DP1

Web mining and Social Network Analysis

Dino Pedreschi
pedre@di.unipi.it

Lecture 2 – Graphs and networks

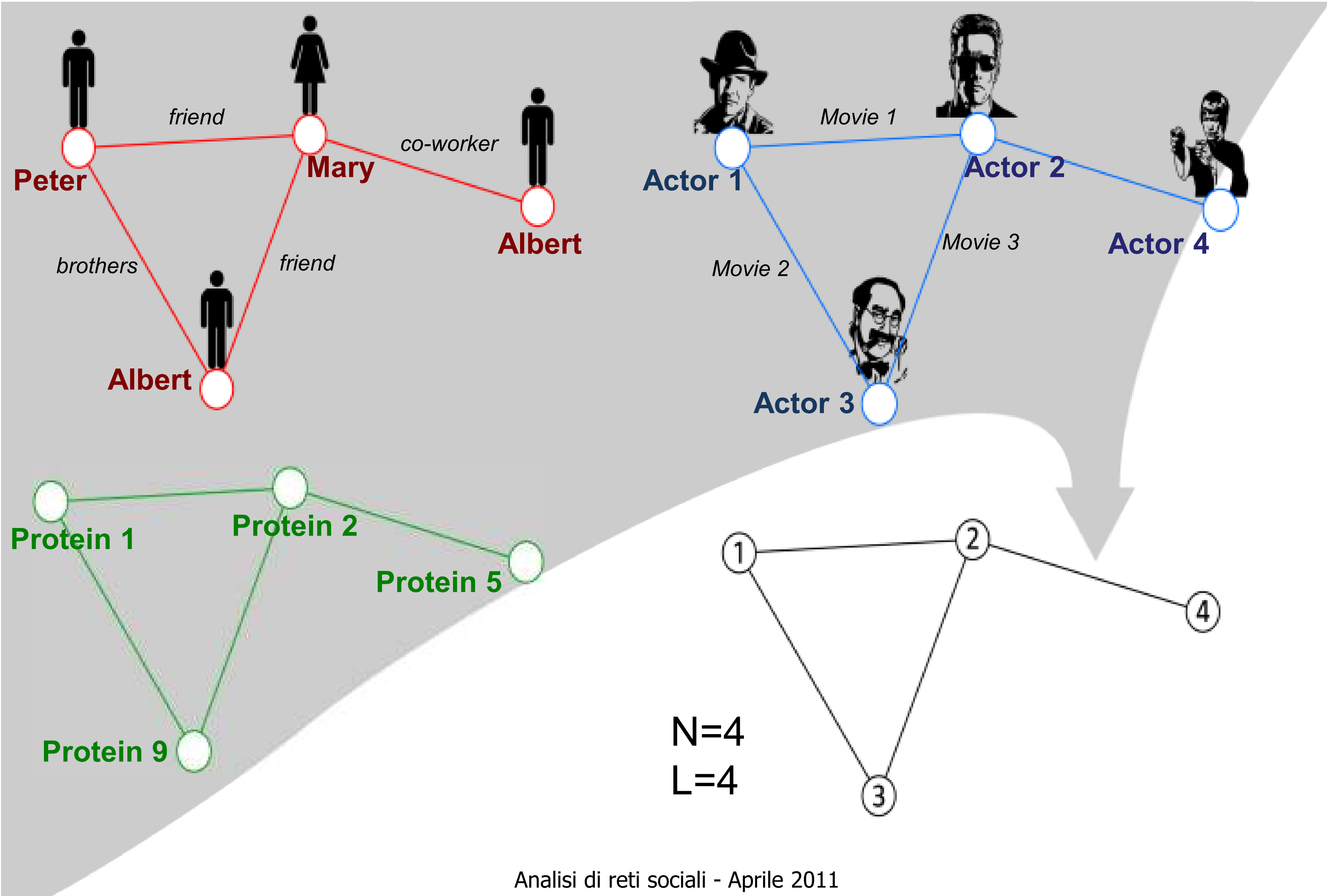
Diapositiva 1

DP1 Dino Pedreschi; 15/04/2011

"Natural" Networks and Universality

- Consider many kinds of networks:
  - social, technological, business, economic, content,…
- These networks tend to share certain *informal* properties:
  - large scale; continual growth
  - distributed, organic growth: vertices "decide" who to link to
  - interaction restricted to links
  - mixture of local and long-distance connections
  - abstract notions of distance: geographical, content, social,…
- Do natural networks share more *quantitative* universals?
- What would these "universals" be?
- How can we make them precise and measure them?
- How can we explain their universality?
- This is the domain of *social network theory*
- Sometimes also referred to as *link analysis*

