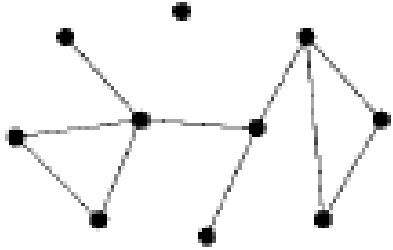


# Lecture 4: Large Scale Structures of Networks and More Metrics

# Degree Distribution



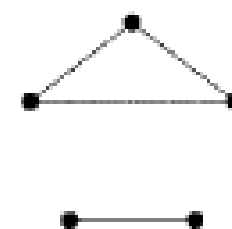
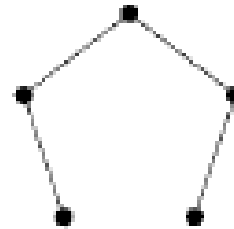
We define  $P(k)$  to be the fraction of vertices in a network that have degree  $k$ . Consider the network on the left.

It has  $n = 10$  vertices, of which 1 has degree 0, 2 have degree 1, 4 have degree 2, 2 have degree 3, and 1 has degree 4. Thus the values of  $P(k)$  for  $k = 0, \dots, 4$  are

$$P(0) = 1/10, P(1) = 2/10, P(2) = 4/10, P(3) = 2/10, P(4) = 1/10.$$

$P(k) = 0$  for all  $k > 4$ . The quantities  $P(k)$  represent the *degree distribution* of the network.

Note that a knowledge of the degree distribution (or degree sequence) does not, in most cases, tell us the complete structure of a network. For example, the two networks below are different but have the same degree distribution.



# Degree Distribution

## Degree sequence and Degree frequency:

■ **Degree sequence:** An ordered list of the (in, out) degree of each node

● In-degree sequence:

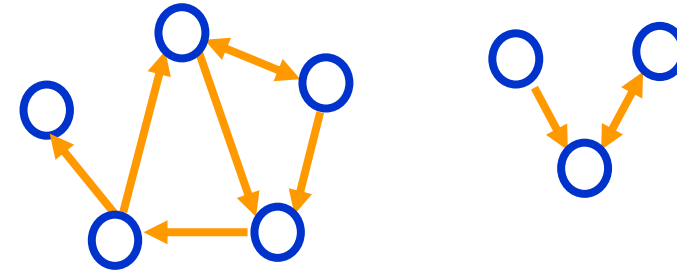
● [2, 2, 2, 1, 1, 1, 1, 0]

● Out-degree sequence:

● [2, 2, 2, 1, 1, 1, 1, 0]

● (undirected) degree sequence:

● [3, 3, 3, 2, 2, 1, 1, 1]



■ **Degree frequency:** A frequency count of the occurrence of each degree

In-degree frequency:

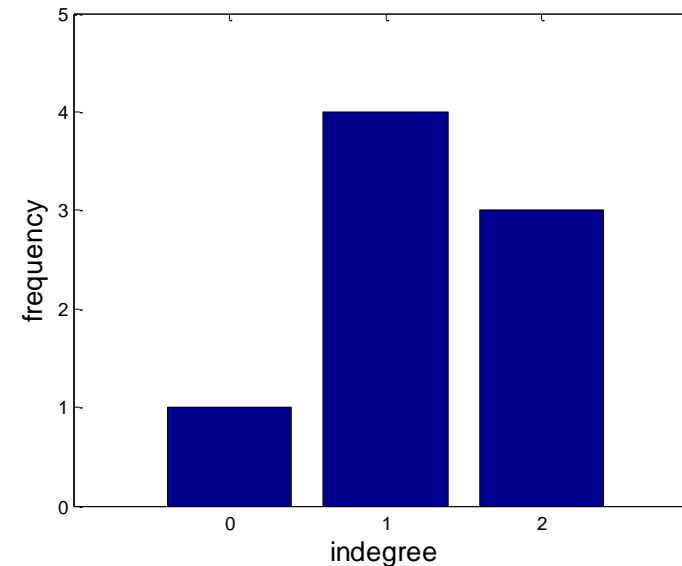
[(2,3) (1,4) (0,1)]

Out-degree frequency :

[(2,3) (1,4) (0,1)]

(undirected) frequency :

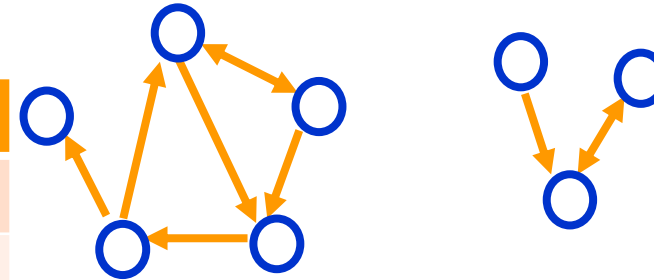
[(3,3) (2,2) (1,3)]



# Degree Distribution

The degree distribution is a function  $P(k)$ , which gives the probability of a randomly chosen node from the graph having degree  $k$

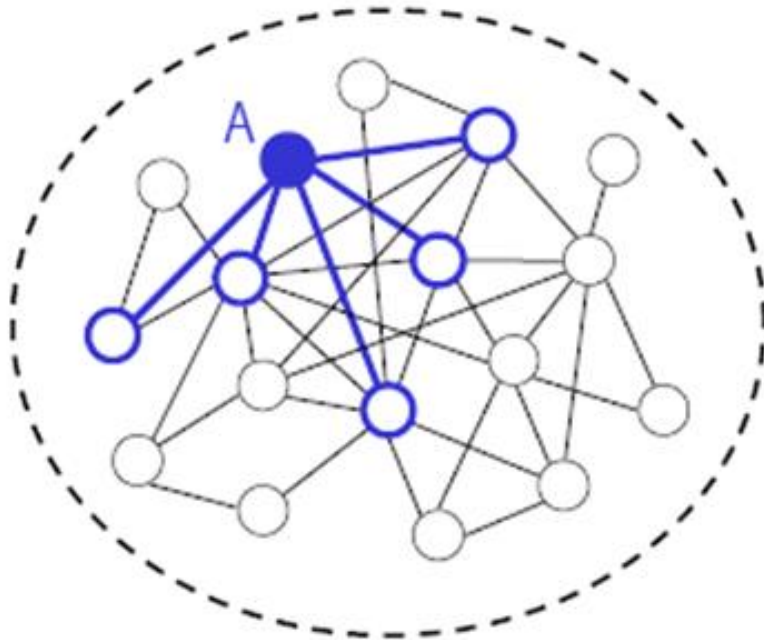
| In-degree    | 0     | 1    | 2     | 3    |
|--------------|-------|------|-------|------|
| Frequency    | 1     | 4    | 3     | 0    |
| Distribution | 0.125 | 0.50 | 0.375 | 0.00 |



| Out-degree   | 0     | 1    | 2     | 3    |
|--------------|-------|------|-------|------|
| Frequency    | 1     | 4    | 3     | 0    |
| Distribution | 0.125 | 0.50 | 0.375 | 0.00 |

