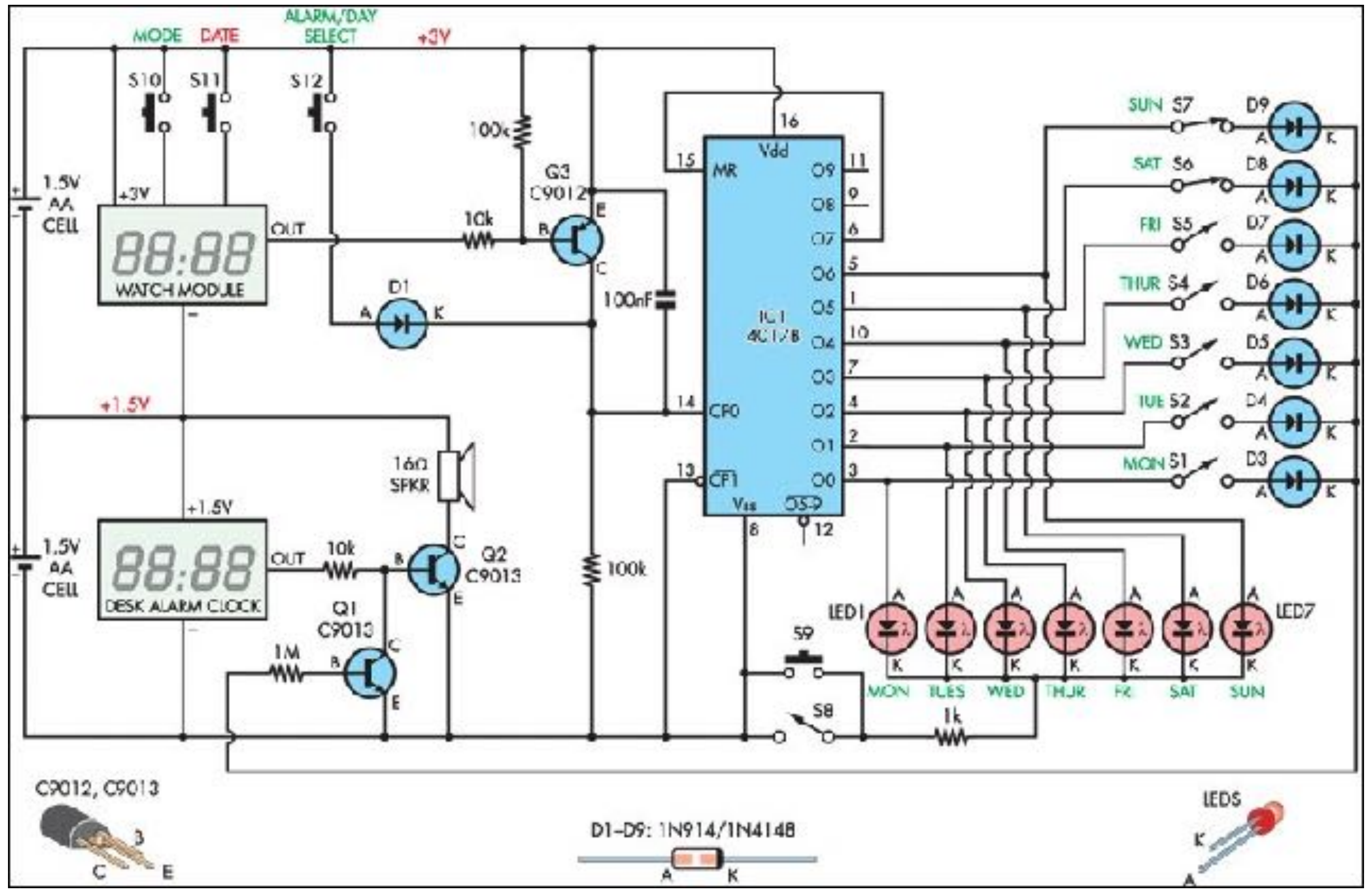
Networks are the backbones of complex systems

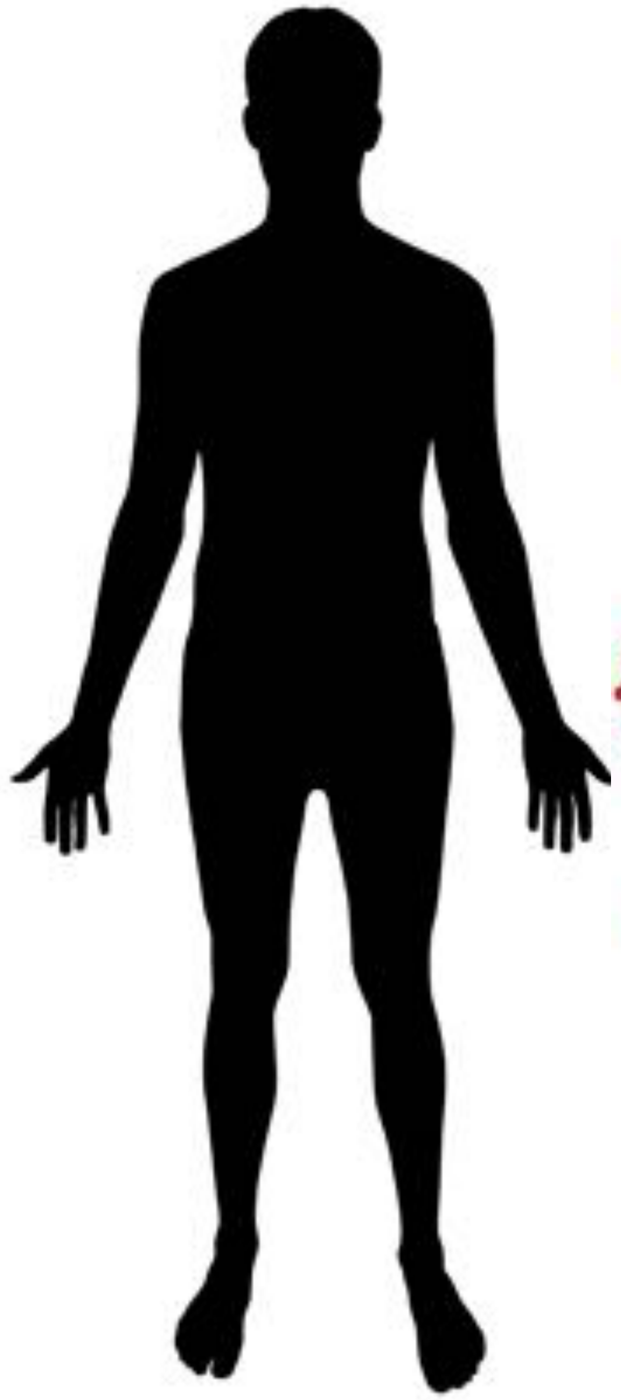
Many single units

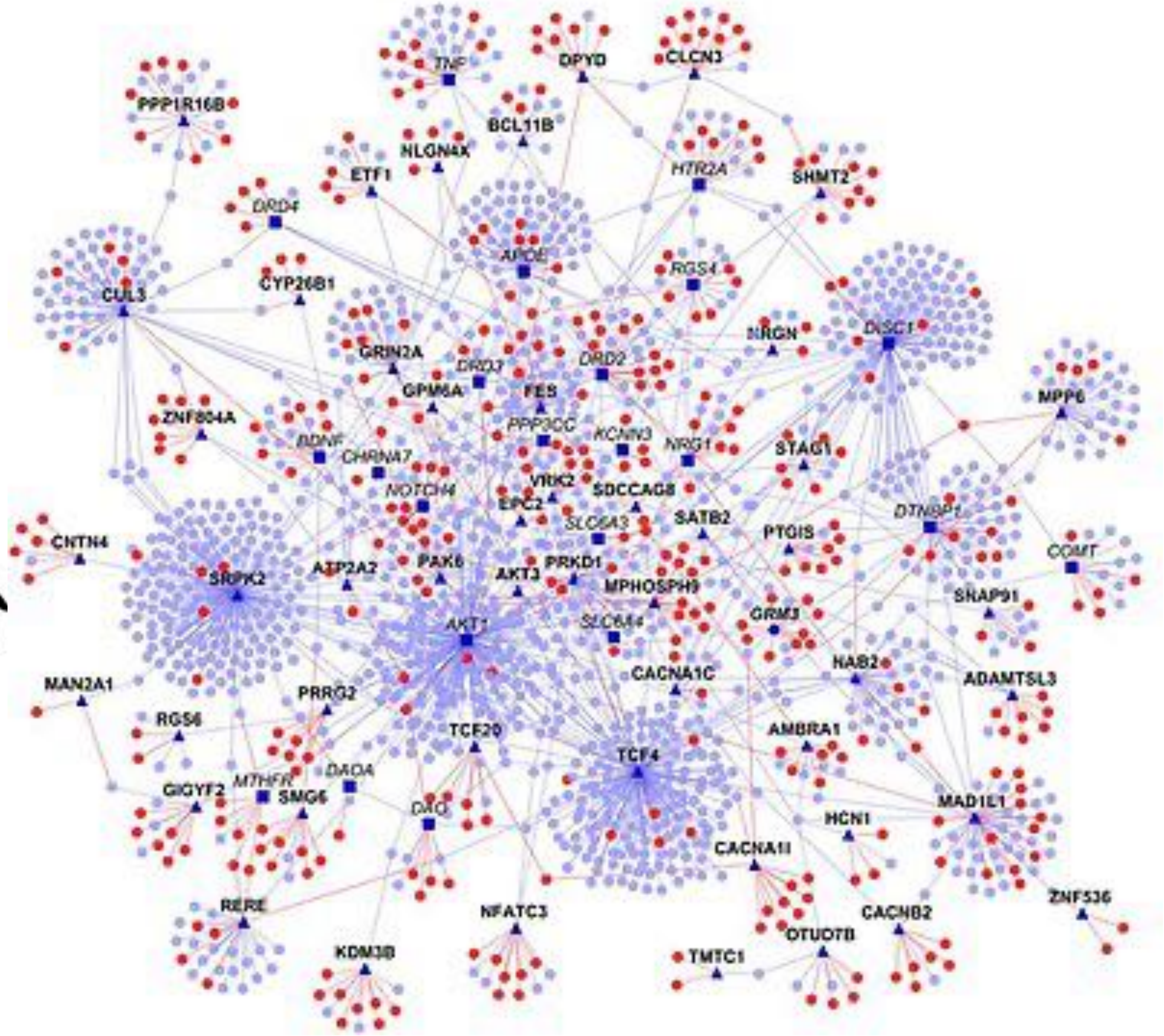
Strong, nonlinear interactions

Emergence of collective behavior

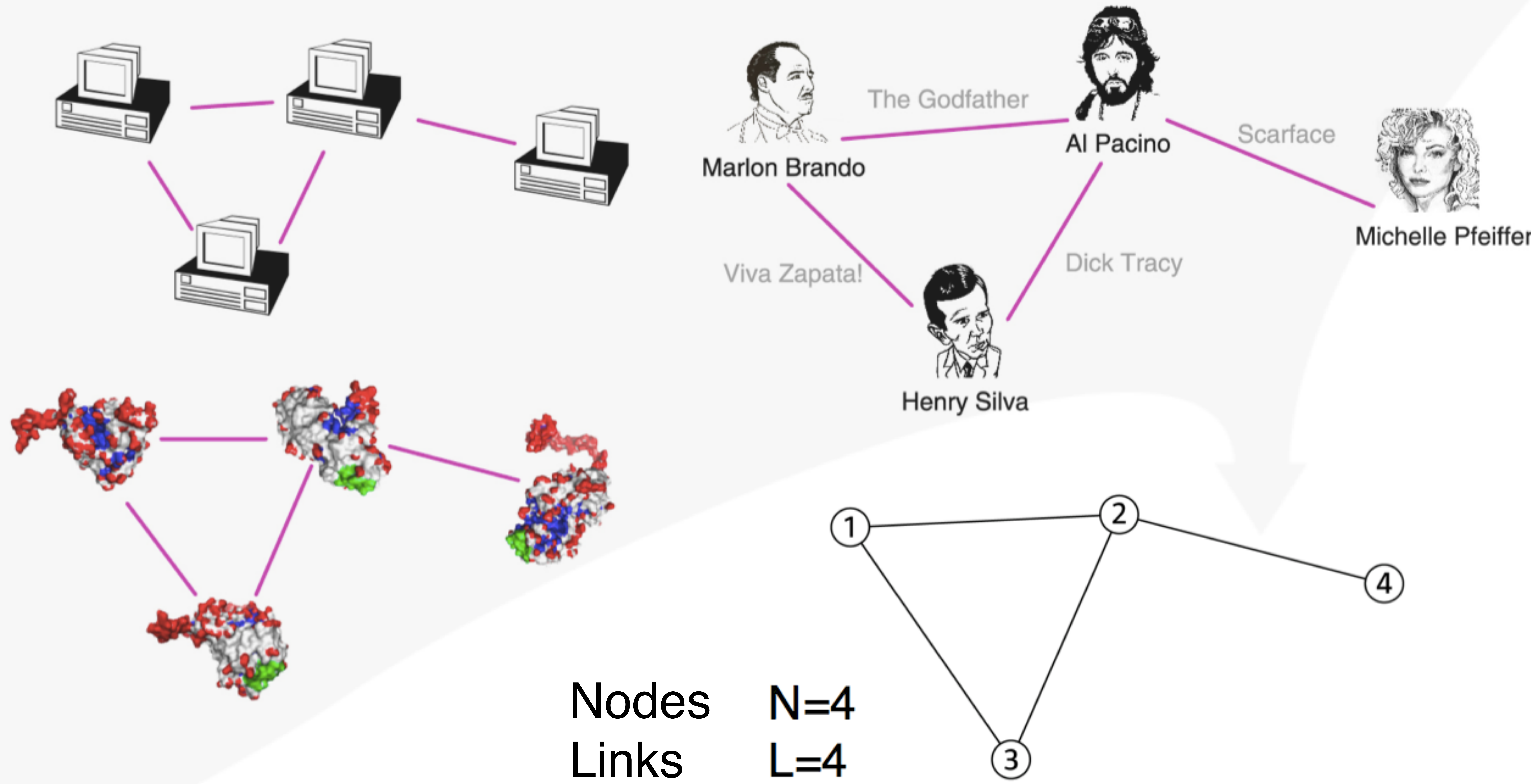
Networks are the maps of complex systems
We now have the data to see and study them







