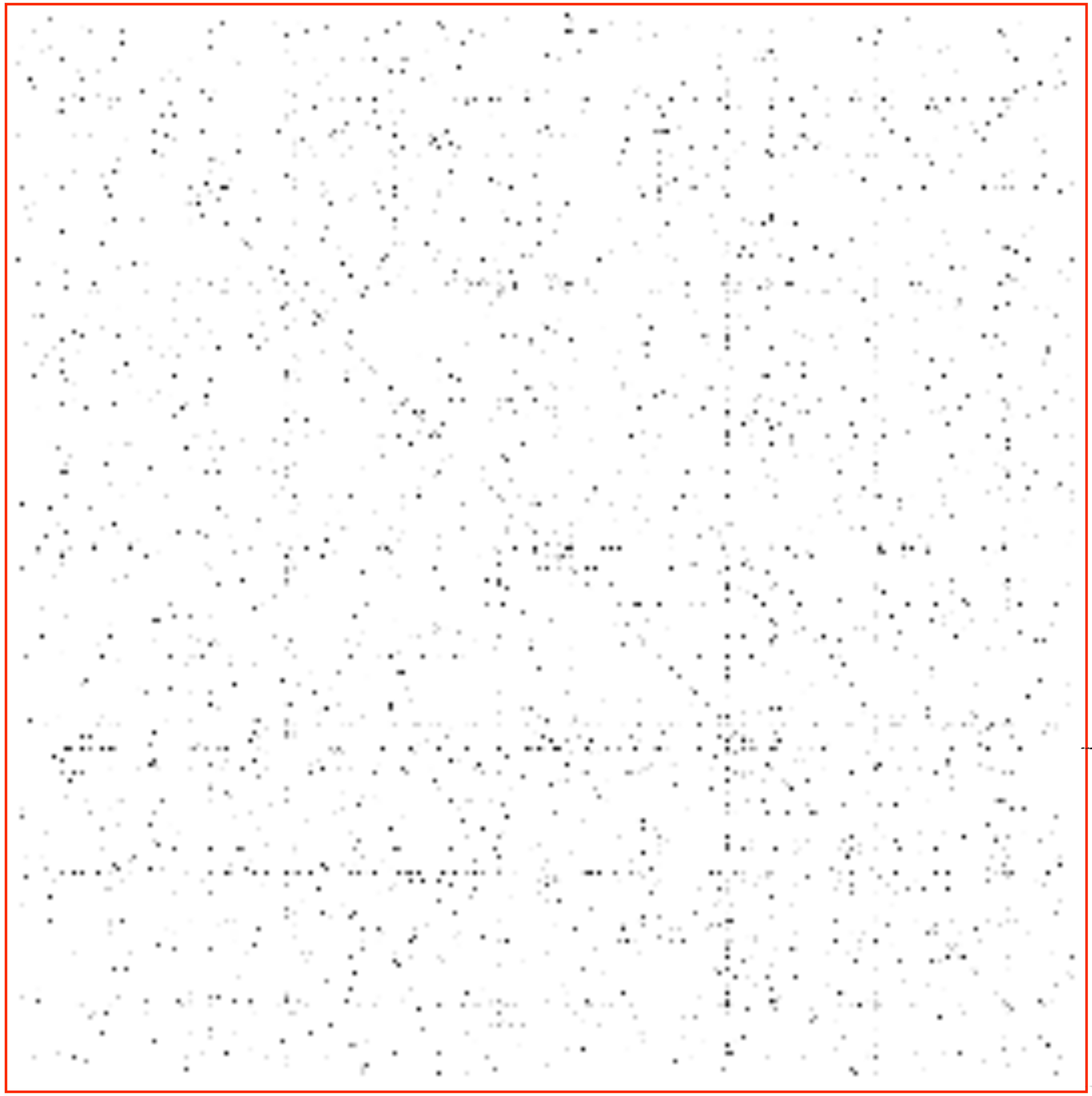
Most networks observed in real systems are **sparse**

$$L \ll L_{max} \qquad \langle k \rangle \ll N - 1$$

WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	<k>=4.51
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^{7}$	<k>=2.39
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	<k>=3.9
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	<k>=28.78

(Source: Albert, Barabasi, RMP2002)

The adjacency matrix of the yeast protein-protein interaction network, consisting of 2,018 nodes, each representing a yeast protein.



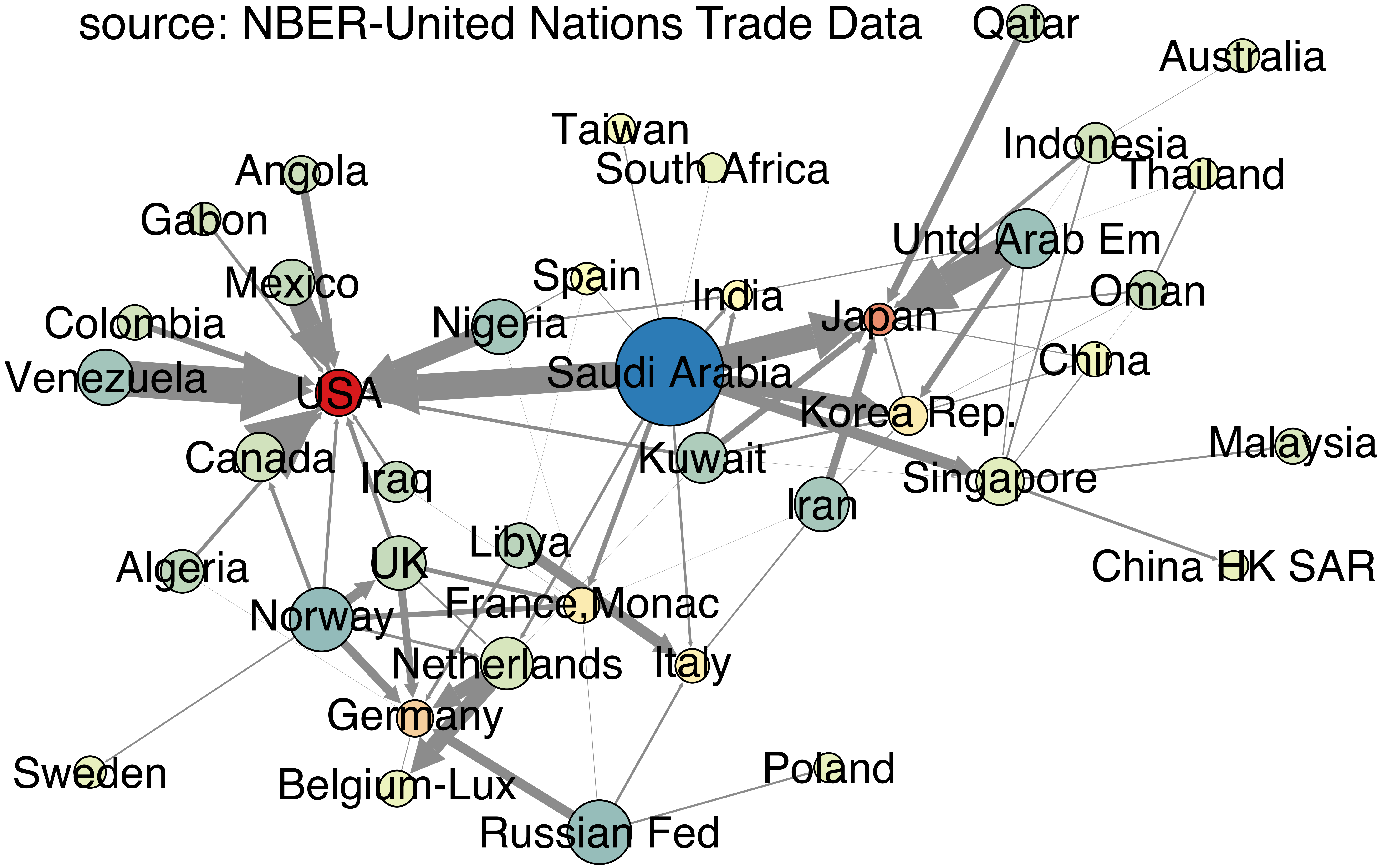
The adjacency matrix is not efficient to store the network

weighted networks

$$A_{ij} = w_{ij}$$

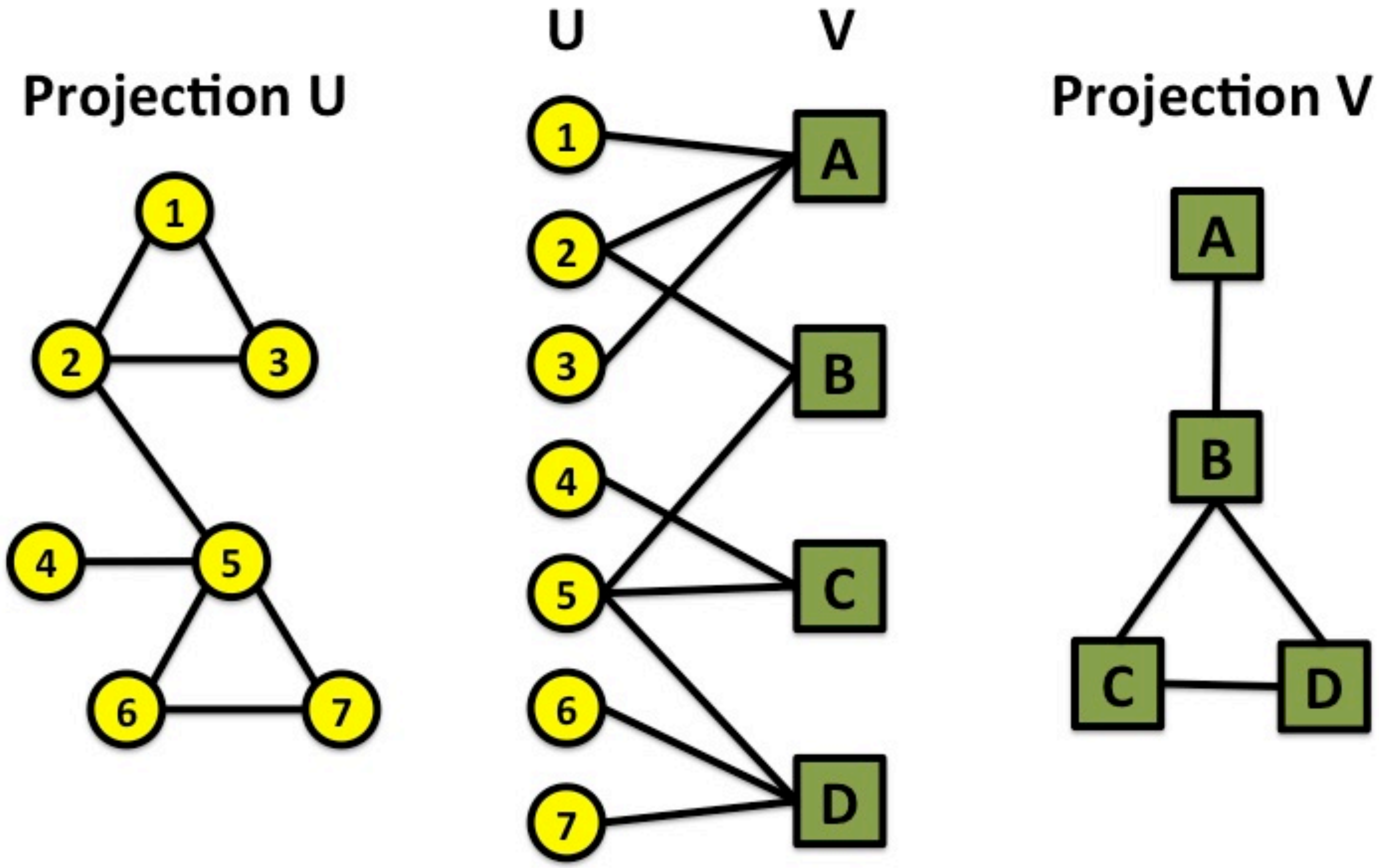
$$s_i = \sum_j w_{ij}$$

trade in petroleum and petroleum products, 1998,
source: NBER-United Nations Trade Data