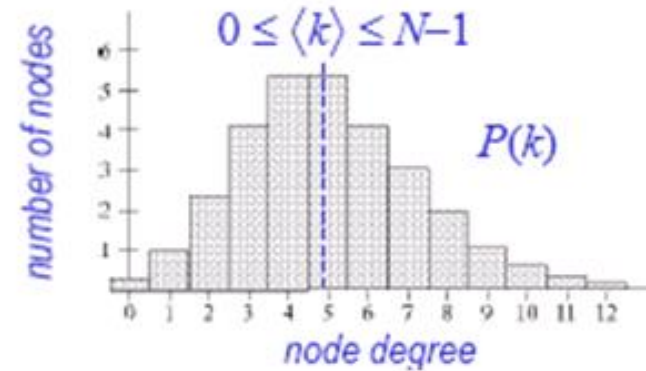
| Degree       | 0    | 1     | 2    | 3     |
|--------------|------|-------|------|-------|
| Frequency    | 0    | 3     | 2    | 3     |
| Distribution | 0.00 | 0.375 | 0.25 | 0.375 |

# Degree Distribution



The degree of A is 5

➤ the *degree distribution* function  $P(k)$  is the histogram (or probability) of the node degrees: it shows their spread around the average value



$$p_k = \frac{N_k}{N} \quad \sum_{k=0}^{\infty} p_k = 1$$

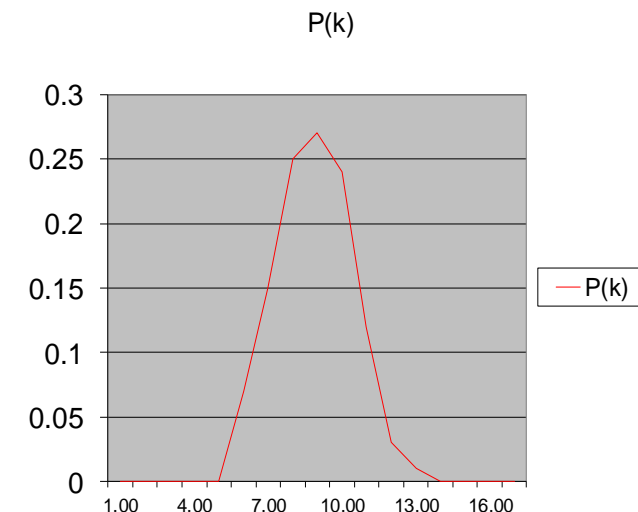
$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

# Degree Distribution

- Let us try to understand degree distribution with a simple case. Imagine I have a graph with 1000 nodes, but no links. Now I start adding links randomly, one by one.
  - After 10 random additions, what do you expect the degree distribution to be?
  - What will the average node degree be after 1000 additions?
- The standard situation in a network where links are added completely at random.
  - If there are  $n$  nodes, and  $m$  edges randomly added, then the peak of this is at  $2m/n$ , the average degree.
  - For a randomly picked node, the most likely degree is the average one.
  - The probabilities then drop quickly either side.

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

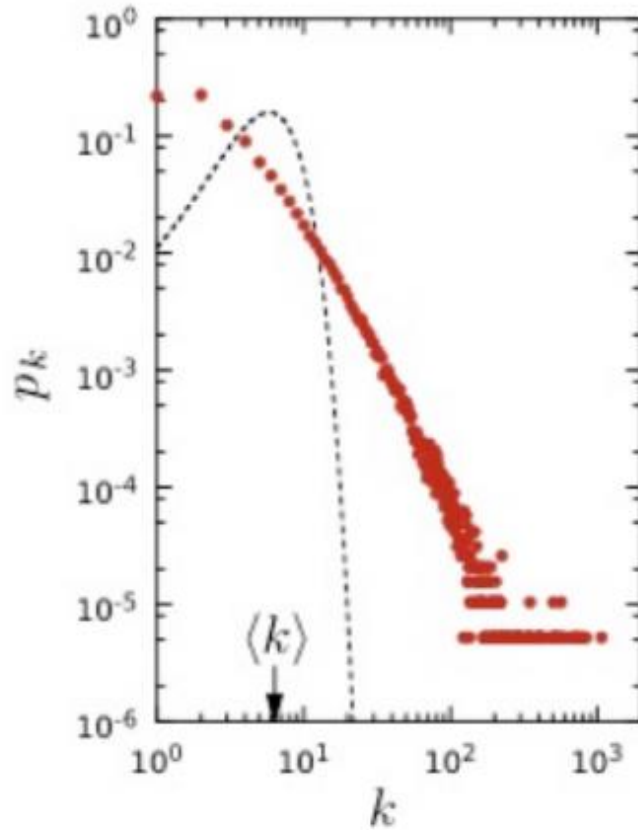
where  $\langle k \rangle$  is the average degree of the network.