Networks are a common language for different applications

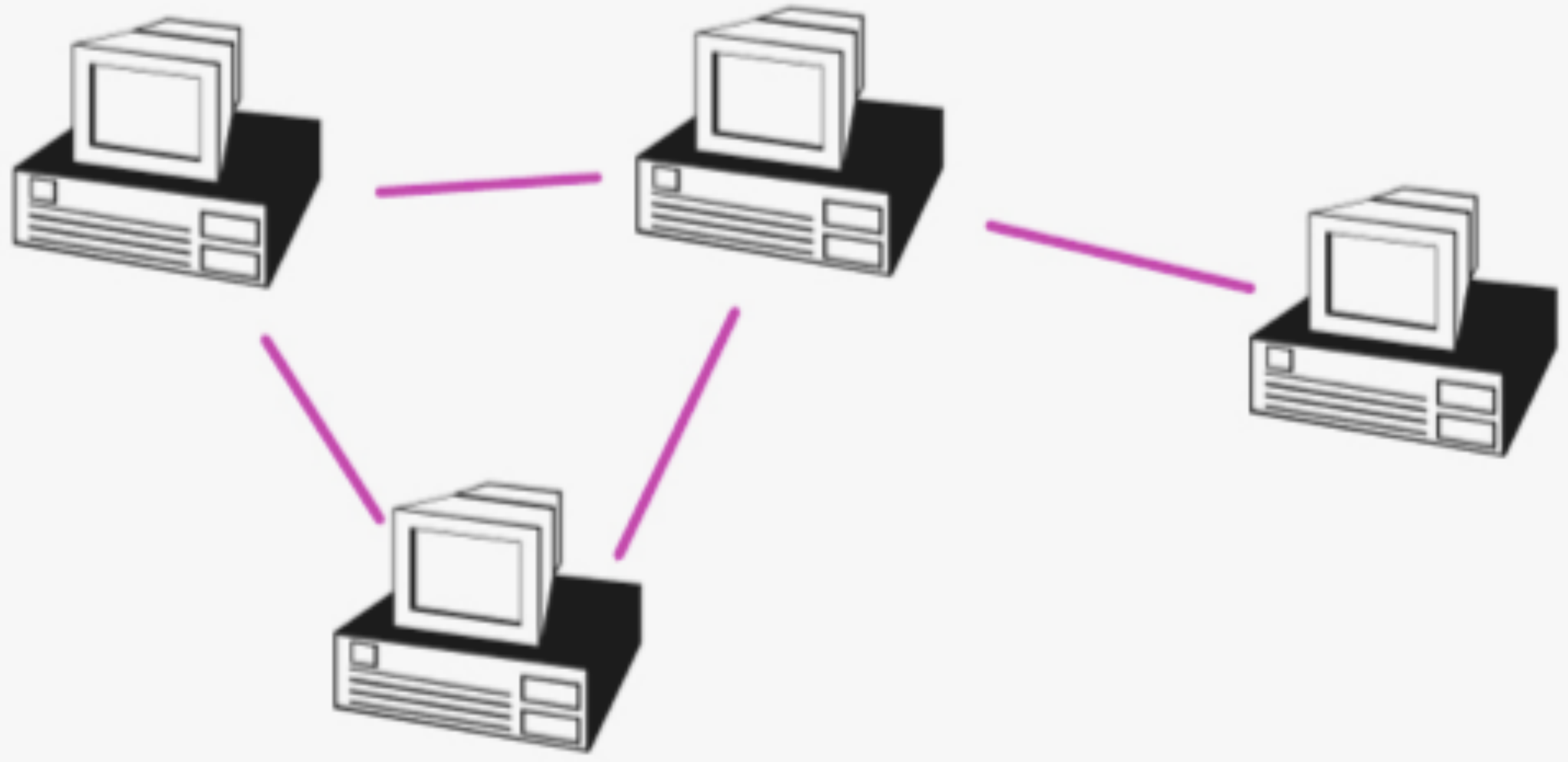


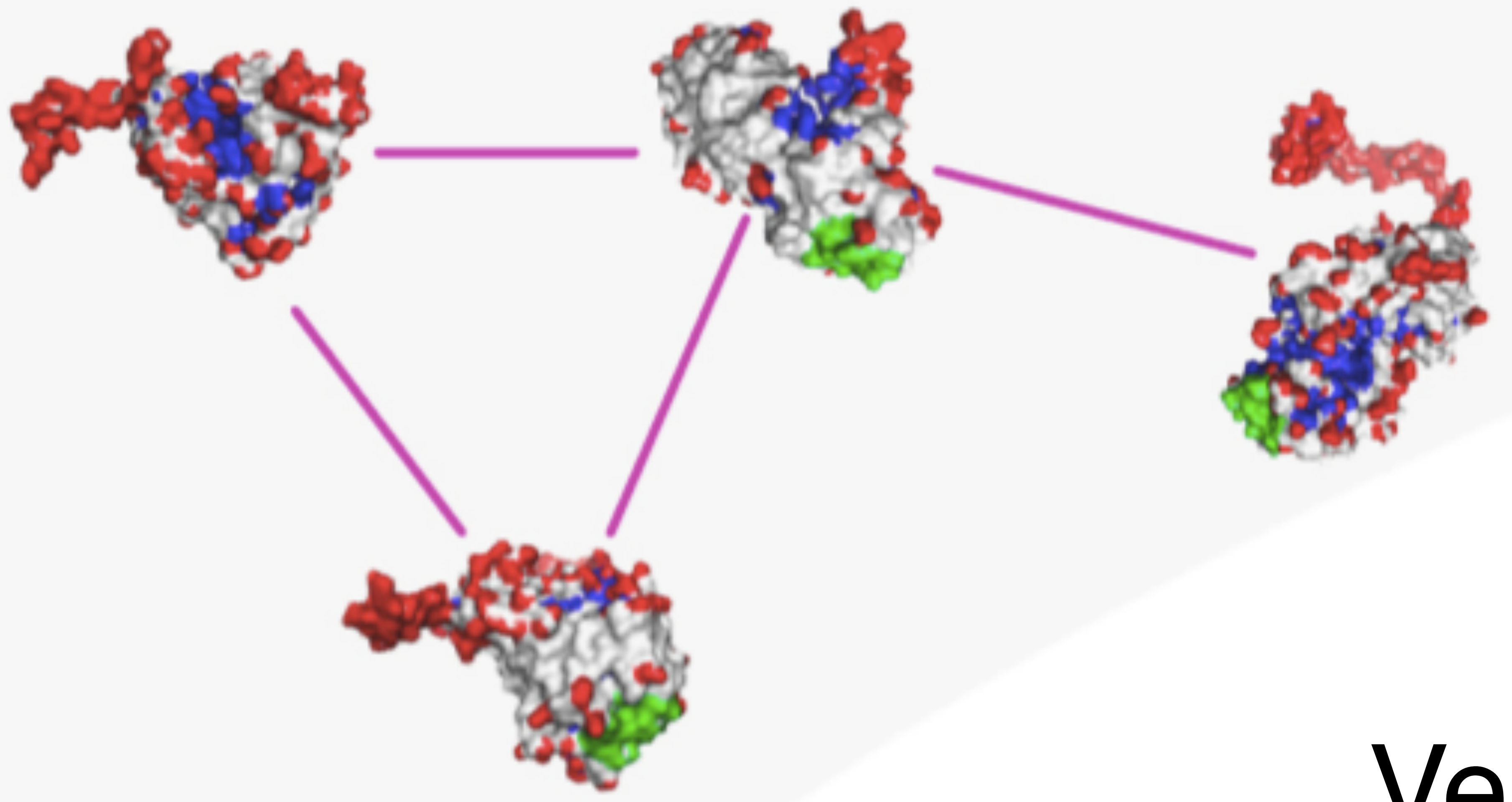
Nodes N=4

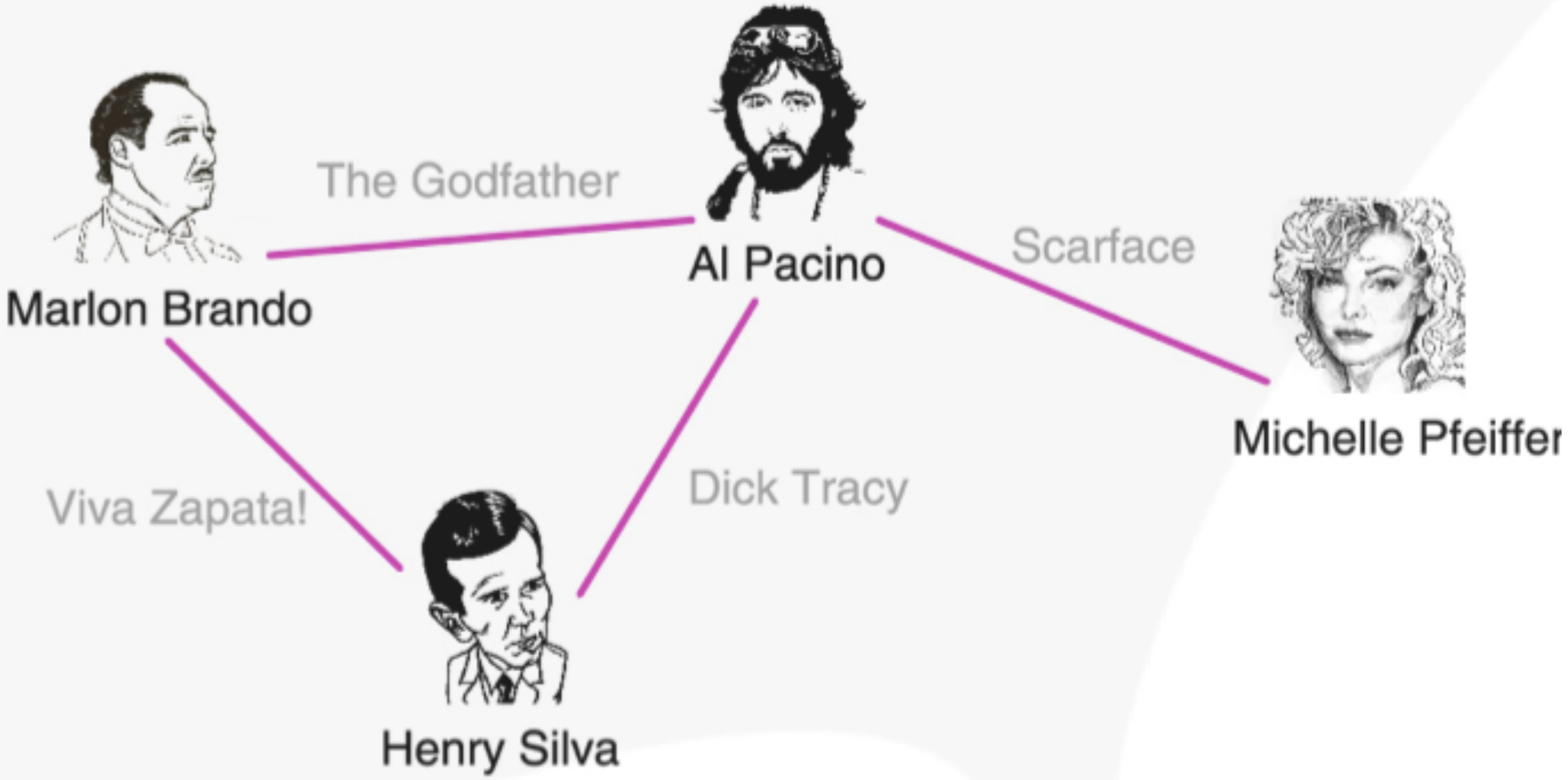
Links L=4

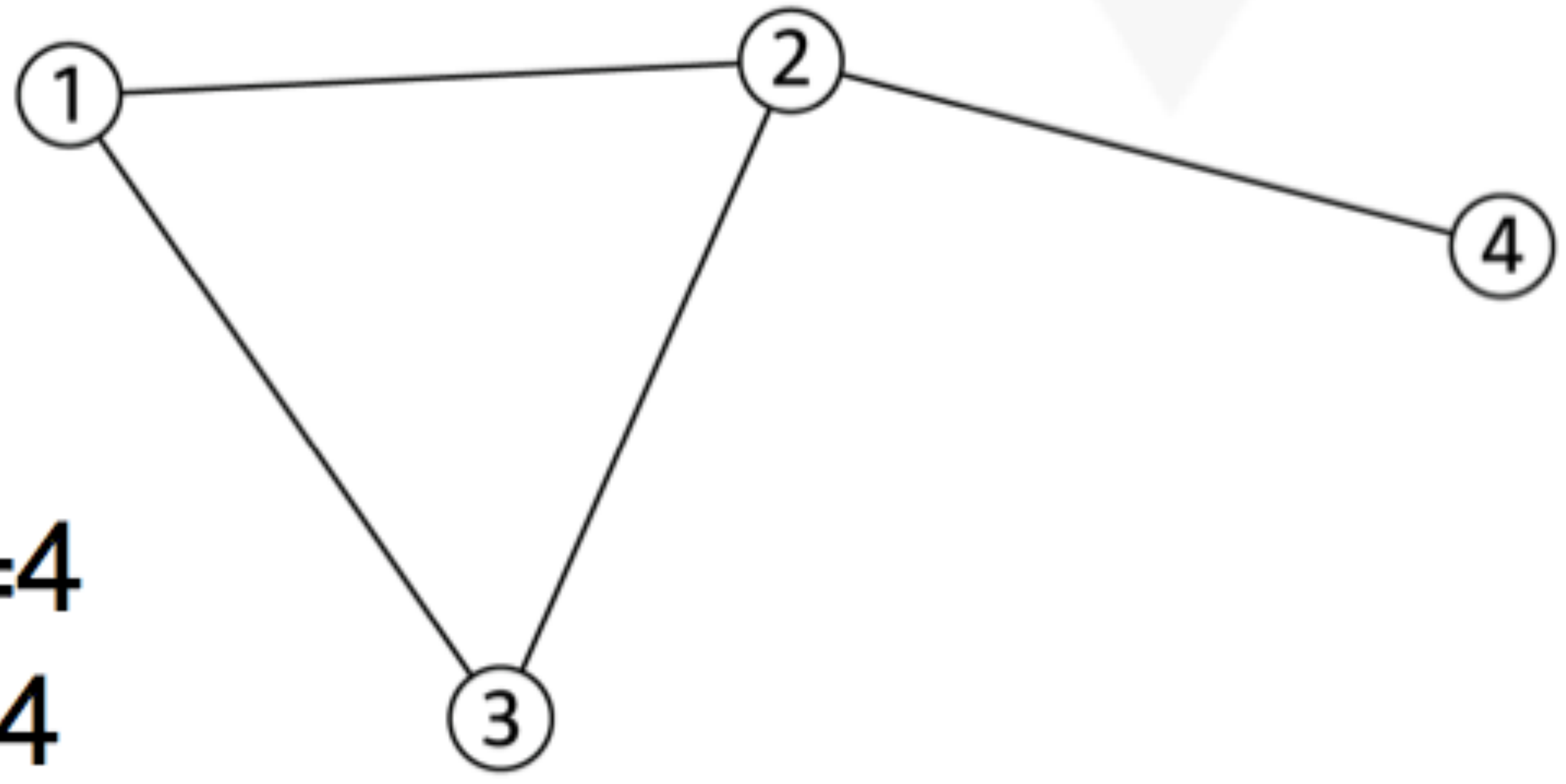
Graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$

Network = Graph + real-world meaning







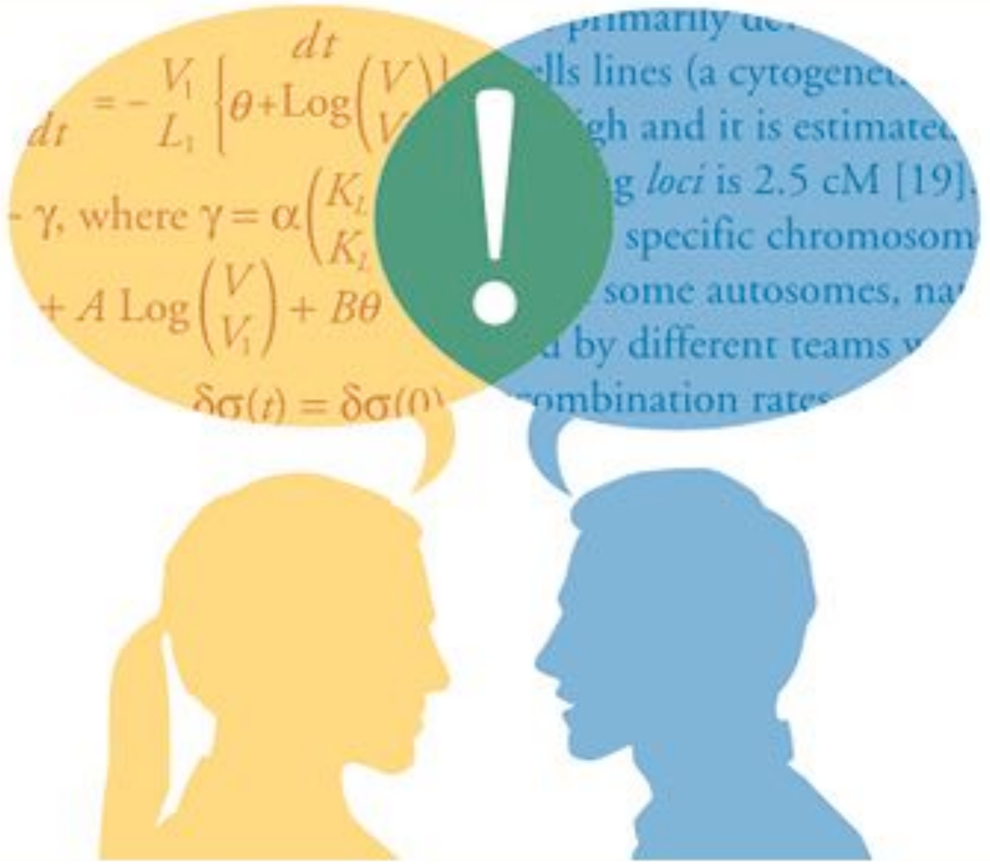


Vertices V=4

Edges E=4

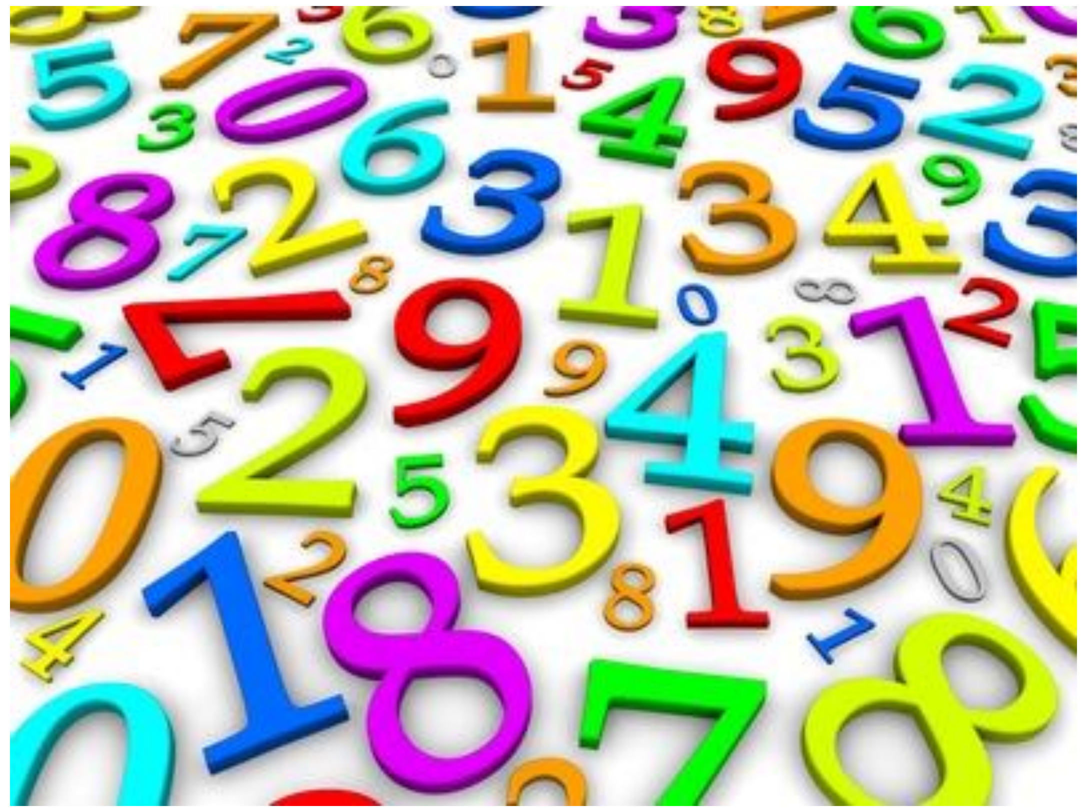
Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Network science is a part of Data science. It is:



Interdisciplinary

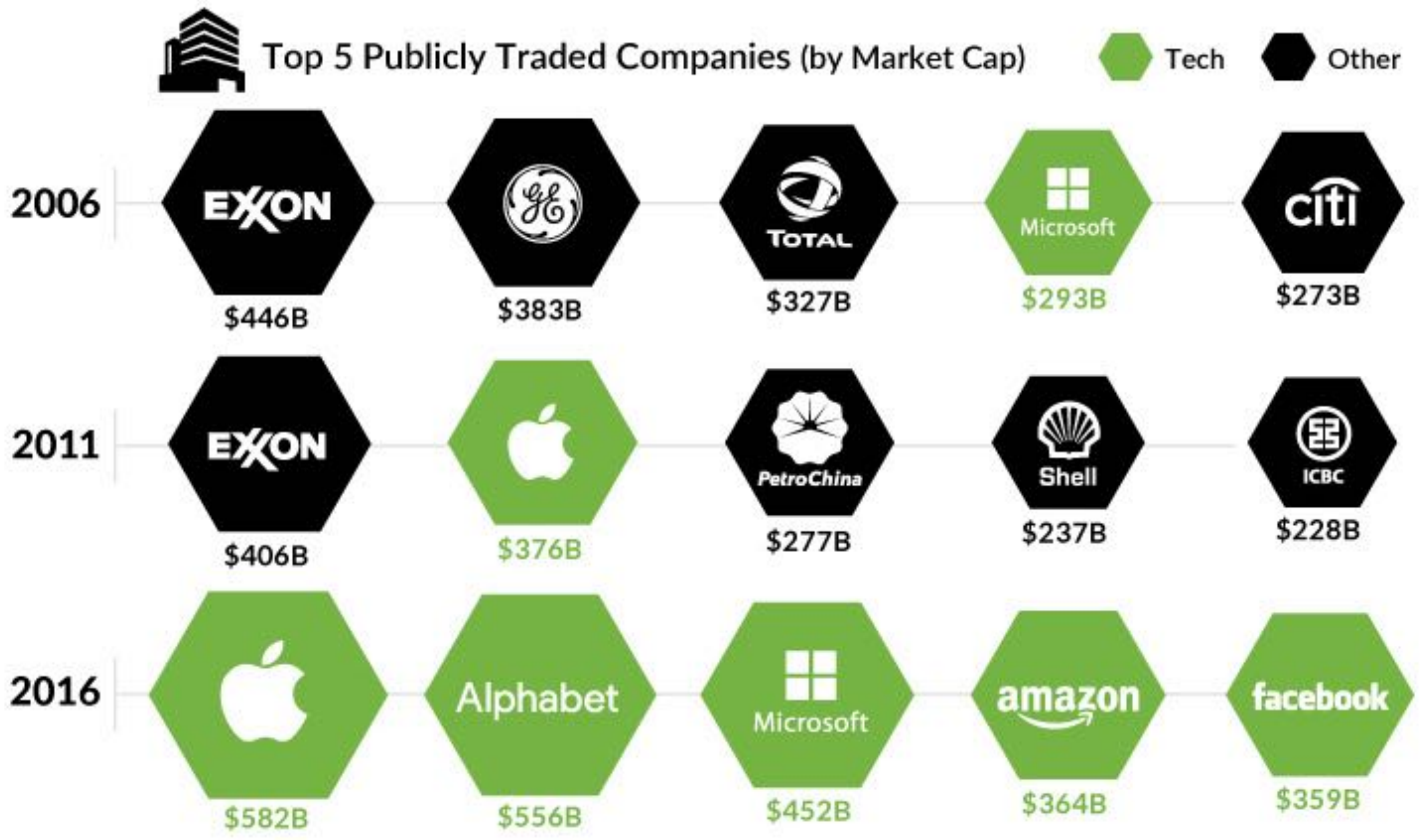
Empirical



Quantitative,
mathematical

Computational

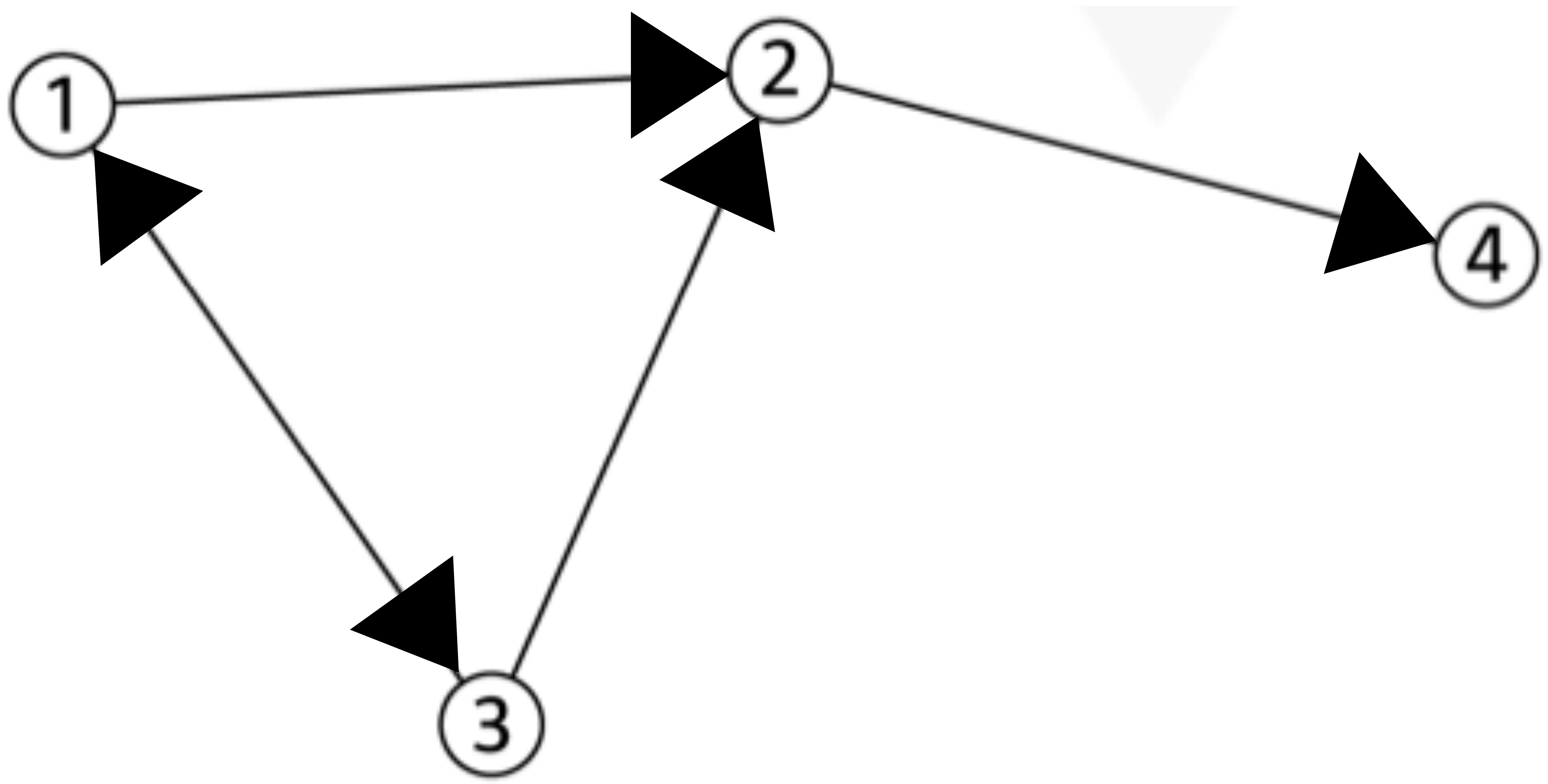
Networks have a huge economic impact



Data is the new oil

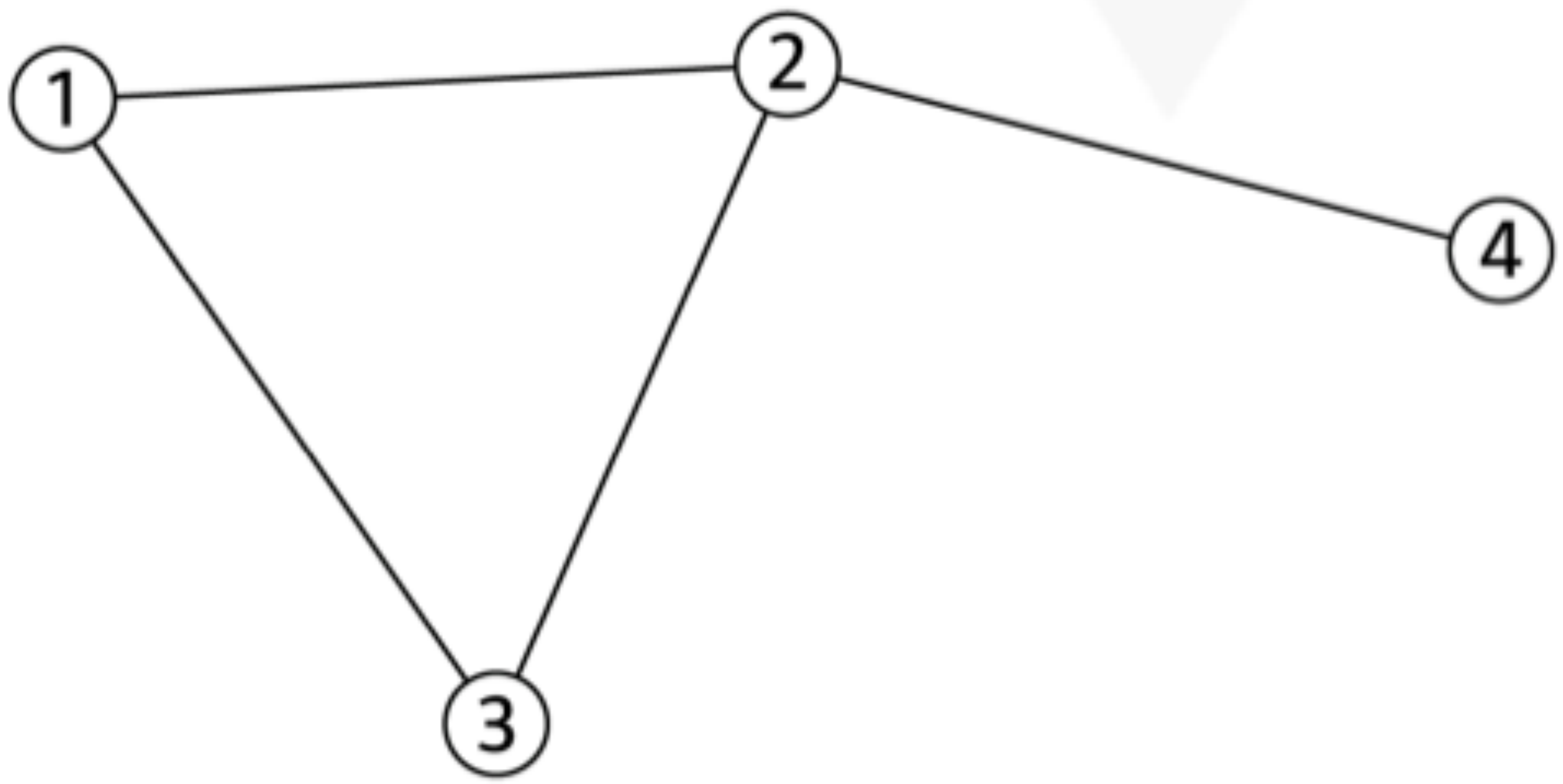
A directed graph (digraph) has links with a direction