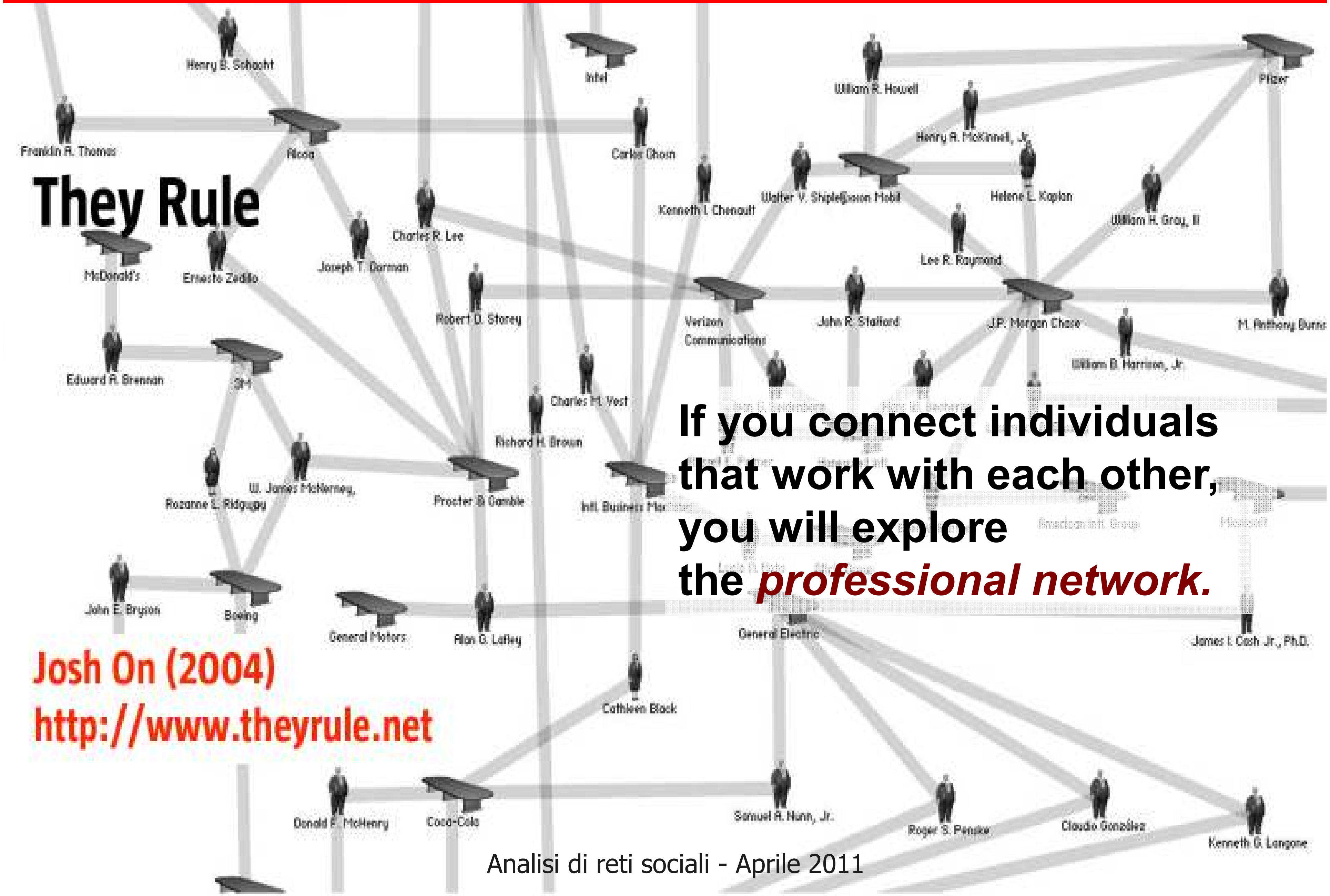
Graphs as common language



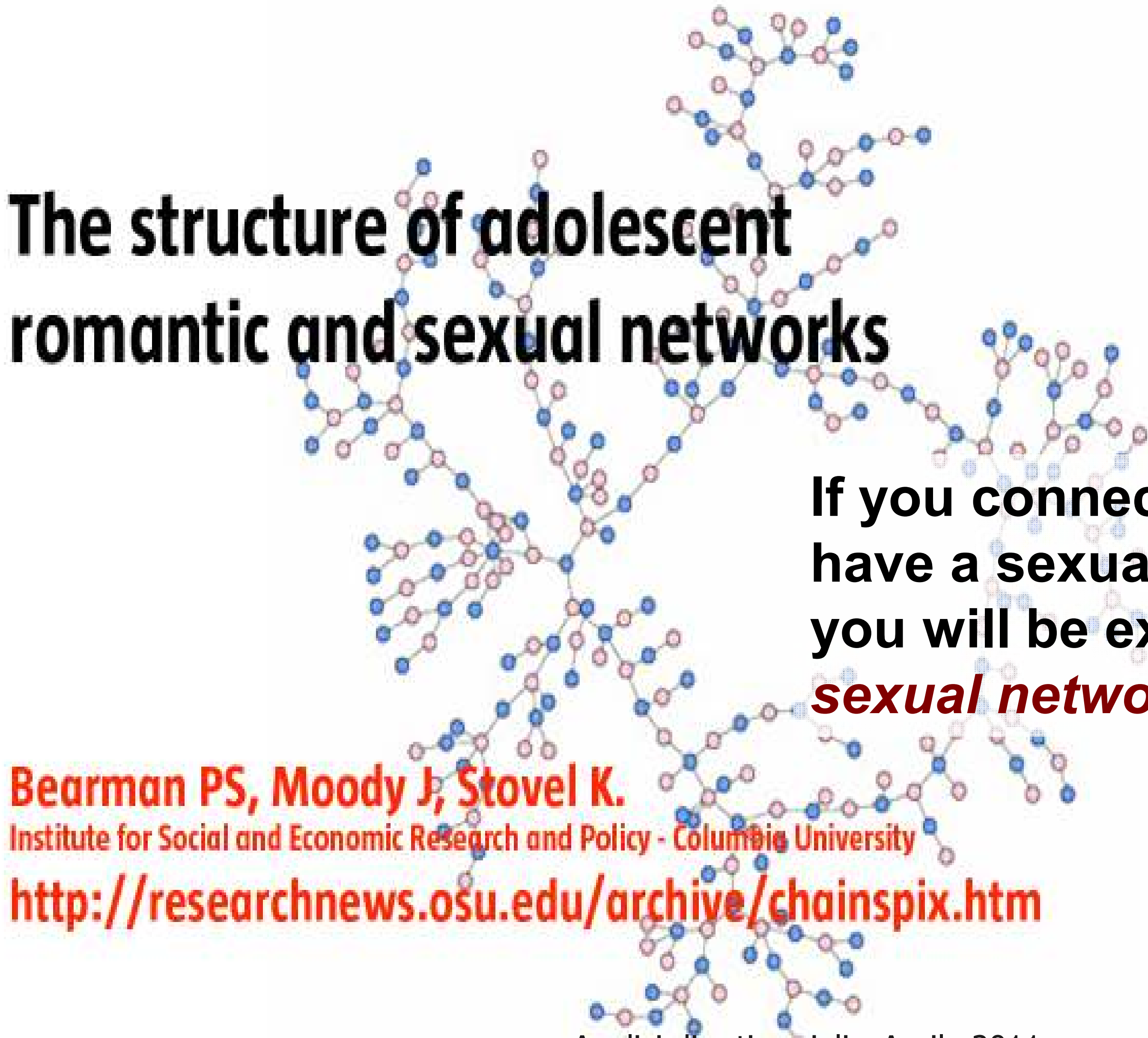
Choosing the proper representation

- The choice of the proper network representation determines our ability to use network theory successfully.
    - In some cases there is a unique, unambiguous representation.
    - In other cases, the representation is by no means unique.
- For example, for a group of individuals, the way you assign the links will determine the nature of the question you can study.

CHOOSING A PROPER REPRESENTATION



If you connect individuals that work with each other, you will explore the *professional network.*



If you connect those that have a sexual relationship, you will be exploring the *sexual networks.*

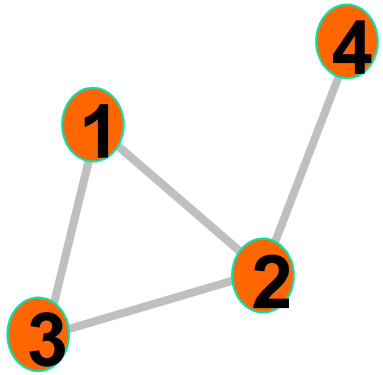
CHOOSING A PROPER REPRESENTATION

If you connect individuals based on their first name (*all Peters connected to each other*), you will be exploring what?

It is a network, nevertheless.