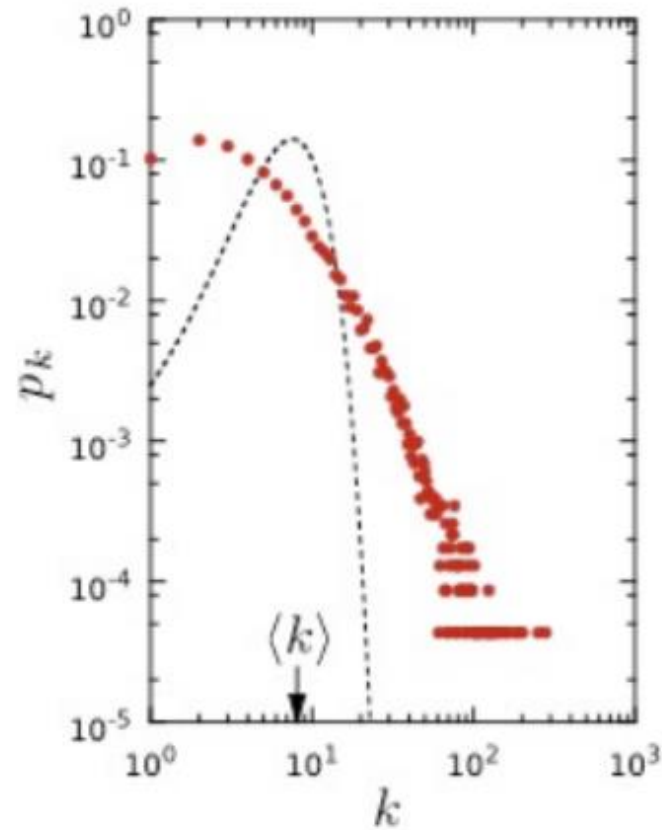
# Degree Distribution

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

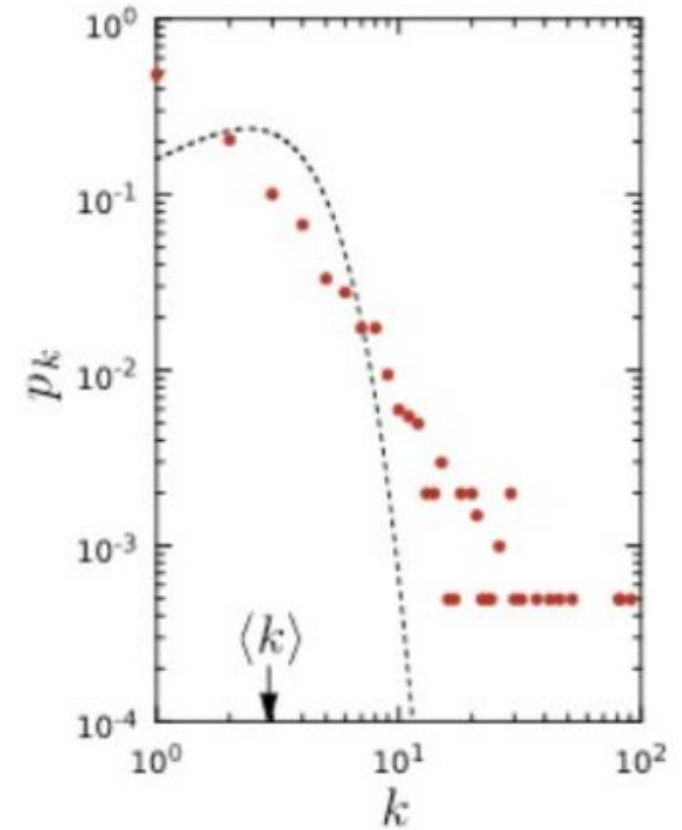
Internet



Science Collaboration

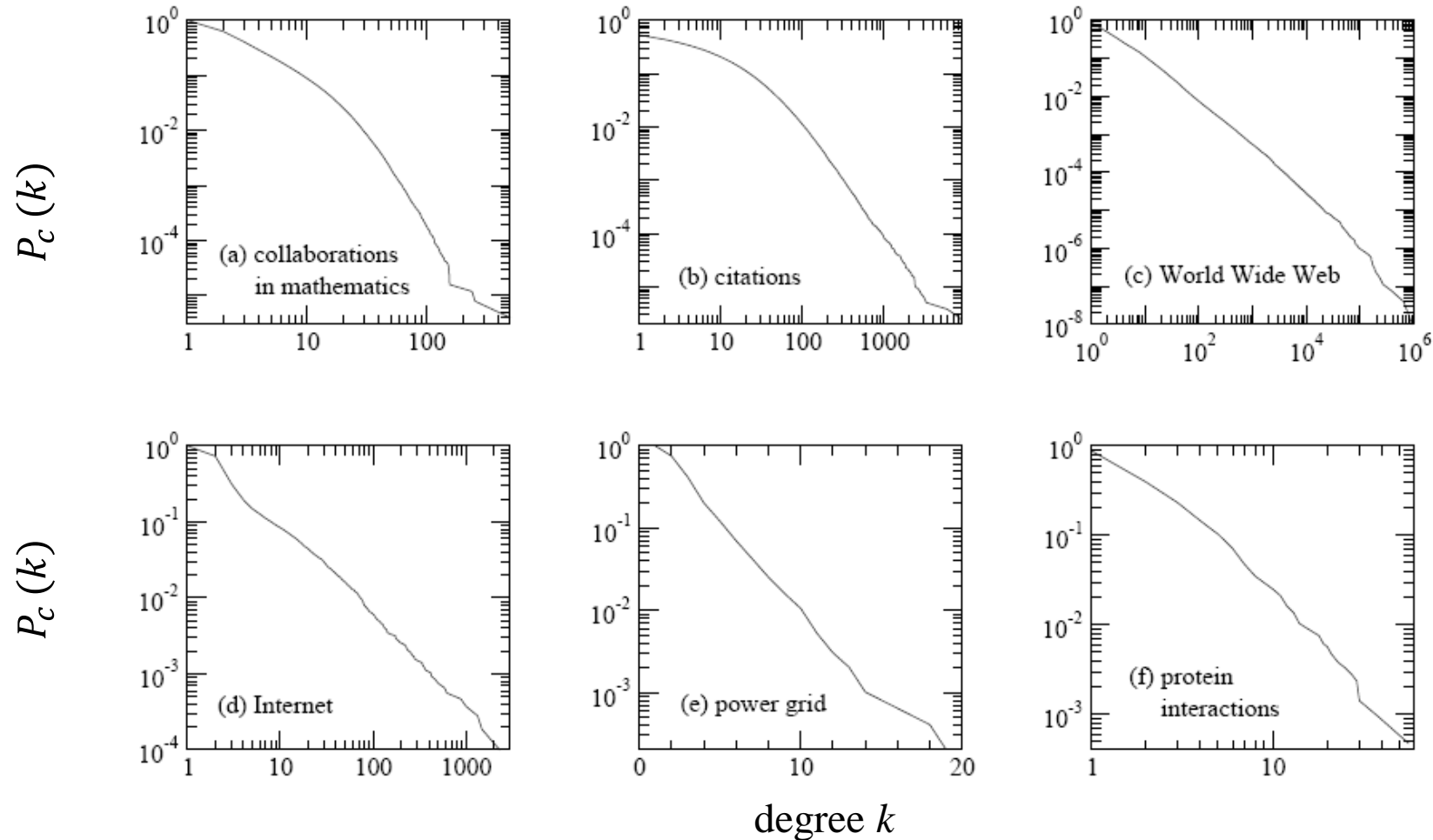


Protein Interactions



# Degree Distribution

The (cumulative) degree distribution ( $P_c(k)$ )  $P(k)$  accounts for the fraction of nodes in the network with a degree (higher than) equal to  $k$ .