bipartite networks

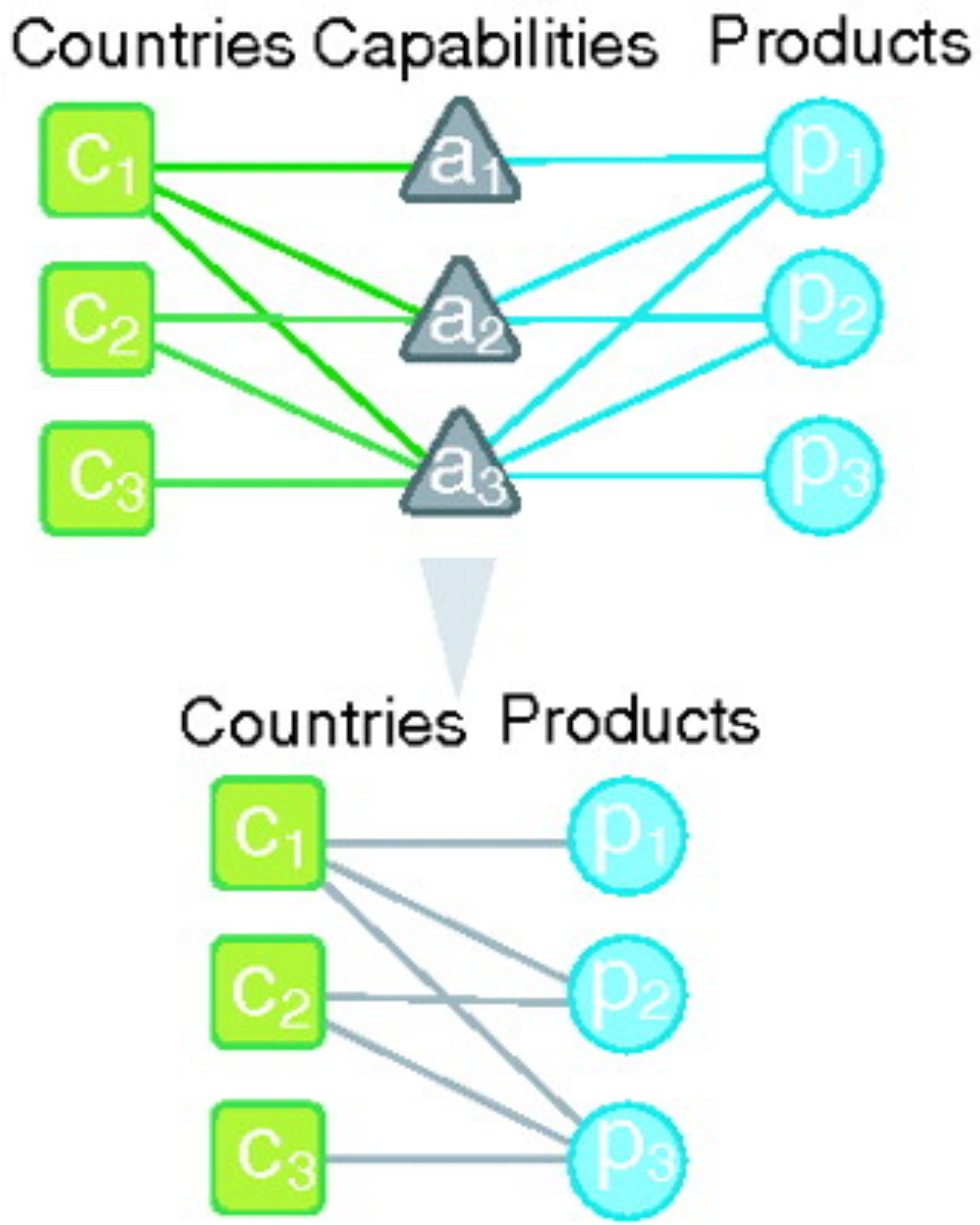
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.