GRAPHOLOGY *1*

Undirected



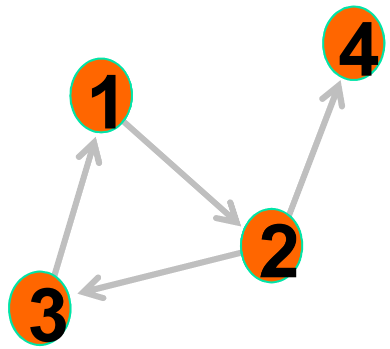
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad <k> = \frac{2L}{N}$$

Actor network, protein-protein interactions

Directed



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

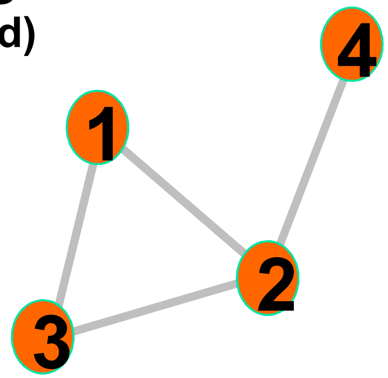
$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^{N} A_{ij} \qquad <k> = \frac{L}{N}$$

WWW, citation networks

Unweighted

(undirected)



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

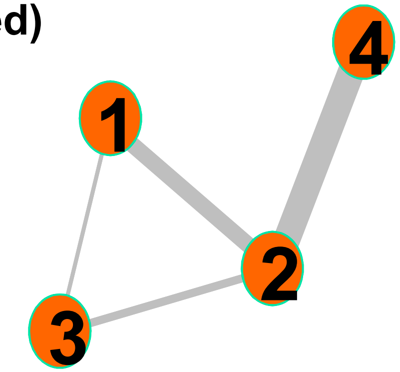
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} \qquad < k > = \frac{2L}{N}$$

protein-protein interactions, www

Weighted

(undirected)



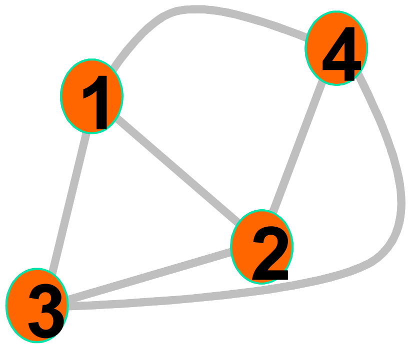
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

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

$$L = \frac{1}{2} \sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad < k > = \frac{2L}{N}$$

Call Graph, metabolic networks

Complete Graph
(undirected)



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

$$A_{ii} = 0 \qquad A_{i \neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \qquad \langle k \rangle = N - 1$$

Actor network, protein-protein interactions

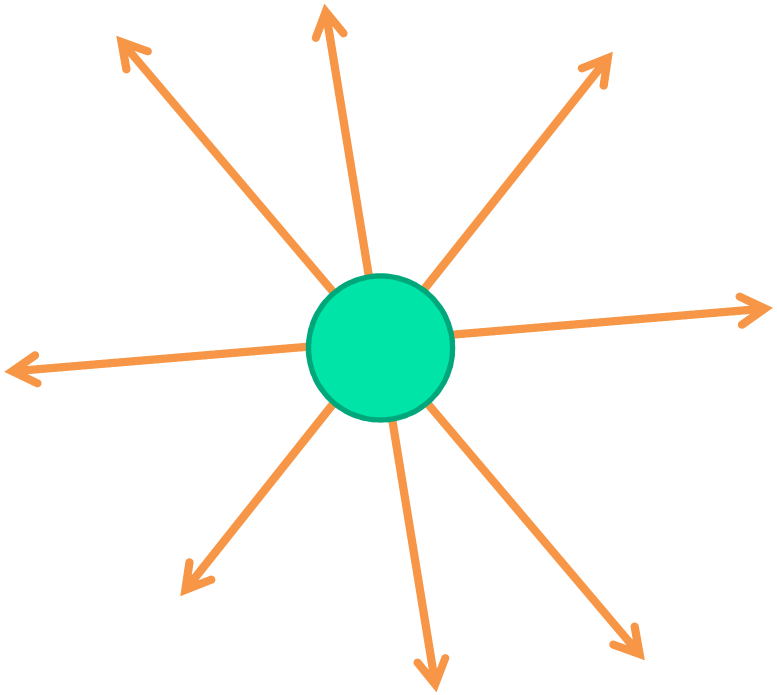
The key basic quantities

- *Degree distribution: about connectivity*
  - what is the typical degree in the network?
  - what is the overall distribution?
- *Network diameter: about social distance*
  - maximum (worst-case) or average?
  - exclude infinite distances? (disconnected components)
  - the small-world phenomenon
- *Clustering : about social transitivity*
  - to what extent that links tend to cluster “locally”?
  - what is the balance between local and long-distance connections?
  - what roles do the two types of links play?
- *Connected components: about social partitioning*
  - how many, and how large?

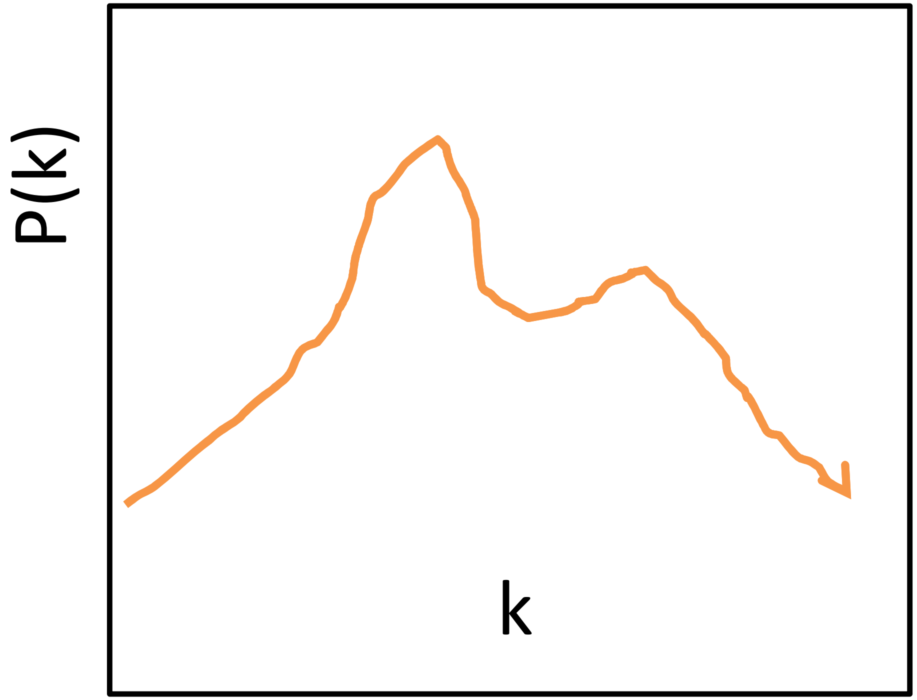
Degree distribution

- The **degree** of a vertex in a network is the number of edges incident on (i.e., connected to) that vertex.
- $\mathbf{p_k}$ = the fraction of vertices in the network that have degree k.
- Equivalently, $\mathbf{p_k}$ = the **probability** that a vertex chosen uniformly at random has **degree k**.
- A plot of $\mathbf{p_k}$ for any given network can be formed by a **histogram** of the degrees of vertices.
- This histogram is the **degree distribution** for the network

Degree (*k*)