Nodes N=4
Links (Arcs) L=5



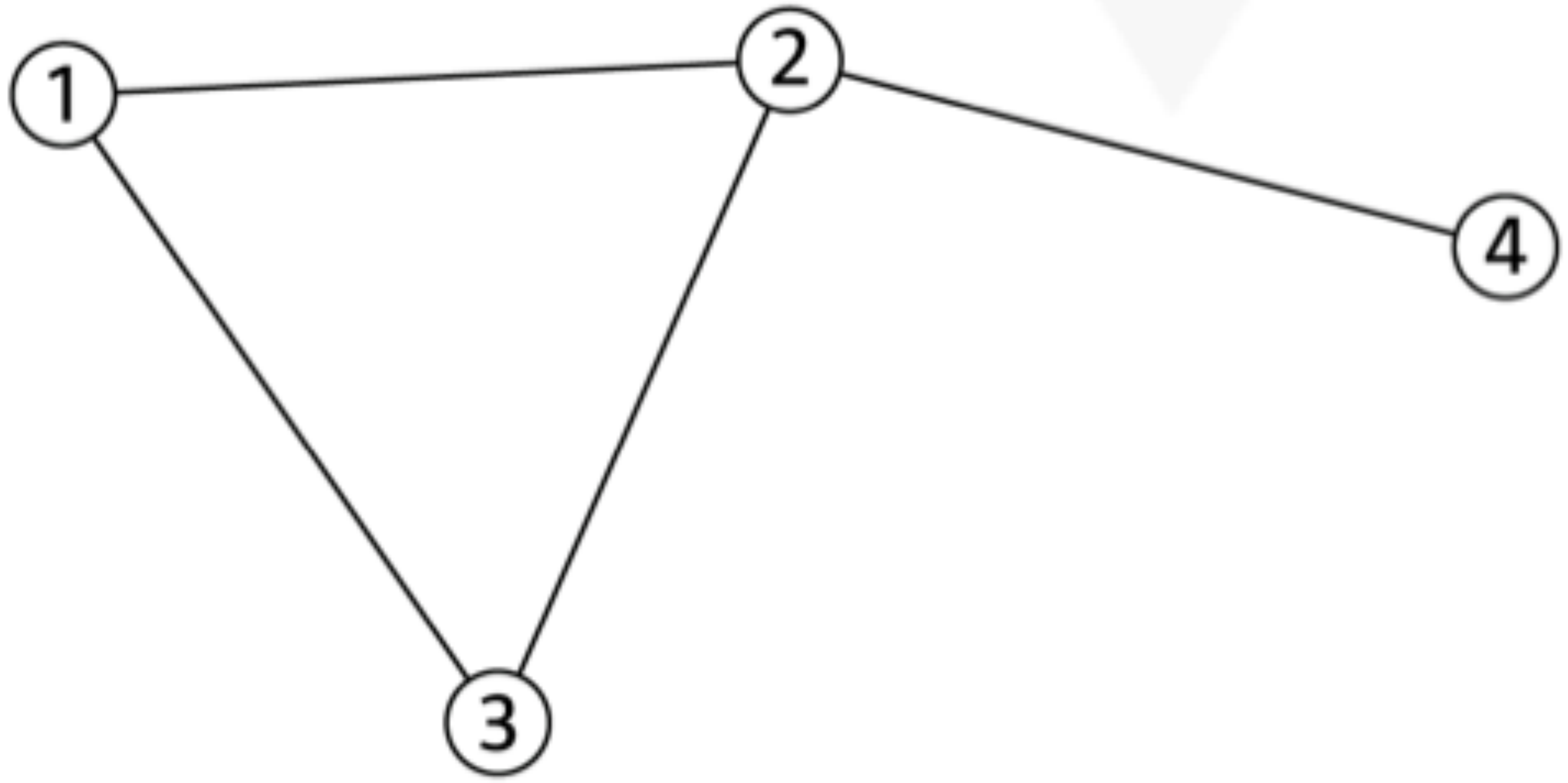
The degree k_i of a node i is the number of incident links

$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$



Every network has an average degree $\langle k \rangle$

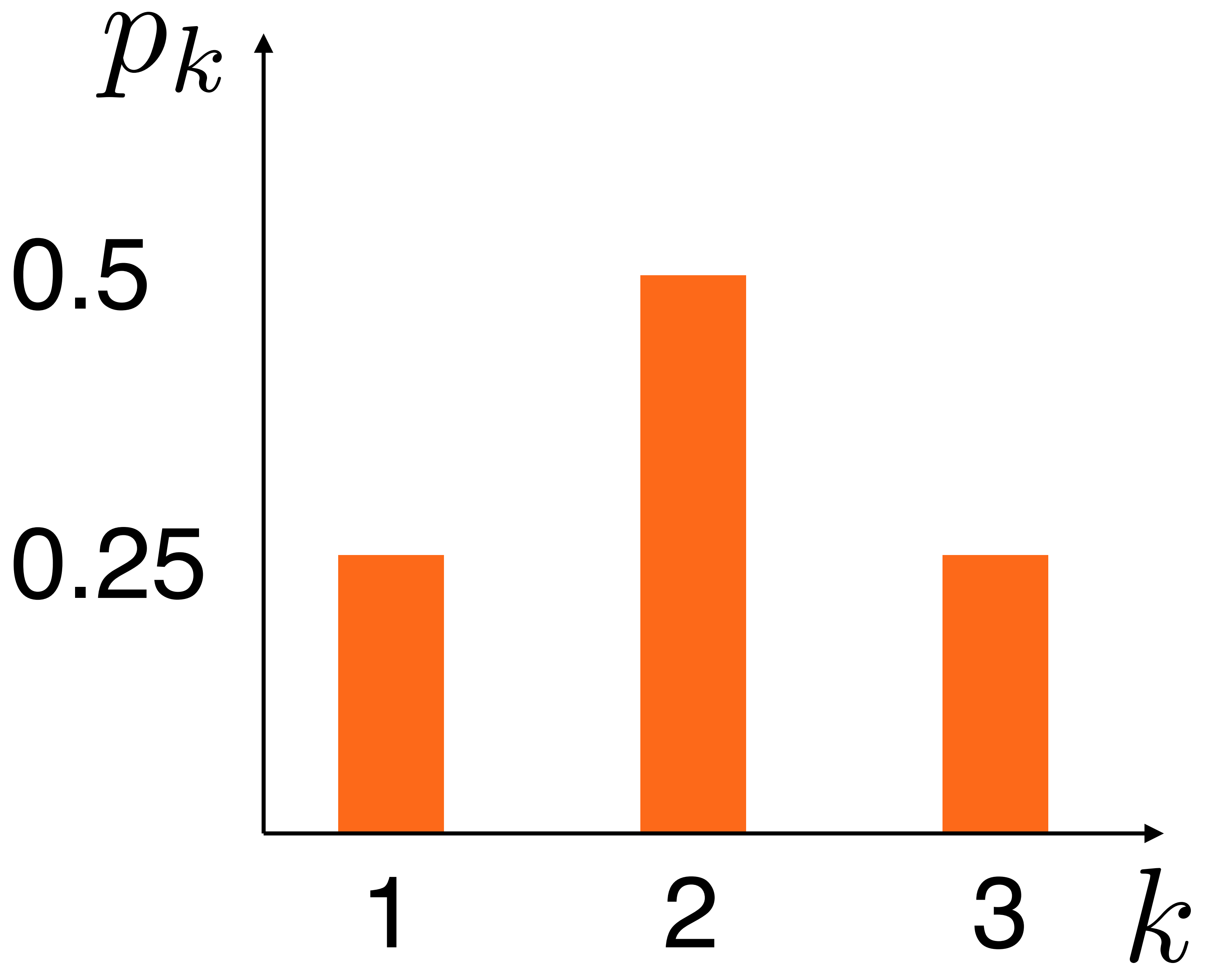
$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$

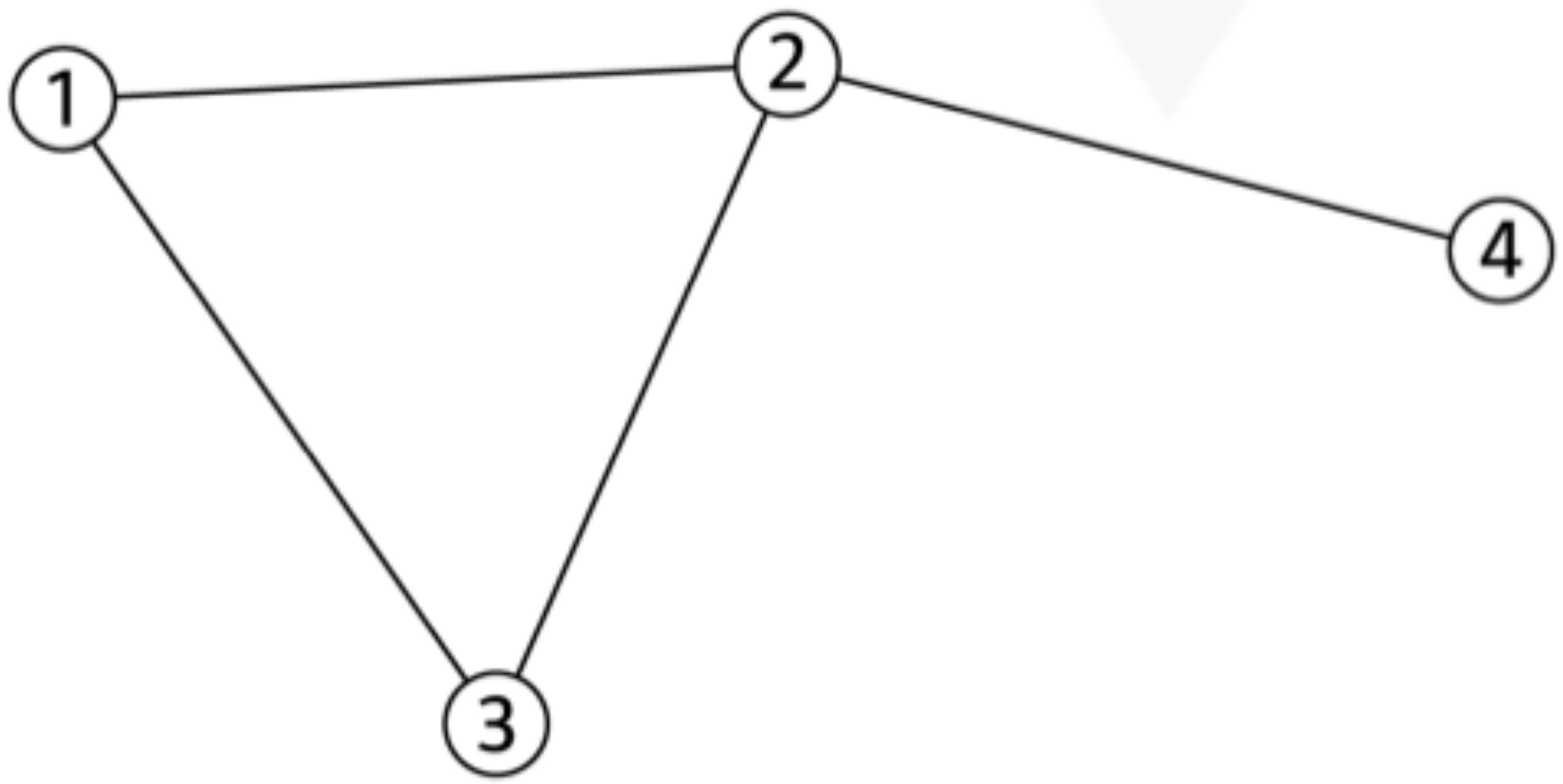


$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle = \frac{2+3+2+1}{4} = 2$$

The degree distribution $p_k = N_k/N$ captures the probabilities that a node has a certain degree

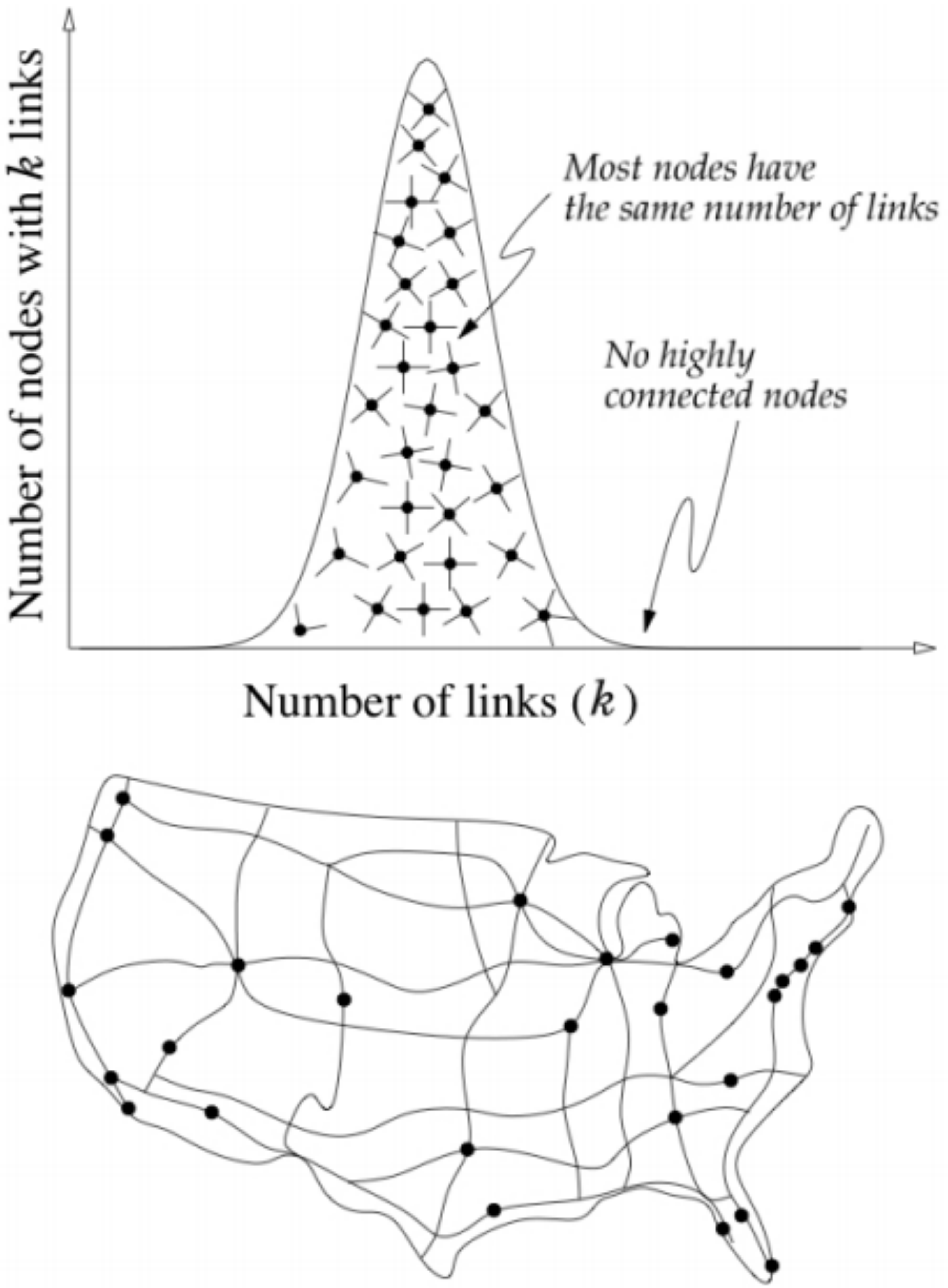
$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$





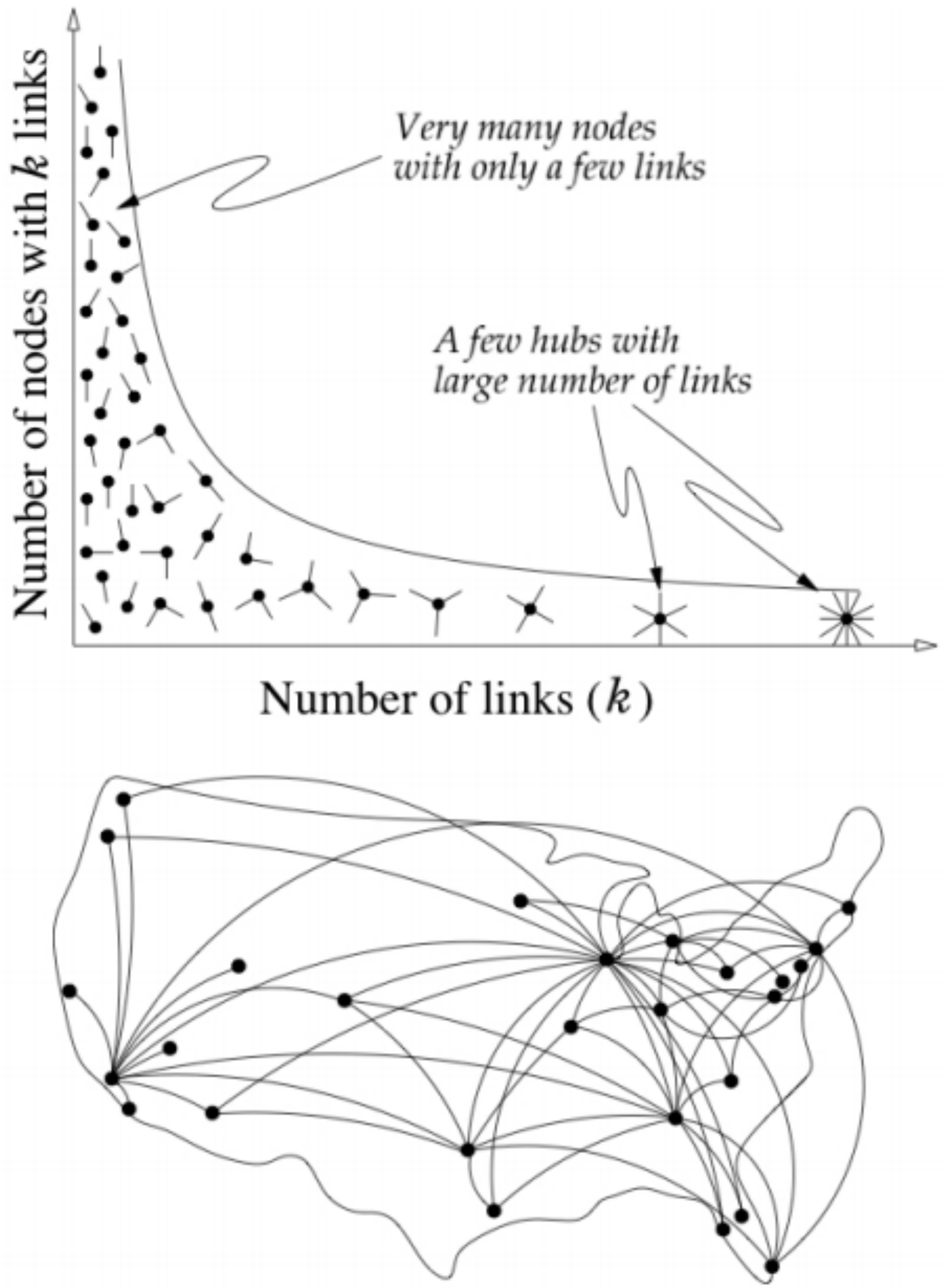
A network's degree distribution tells us something fundamental about how individuals in the system connect