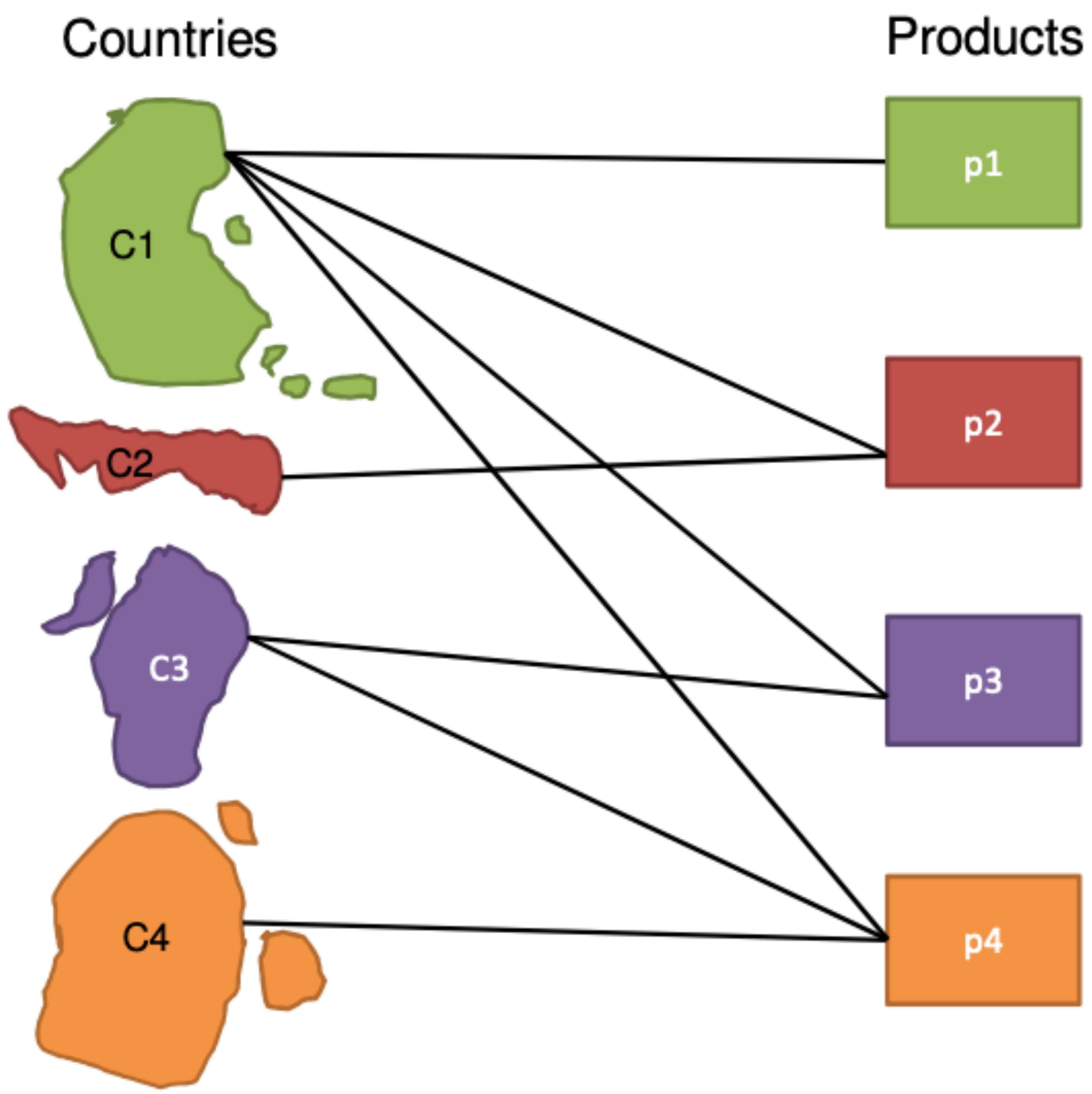


Examples

- actor network
- collaboration network
- host-pathogen networks

bipartite networks



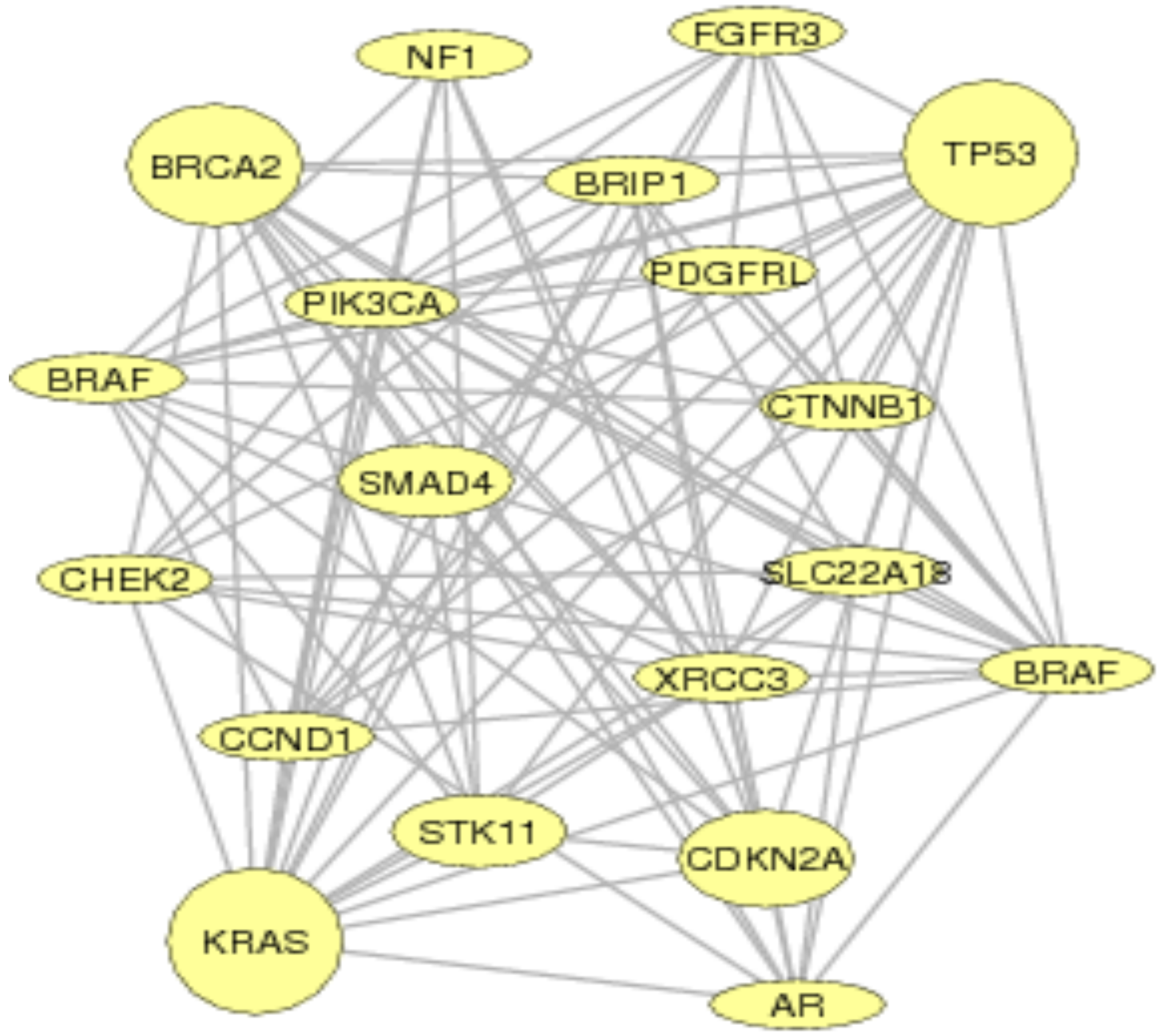


The Atlas of Economic Complexity

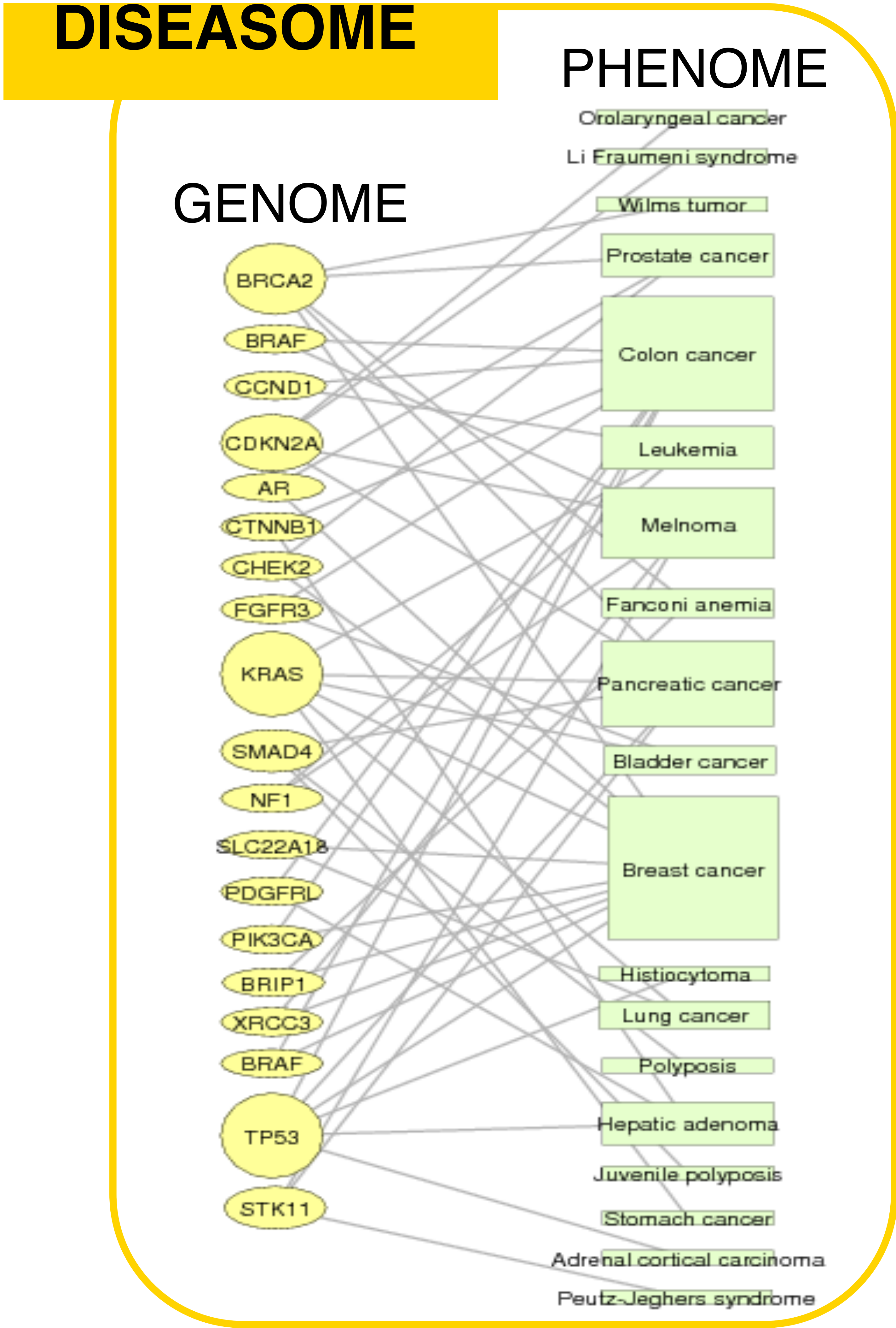
C. Hidalgo

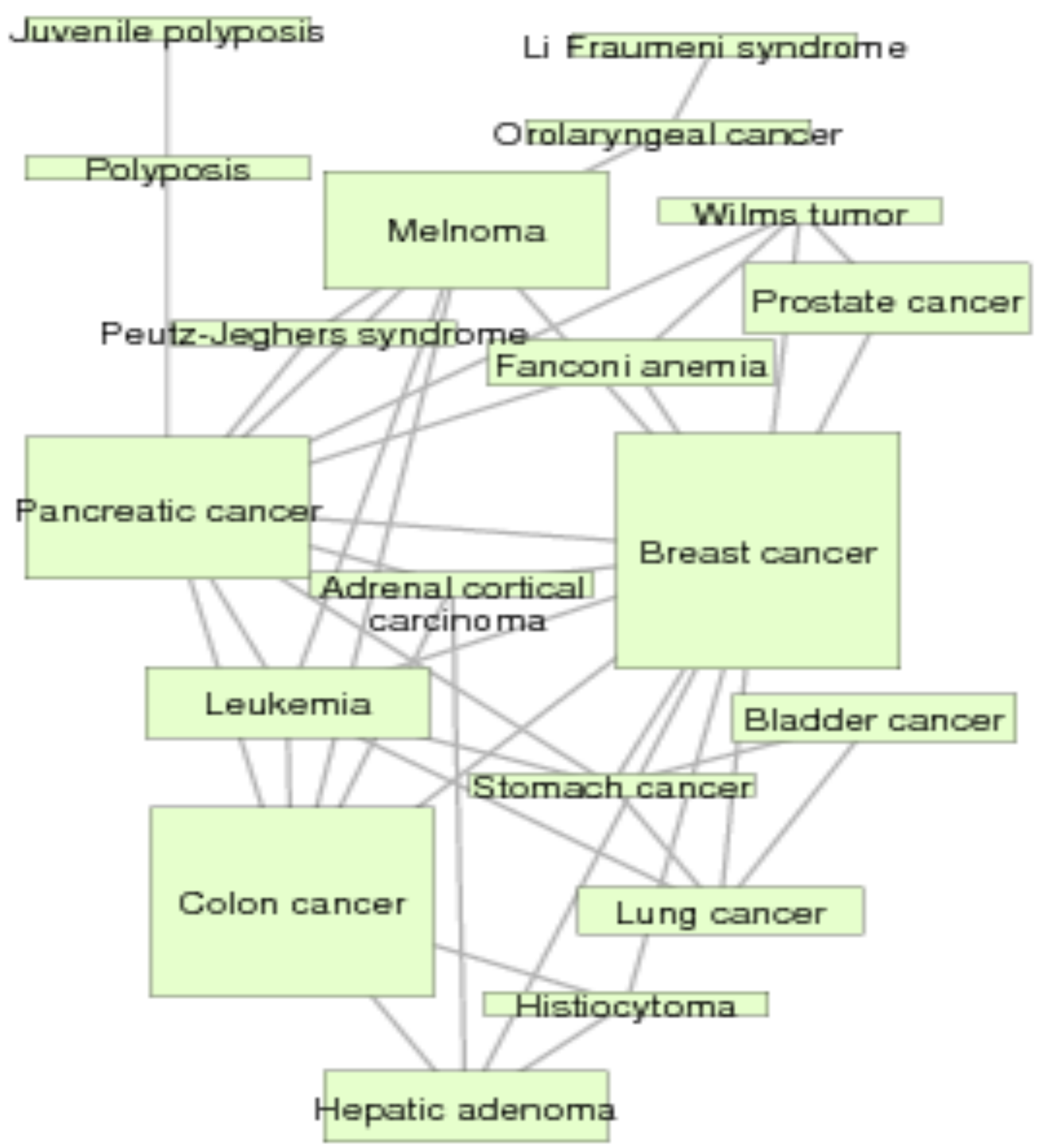
http://atlas.cid.harvard.edu/

bipartite networks



Gene network





Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

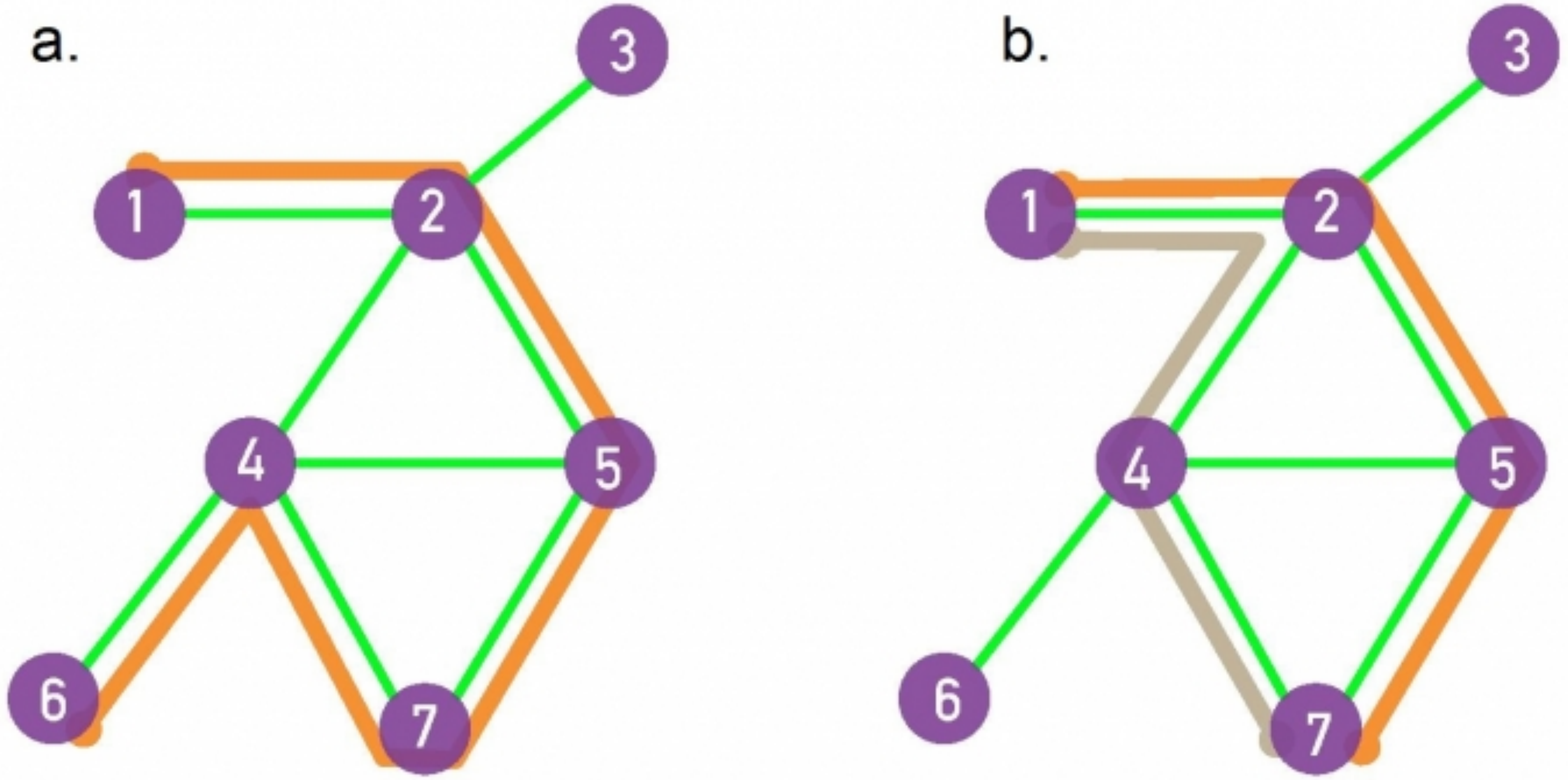
paths

A *path is* a sequence of nodes in which each node is adjacent to the next one

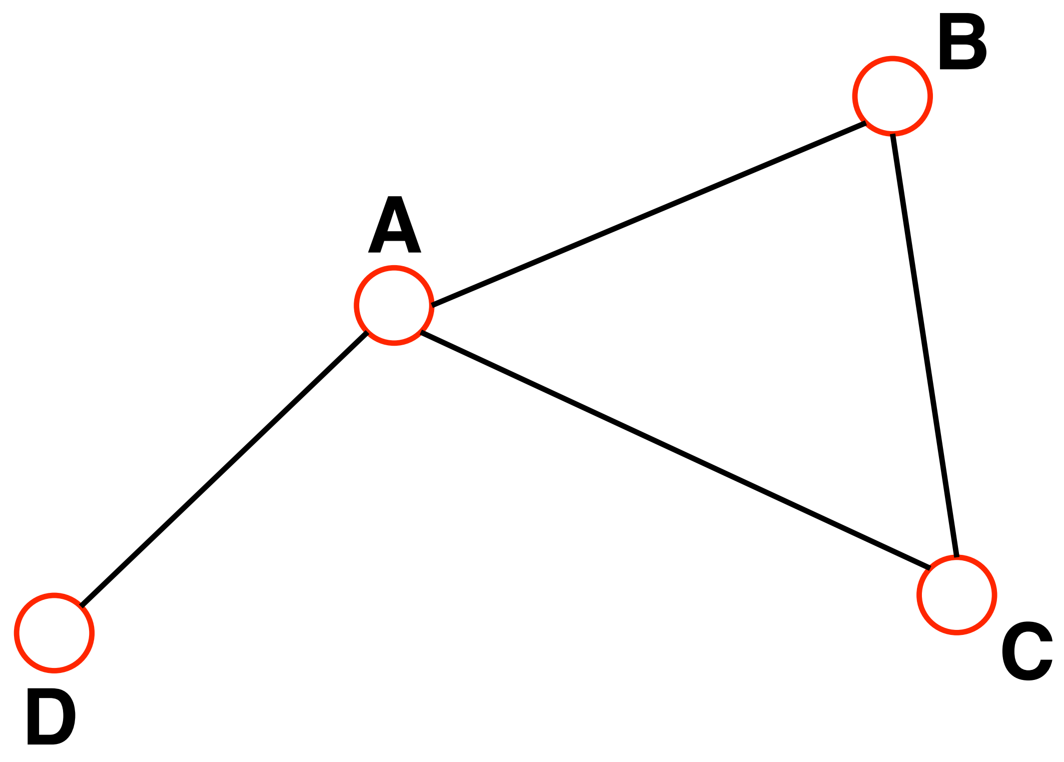
$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

(a) path of length 5
(b) two paths of equal length

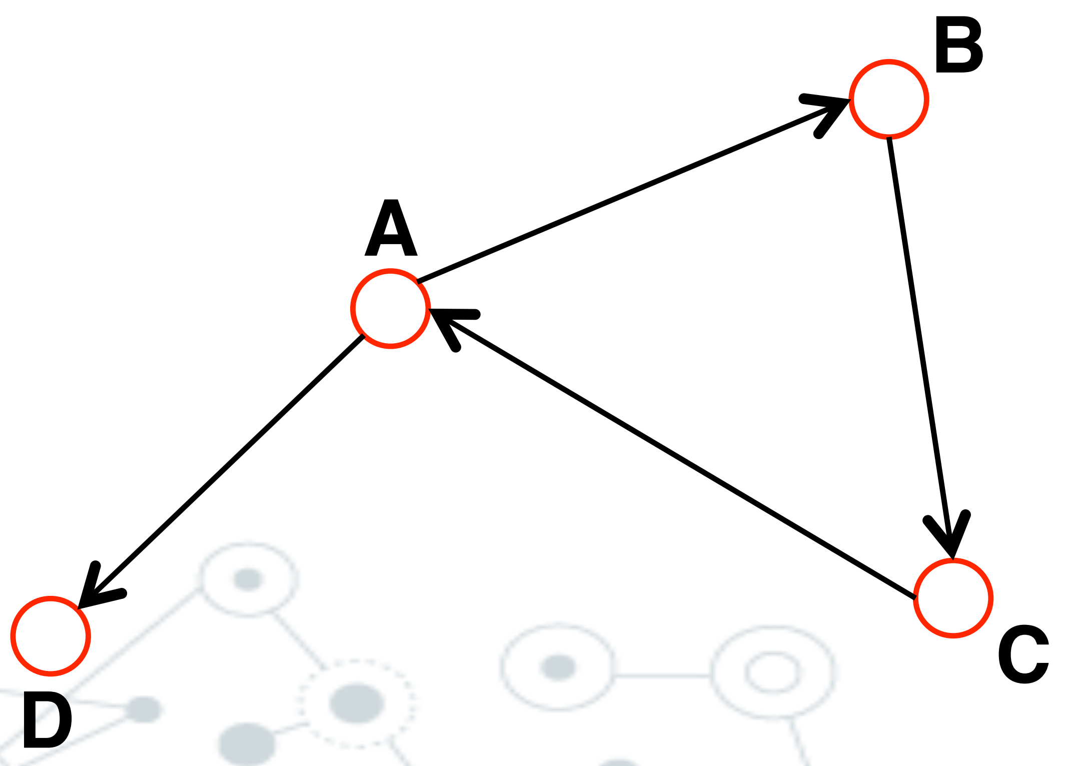


distance



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them. The *diameter of a graph* is the length of the longest geodesic path between any pair of vertices in the network for which a path actually exists.