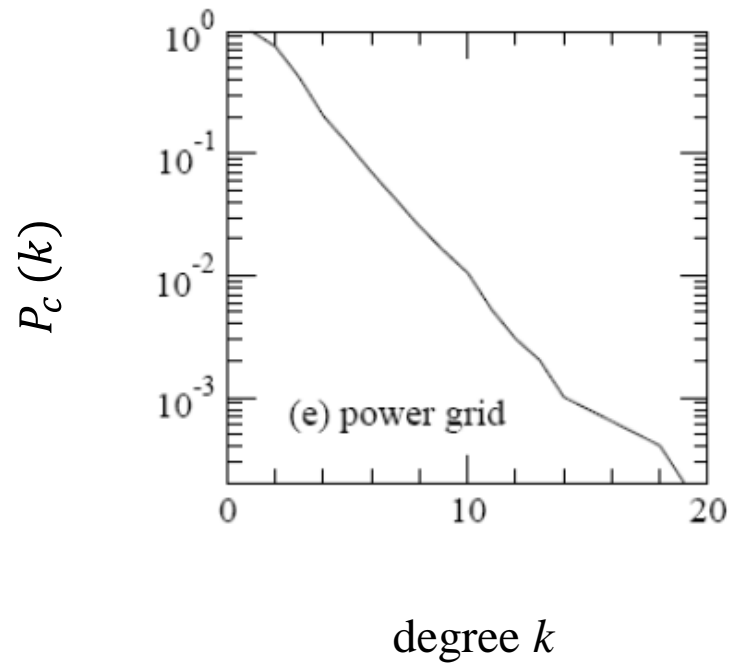
$$P_c(k) \sim k^{-\gamma}$$



# Degree Distribution

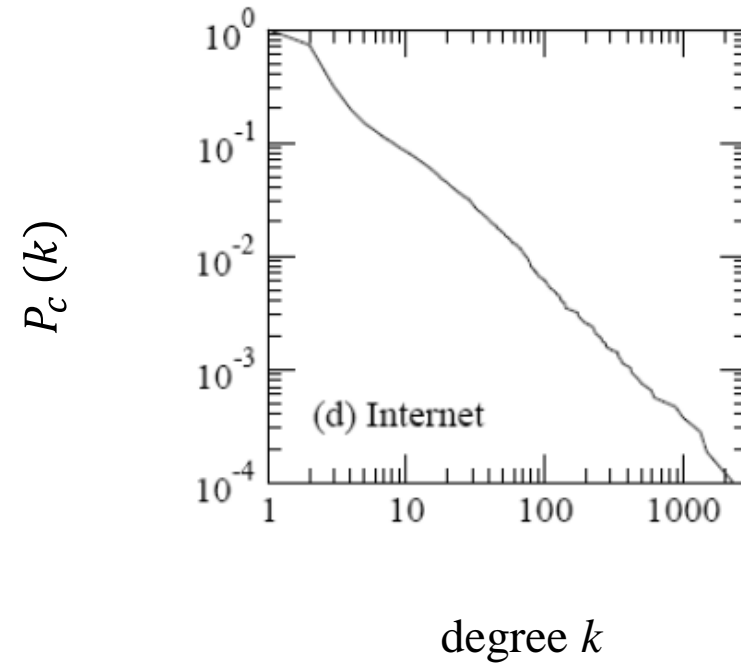
Two types of degree distribution appear more frequently in real networks:

Exponential decay:  $P_c(k) \sim e^{-\alpha k}$



Typical in random networks

Power-law decay:  $P_c(k) \sim k^{-\gamma}$

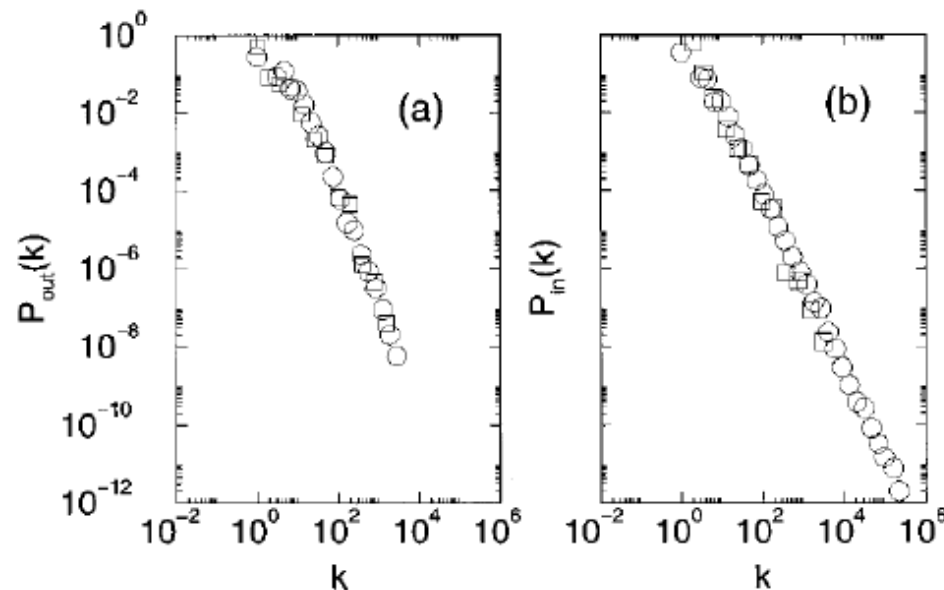


Networks with power-law decay  
are called scale-free networks

# Degree Distribution

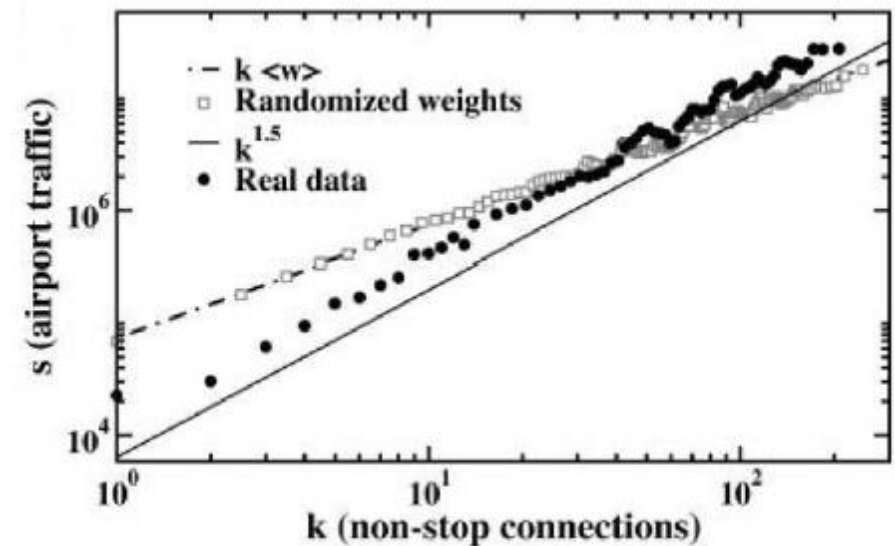
*Other related distributions:*

In/out degree distributions  
(directed networks)



In/out degree distributions of WWW (R. Albert et al., Rev. Mod. Phys. 74, 47(2002))

Strength distribution  
(weighted networks)



Strength distribution of the International Air Transportation Network ([www.iata.org](http://www.iata.org)).  
From A. Barrat et al., PNAS, 101, 3747 (2004).