Degree Distribution

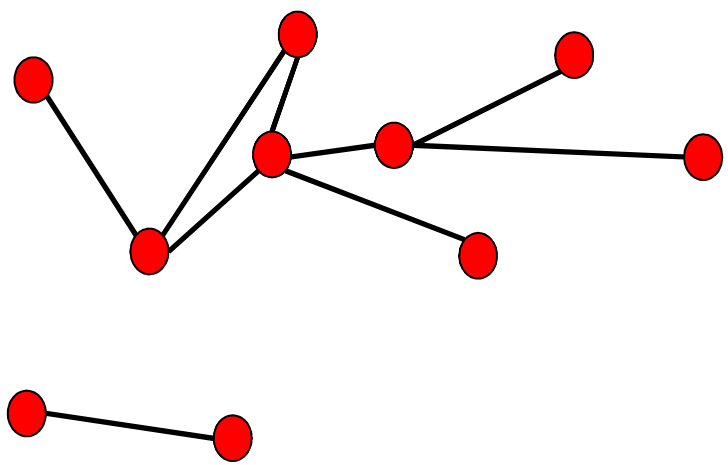


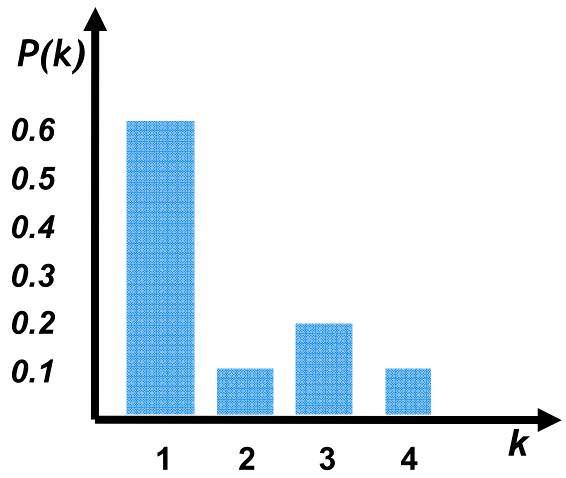
Degree distribution

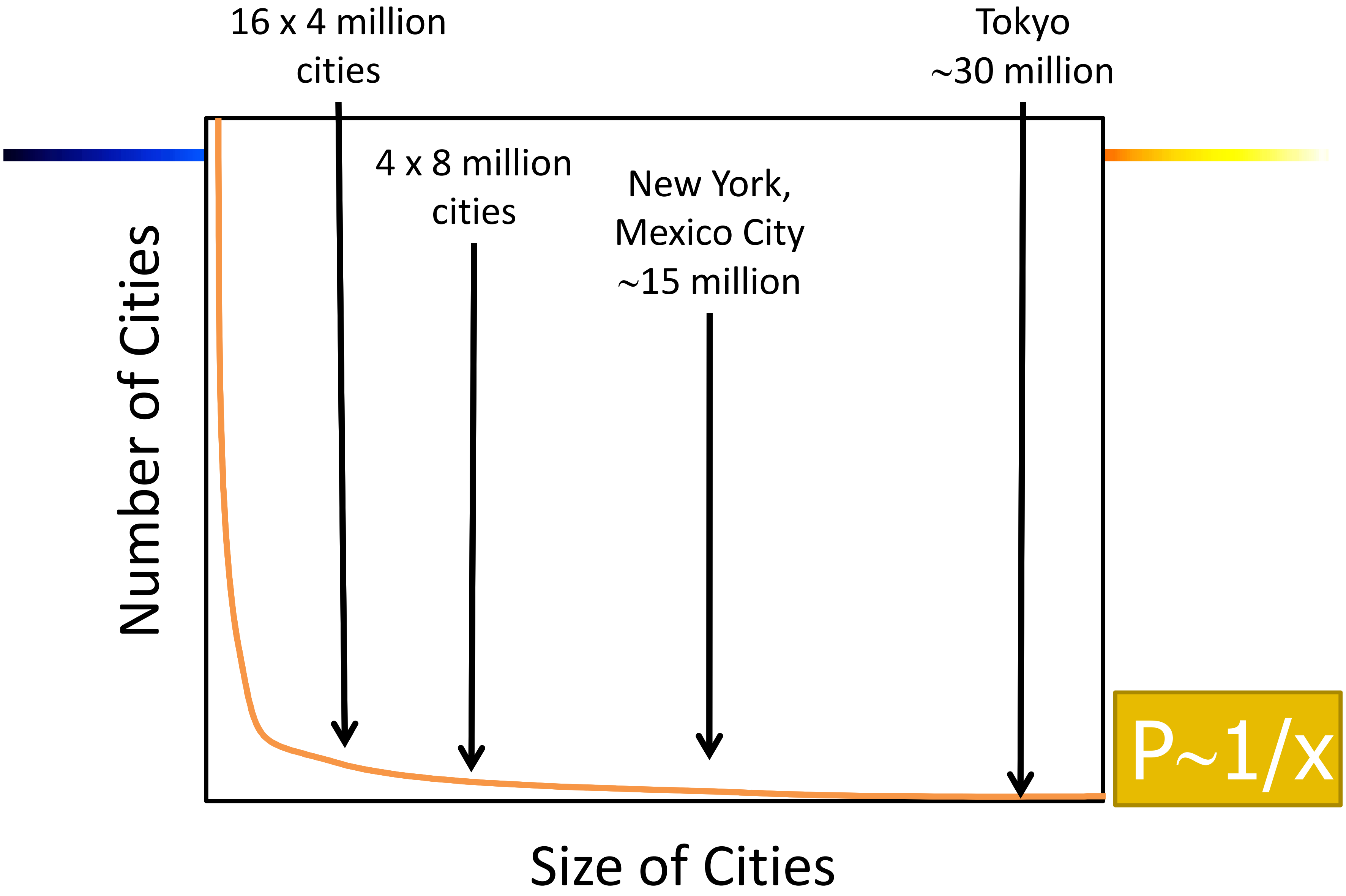
Degree distribution P(k): probability that a randomly chosen vertex has degree k