*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

paths

N_{ij},number of paths between any two nodes *i* and *j*:

***Length n=1*:** If there is a link between *i* and *j*, then A_{ij}=1 and A_{ij}=0 otherwise.

Length n=2: If there is a path of length two between *i* and *j*, then $A_{ik}A_{kj}$=1, and $A_{ik}A_{kj}$=0 otherwise.
The number of paths of length 2:

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik}A_{kj} = [A^2]_{ij}$$

Length n: In general, if there is a path of length *n* between *i* and *j*, then $A_{ik}\ldots A_{lj}$=1 and $A_{ik}\ldots A_{lj}$=0 otherwise.
The number of paths of length *n* between *i* and *j* is*

$$N_{ij}^{(n)} = [A^n]_{ij}$$

paths

Diameter: d_{max} the maximum distance between any pair of nodes in the graph.

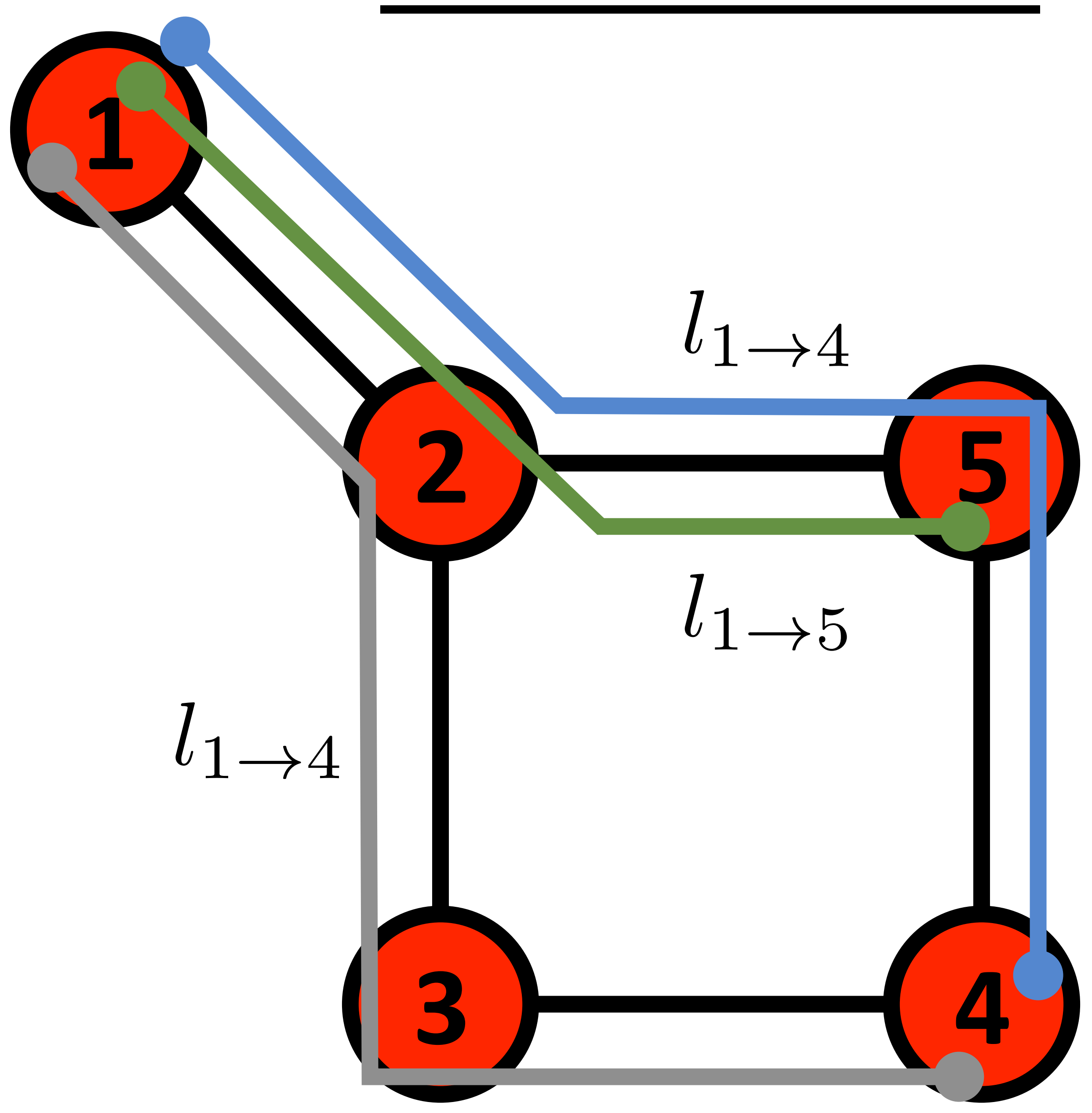
Average path length/distance, <d>, for a connected graph: average distance between all pairs of nodes in the network, where d_{ij} is the distance from node *i* to node j

$$\langle d \rangle = \frac{1}{2L_{max}} \sum_{i,j \neq i} d_{ij}$$

In an *undirected graph* $d_{ij} = d_{ji}$, *so* we only need to count them once: $\langle d \rangle = \frac{1}{L_{max}} \sum_{i,j>i} d_{ij}$

paths: summary

<u>Shortest Path</u>



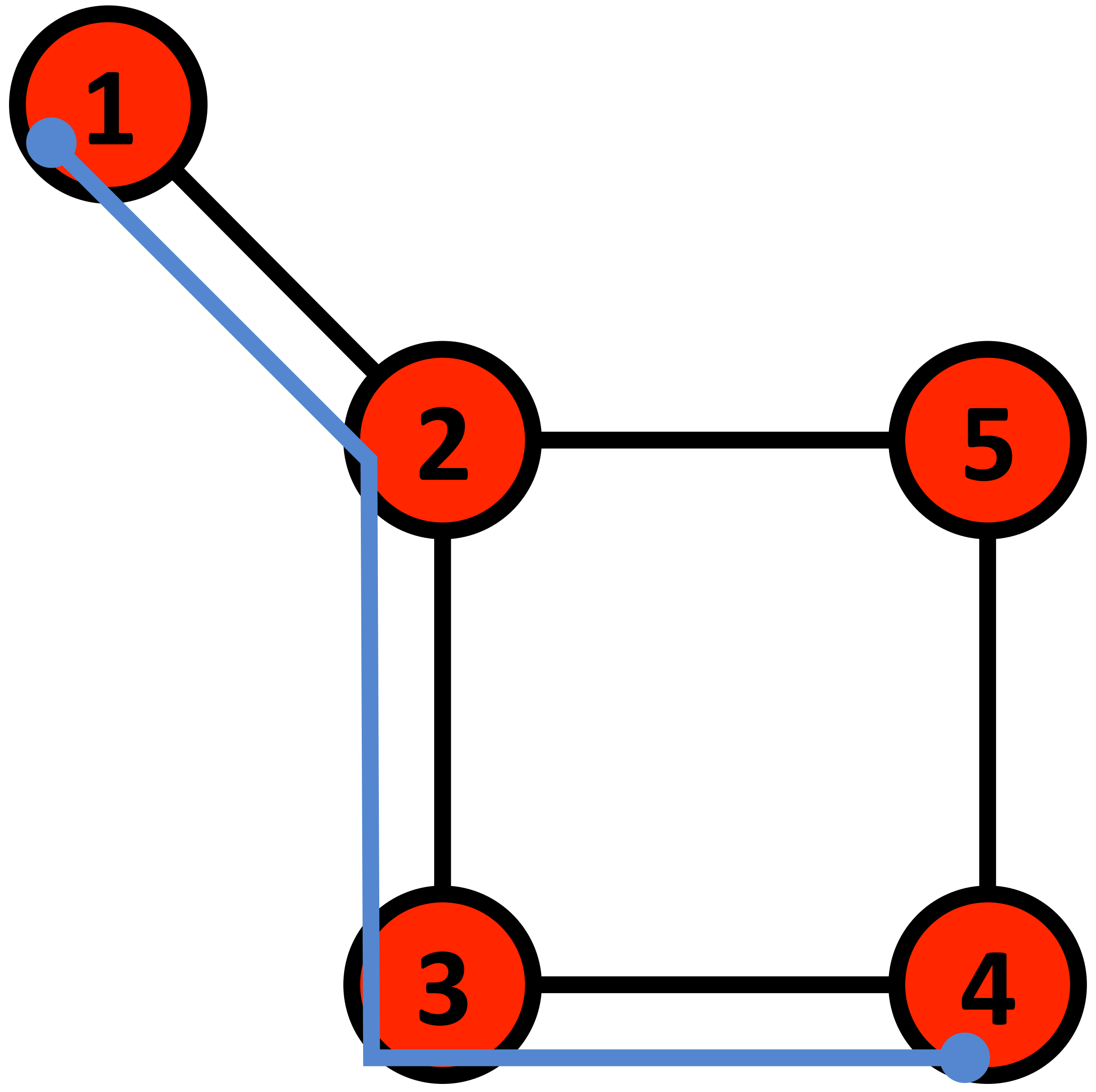
$l_{1\to4} = 3$

$l_{1\to5} = 2$

The path with the shortest length between two nodes (distance).