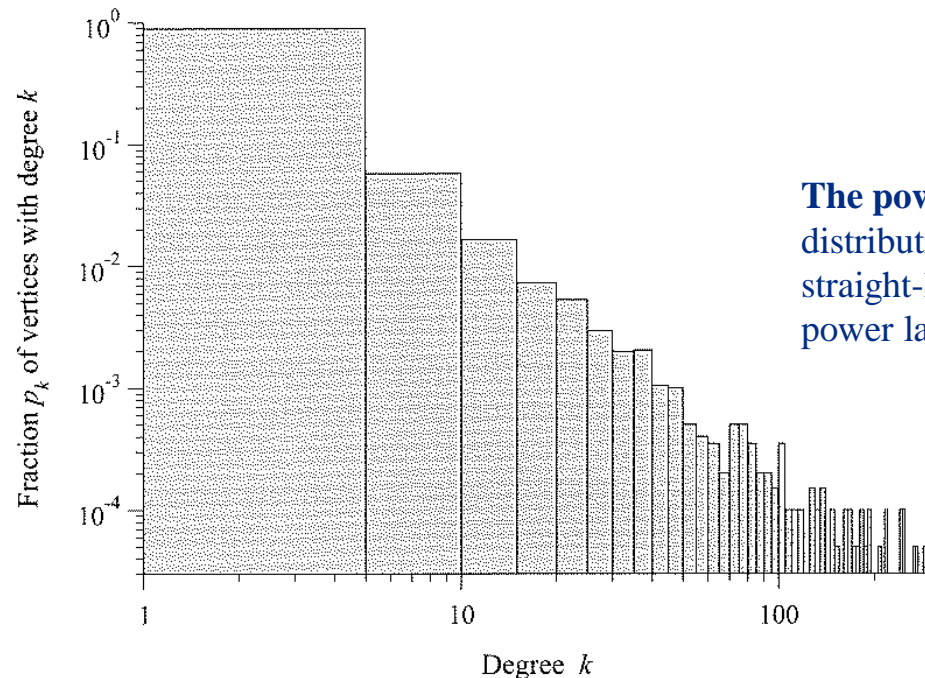
# Degree Distribution

| Network                     | Size             | $\langle k \rangle$ | $\kappa$ | $\gamma_{out}$ | $\gamma_{in}$ |
|-----------------------------|------------------|---------------------|----------|----------------|---------------|
| WWW                         | 325 729          | 4.51                | 900      | 2.45           | 2.1           |
| WWW                         | $4 \times 10^7$  | 7                   |          | 2.38           | 2.1           |
| WWW                         | $2 \times 10^8$  | 7.5                 | 4000     | 2.72           | 2.1           |
| WWW, site                   | 260 000          |                     |          |                | 1.94          |
| Internet, domain*           | 3015–4389        | 3.42–3.76           | 30–40    | 2.1–2.2        | 2.1–2.2       |
| Internet, router*           | 3888             | 2.57                | 30       | 2.48           | 2.48          |
| Internet, router*           | 150 000          | 2.66                | 60       | 2.4            | 2.4           |
| Movie actors*               | 212 250          | 28.78               | 900      | 2.3            | 2.3           |
| Co-authors, SPIRES*         | 56 627           | 173                 | 1100     | 1.2            | 1.2           |
| Co-authors, neuro.*         | 209 293          | 11.54               | 400      | 2.1            | 2.1           |
| Co-authors, math.*          | 70 975           | 3.9                 | 120      | 2.5            | 2.5           |
| Sexual contacts*            | 2810             |                     |          | 3.4            | 3.4           |
| Metabolic, <i>E. coli</i>   | 778              | 7.4                 | 110      | 2.2            | 2.2           |
| Protein, <i>S. cerev.</i> * | 1870             | 2.39                |          | 2.4            | 2.4           |
| Ythan estuary*              | 134              | 8.7                 | 35       | 1.05           | 1.05          |
| Silwood Park*               | 154              | 4.75                | 27       | 1.13           | 1.13          |
| Citation                    | 783 339          | 8.57                |          |                | 3             |
| Phone call                  | $53 \times 10^6$ | 3.16                |          | 2.1            | 2.1           |
| Words, co-occurrence*       | 460 902          | 70.13               |          | 2.7            | 2.7           |
| Words, synonyms*            | 22 311           | 13.48               |          | 2.8            | 2.8           |

# Degree Distribution

## Visualization of Power-law Decay:

**Problem:** The statistics of the histogram are poor in the tail of the distribution, the *large- $k$*  region, which is precisely the region in which the power law is normally followed most closely. Each bin of the histogram in this region contains only a few samples, which means that statistical fluctuations in the number of samples from bin to bin are large. This is visible as a "noisy signal" at the right hand end.

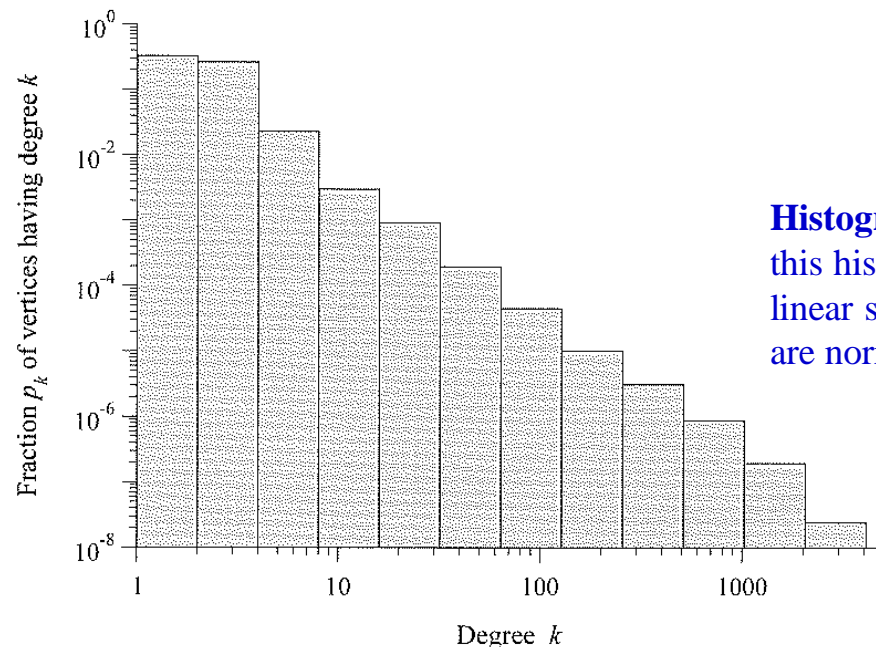


**The power-law degree distribution of the Internet.** Histogram of the degree distribution of the Internet graph, plotted this time on logarithmic scales. The approximate straight-line form of the histogram indicates that the degree distribution roughly follows a power law.

# Degree Distribution

## Visualization of Power-law Decay:

**Solution:** One commonly used version of this idea is called *logarithmic binning*. In this scheme, each bin is made wider than its predecessor by a constant factor  $a$ . For instance, if the first bin in a histogram covers the interval  $1 \leq k < 2$  (meaning that all vertices of degree 1 fall in this bin) and  $a = 2$ , then the second would cover the interval  $2 \leq k < 4$  (vertices of degrees 2 and 3), the third the interval  $4 \leq k < 8$ , and so forth. In general the  $n$ -th bin would cover the interval  $a^{n-1} \leq k < a^n$  and have width  $a^n - a^{n-1} = (a - 1)a^{n-1}$ . The most common choice for  $a$  is  $a = 2$ , since larger values tend to give bins that are too coarse while smaller ones give bins with non-integer limits.



**Histogram of the degree distribution of the Internet, created using logarithmic binning.** In this histogram the widths of the bins are constant on a logarithmic scale, meaning that on a linear scale each bin is wider by a constant factor than the one to its left. The counts in the bins are normalized by dividing by bin width to make counts in different bins comparable.

# *About Power-Law Decay Behavior*



Vilfredo  
Pareto



## **Pareto Principle (also known as the 80-20 rule)**

For many events, roughly 80% of the effects come from 20% of the causes. It is named after the Italian economist Vilfredo Pareto, who observed in 1906 that 80% of the land in Italy was owned by 20% of the population; he developed the principle by observing that 20% of the pea pods in his garden contained 80% of the peas.



# *About Power-Law Decay Behavior*

## *Pareto Distribution (1896)*

If  $X$  is a random variable with a Pareto distribution, then the probability that  $X$  is greater than some number  $x$  is given by

$$\Pr[X \geq x] = \left( \frac{x^{-\alpha}}{k} \right)$$

where  $\alpha$  and  $k$  are positive parameters.

Vilfredo Pareto. *Cours d'économie politique professé à l'université de Lausanne*. Vol. I' 1896; Vol. II, 1897.