N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot





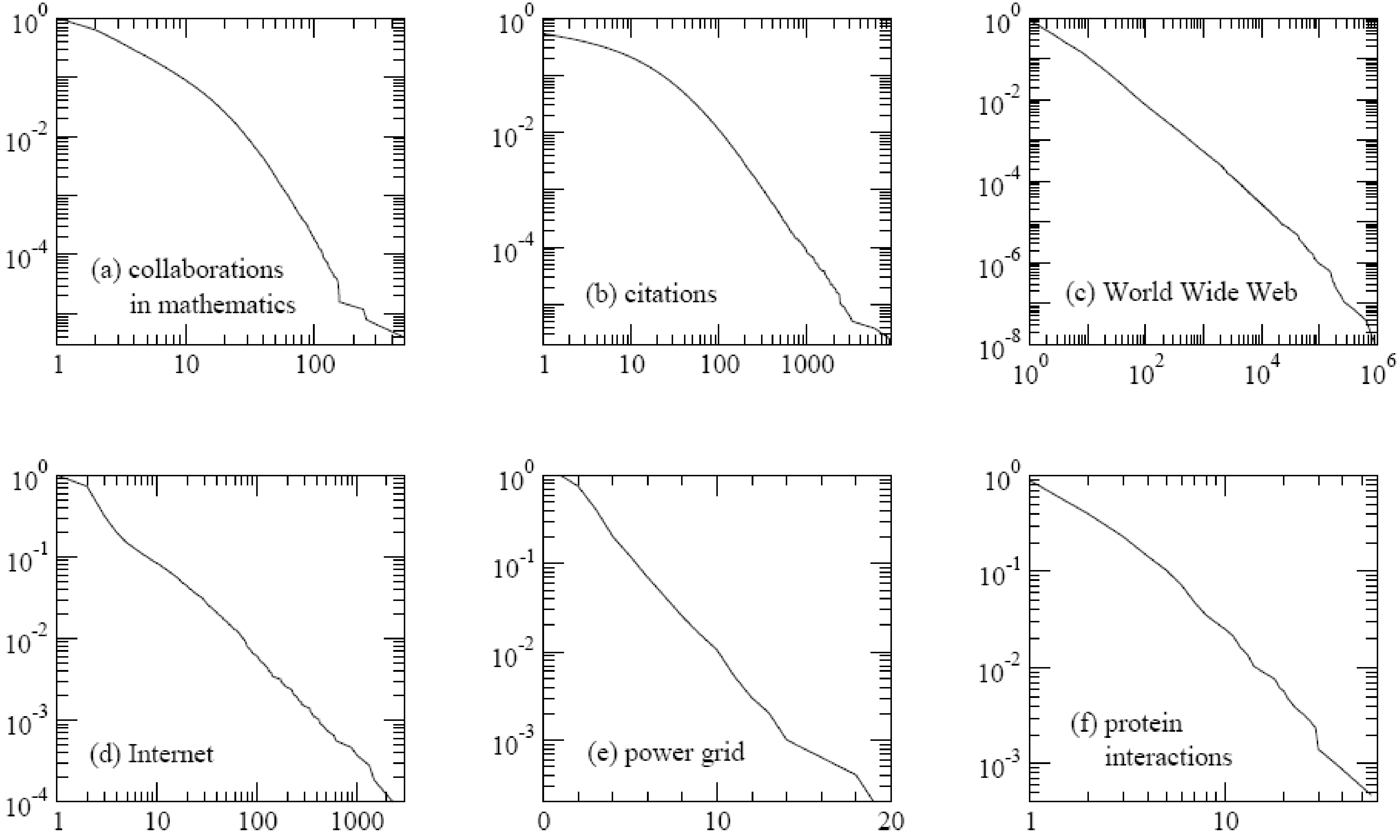


There is an equivalent number of people living in cities of all sizes!

After Bill enters the arena the average income of the public ~ USD $1,000,000

~ $50 billion

Degree distributions for six networks



Actor Connectivity (power law)

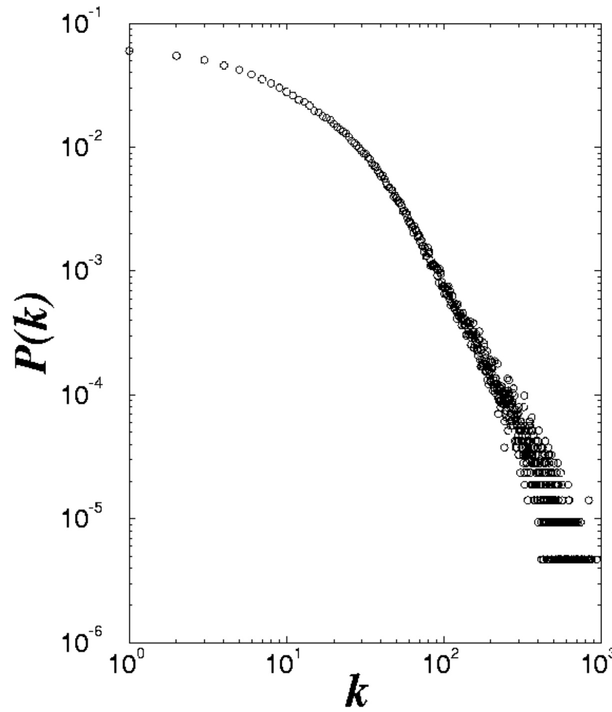
Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