paths: summary

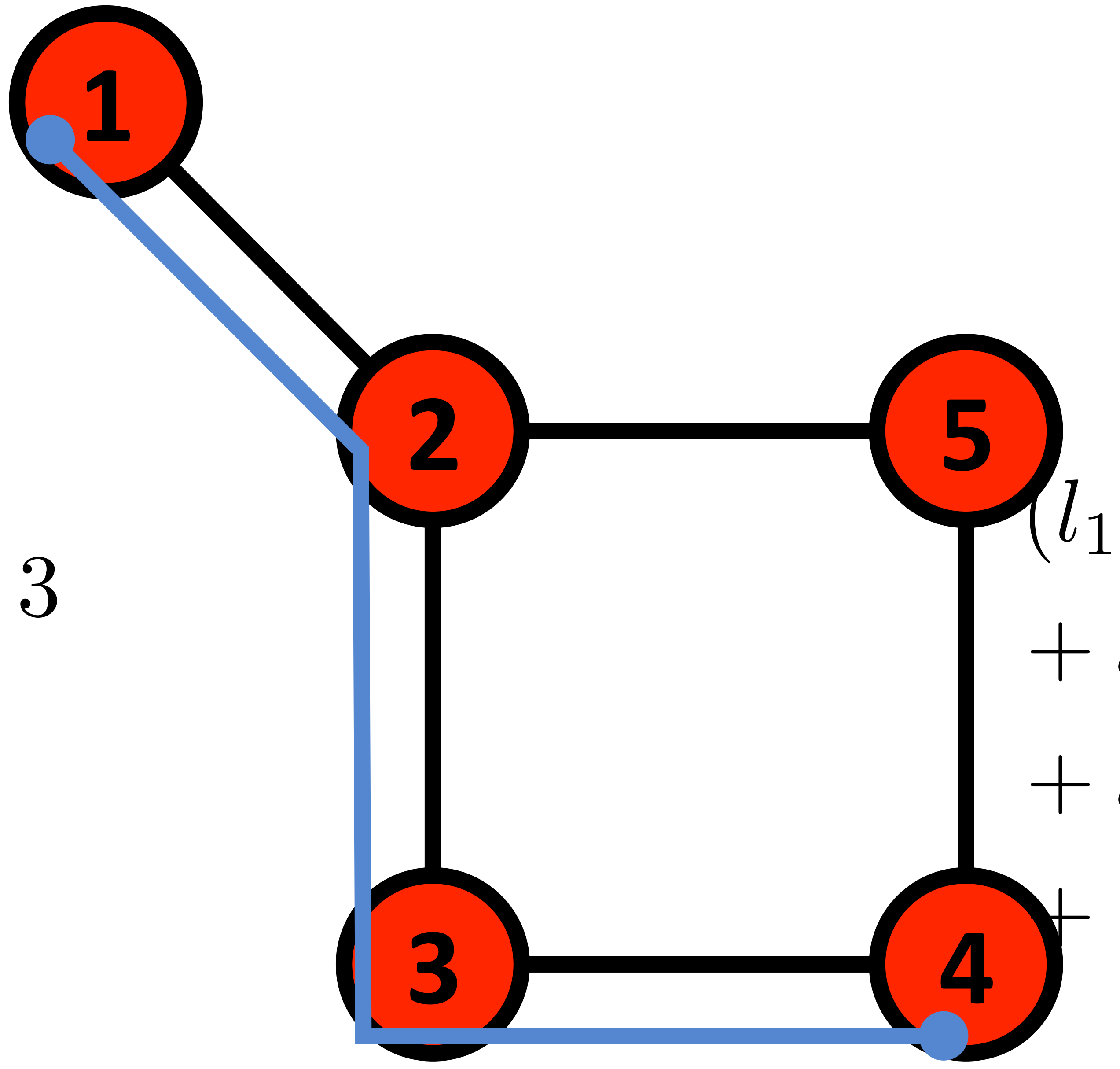
Diameter



$$l_{1\to 4} = 3$$

The longest shortest path in a graph

Average Path Length

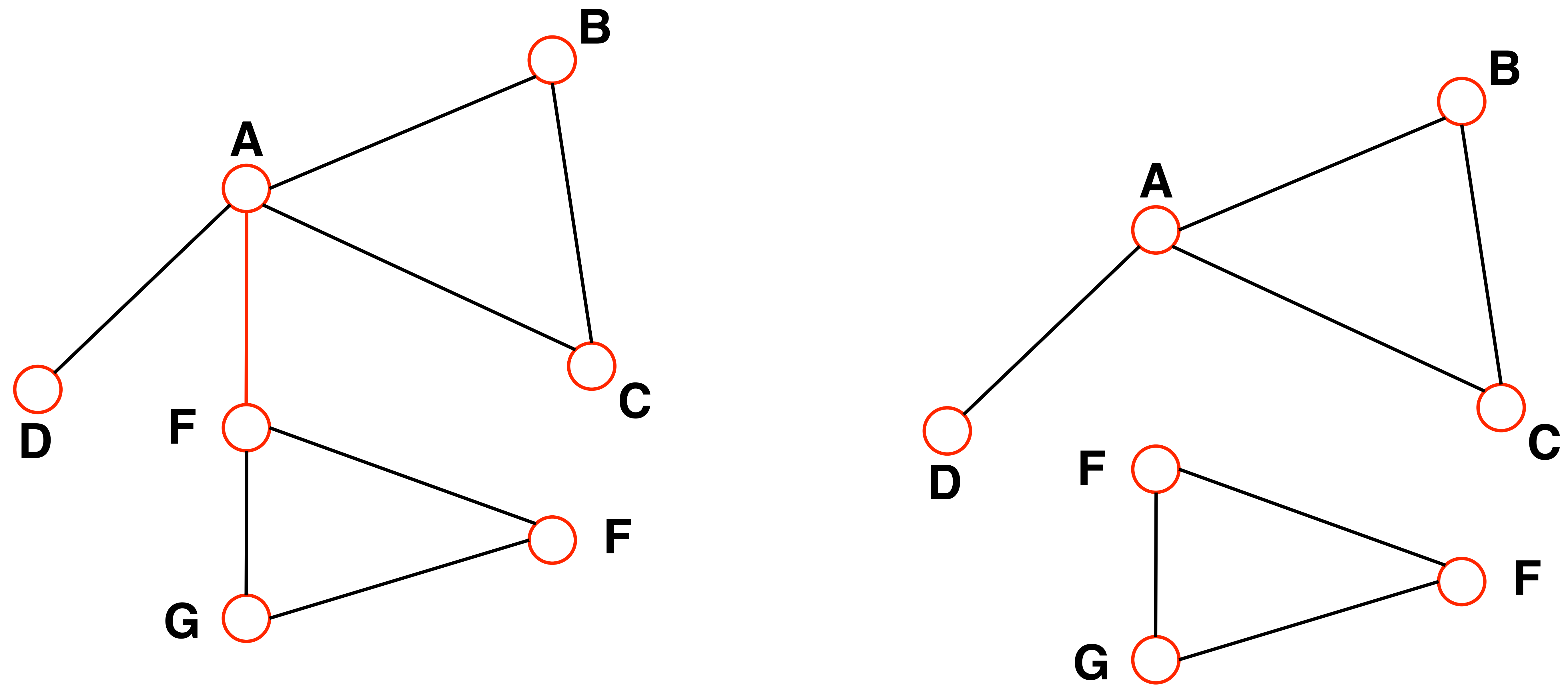


$$(l_{1\to 2} + l_{1\to 3} + l_{1\to 4} + \\ + l_{1\to 5} + l_{2\to 3} + l_{2\to 4} + \\ + l_{2\to 5} + l_{3\to 4} + l_{3\to 5} + \\ + l_{4\to 5})\,/10 = 1.6$$

The average of the shortest paths for all pairs of nodes.

connectivity

Connected (undirected) graph: any two vertices can be joined by a path.
A disconnected graph is made up by two or more connected components.



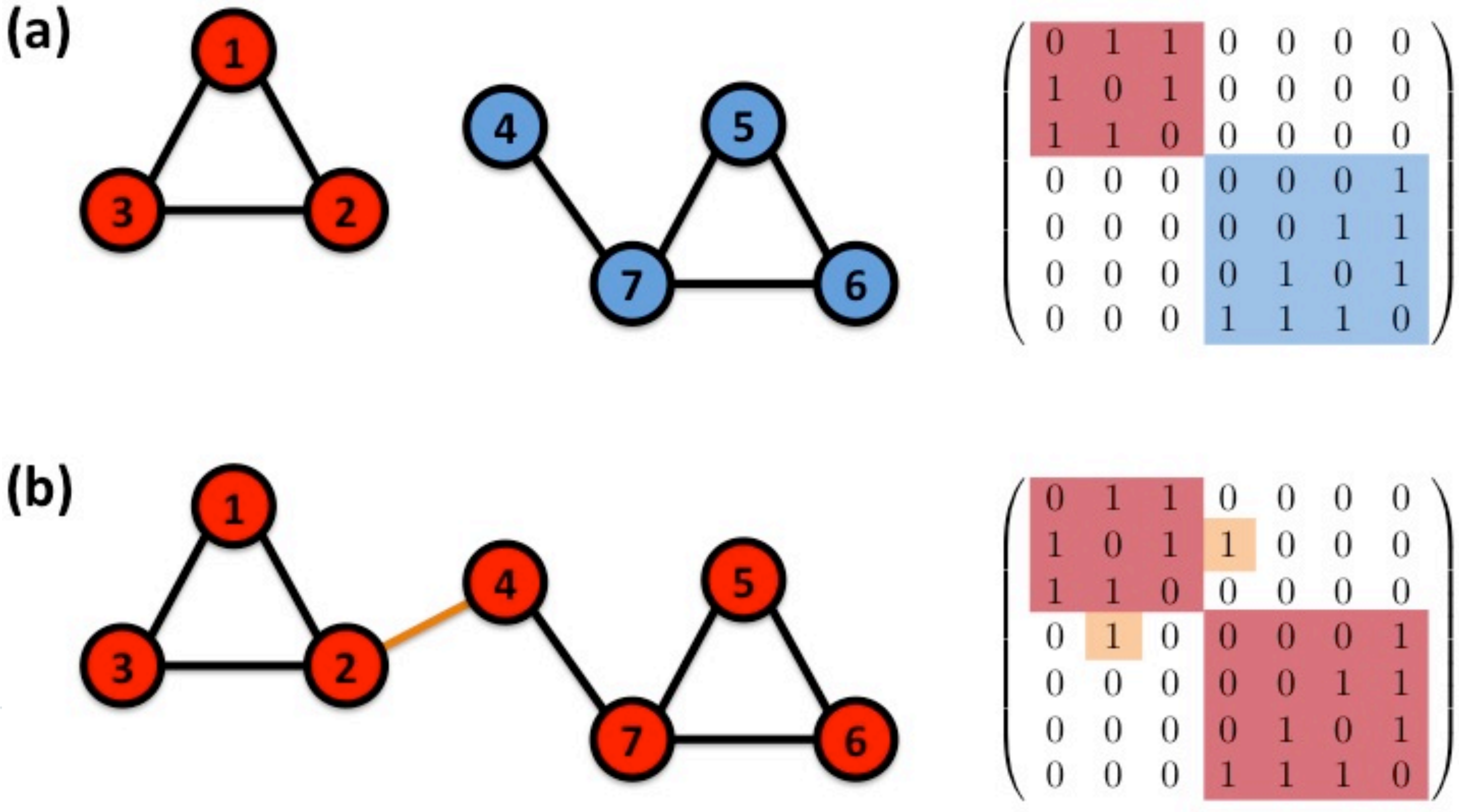
Largest Component:
Giant Component

The rest: **Isolates**

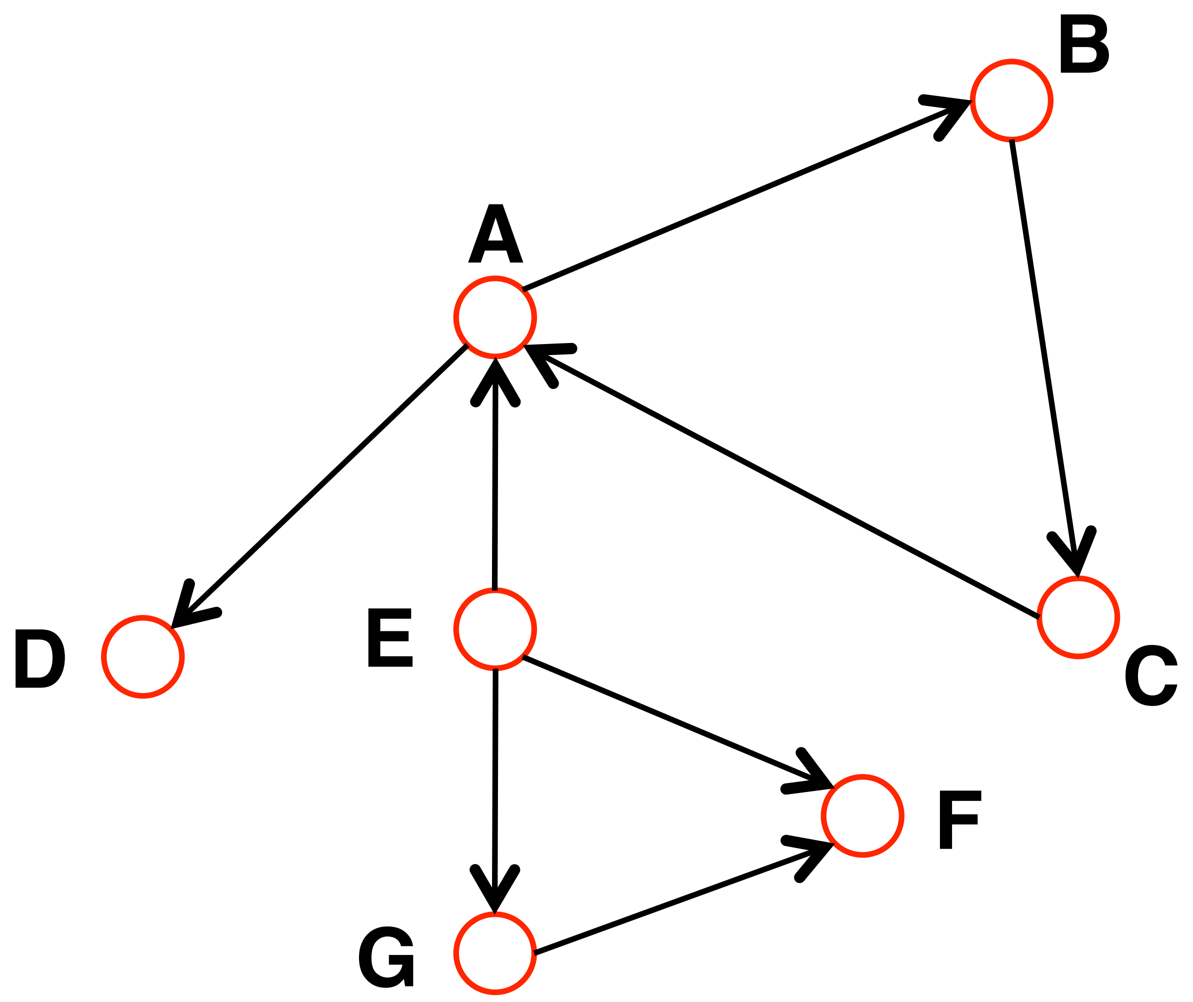
Bridge: if we erase it, the graph becomes disconnected.