# *About Power-Law Decay Behavior*

## *Examples of Pareto Principle in Business*

- 80 % of a company's revenues are generated by 20 % of its customers
- 80 % of complaints come from 20 % of customers
- 80 % of quality issues impact 20 % of a company's products
  
- 20 % of the work on a project consumes 80 % of time and resources
- 20 % of investors provide 80 % of funding
- 20 % of employees use 80 % of all sick days
- 20 % of a blog's posts generate 80 % of its traffic



# *About Power-Law Decay Behavior*

## *Zipf's Law*

*Zipf's law* is an empirical law formulated using mathematical statistics, refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution. It states that, in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table. So, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.

$$p(r) \sim r^{-\alpha} \ ; \ \alpha \sim 1$$

K.G. Zipf, *The Psycho-Biology of Language*, Cambridge (Mass), 1935; *Human Behavior and the Principle of least Effort*, 1949.

# About Power-Law Decay Behavior

## Zipf's Law and city sizes (~1930)

| Rank(k) | City           | Population<br>(1990) | Zipf's Law<br>$10,000,000/k$ | Modified Zipf's law:<br>(Mandelbrot)<br>$5,000,000 / (k - 2/5)^{3/4}$ |
|---------|----------------|----------------------|------------------------------|-----------------------------------------------------------------------|
| 1       | New York       | 7,322,564            | 10,000,000                   | 7,334,265                                                             |
| 7       | Detroit        | 1,027,974            | 1,428,571                    | 1,214,261                                                             |
| 13      | Baltimore      | 736,014              | 769,231                      | 747,693                                                               |
| 19      | Washington DC  | 606,900              | 526,316                      | 558,258                                                               |
| 25      | New Orleans    | 496,938              | 400,000                      | 452,656                                                               |
| 31      | Kansas City    | 434,829              | 322,581                      | 384,308                                                               |
| 37      | Virginia Beach | 393,089              | 270,270                      | 336,015                                                               |
| 49      | Toledo         | 332,943              | 204,082                      | 271,639                                                               |
| 61      | Arlington      | 261,721              | 163,932                      | 230,205                                                               |
| 73      | Baton Rouge    | 219,531              | 136,986                      | 201,033                                                               |
| 85      | Hialeah        | 188,008              | 117,647                      | 179,243                                                               |
| 97      | Bakersfield    | 174,820              | 103,270                      | 162,270                                                               |

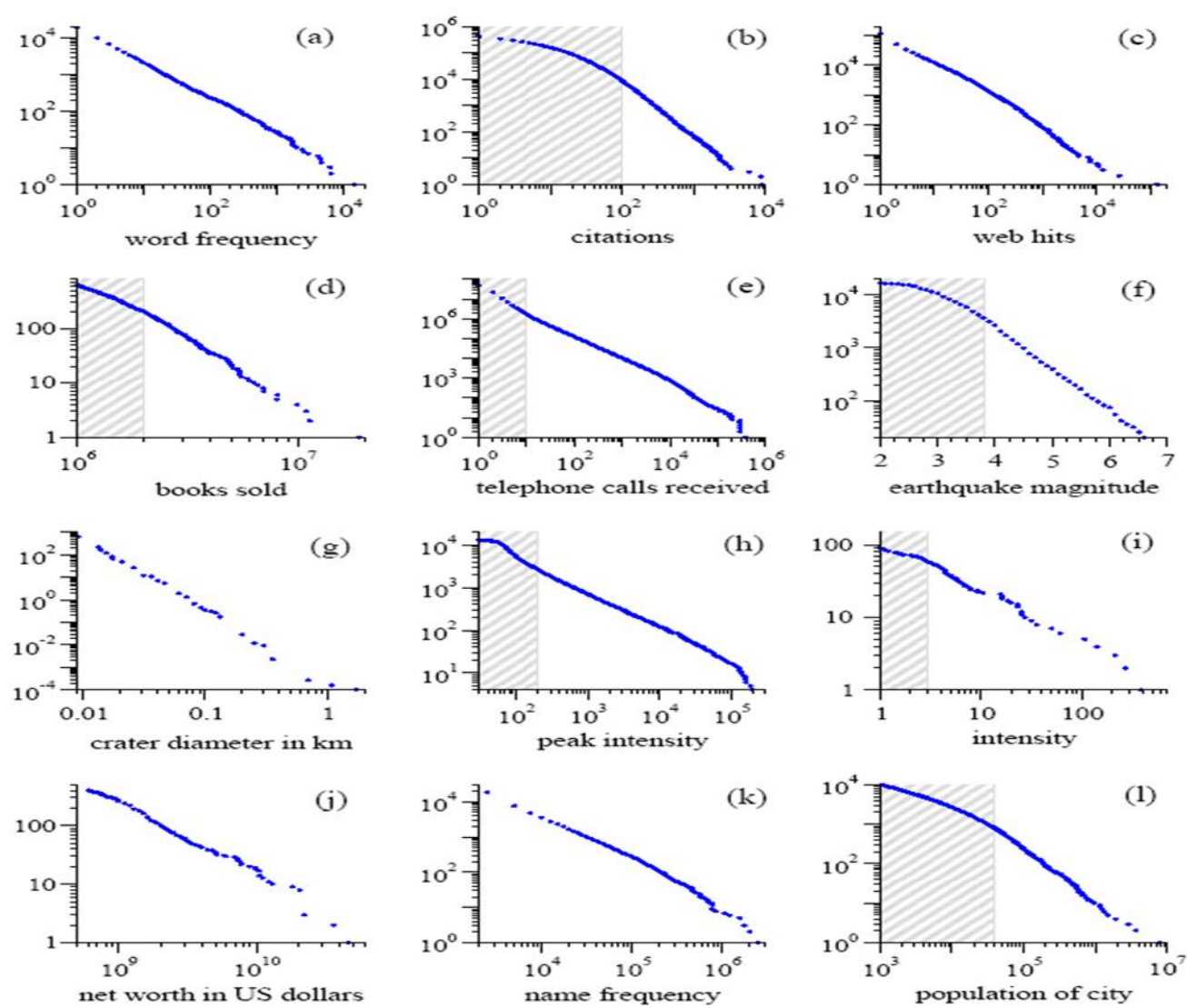
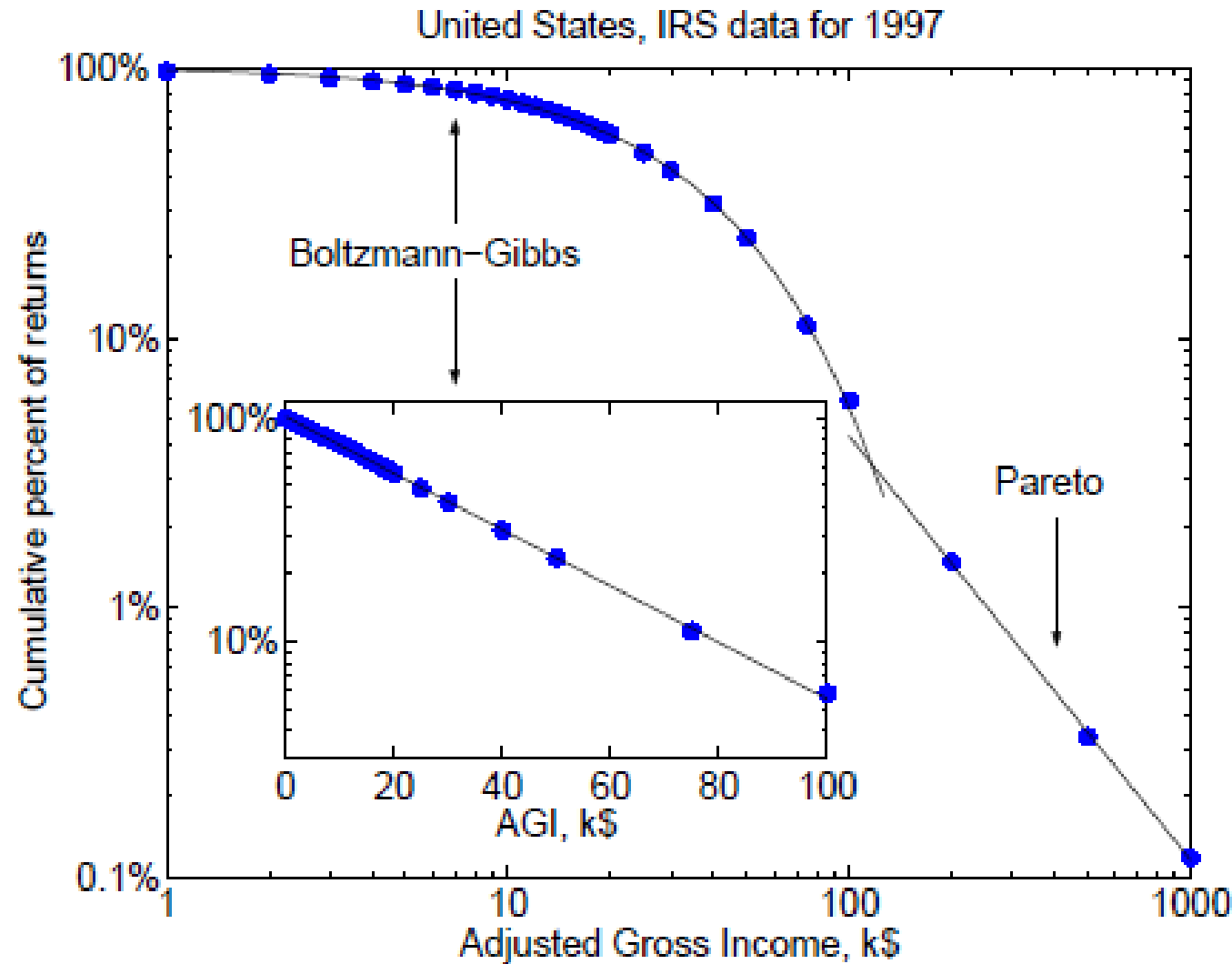