N = 212,250 actors
$\langle k \rangle = 28.78$

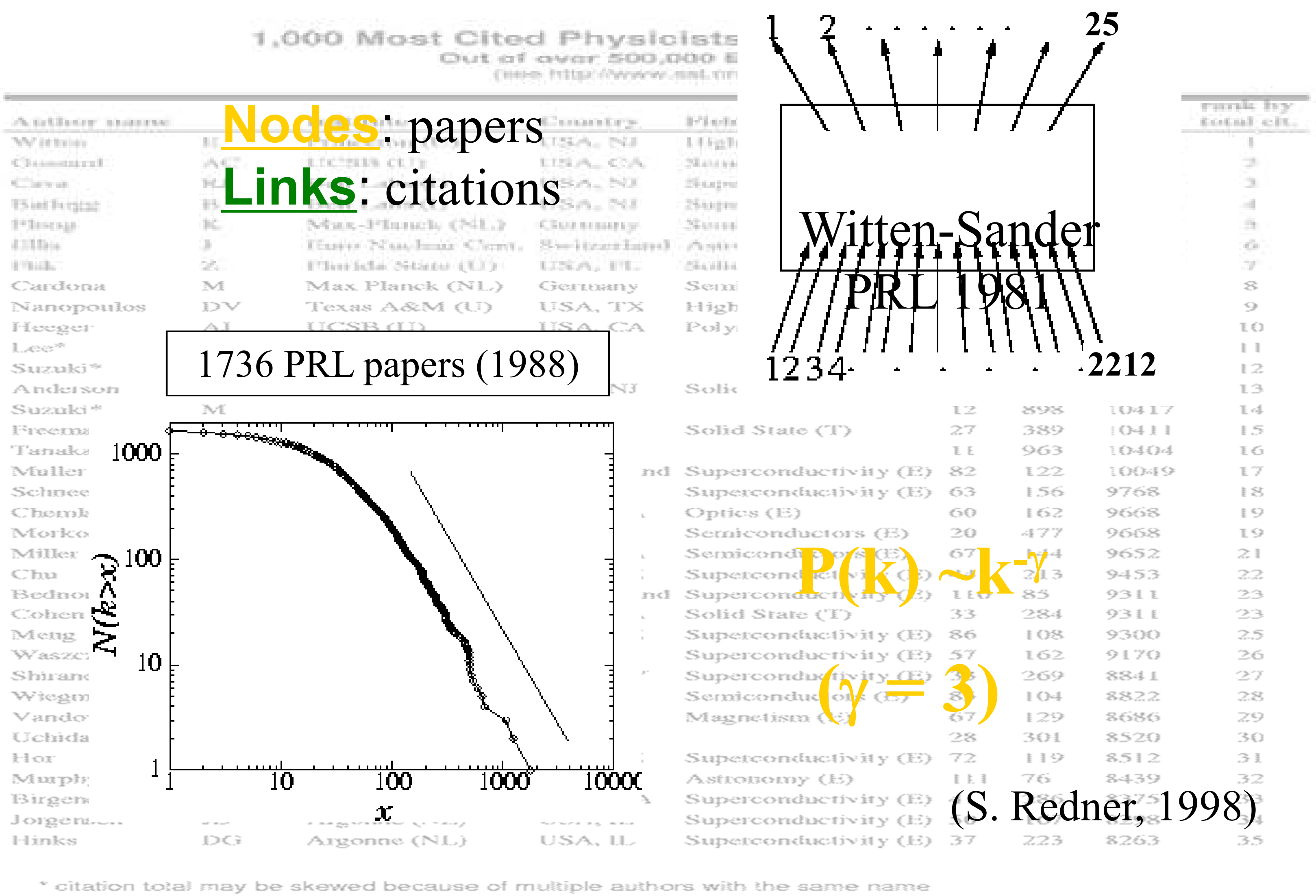
$P(k) \sim k^{-\gamma}$

$\gamma = 2.3$

Nodes: actors
Links: cast jointly

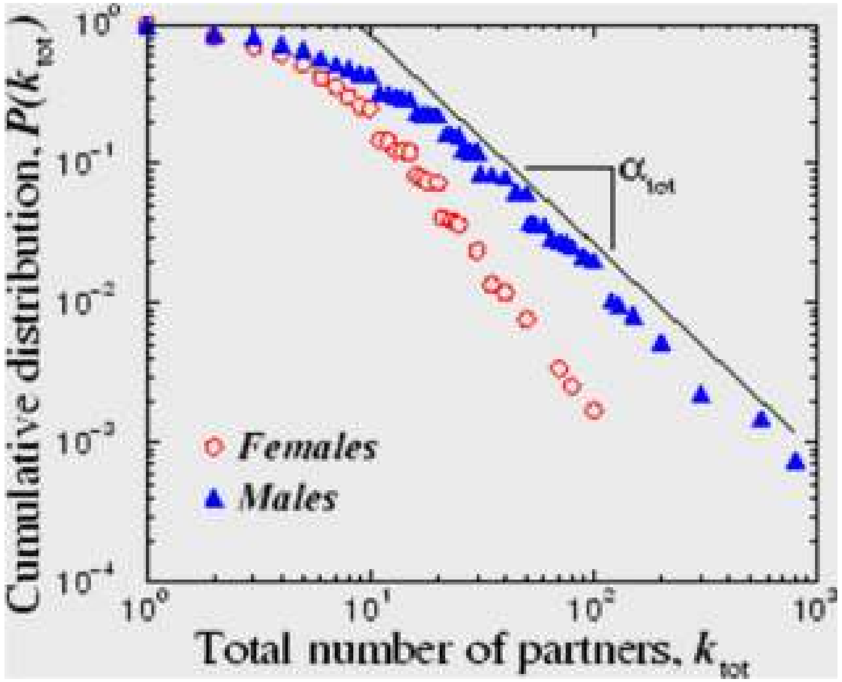


Science Citation Index (power law)



Sex-Web (power law)

Nodes: people (Females; Males)
Links: sexual relationships



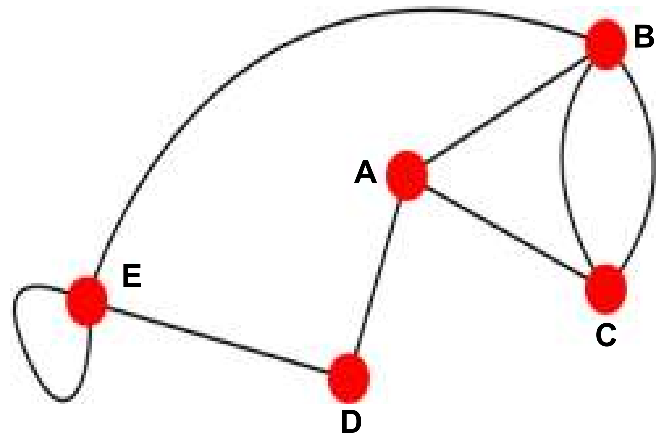
4781 Swedes; 18-74;
59% response rate.
Liljeros et al. Nature 2001

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

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

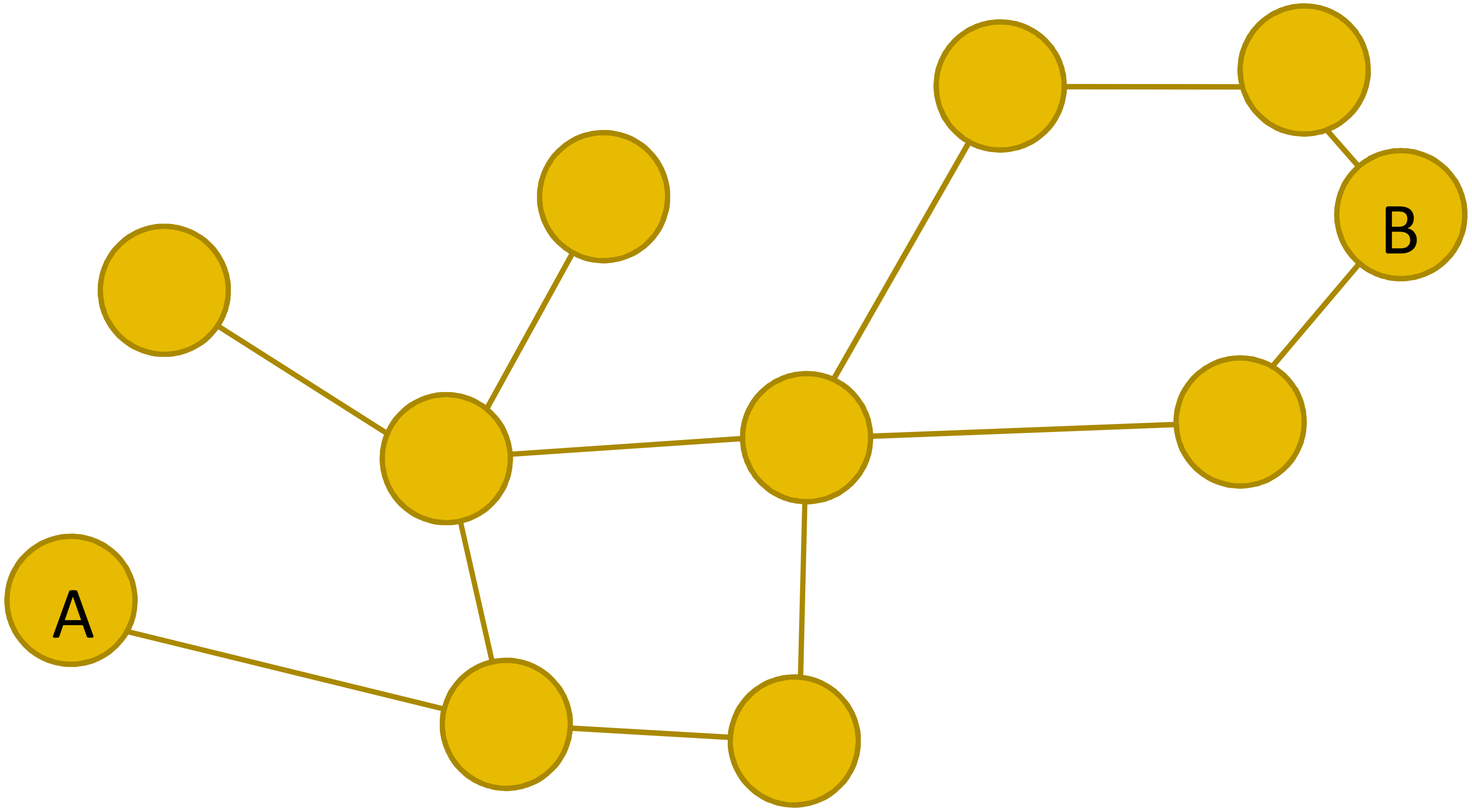


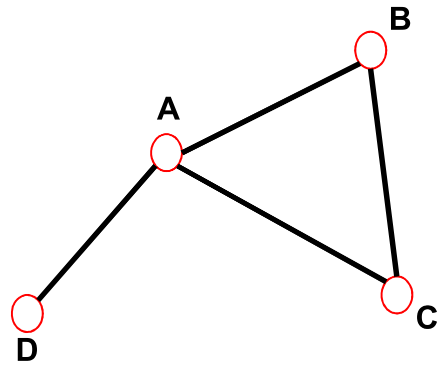
- A path can intersect itself and pass through the same link repeatedly. Each time a link is crossed, it is counted separately

- A legitimate path on the graph on the right: **ABCBCADEEBA**

- In a directed network, the path can follow only the direction of an arrow.