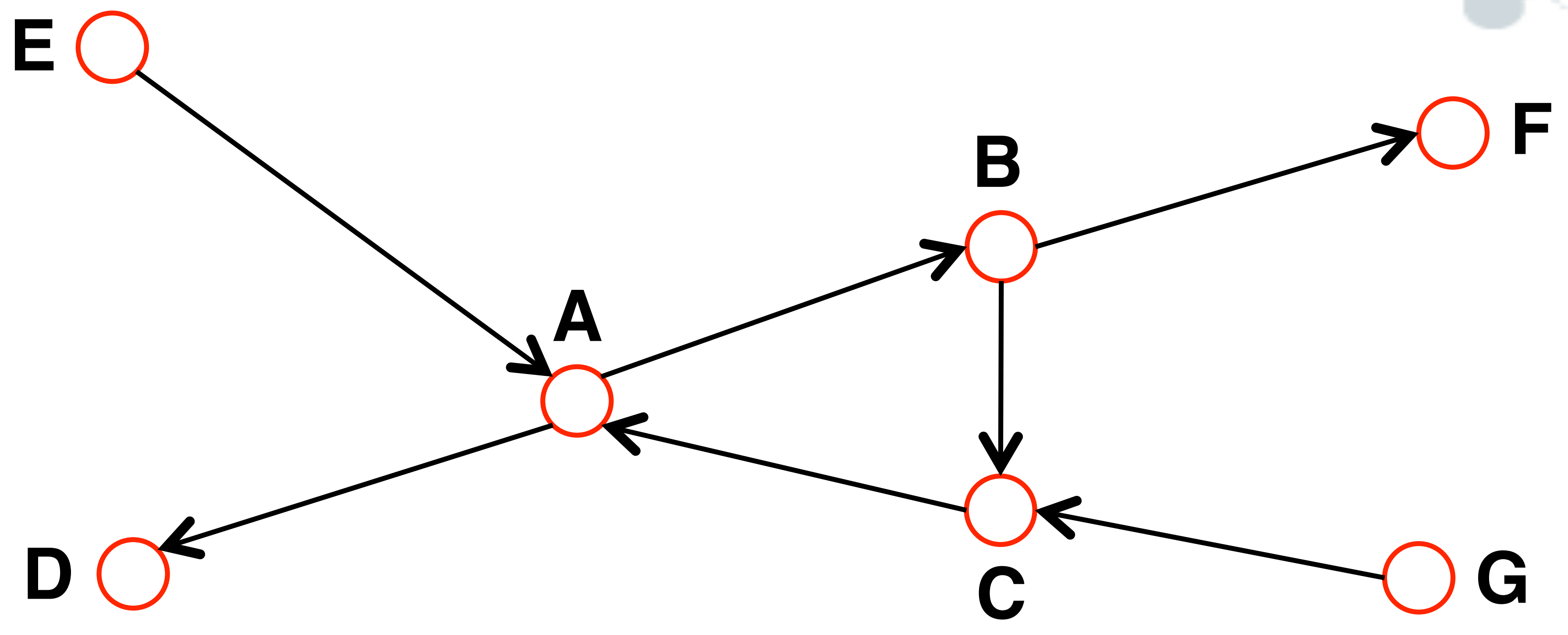
connectivity

The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:



connectivity in directed graphs

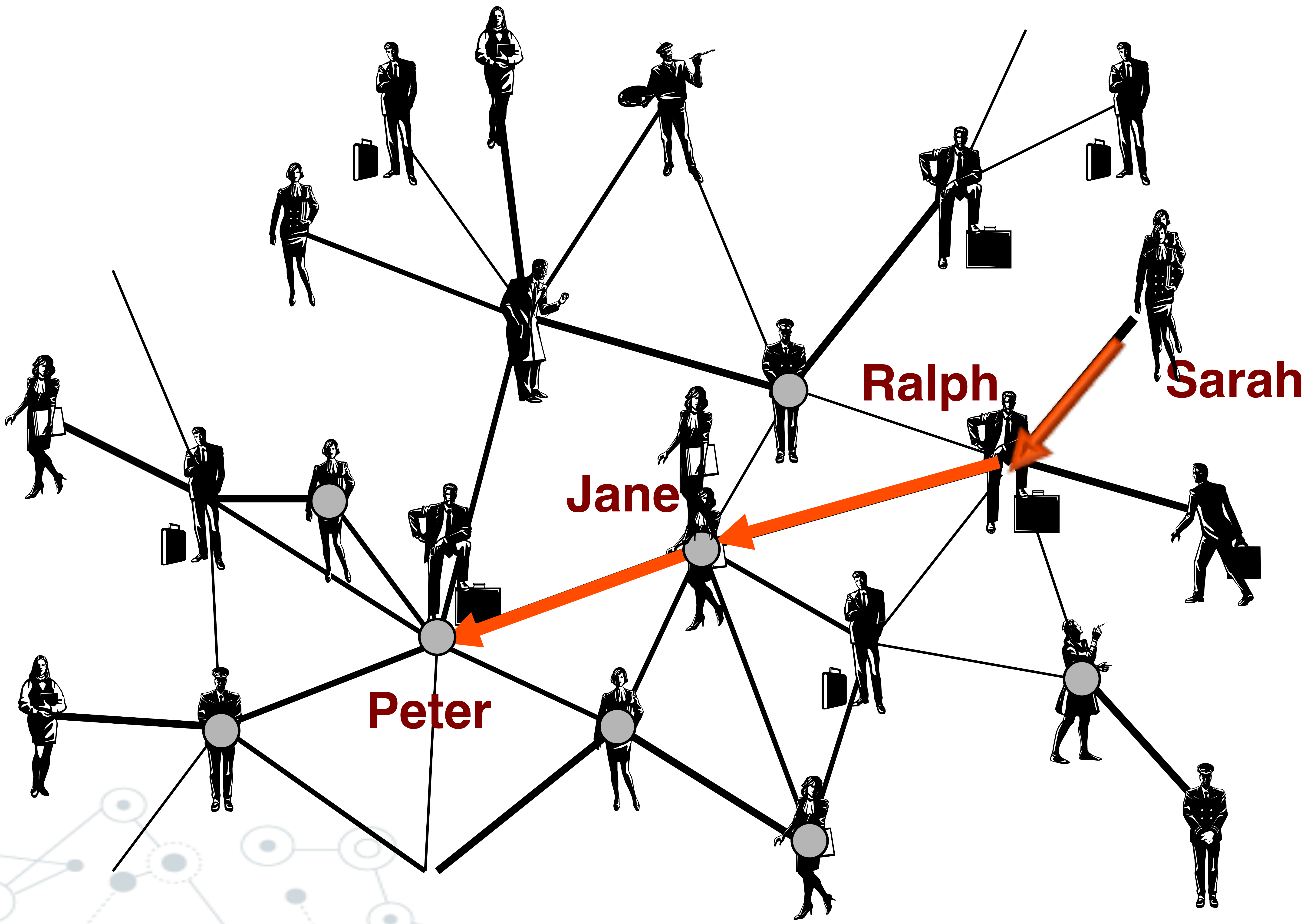




Strongly connected directed graph: has a path from each node to every other node and vice versa

Weakly connected directed graph: it is connected if we disregard the edge directions.

real world networks



Small world effect

Frigyes Karinthy, 1929
Stanley Milgram, 1967

six degrees of separation

Stanley Milgram (1967)

Two targets in Boston and Sharon, MA.

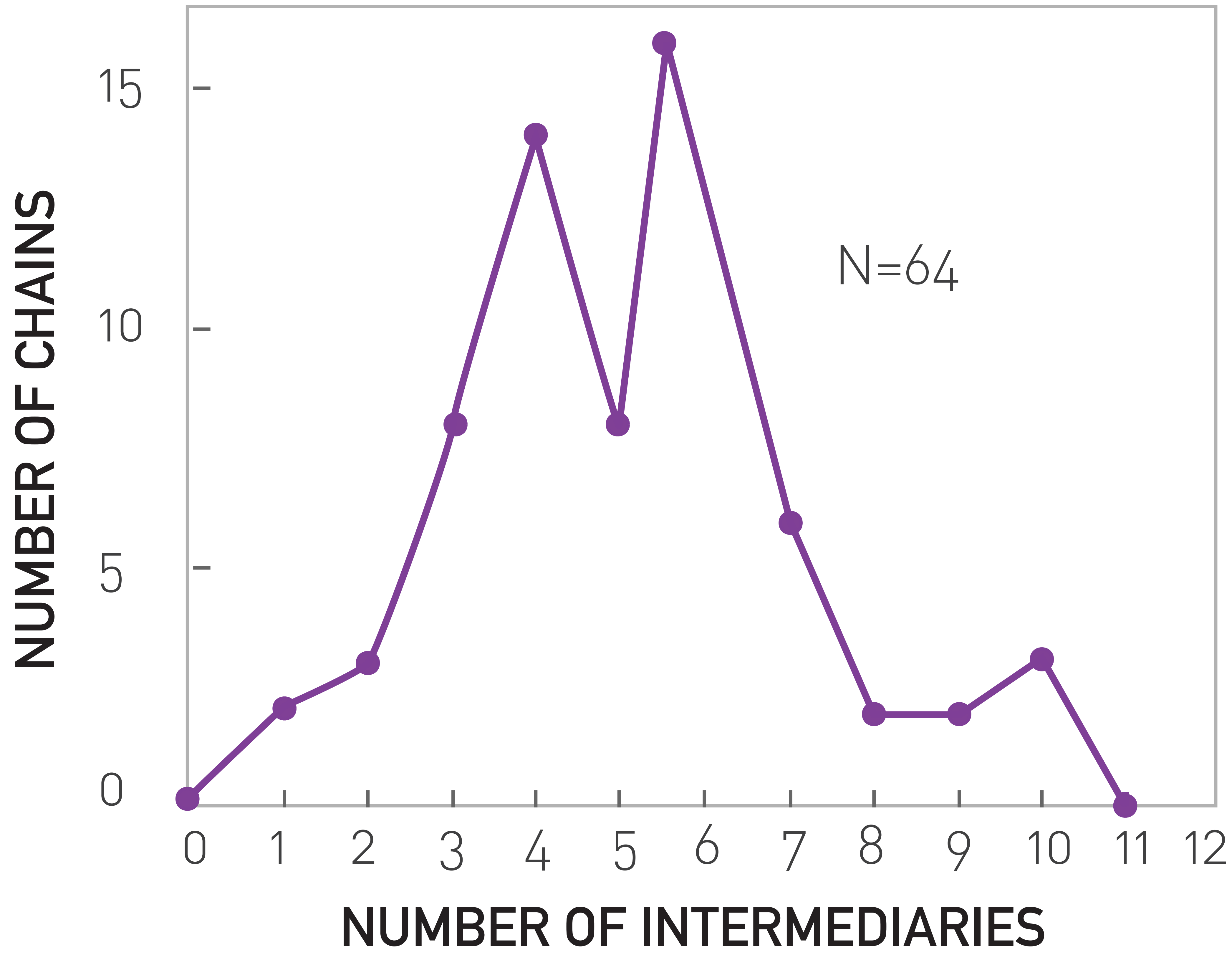
Randomly selected residents of Wichita and Omaha were asked to forward a letter to someone who is most likely to know the target person.

six degrees of separation

Stanley Milgram (1967)

Two targets in Boston and Sharon, MA.

Randomly selected residents of Wichita and Omaha were asked to forward a letter to someone who is most likely to know the target person.



clustering coefficient

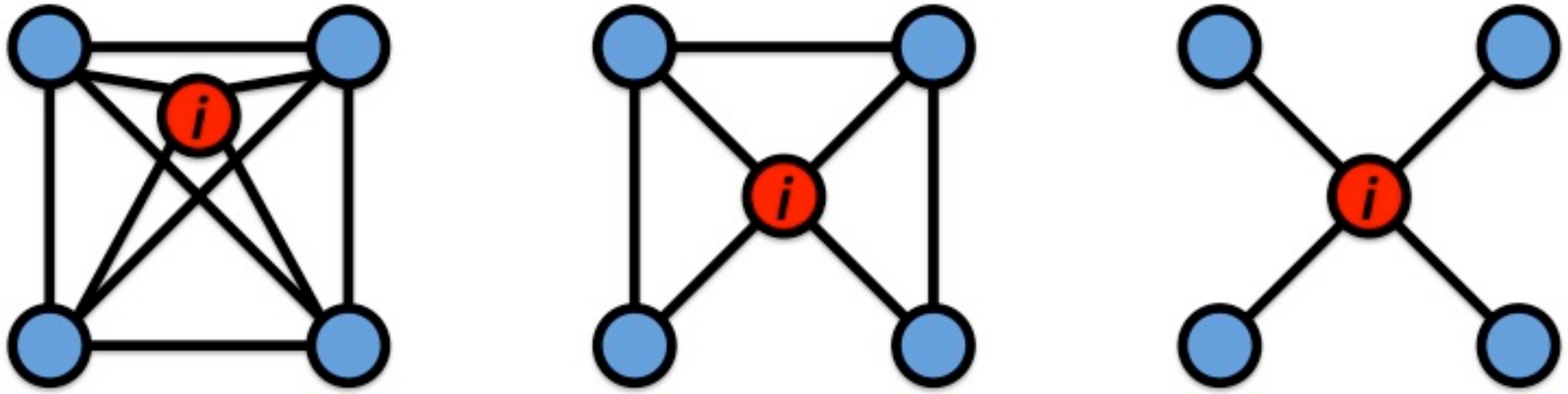
The clustering coefficient of a node captures the degree to which the neighbors of a given node link to each other, i.e. **what fraction of your neighbors are connected?**

Watts & Strogatz, Nature 1998.

clustering coefficient

The clustering coefficient of a node captures the degree to which the neighbors of a given node link to each other, i.e.
what fraction of your neighbors are connected?