FIG. 4 Cumulative distributions or “rank/frequency plots” of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponent in Table I. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel *Moby Di* by Hermann Melville. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1999. (c) Numbers of hits on web sites by 60 000 users of the America Online Internet service for the day of 1 December 1997. (d) Number of copies of bestselling books sold in the US between 1895 and 1965. (e) Number of calls received by AT&T telephone customers in the US for a single day. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometre. (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10 000 of the population of the participating countries. (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. (k) Frequency of occurrence of family names in the US in the year 1990. (l) Populations of US cities in the year 2000.

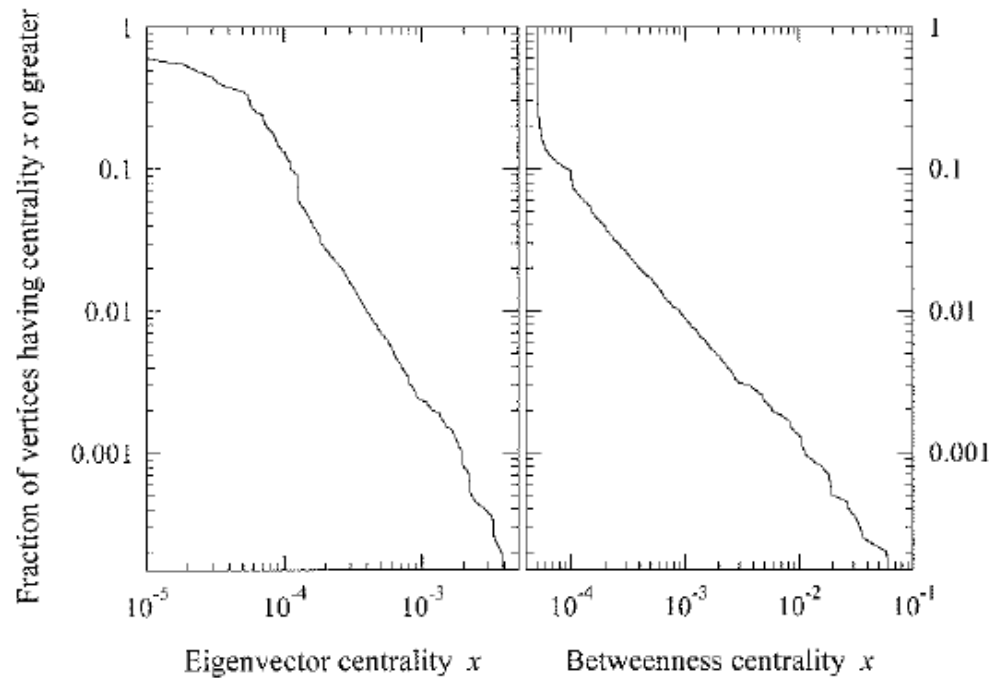
# About Power-Law Decay Behavior



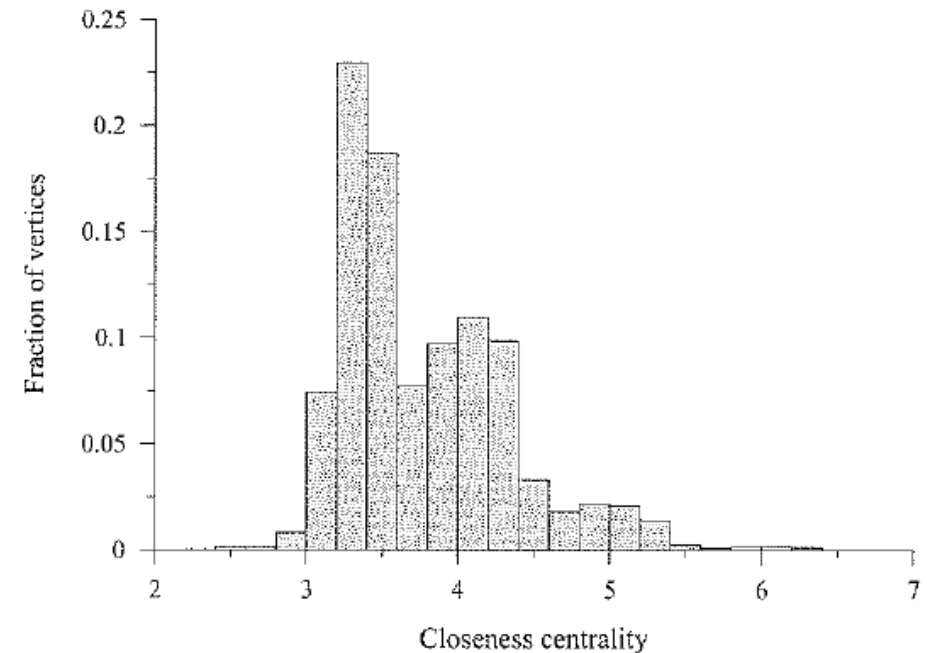
Cumulative probability distribution of US individual income for 1997 in log-log scale, with points (raw data) and solid lines (exponential and power-law fit). The inset shows the exponential regime and the fit with a Boltzmann-Gibbs distribution in the log-linear scale.<sup>20</sup>

# Distribution of other centrality measures

Since *eigenvector centrality* can be thought of as an extended form of *degree centrality*, in which we take into account not only how many neighbors a vertex has but also how central those neighbors themselves are. It is not surprising to learn that *eigenvector centrality* often has a highly right-skewed distribution. *Betweenness centrality* also tends to have right-skewed distributions on most networks. *Closeness centrality*, on the other hand, shows very different behavior.