Thin-tailed network



Streets

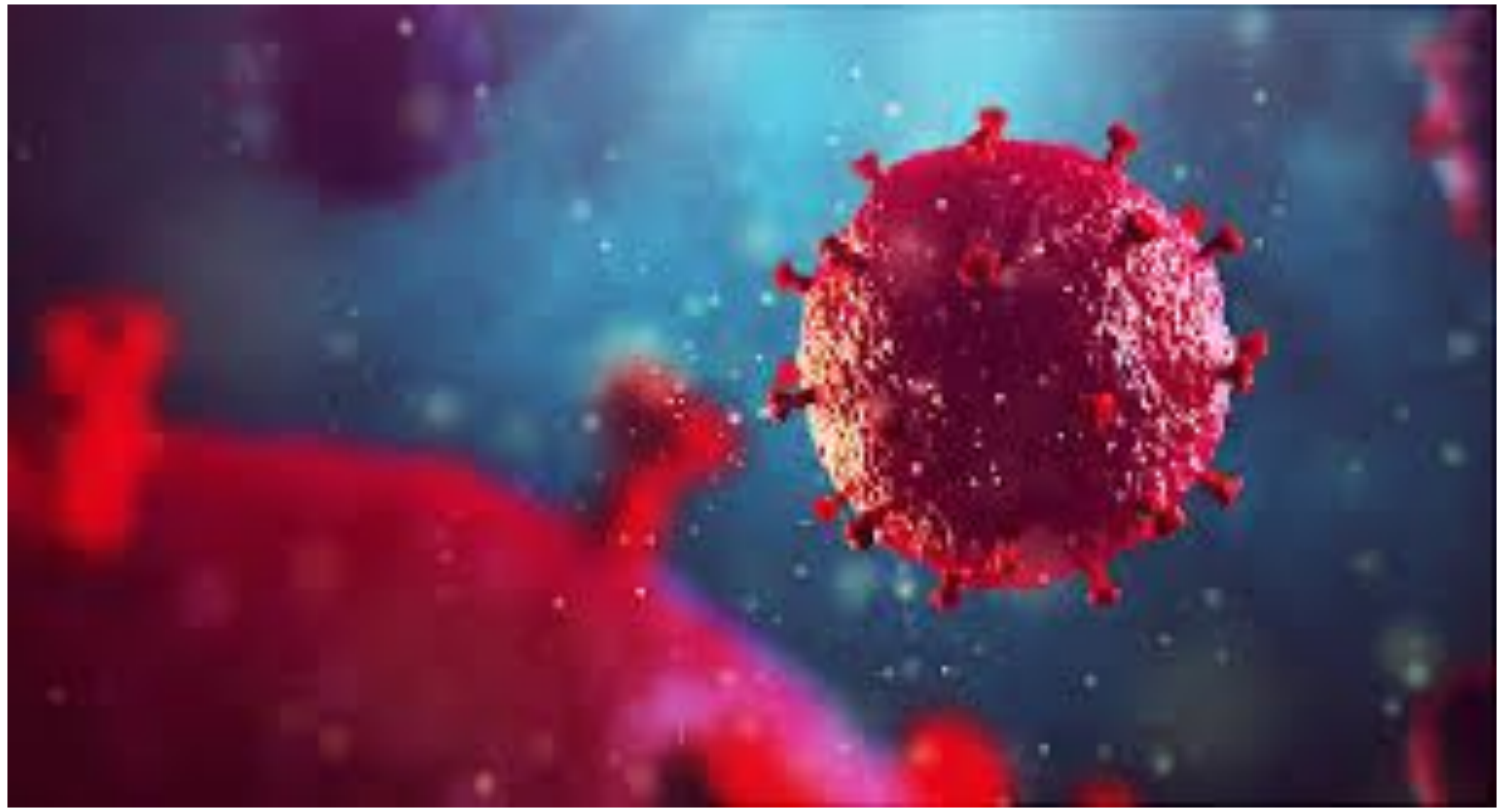
Heavy-tailed network



Airlines

A heavy-tailed degree distribution means we have hubs

Hubs completely change network processes
like epidemic spreading



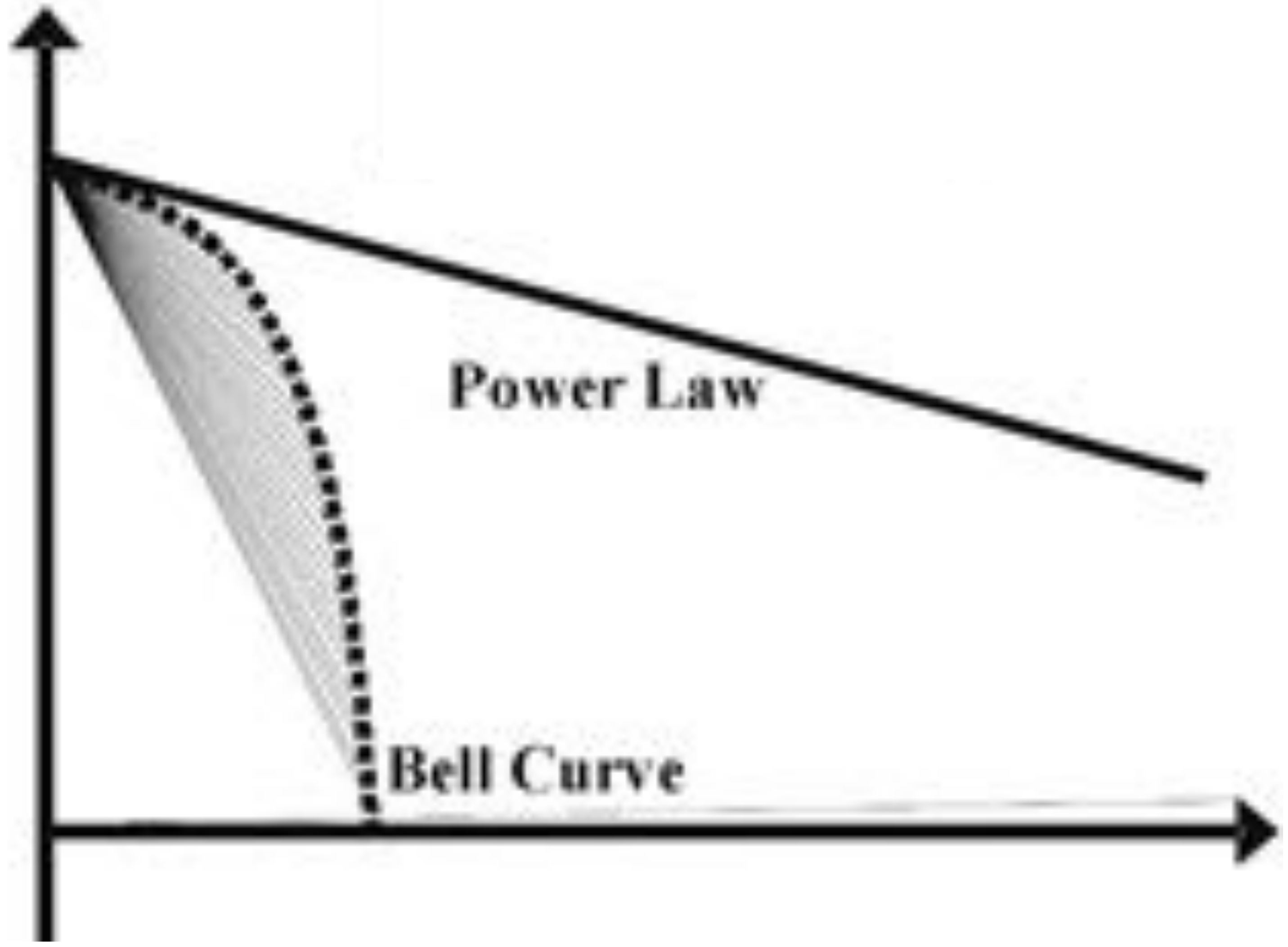
Heavy-tailed distributions govern the world

20th century statistics

Strange outliers

Focus on the head/center

21th century statistics add:



Not outliers but part of the system

Focus on the tail

Organizing principles of networks

Many networks are:

1) Heavy-tailed

2) Sparse

3) Small-world

4) Clustered

Organizing principles of networks

Many networks are:

1) Heavy-tailed

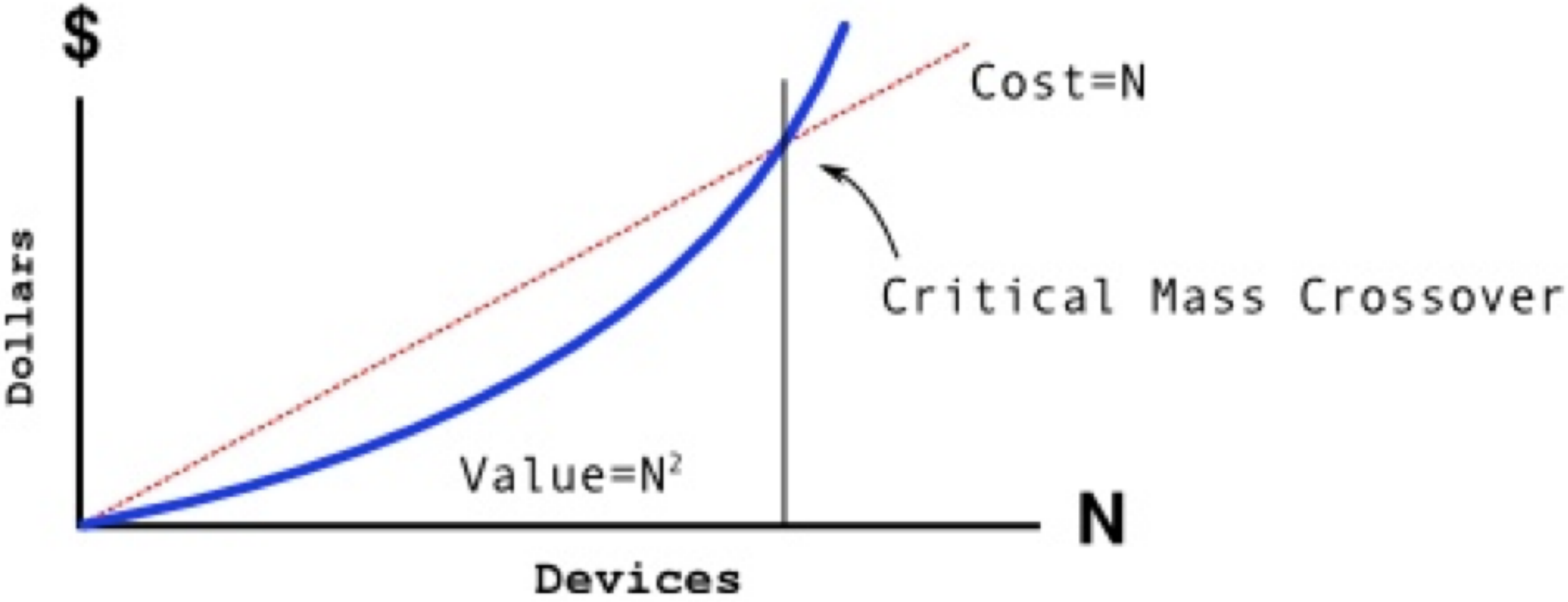
2) Sparse

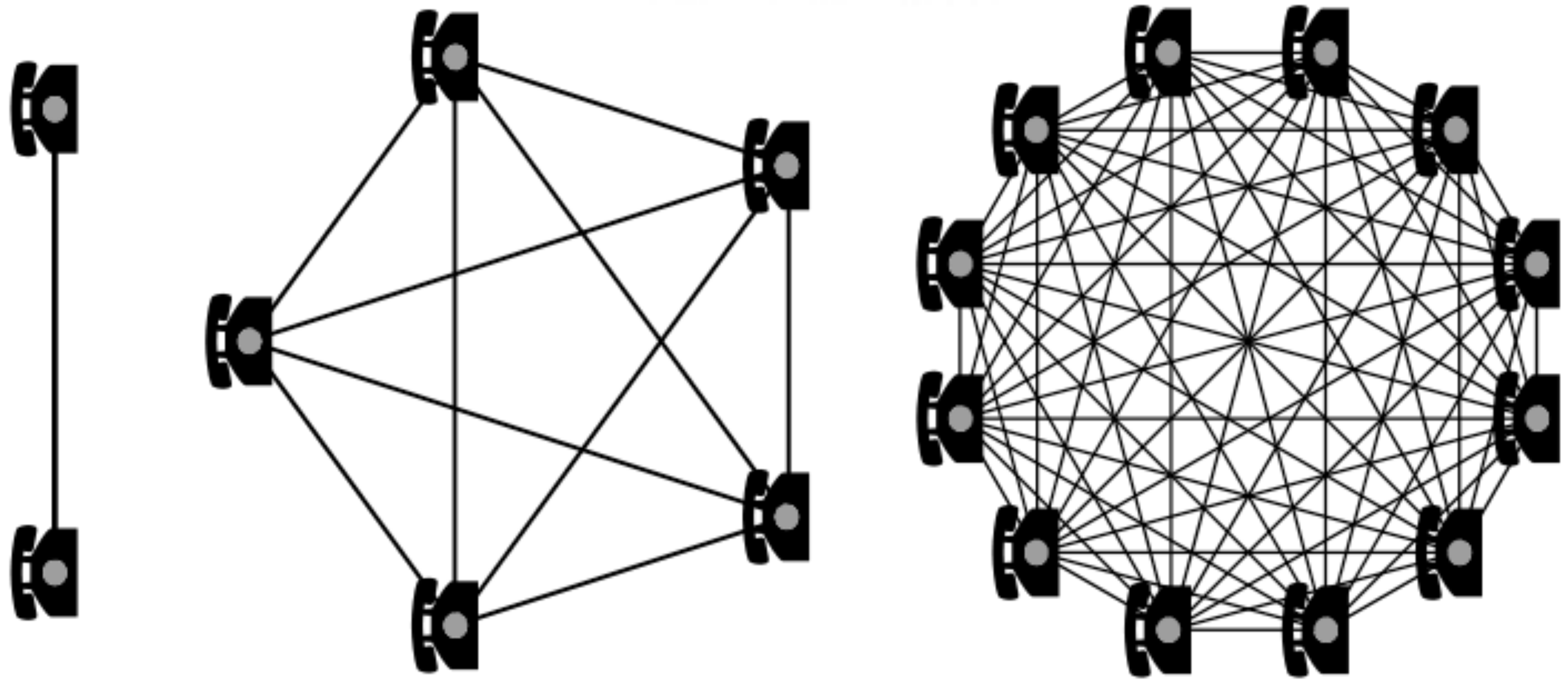
3) Small-world

4) Clustered

The person who bought the first fax, what was he/she thinking???

Metcalfe's law states that the value of a communication network increases with the square of users





$$L_{\max} = \frac{N(N-1)}{2} \sim N^2$$

Sparsity means: Although a lot of links are possible, only very few are actually there: $L \ll L_{\max}$

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
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-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	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

Sparsity means: Although a lot of links are possible, only very few are actually there: $L \ll L_{\max}$

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
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-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	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

Also: $\langle k \rangle \ll \langle k \rangle_{\max} = N - 1$

Organizing principles of networks

Many networks are:

1) Heavy-tailed

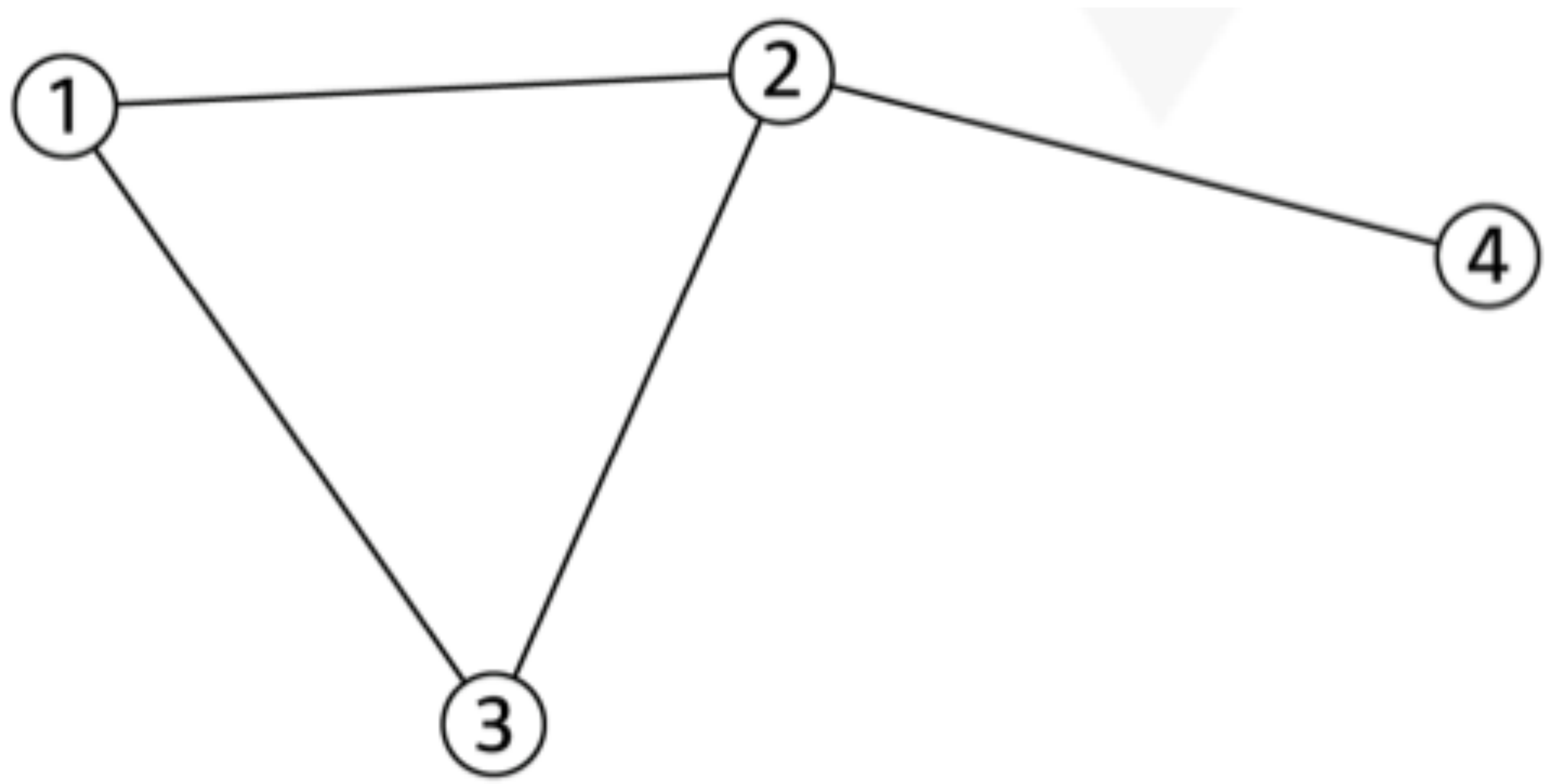
2) Sparse

3) Small-world

4) Clustered

To formalize this idea, we need some definitions:

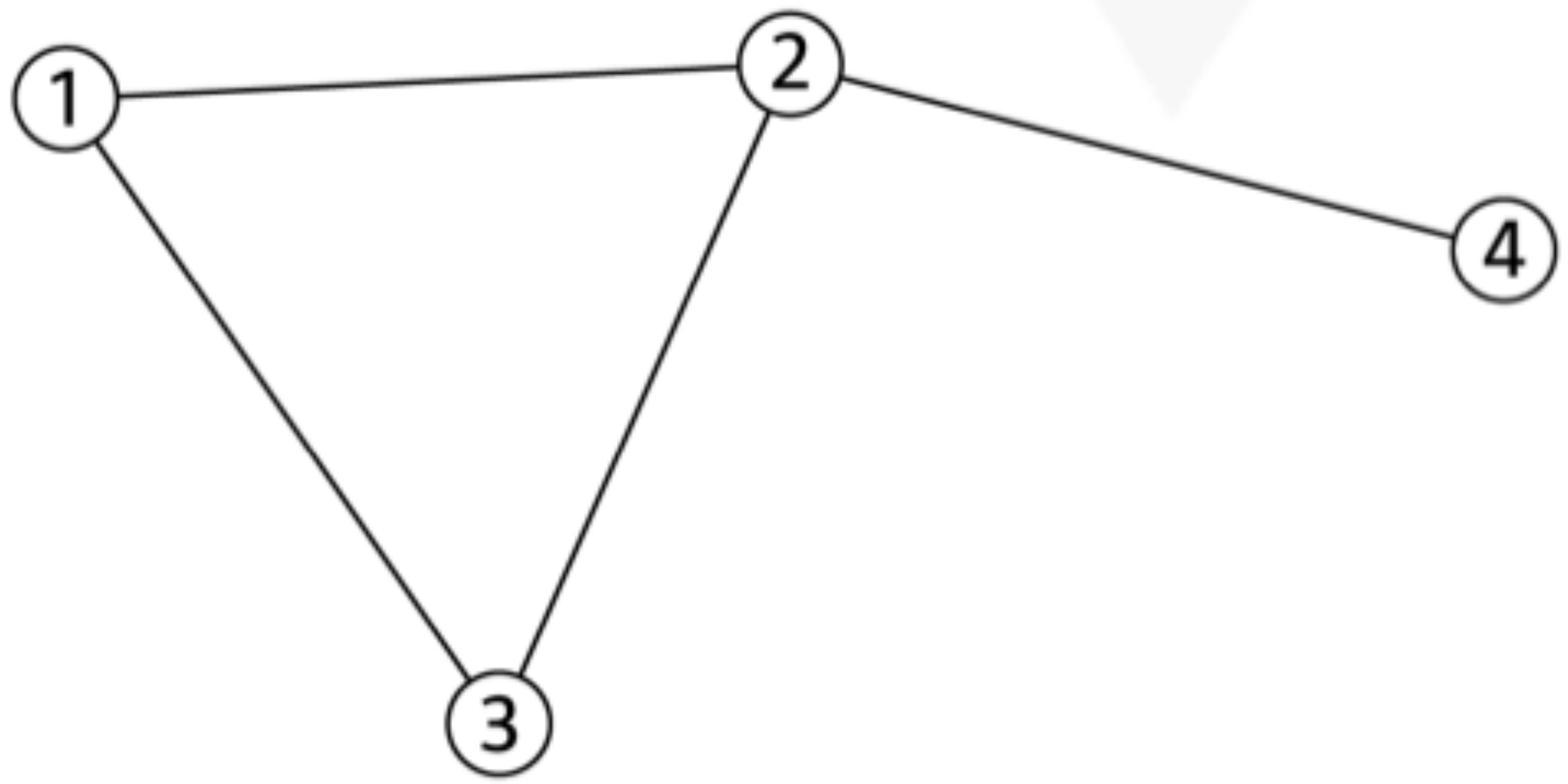
Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$