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$



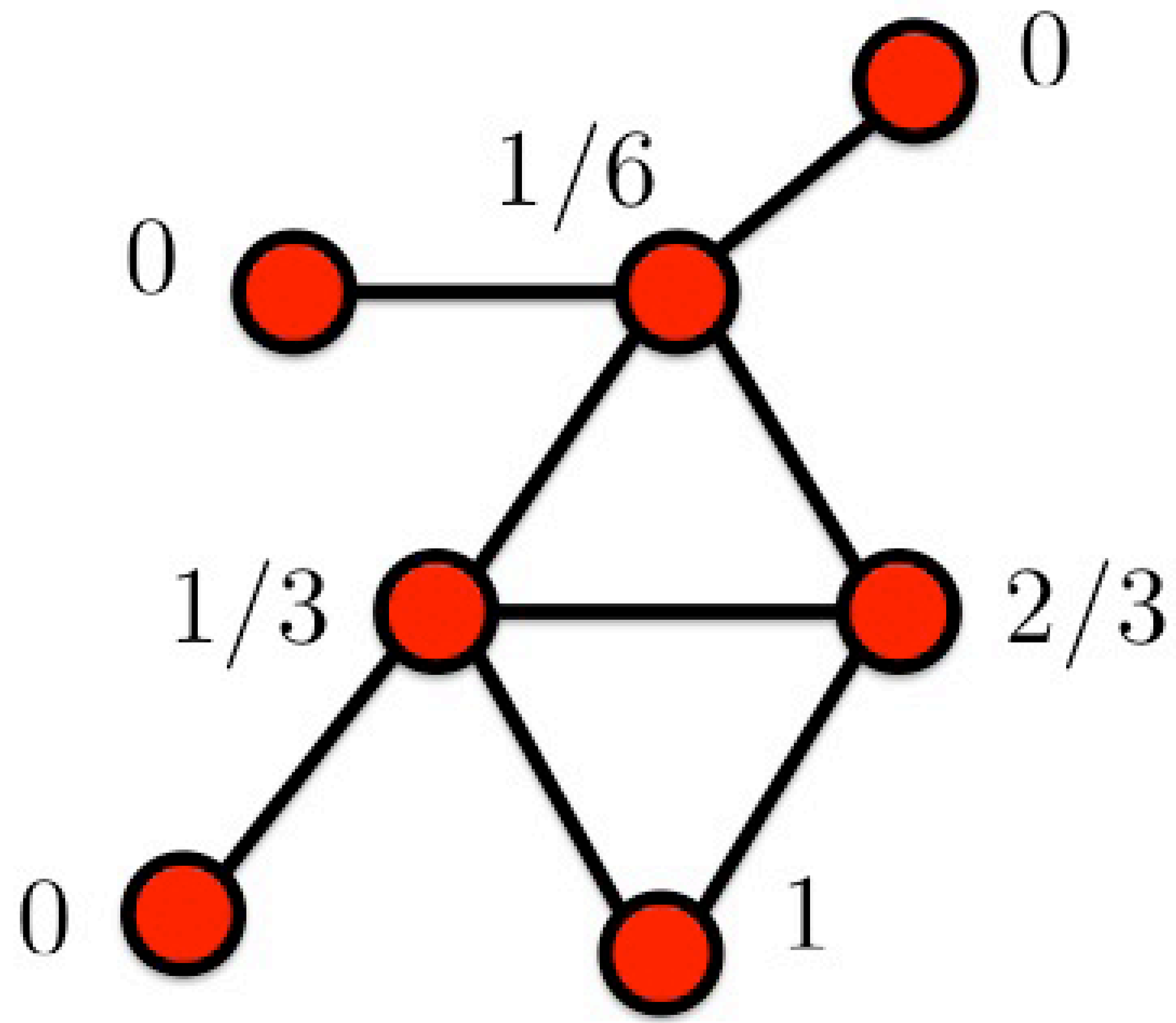
$C_i = 1$ $C_i = 1/2$ $C_i = 0$

Watts & Strogatz, Nature 1998.

clustering coefficient

The degree of clustering of a whole network is captured by the **average clustering coefficient**, representing the average of C over all nodes i = 1,…,N

$$\langle C_i \rangle = \frac{1}{N}\sum_i C_i$$



$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

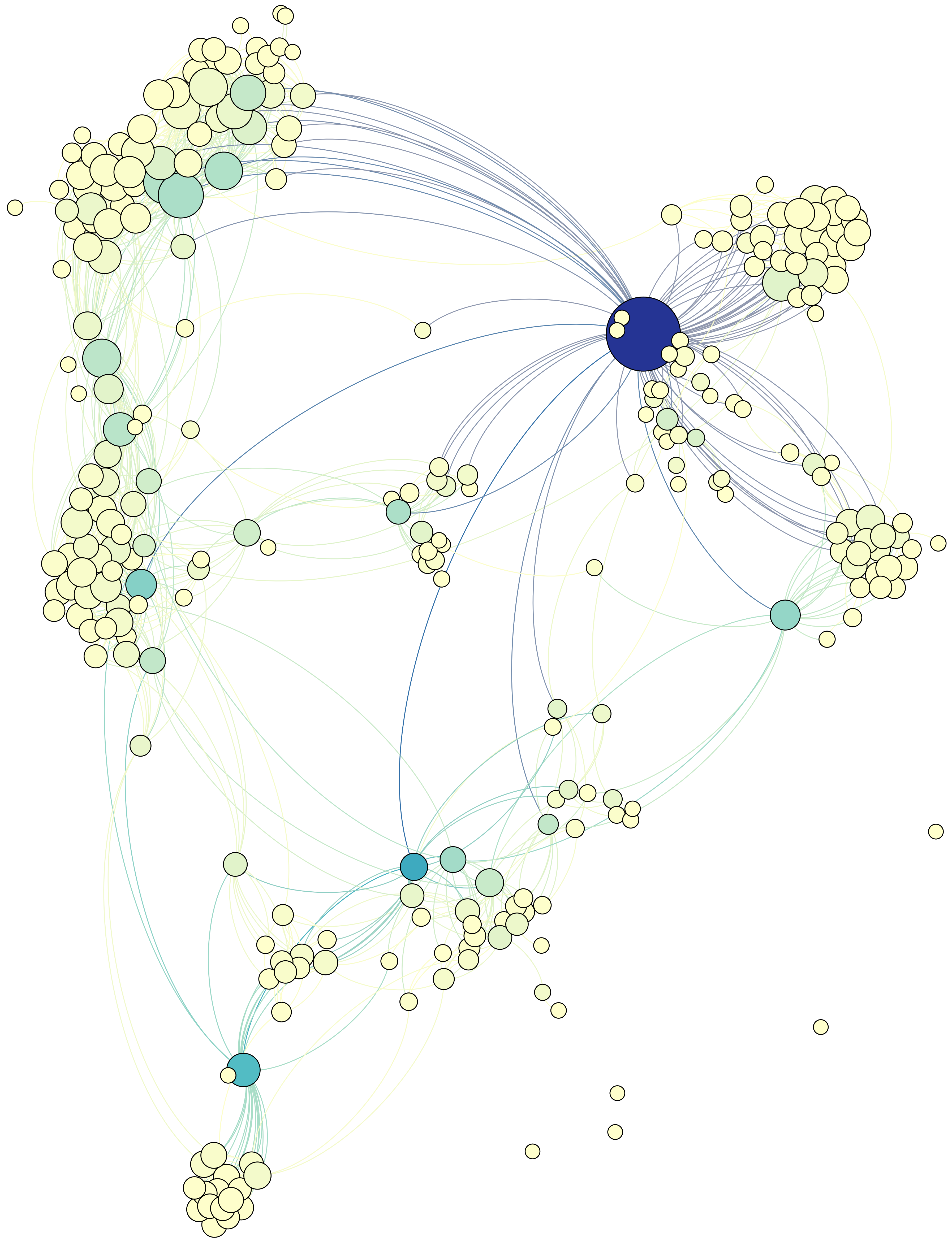
clustering coefficient

the ***global clustering coefficient*** measures the total number of closed triangles in a network. Indeed, L_i in the previous equation is the number of triangles that node *i* participates in, as each link between two neighbors of node *i* closes a triangle.

$$C_\Delta = \frac{3 \times Number\ Of\ Triangles}{Number\ Of\ Connected\ Triples}$$

real world networks

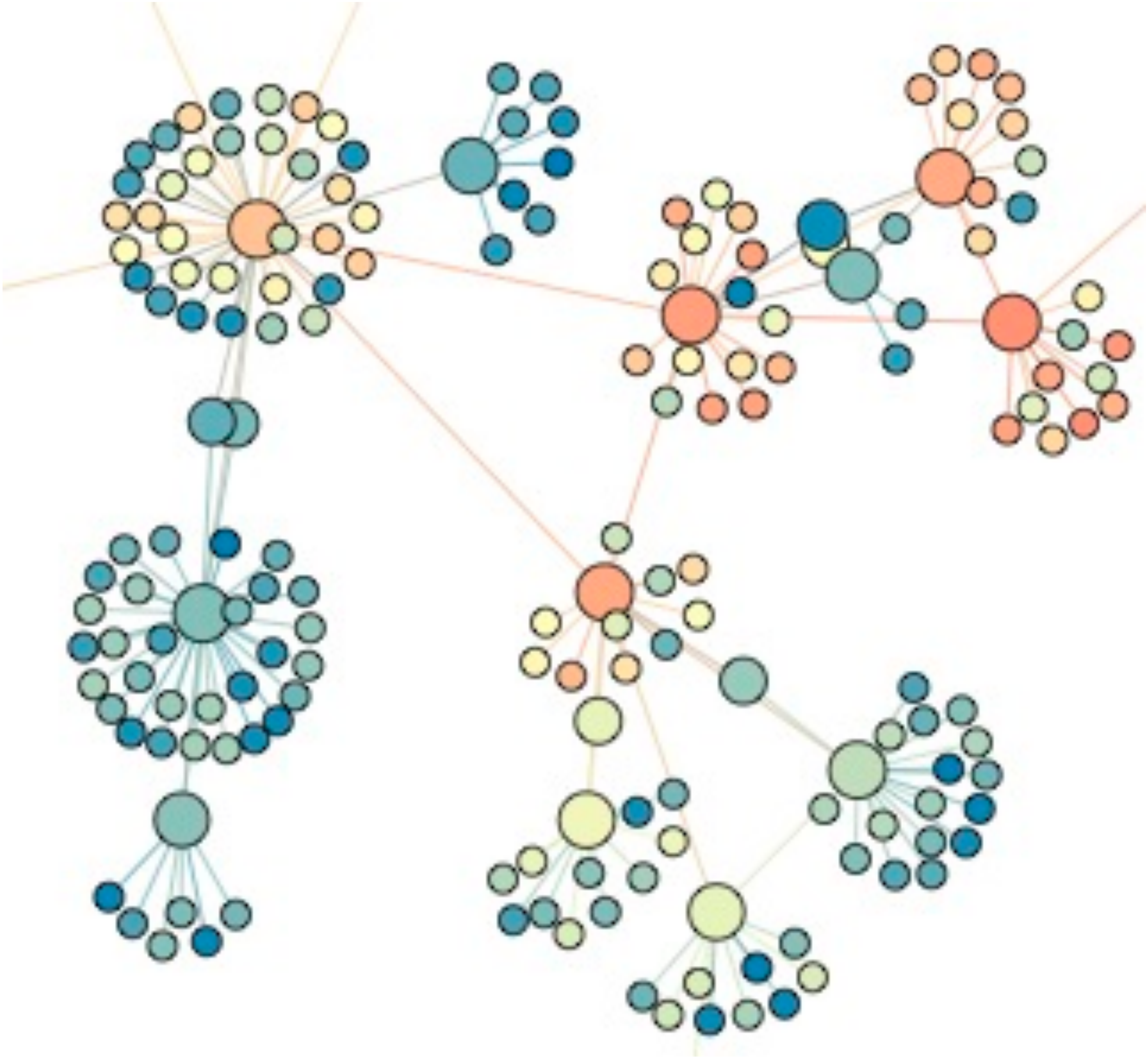
- Real world networks are highly clustered
- Average clustering coefficient can have values >0.5
- **Triadic closure** in social networks is a common phenomenon



measures of centrality

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality
- Katz centrality
- Pagerank

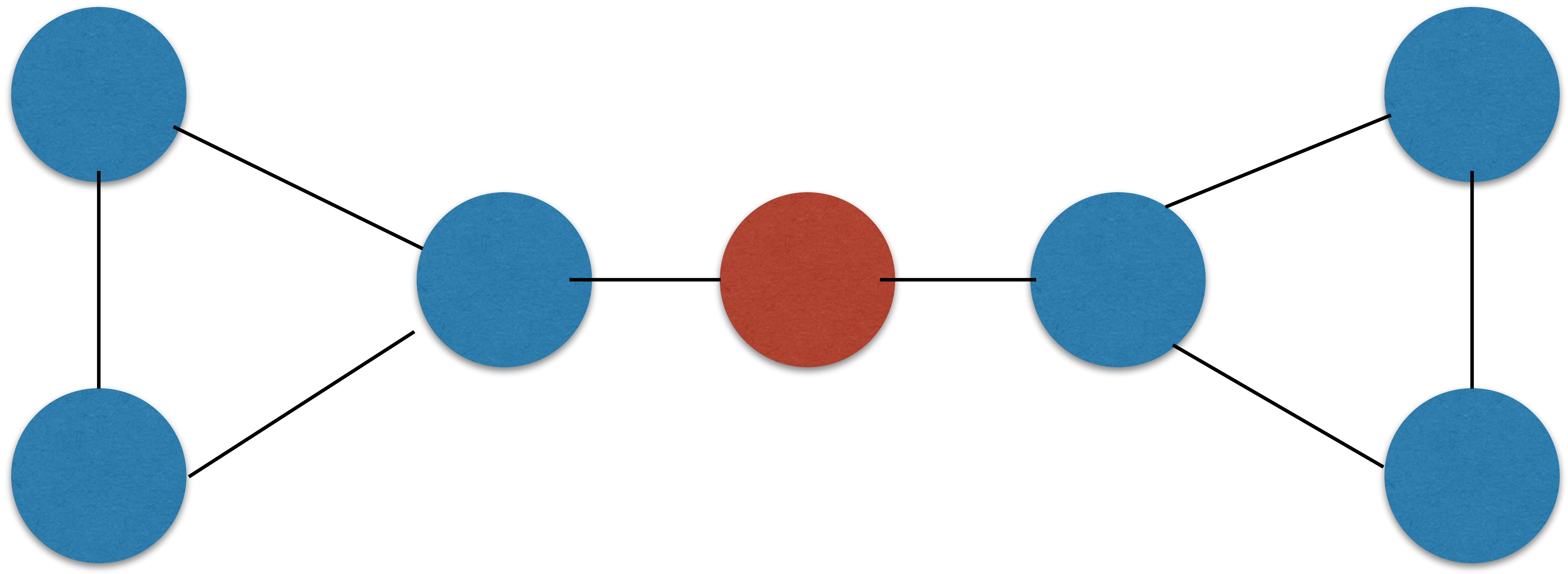
…



betweeness centrality

Betweeness captures a node's brokerage

intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?