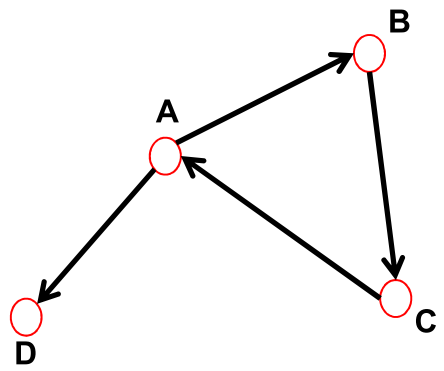
Distance Between A and B?





The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

*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).

NETWORK DIAMETER AND AVERAGE DISTANCE

***Diameter*:** the maximum distance between any pair of nodes in the graph.

***Average path length/distance* for a direct** connected graph (component) or a strongly connected (component of a) digraph.

where l_{ij} is the distance from node i to node j

$$\langle l \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} l_{ij}$$

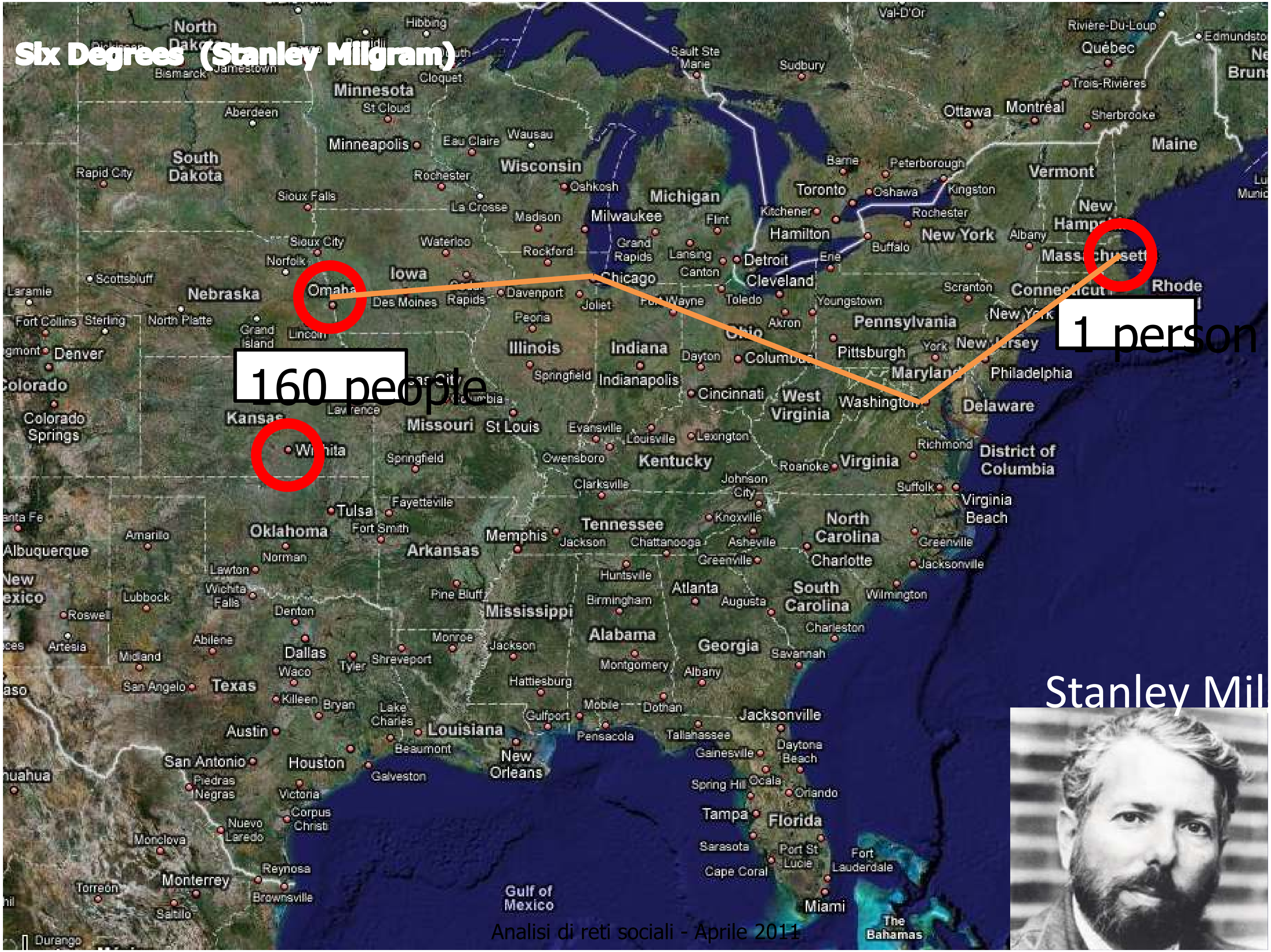
In an undirected (symmetrical) graph $l_{ij} = l_{ji}$, we only need to count them once

$$\langle l \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j>i} l_{ij}$$

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

IT IS A SMALL WORLD

Six Degrees (Stanley Milgram)



Stanley Mil

Stanley Milgram found that the average length
of the chain connecting the sender and receiver was of length 5.5.

But only a few chains were ever completed!

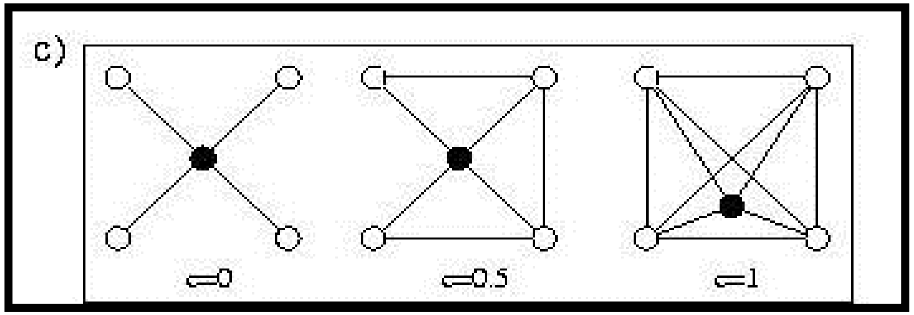
CLUSTERING COEFFICIENT

- **Clustering coefficient:**

  what portion of your neighbors are connected?