To formalize this idea, we need some definitions:

Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$

Path: A walk where no node is repeated $\{n_1, n_3, n_2, n_4\}$

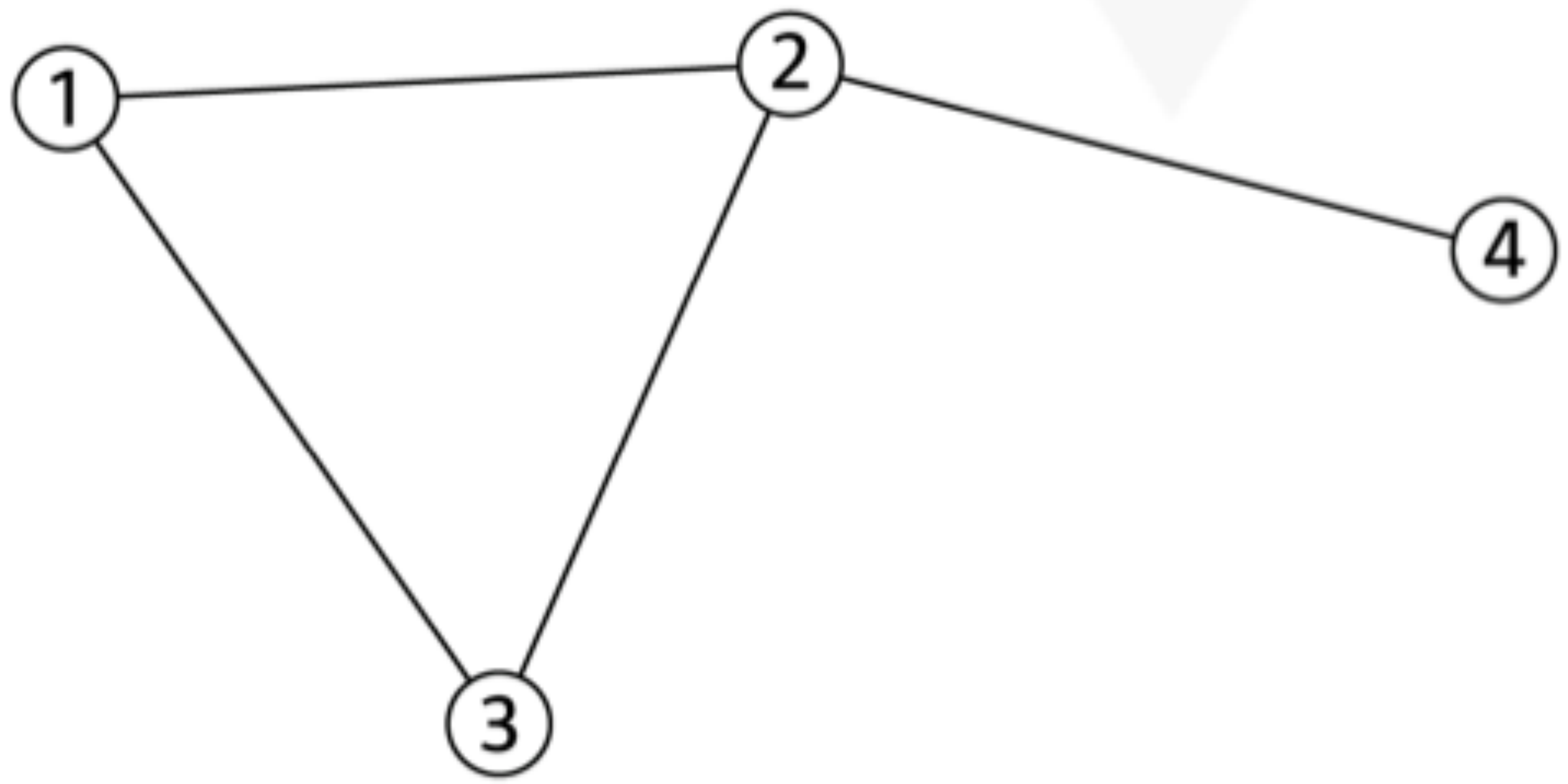


To formalize this idea, we need some definitions:

Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$

Path: A walk where no node is repeated $\{n_1, n_3, n_2, n_4\}$

Shortest path: A path of minimal length $\{n_1, n_2, n_4\}$



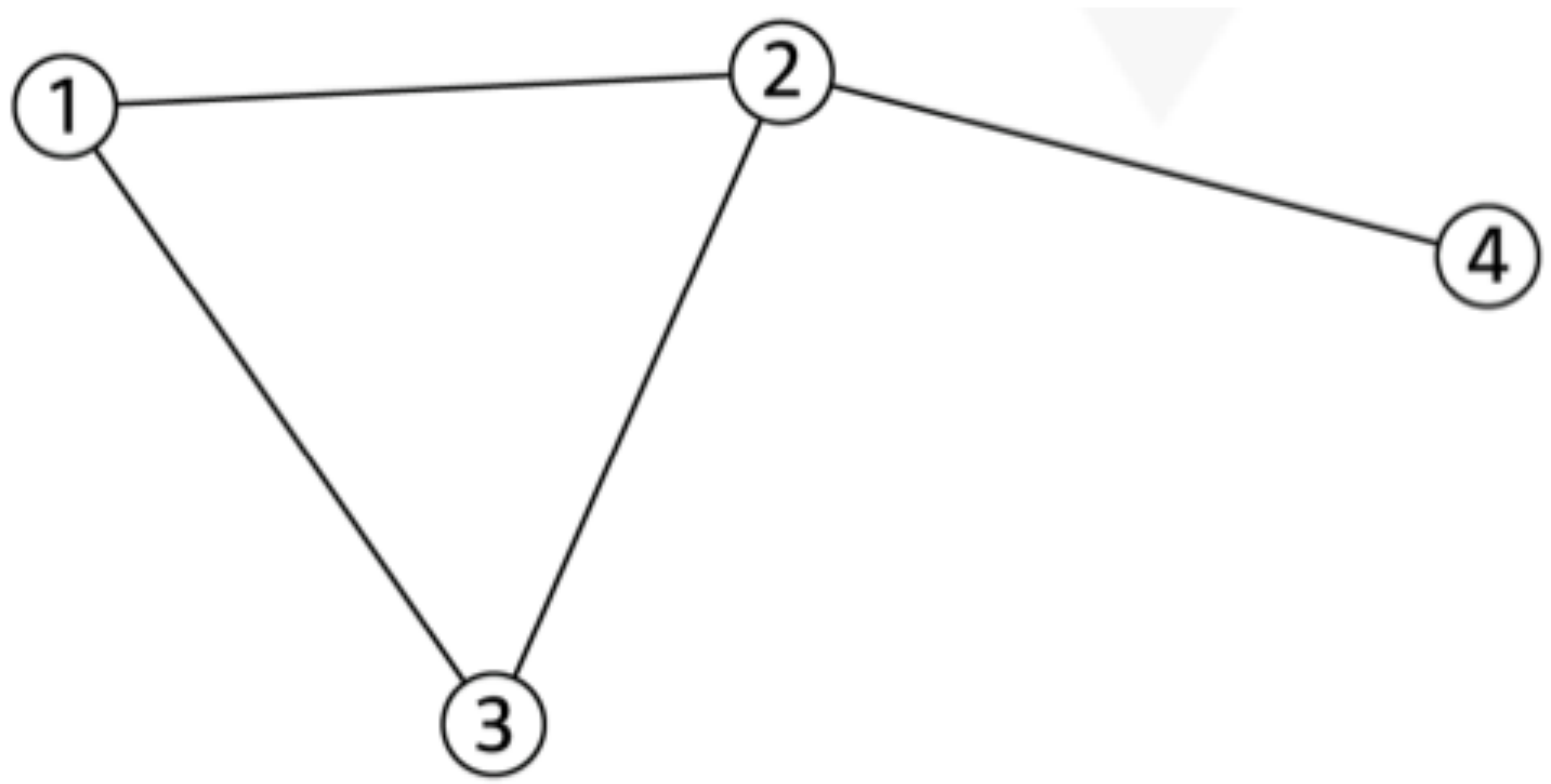
To formalize this idea, we need some definitions:

Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$

Path: A walk where no node is repeated $\{n_1, n_3, n_2, n_4\}$

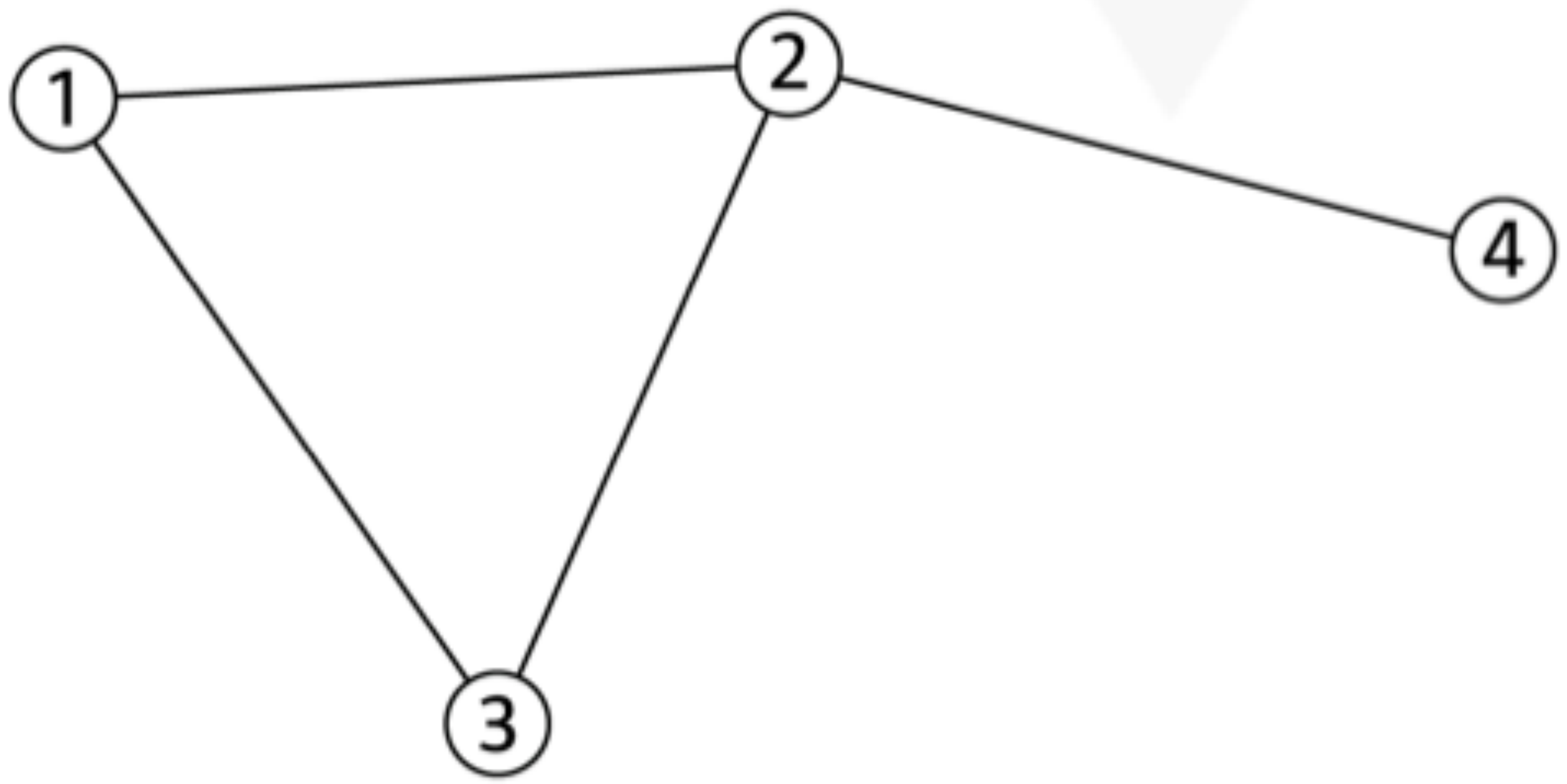
Shortest path: A path of minimal length $\{n_1, n_2, n_4\}$

Graph distance: Length of shortest path $d_{1,4} = 2$



The average path length ℓ is the mean graph distance over all pairs of nodes

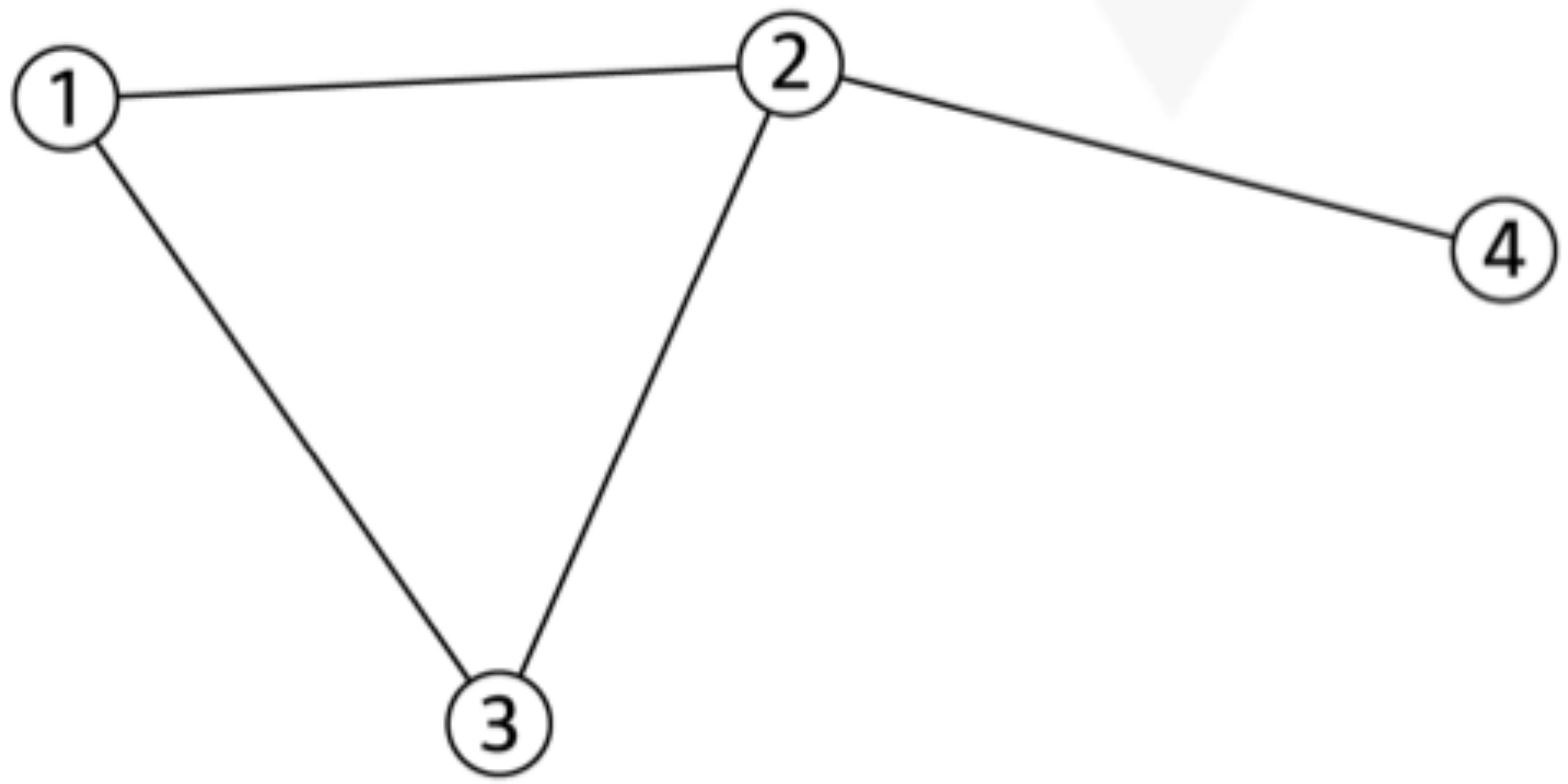
$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$



The average path length ℓ is the mean graph distance over all pairs of nodes

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

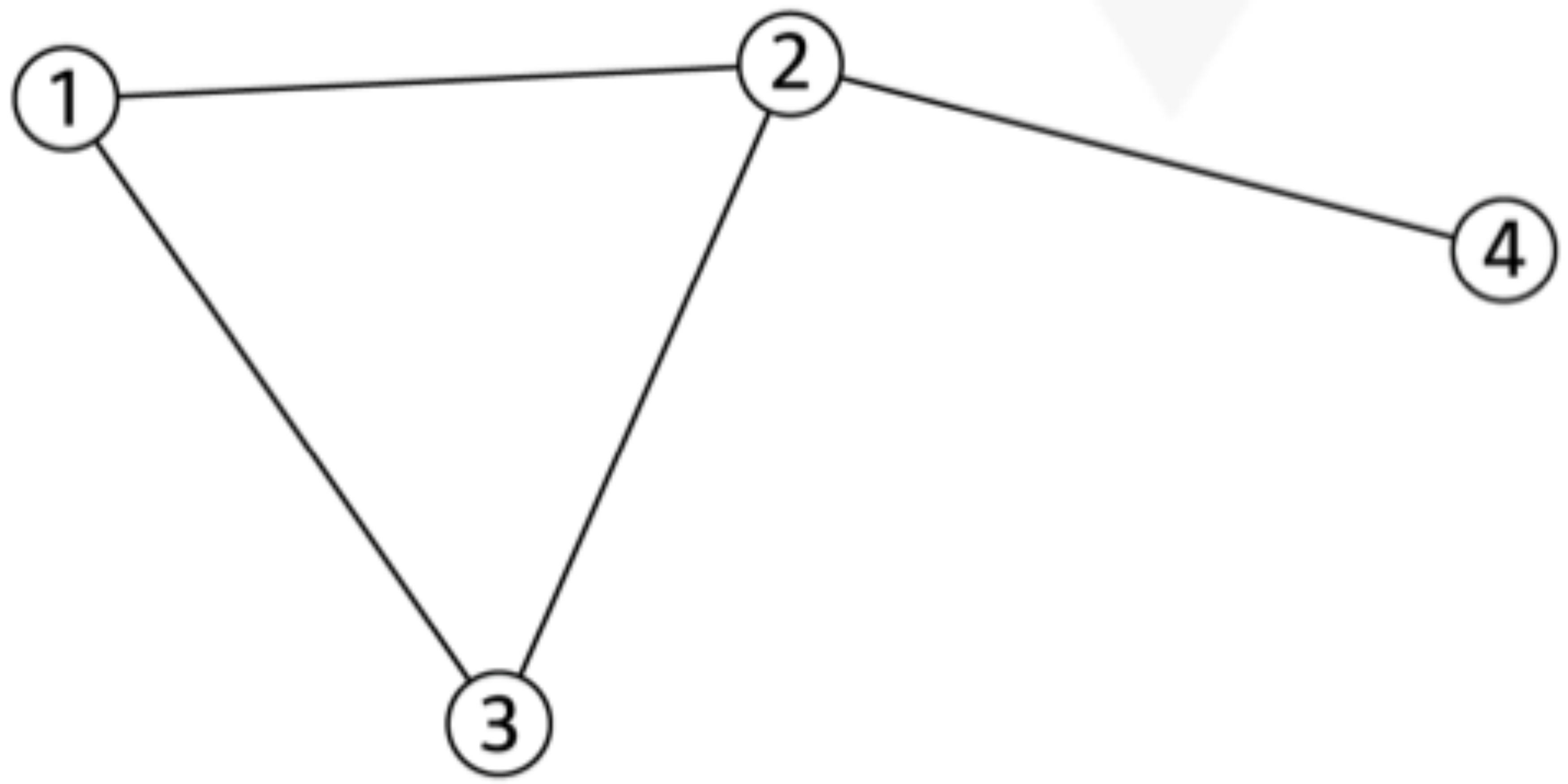
$$\ell = \frac{4+3+4+5}{4(4-1)} = \frac{4}{3}$$