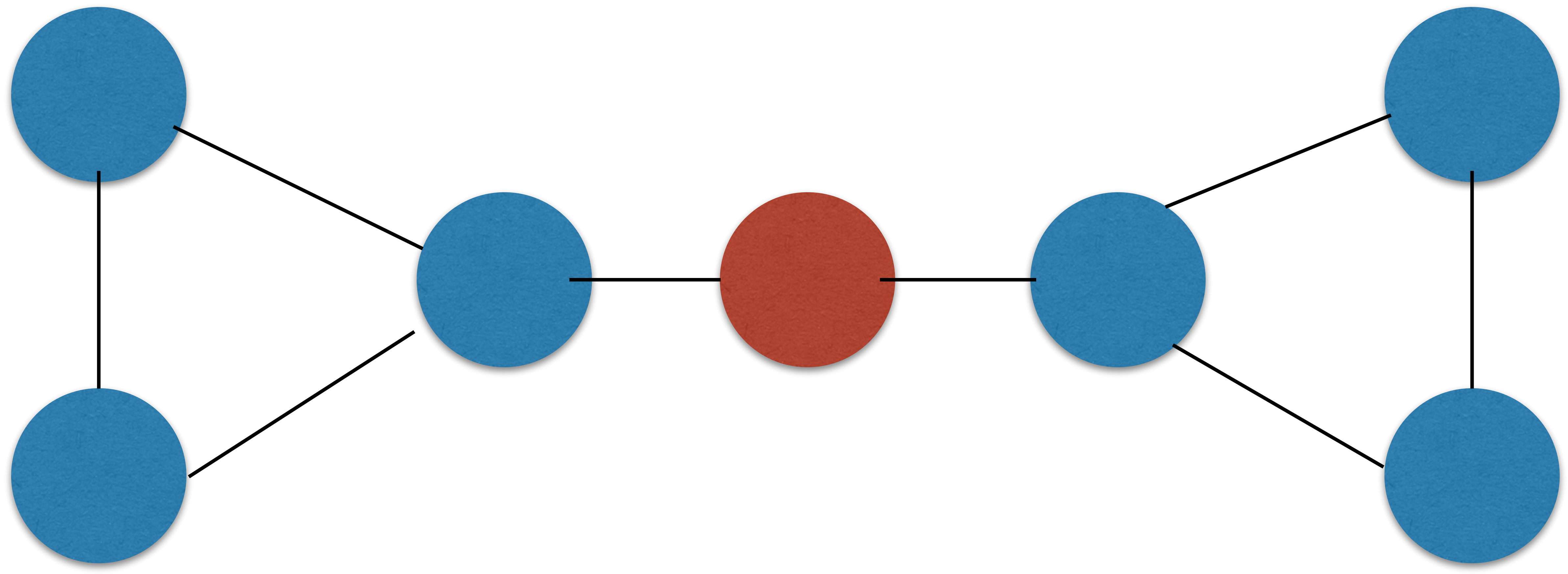


betweeness centrality

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$

g_{jk} = #shortest paths connecting j and k

$g_{jk}(i)$ = #shortest paths connecting j and k through i

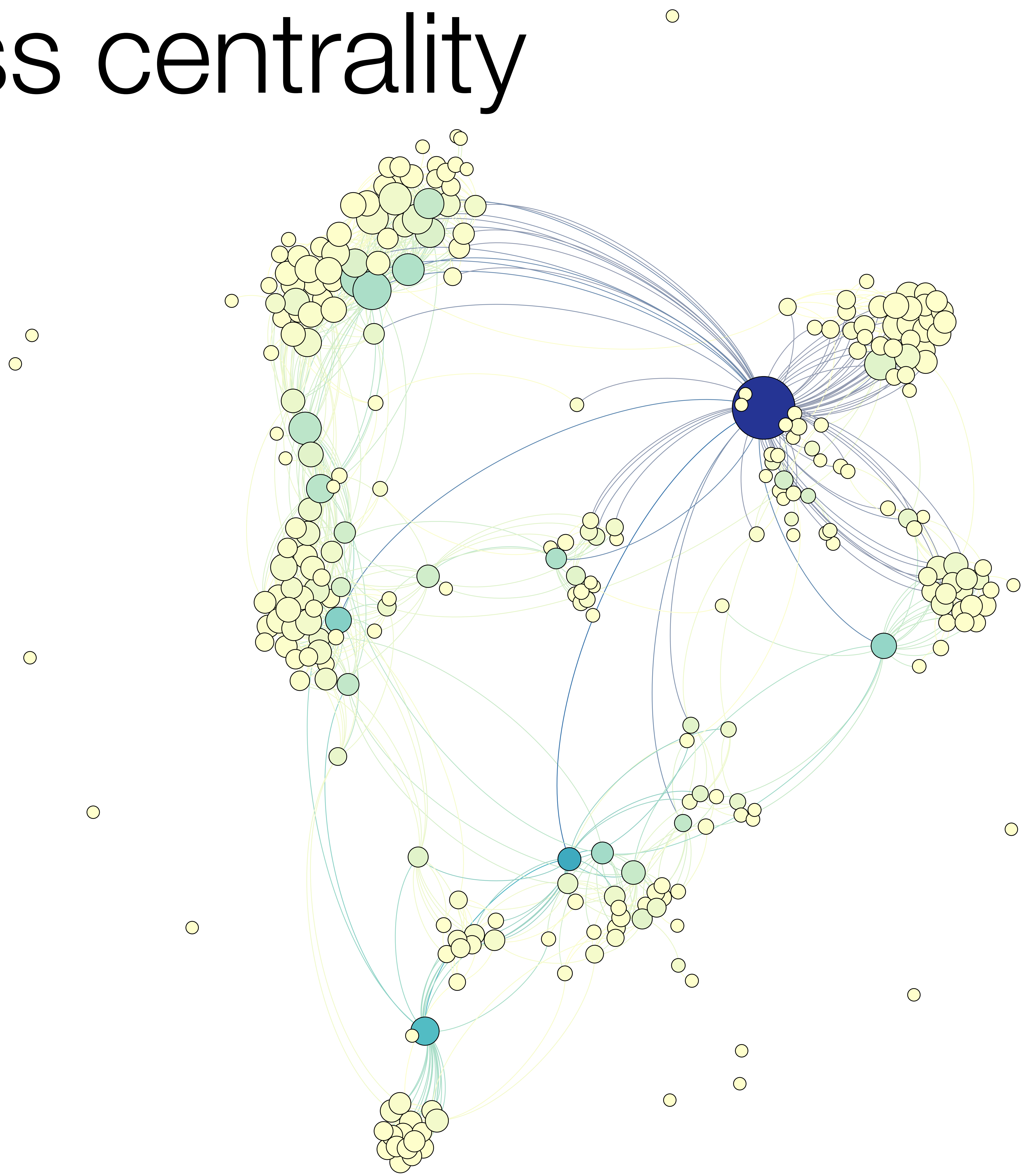


betweeness centrality

My Facebook graph

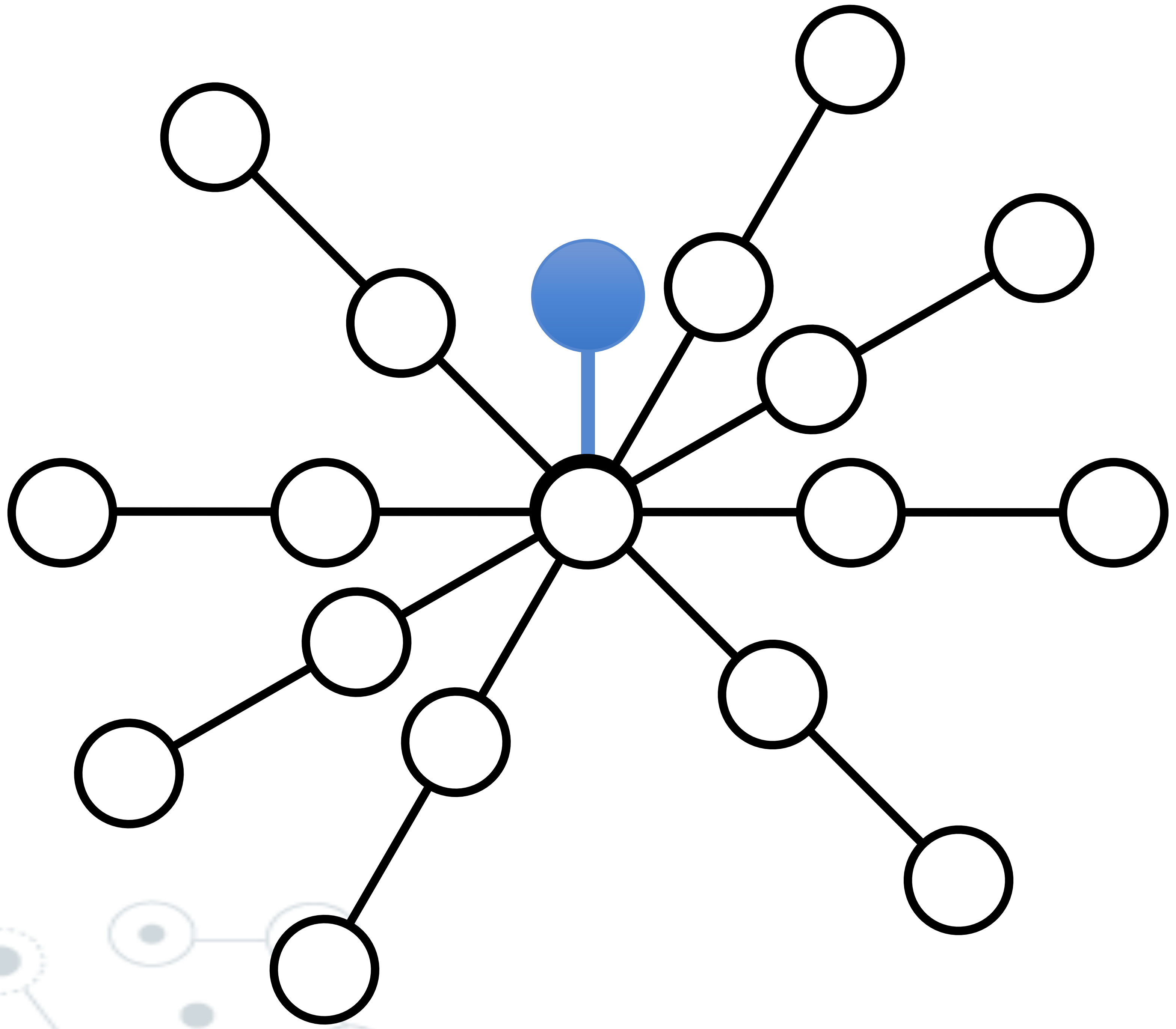
Node size is proportional to the degree

Node color is proportional to the betweenness



closeness centrality

Closeness is based on the length of the average shortest path between a node and all other nodes in the network. It quantifies the reachability of a node.



$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Katz centrality

$$C_{KZ}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$$

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors and also all other nodes in the network that connect to the node through these immediate neighbors. Connections made with distant neighbors are, however, penalized by an attenuation factor

Katz centrality

$$C_{KZ}(i) = \alpha \sum_{j=1}^{n} A_{ij} C_{KZ}(j)$$

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors and also all other nodes in the network that connect to the node through these immediate neighbors. Connections made with distant neighbors are, however, penalized by an attenuation factor