  - Node i with degree k_i
  - C_i in [0,1]

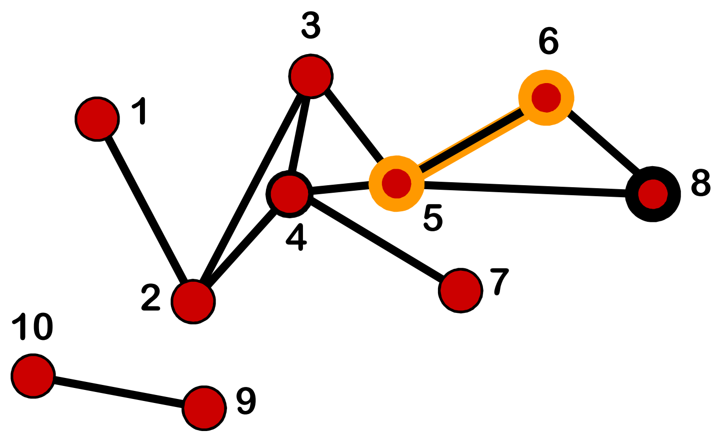
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



CLUSTERING COEFFICIENT

- **Clustering coefficient:** what portion of your neighbors are connected?
  - Node i with degree k_i

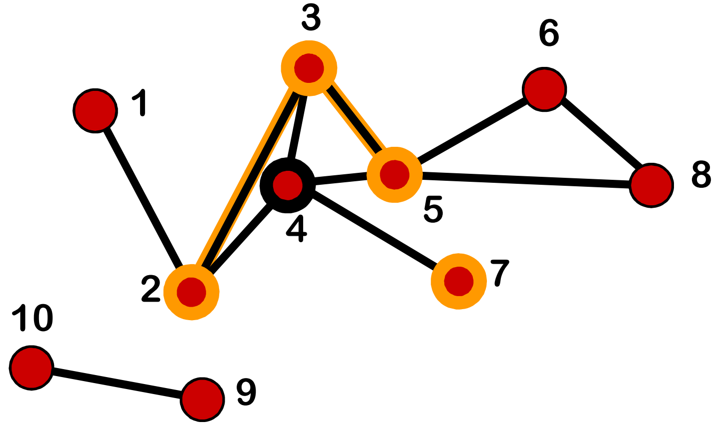
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



i=8: $k_8=2$, $e_8=1$, *TOT=2*1/2=1* → $C_8=1/1=1$

- **Clustering coefficient:** what portion of your neighbors are connected?
  - Node i with degree k_i

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



i=4: k_4=4, e_4=2, TOTAL=4*3/2=6 → C_4=2/6=1/3

KEY MEASURES

Degree distribution: **P(k)**

Path length: ***l***

Clustering coefficient:

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

Transitivity – the clustering coefficient

An alternative definition of the clustering coefficient, also widely used, has been given by Watts and Strogatz [416], who proposed defining a local value

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}. \quad (5)$$

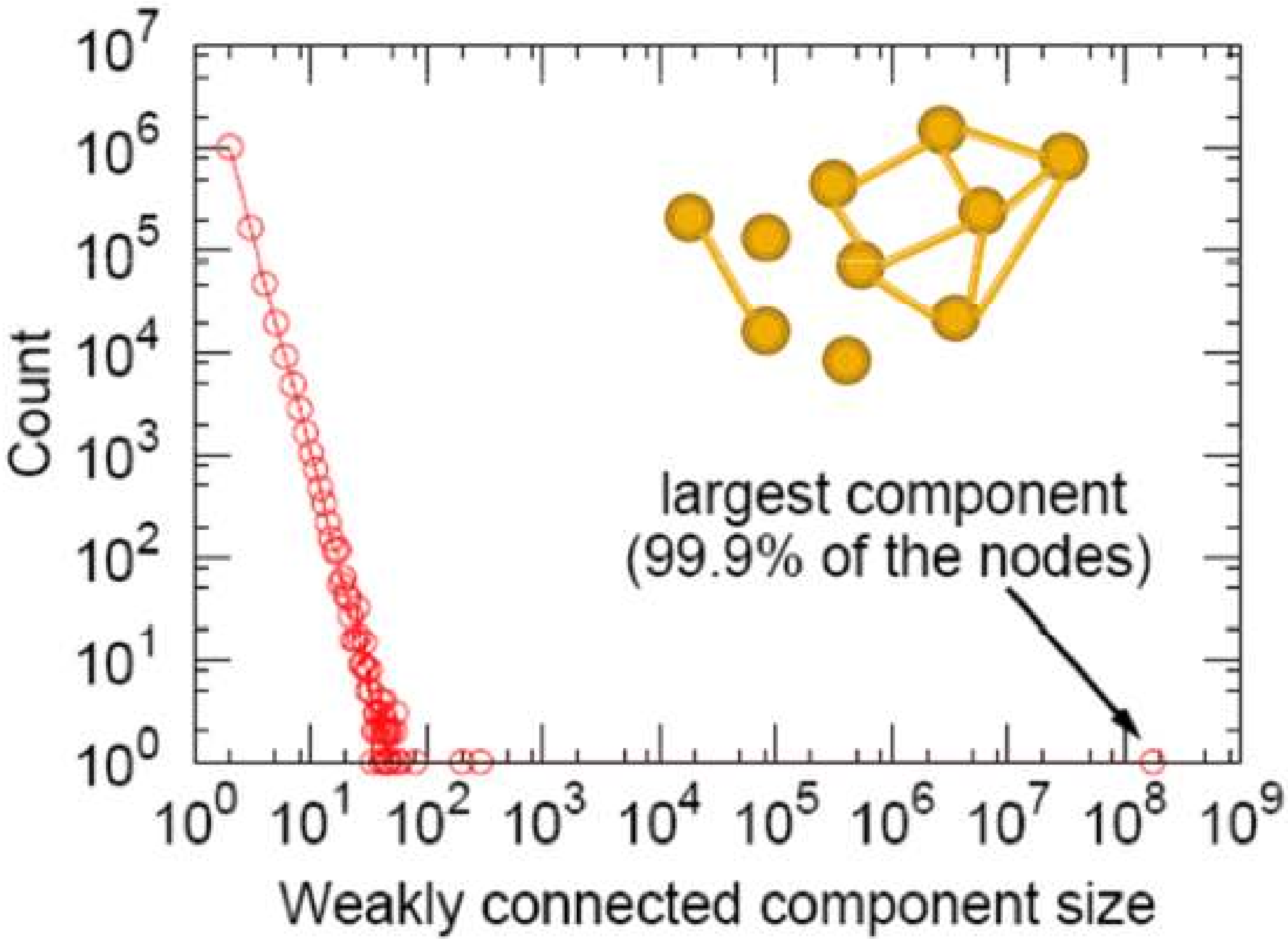
For vertices with degree 0 or 1, for which both numerator and denominator are zero, we put $C_i = 0$. Then the clustering coefficient for the whole network is the average

$$C = \frac{1}{n} \sum_i C_i. \quad (6)$$

Basic statisics for some published networks

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1			
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16	
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029
	sexual contacts	undirected	2 810				3.2			
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7			
	citation network	directed	783 339	6 716 198	8.57		3.0/–			
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44	
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226

The giant connected component



A "Canonical" Natural Network has…

- *Few* connected components:
  - often only 1 or a small number, indep. of network size
- *Small* diameter:
  - often a constant independent of network size (like 6)
  - or perhaps growing only logarithmically with network size or even shrink?
  - typically exclude infinite distances
- A *high* degree of clustering:
  - considerably more so than for a random network
  - in tension with small diameter
- A *heavy-tailed* degree distribution:
  - a small but reliable number of high-degree vertices
  - often of *power law* form