The diameter D is the maximum length of shortest paths

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

$$\ell = \frac{4+3+4+5}{4(4-1)} = \frac{4}{3}$$

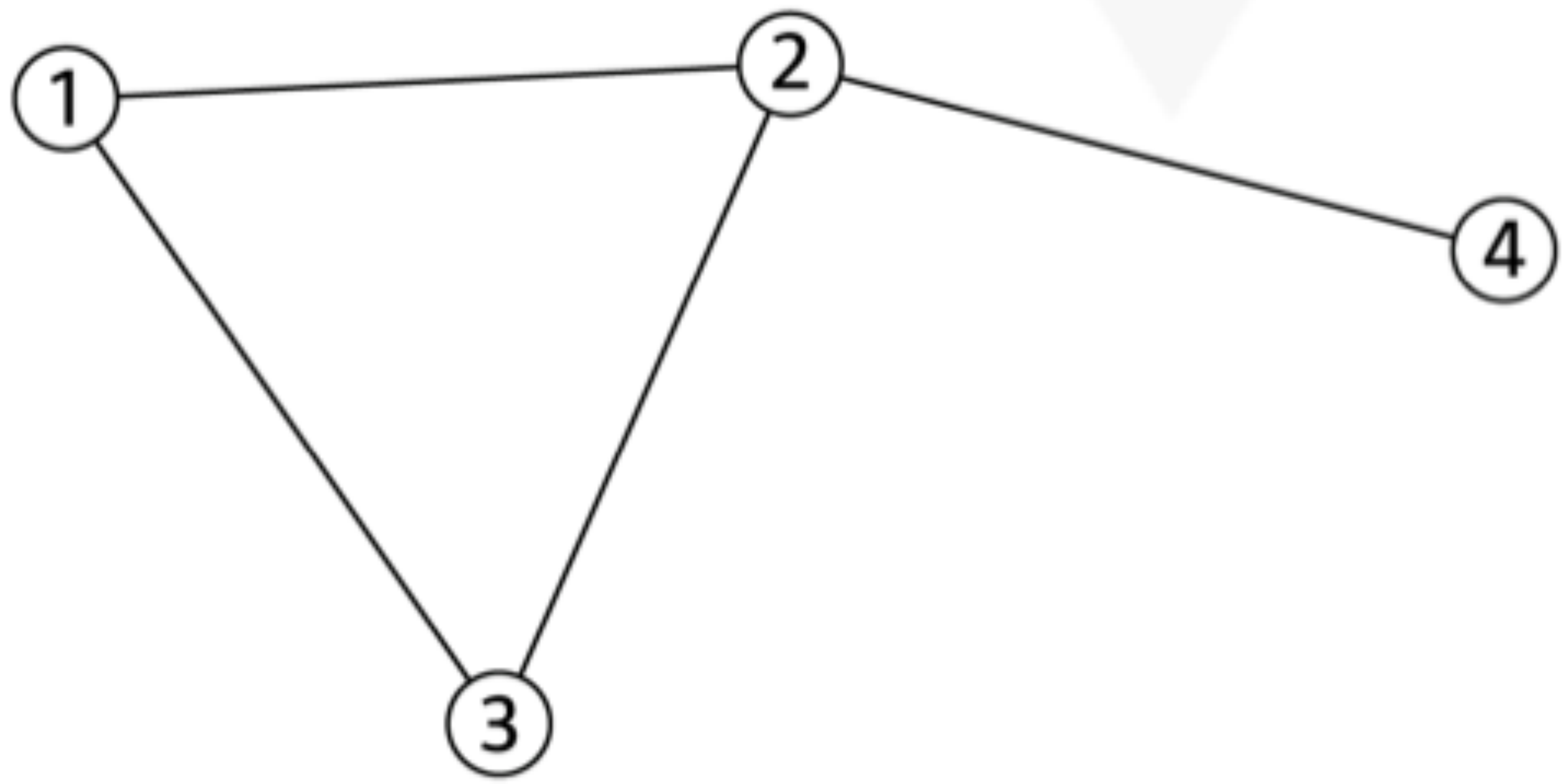


The diameter D is the maximum length of shortest paths

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

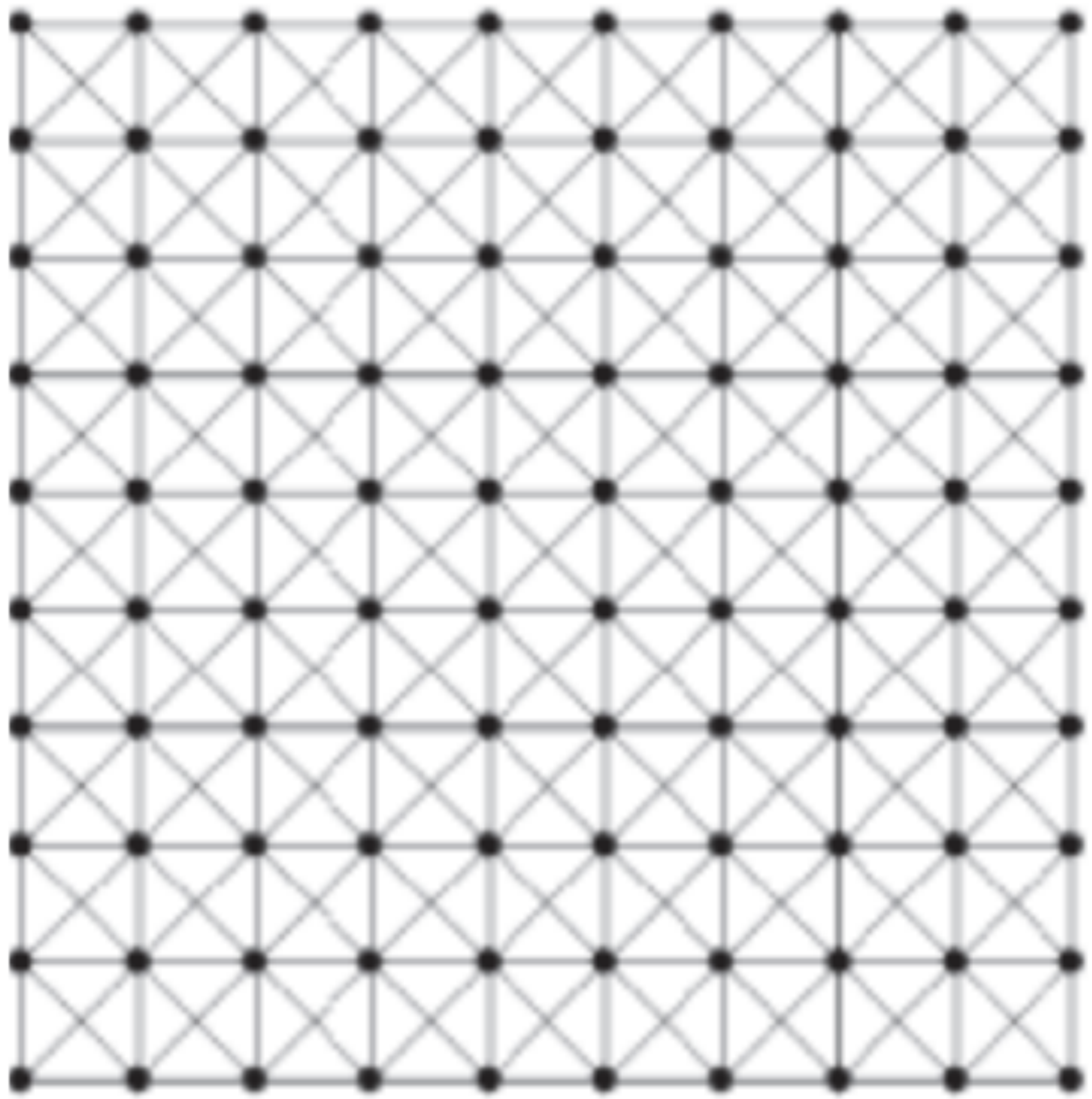
$$\ell = \frac{4+3+4+5}{4(4-1)} = \frac{4}{3}$$

$$D = 2$$

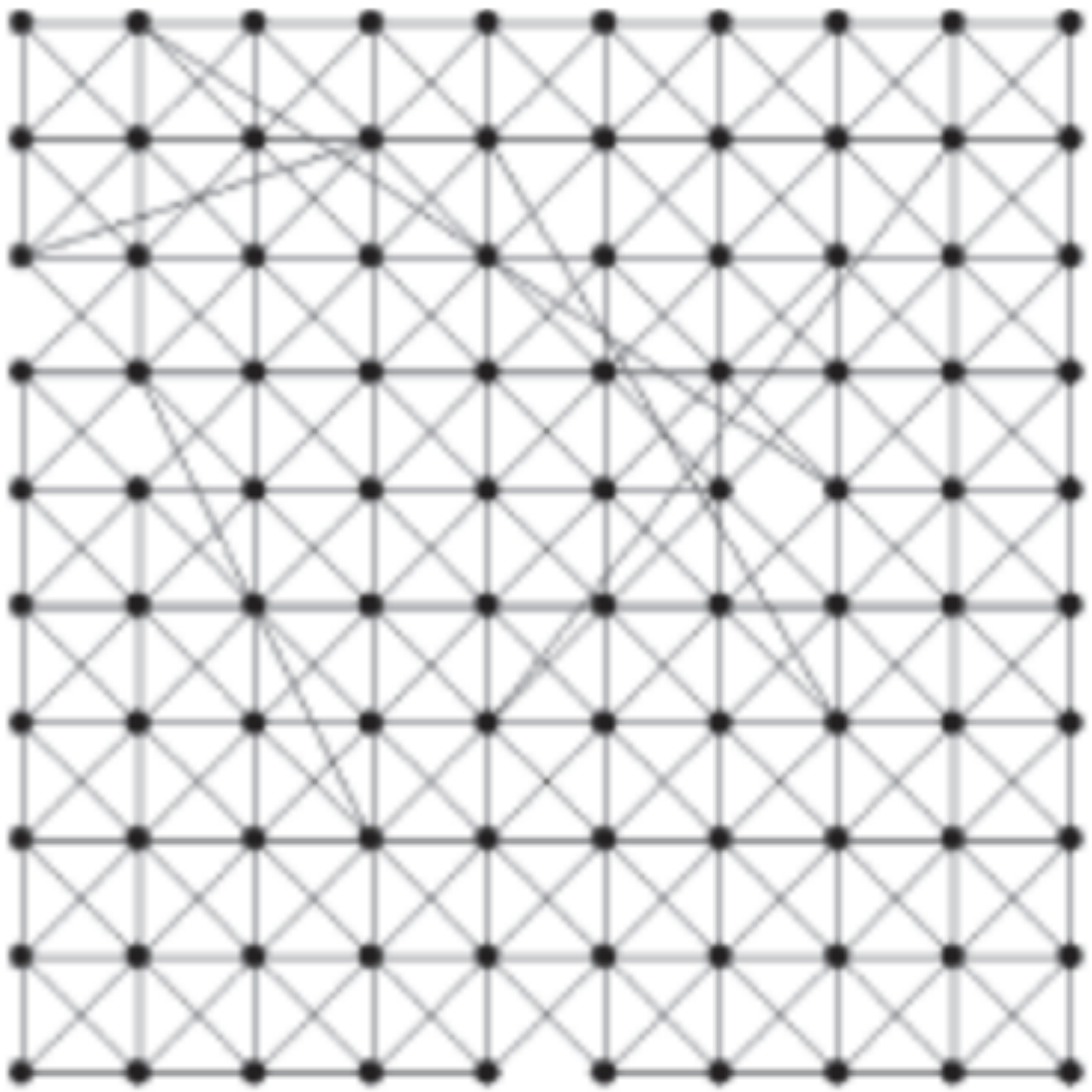


A small world network has short average path length

$$\ell \sim N \qquad\qquad \ell \sim \log N$$



Regular network ϕ=0



Small-world network ϕ=0.01

Organizing principles of networks

Many networks are:

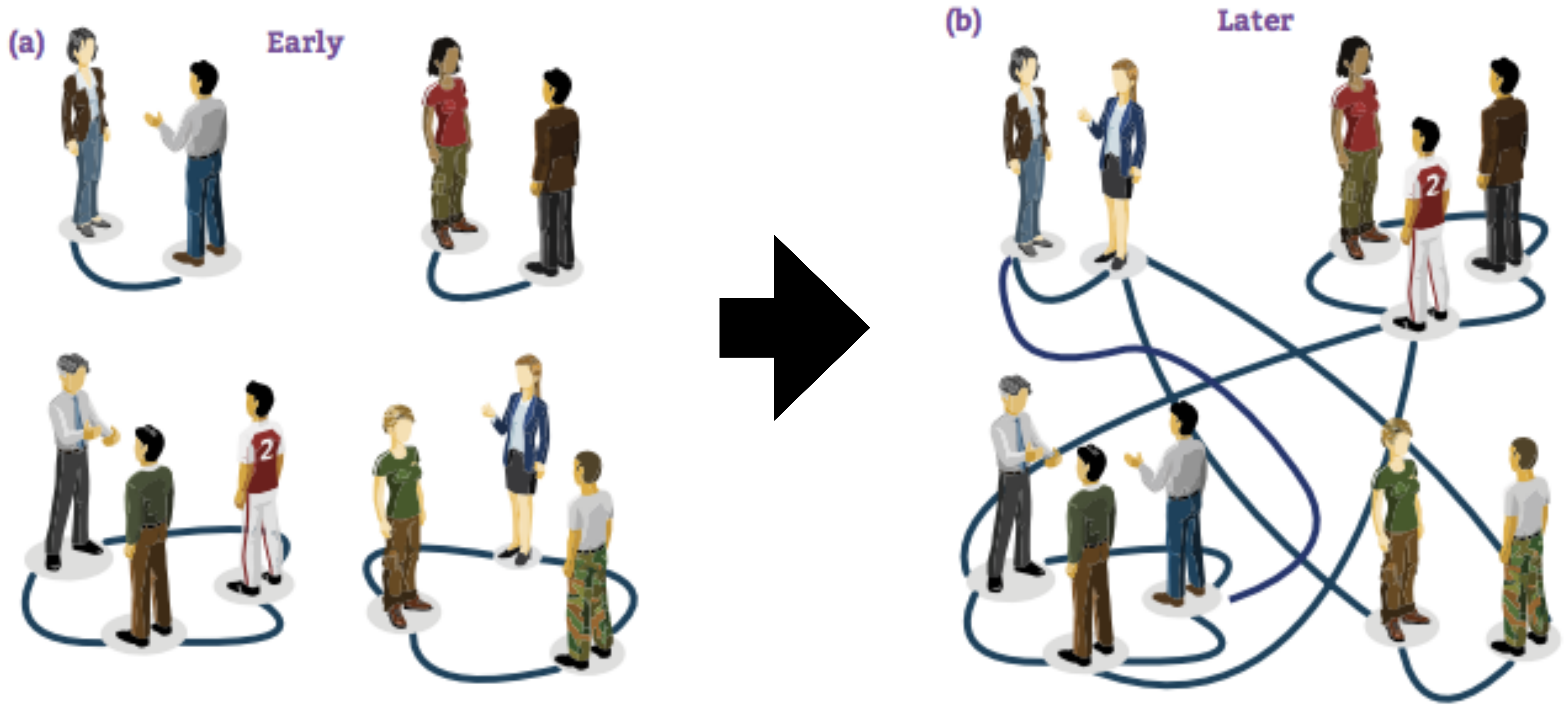
1) Heavy-tailed

2) Sparse

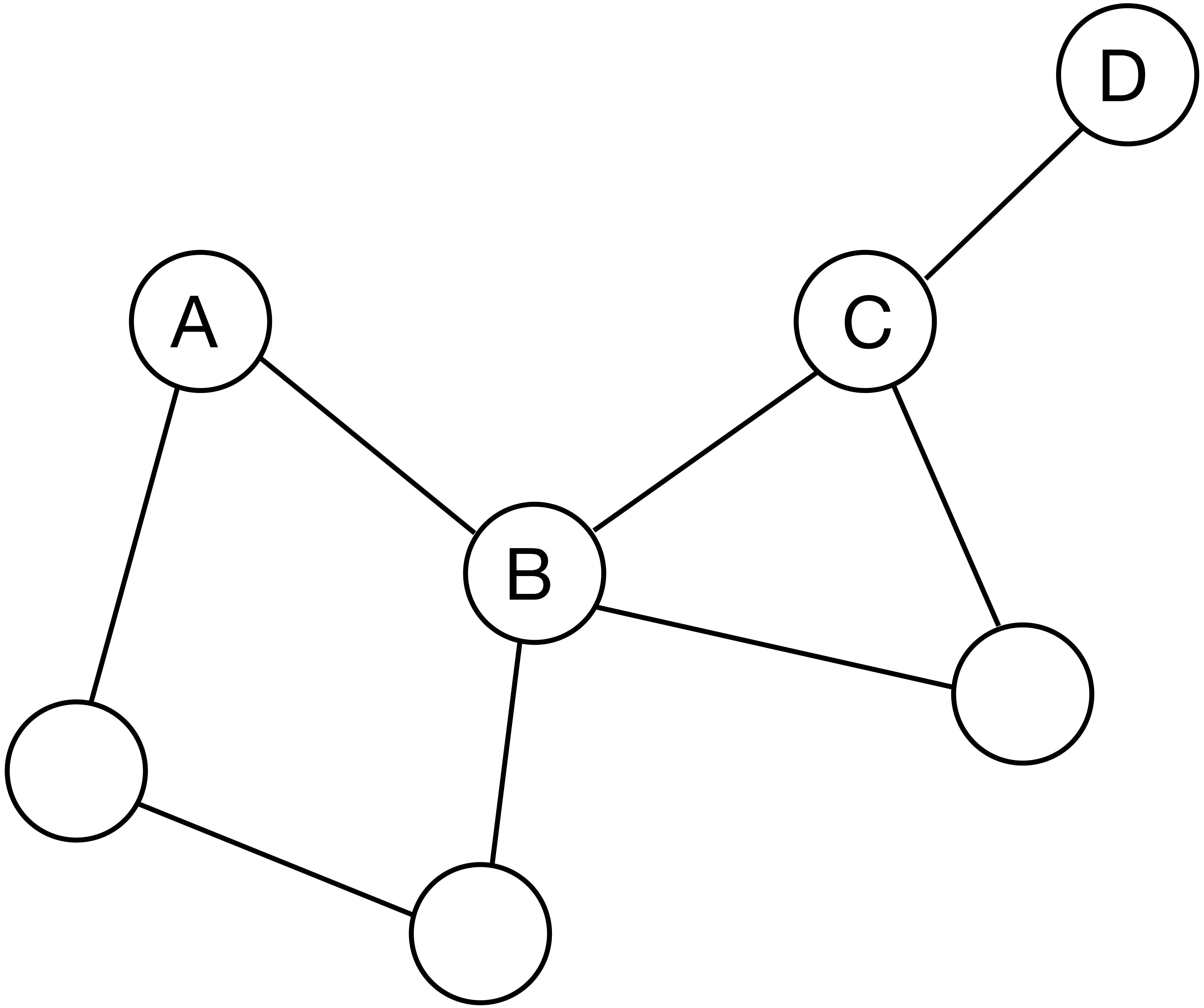
3) Small-world

4) Clustered

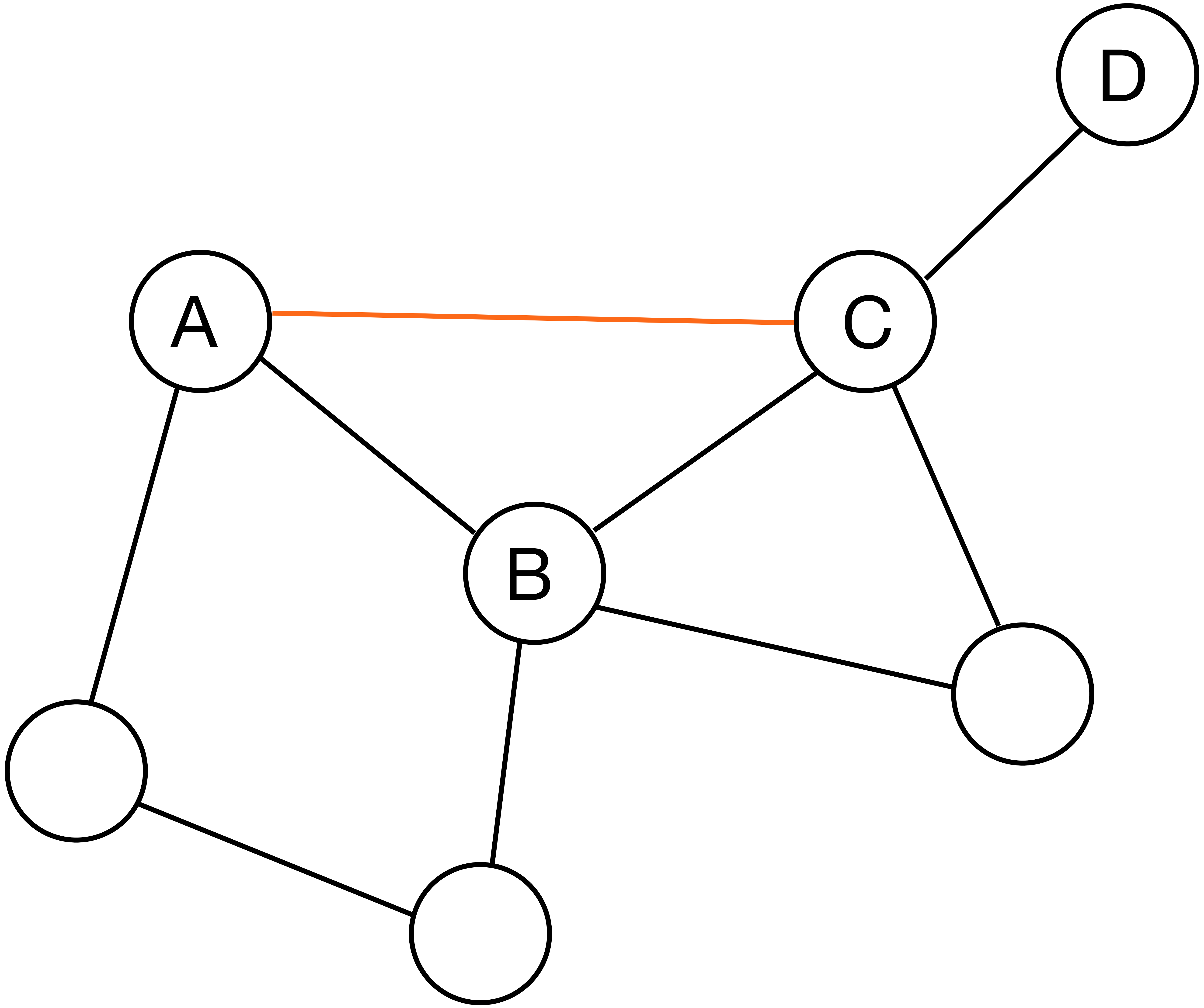
In a cocktail party you introduce each other to new people



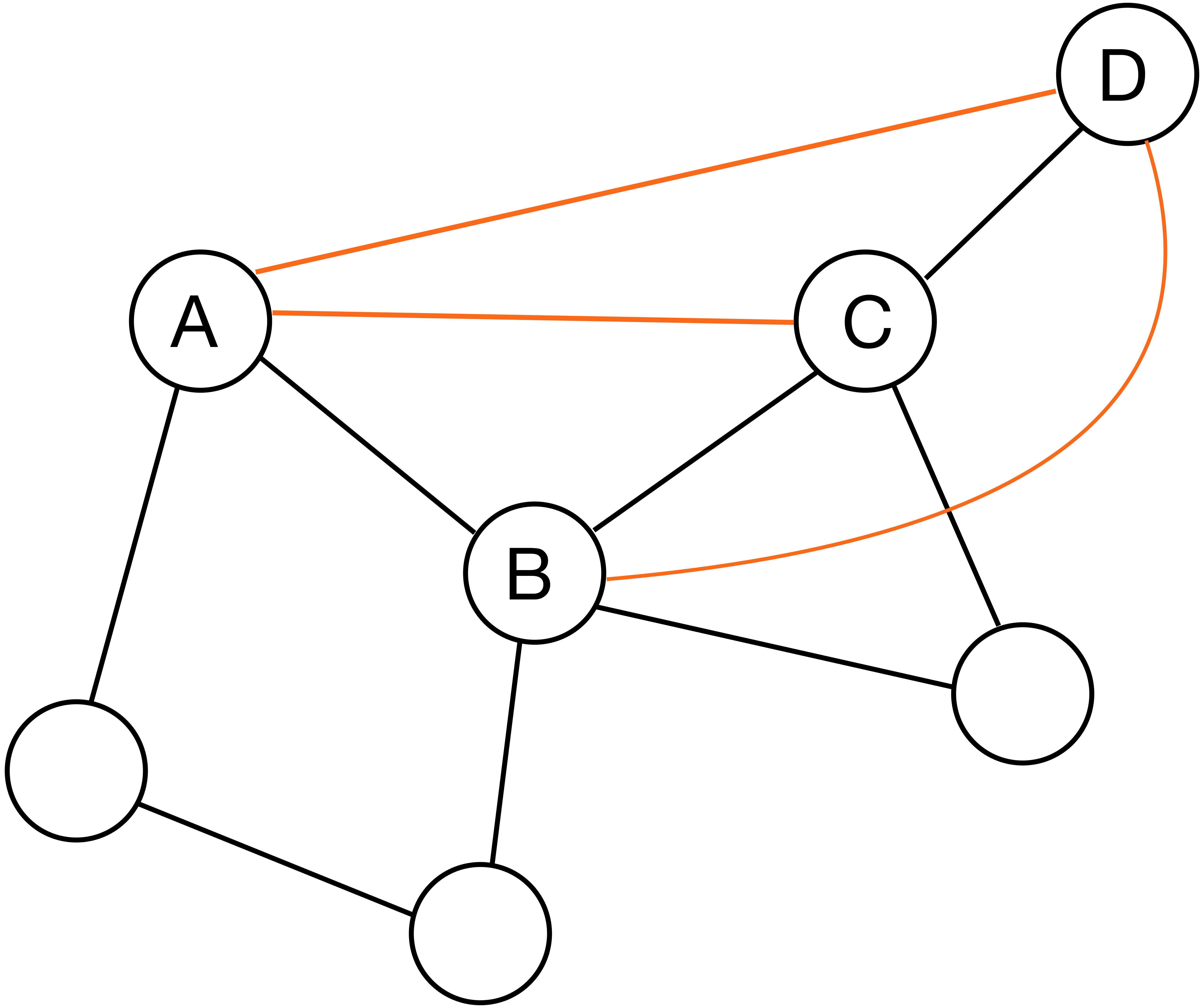
Who are more likely to connect: A—C or A—D?



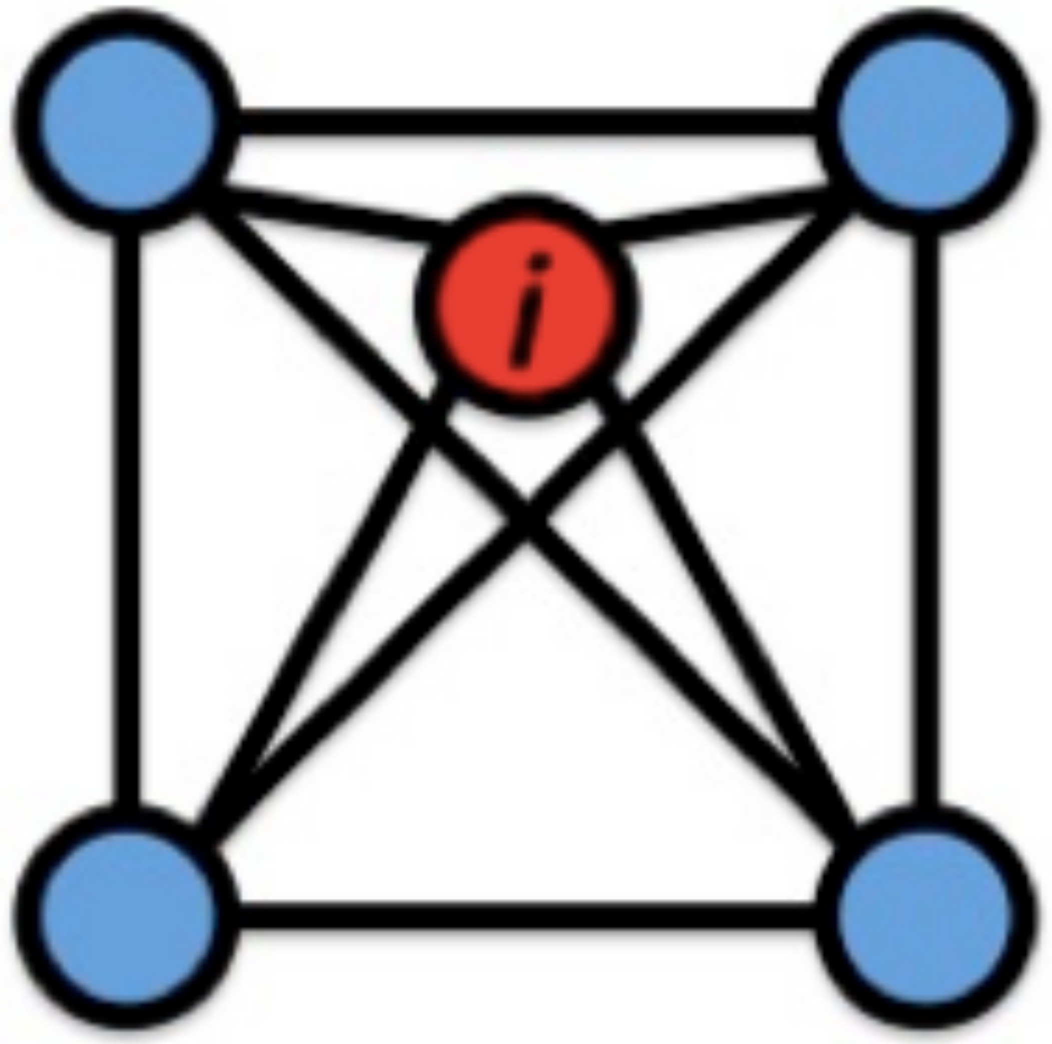
A and C already have a common friend, B



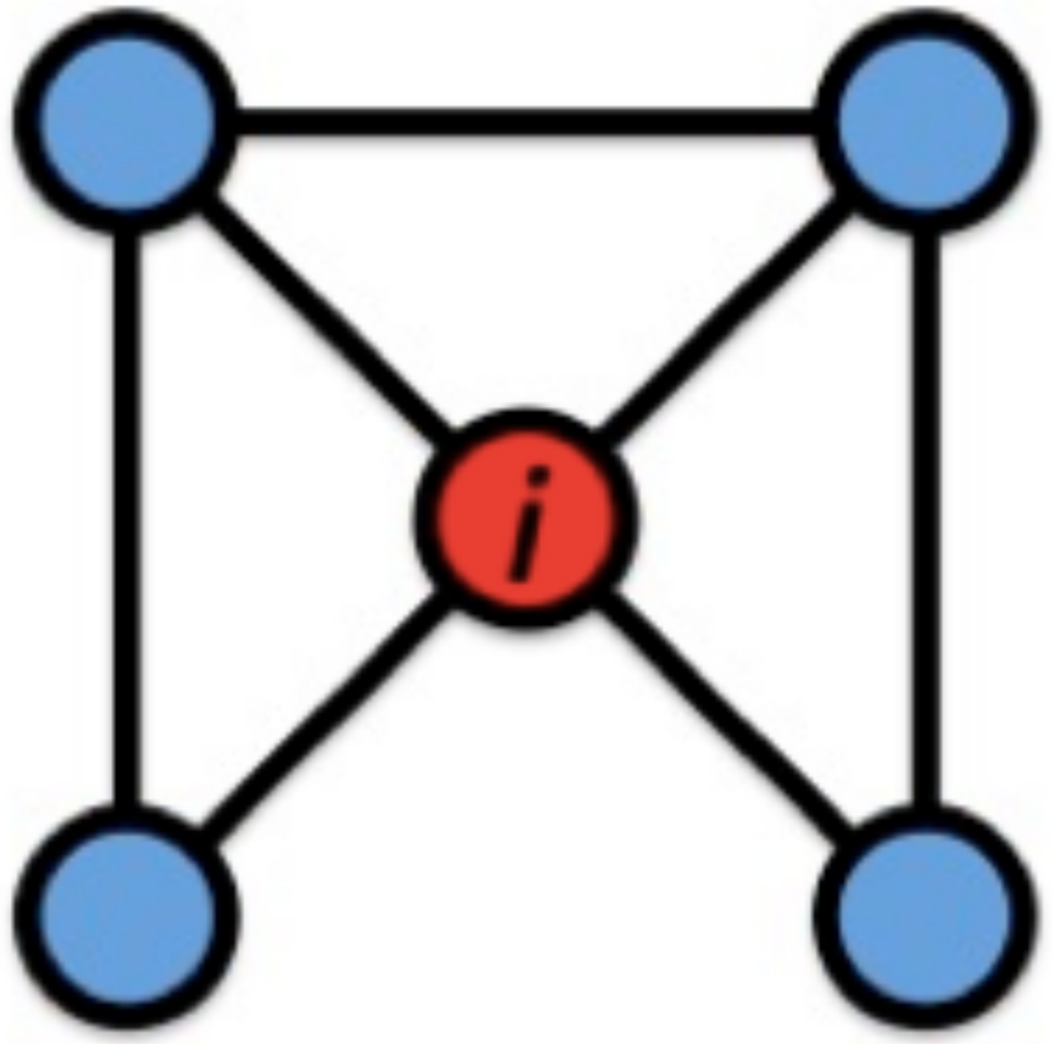
Closing open triangles is called triadic closure



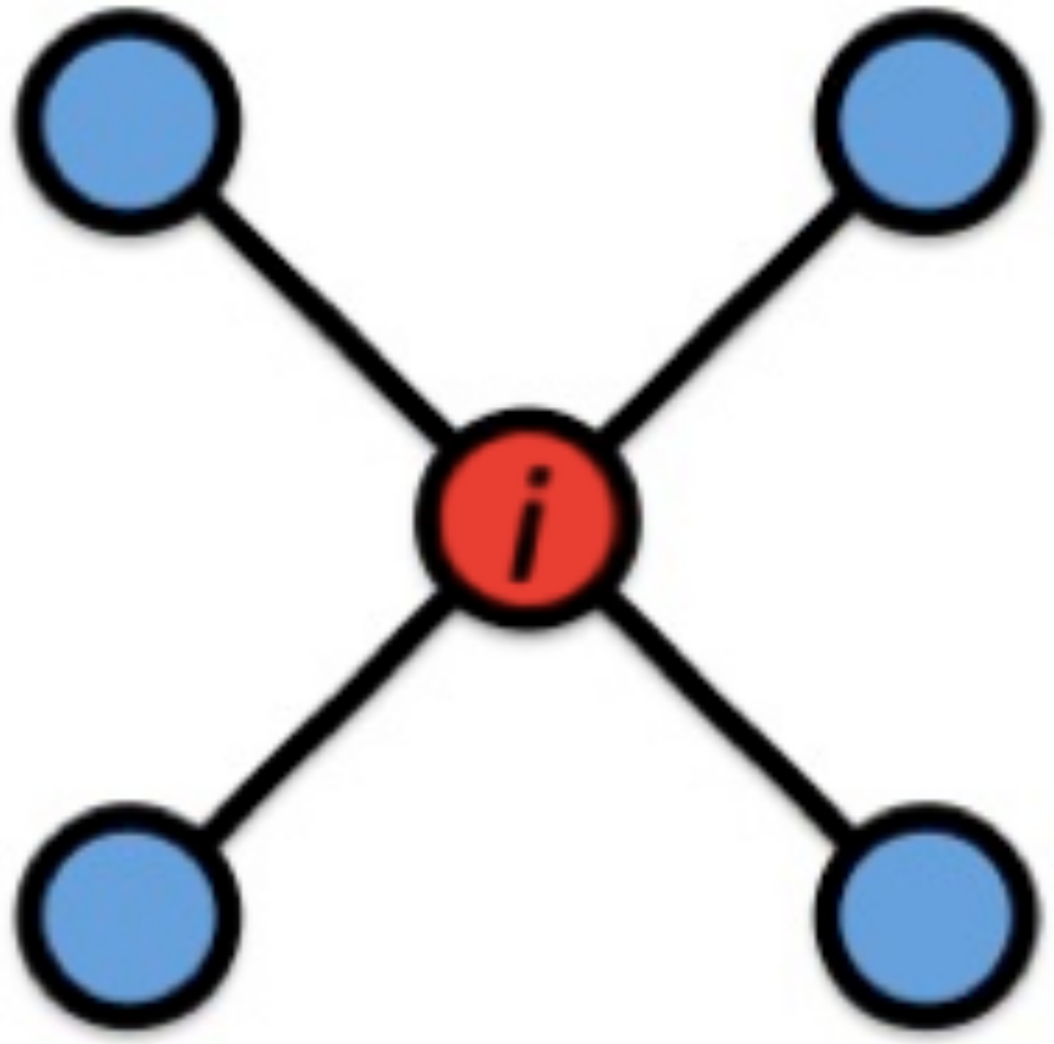
The clustering coefficient measures the fraction of your neighbor pairs who are linked



$$c_i = 1$$



$$c_i = 1/2$$



$$c_i = 0$$

By definition, $0 \leq c_i \leq 1$

Organizing principles of networks

Many networks are:

1) Heavy-tailed

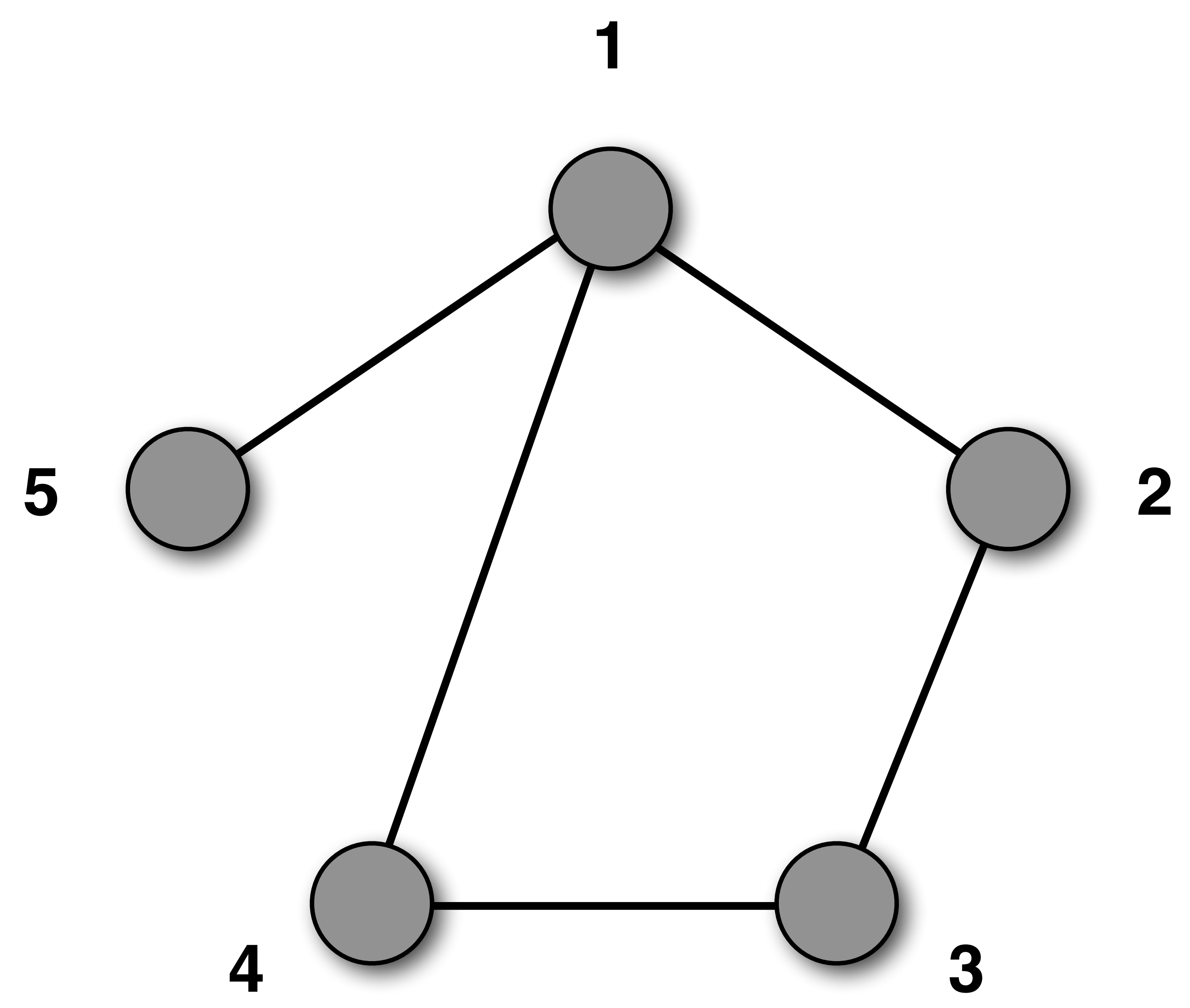
2) Sparse

3) Small-world

4) Clustered

Network data structures

The adjacency matrix stores all possible connections
Realized connections get a 1

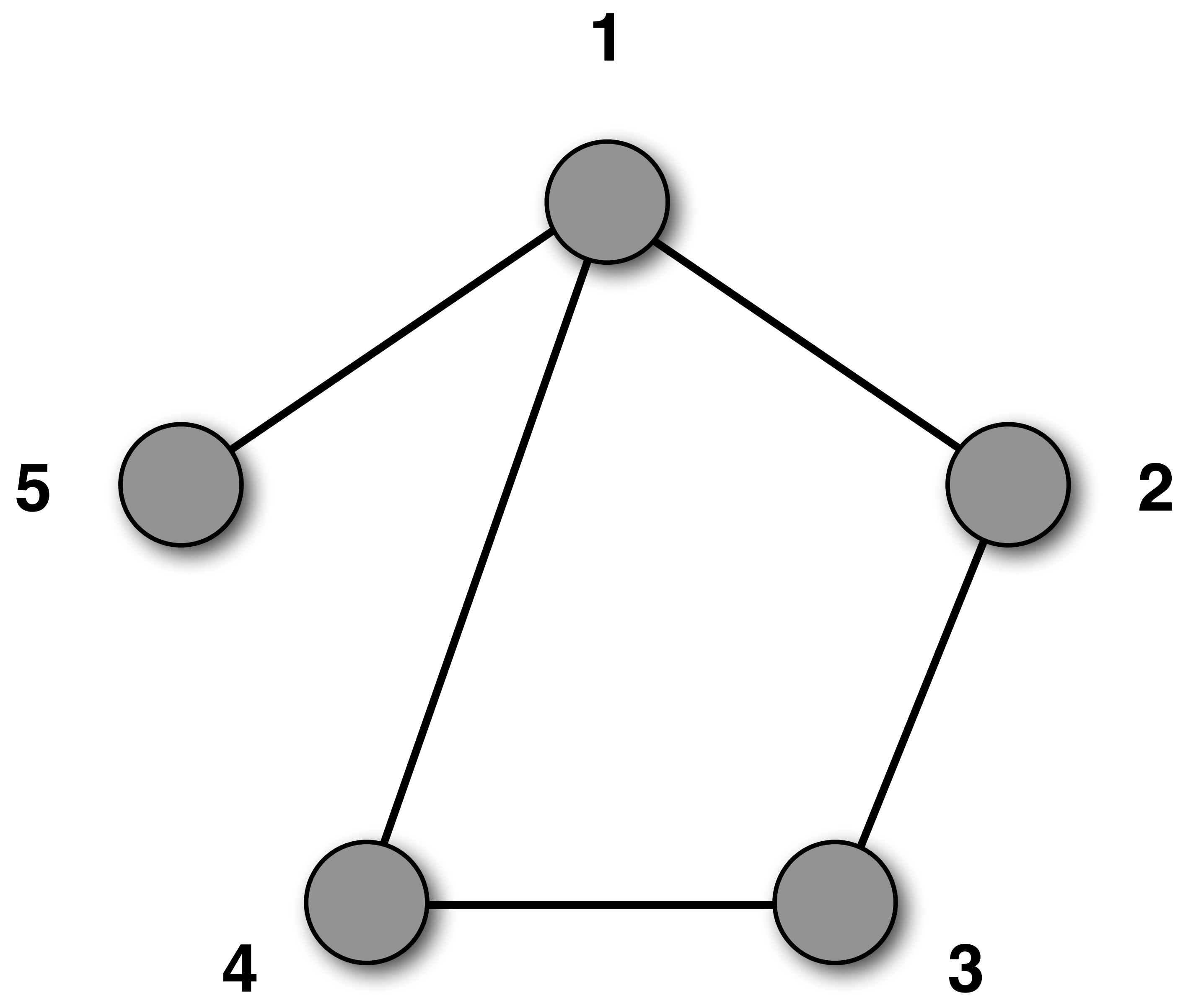


$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ A_{N1} & \dots & \dots & A_{NN} \end{pmatrix}$$

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

$A_{ij} = A_{ji} = 0$ If node *i* and node *j* are not connected

The adjacency matrix stores all possible connections
Realized connections get a 1



$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

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

$A_{ij} = A_{ji} = 0$ If node *i* and node *j* are not connected

The adjacency matrix is good for dense networks, but not practical for real networks that are large and sparse

Easy to use analytical formulas

Easy to find/remove/add a link: $\mathcal{O}(1)$

Useful for dense networks

Needs a lot of memory: $\mathcal{O}(N^2)$

Inconvenient for many numerical calculations

The adjacency matrix is not practical for large networks

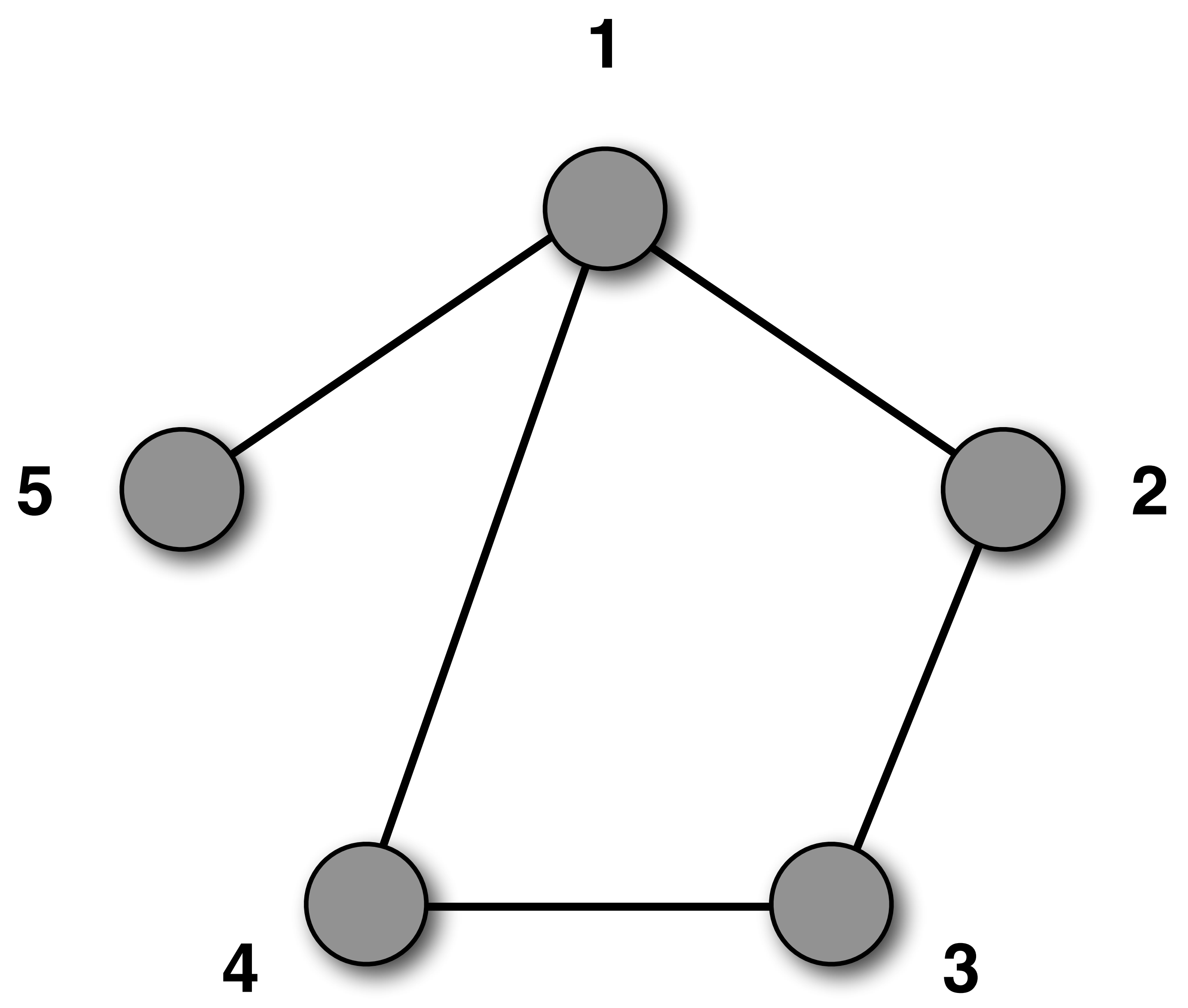
Needs a lot of memory: $\mathcal{O}(N^2)$

A matrix takes $4N^2$ bytes.
Assuming 10 GB = 10^{10} bytes of RAM, the largest possible network satisfies $4N^2 = 10^{10}$, which means $N = 50000$.

The edge list stores the node IDs of connected links

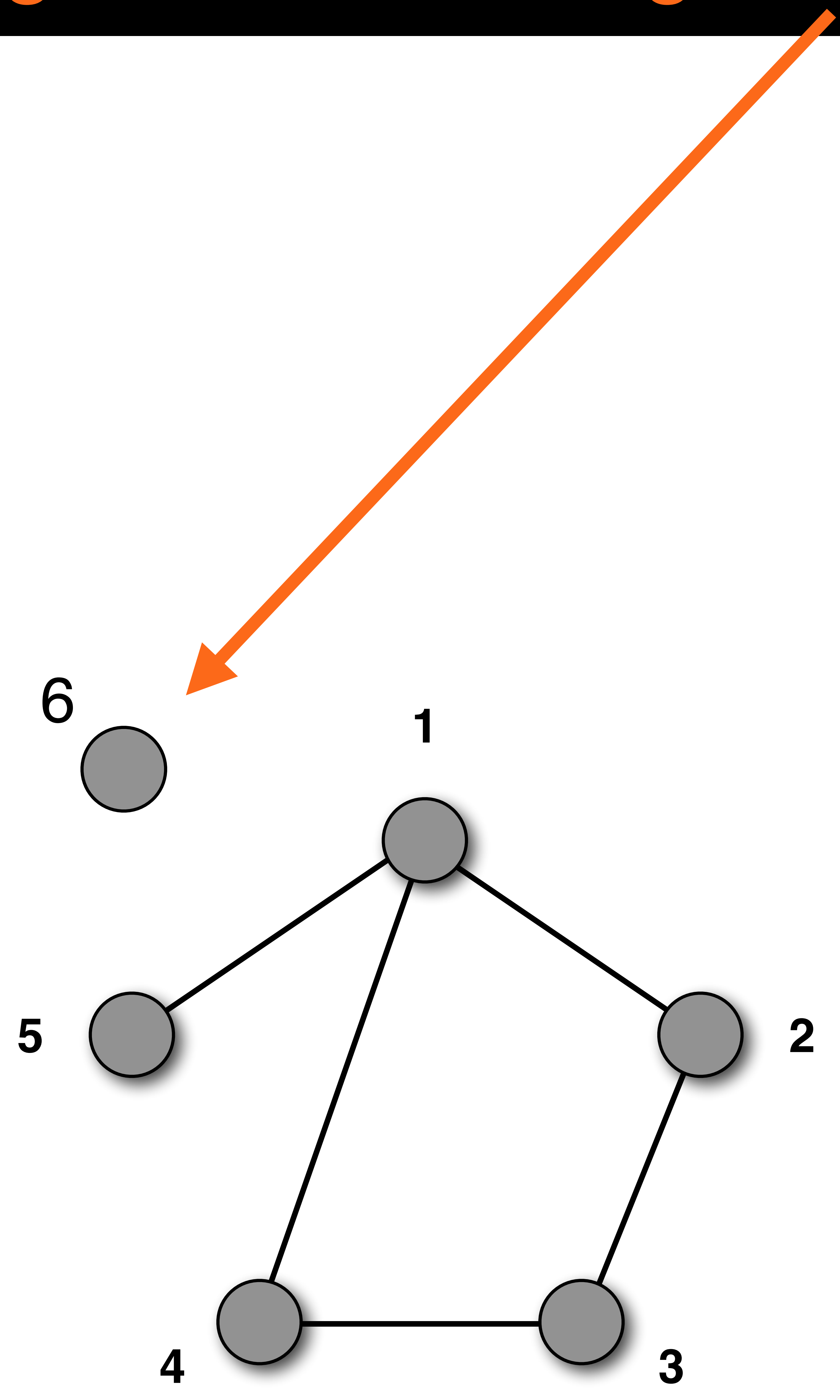
$$I = \begin{pmatrix} \dots \\ i \\ \dots \end{pmatrix} J = \begin{pmatrix} \dots \\ j \\ \dots \end{pmatrix}$$

Only if $A_{ij} = 1$



$$I = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 3 \end{pmatrix} \quad J = \begin{pmatrix} 2 \\ 4 \\ 5 \\ 3 \\ 4 \end{pmatrix}$$

The edge list is missing isolated nodes



$$I = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 3 \end{pmatrix} \quad J = \begin{pmatrix} 2 \\ 4 \\ 5 \\ 3 \\ 4 \end{pmatrix}$$

The edge list needs less space, but has disadvantages for calculations

Needs less memory: $2L$

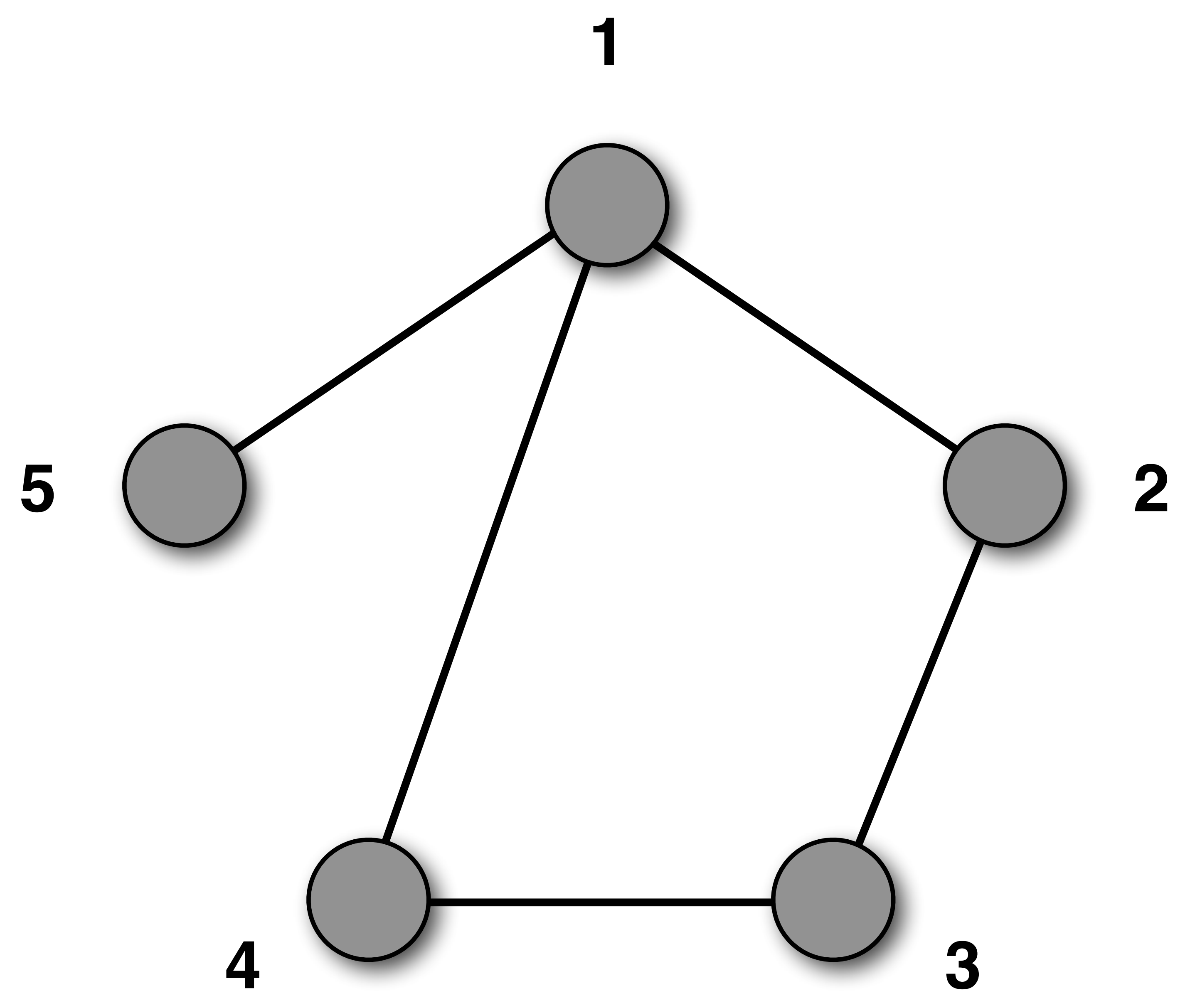
Not fast to find edges: $\mathcal{O}(L)$

Convenient for data collection

Inconvenient for calculations involving node neighbors

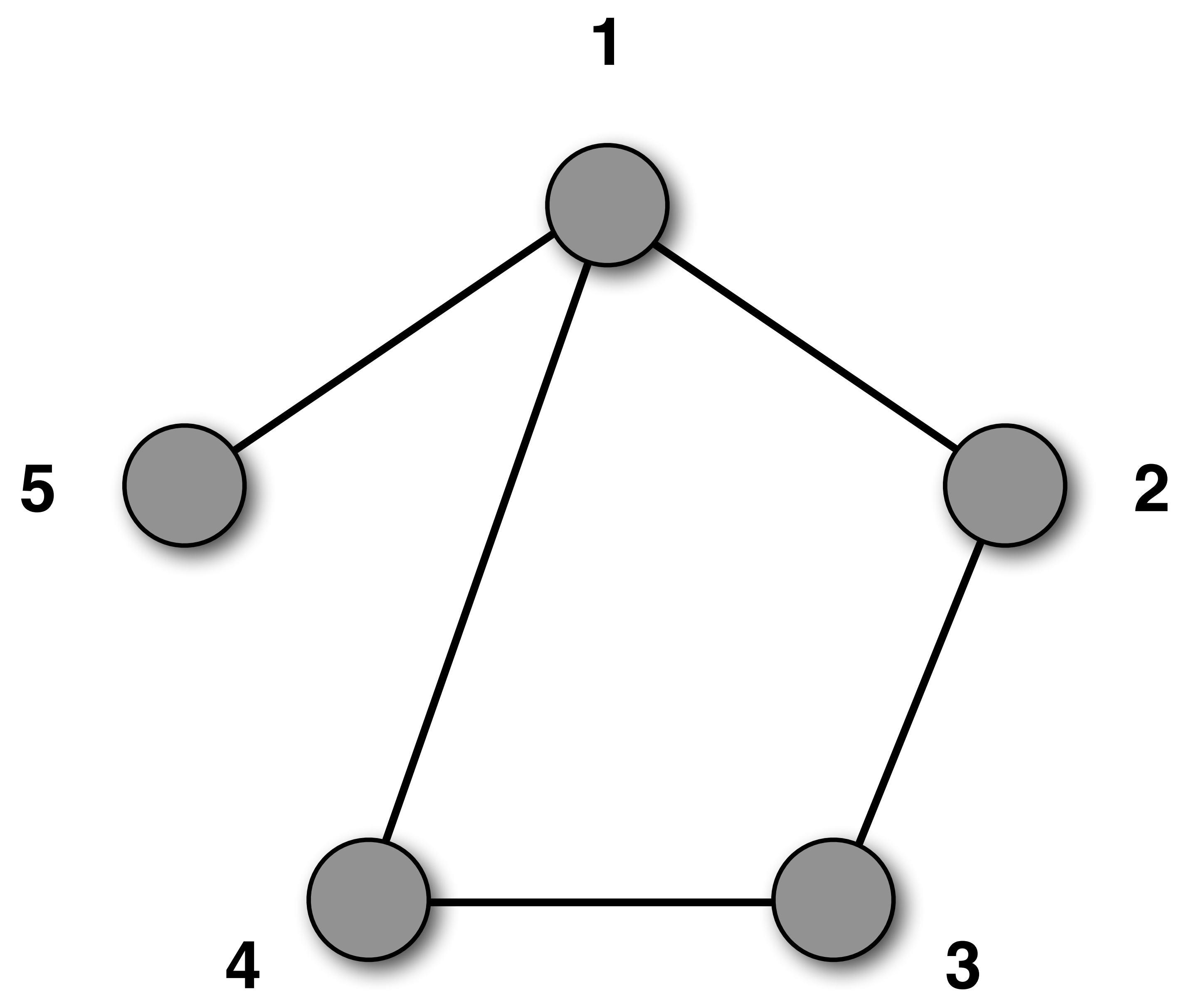
Convenient for data storage
(See: snap.stanford.edu/data/index.html)

The adjacency list stores nodes with their lists of neighbors



Node	Neighbors
1	$\{k_1 \ \ \text{neighbors}\}$
$\ldots$	
i	$\{k_i \ \ \text{neighbors}\}$
$\ldots$	
N	$\{k_N \ \ \text{neighbors}\}$

The adjacency list stores nodes with their lists of neighbors



Node	Neighbors
1	2,4,5
2	1,3
3	2,4
4	1,3
5	1

The adjacency list needs little space and is convenient for many calculations

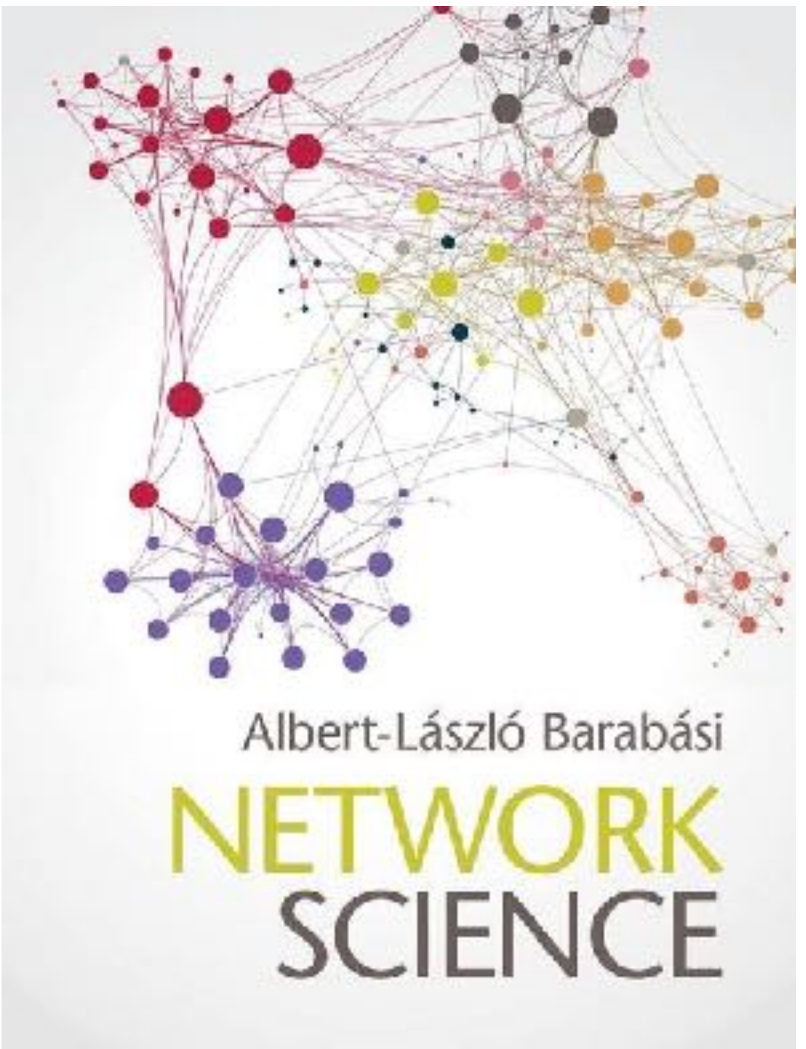
Needs little memory: $2L$

Convenient for many calculations involving neighbors (BFS, spreading processes,...)

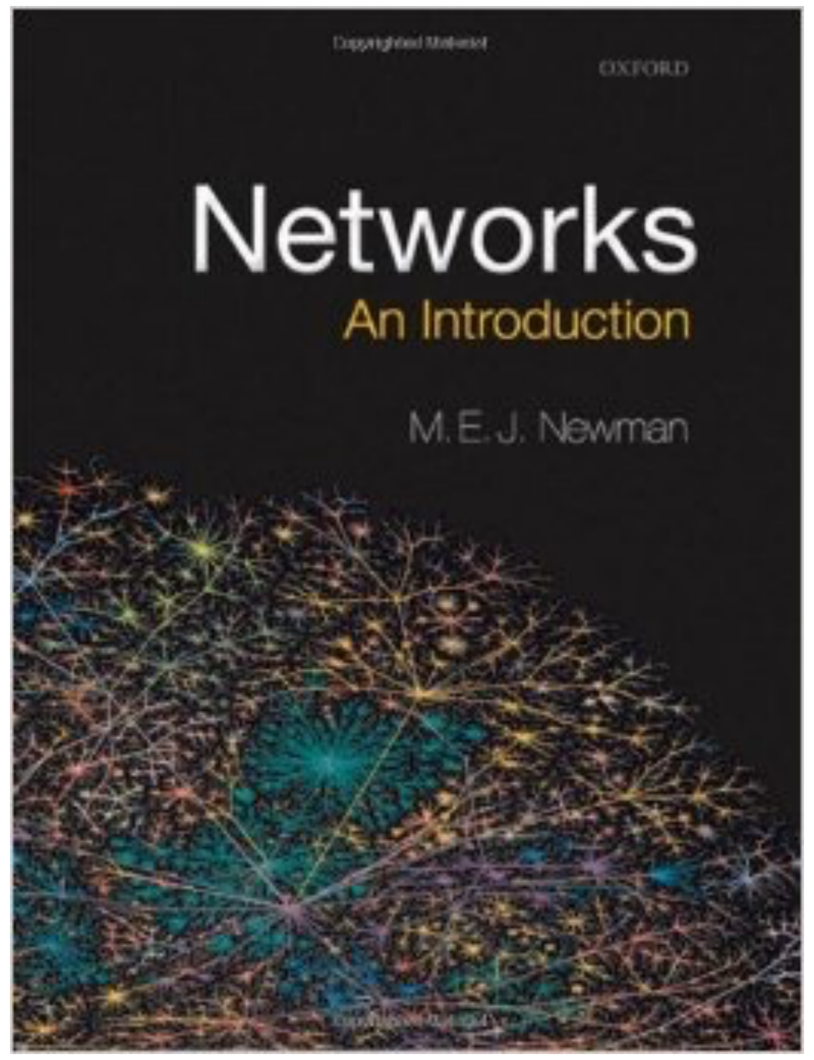
Fast to add elements: $\mathcal{O}(1)$

Not convenient to find/remove elements: $\mathcal{O}(L/N)$

Sources and further materials for today's class



A.-L. Barabási.
Network Science.
Cambridge University Press (2016)
http://barabasi.com/networksciencebook/



M.E.J. Newman.
Networks: An Introduction.
Oxford University Press (2010)