**Cumulative distribution functions for centralities of vertices on the Internet.** Left panel: eigenvector centrality. Right panel: betweenness centrality.



**Histogram of closeness centralities of vertices on the Internet.** This is a normal non-cumulative histogram showing the actual distribution of closeness centralities. This distribution does not follow a power<sup>21</sup> law.

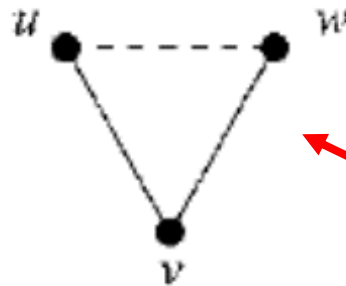
# Transitivity

➤  $\circ$  is said to be transitive if  $a \circ b$  and  $b \circ c$  together imply  $a \circ c$

➤ Perfect transitivity in network  $\rightarrow$  cliques

➤ Partial transitivity

➤  $u$  knows  $v$  and  $v$  knows  $w \rightarrow$



The path  $uvw$  (solid edges) is said to be closed if the third edge directly from  $u$  to  $w$  is present (dashed edge).

# Transitivity

- **Triple**: an *ordered* set of three nodes
  - connected by two (open triple) edges or
  - three edges (closed triple)

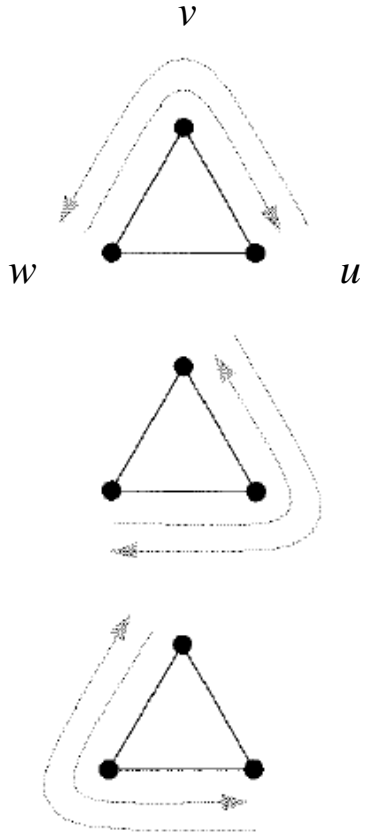
$v_i v_j v_k$  and  $v_j v_k v_i$  are different triples

- The **same members**
- First missing edge  $e(v_k, v_i)$  and second missing  $e(v_i, v_j)$

- A triangle can miss any of its three edges
  - A triangle has **3 Triples**

$v_i v_j v_k$  and  $v_k v_j v_i$  are the same triple

# Transitivity



The triangle on the left contains six distinct paths of length two, all of them closed. Then there are six paths of length two in it:  $uvw$ ,  $vuw$ ,  $wuv$ ,  $vwu$ ,  $vuw$ , and  $uwv$ . Each of these six is closed, so the number of closed paths is six times the number of triangles, giving a global clustering coefficient  $C$

$$C = \frac{\text{number of closed paths of length two}}{\text{number of paths of length two}} = \frac{6 \times (\text{number of triangles})}{(\text{number of paths of length two})}$$

Another way to write the clustering coefficient would be to note that if we have a path of length two,  $uvw$ , then it is also true to say that vertices  $u$  and  $w$  have a common neighbor in  $v$  – they share a mutual acquaintance in social network terms. Therefore, the clustering coefficient can be thought of as the fraction of pairs of people with a common friend who are themselves friends or equivalently as the mean probability that two people with a common friend are themselves friends.

$$C = \frac{\text{number of closed paths of length two}}{\text{number of paths of length two}} = \frac{3 \times (\text{number of triangles})}{(\text{number of connected triples})}$$

$C = 1$  implies perfect transitivity, i.e., a network whose components are all cliques.  $C = 0$  implies no closed triads, which happens for various topologies, such as a tree (which has no closed loops of any kind) or a square lattice.



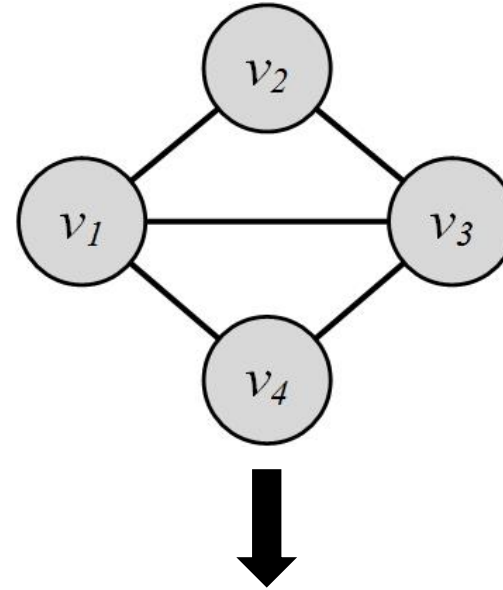
# Clustering Coefficients

## Global Clustering Coefficient

- Measures ratio of transitive triples
- $C_{\Delta} = \frac{3 \times \text{Number of Triangles}}{\text{Number of Connected Triples}}$ 
  - A *connected triple* is an ordered set of three nodes ABC such that A connects to B and B connects to C
- Average clustering coefficient and global clustering coefficient are different.
  - In some extreme cases they could differ considerably.

# Clustering Coefficients

## Global Clustering Coefficient: Example



$$C = \frac{(\text{Number of Triangles}) \times 3}{\text{Number of Connected Triples of Nodes}}$$

$$= \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2 v_1 v_4, v_2 v_3 v_4}} = 0.75.$$

# Clustering Coefficients

The (*local*) *clustering coefficient* measures the average probability that two neighbors of a vertex are themselves neighbors. In effect it measures the density of triangles in the networks and it is of interest because in many cases it is found to have values sharply different from what one would expect on the basis of chance.

Node  $i$  with degree  $k_i$

$e_i$  , the number of edges among the neighbors of  $i$

$C_i$  is the fraction of pairs of neighbors (of the same vertex) that are also neighbors of each other.

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

$$0 \leq C_i \leq 1$$

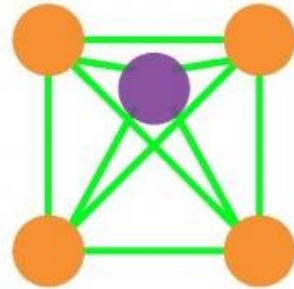
$C_i = 0$  if none of the neighbors of node  $i$  link to each other.

$C_i = 1$  if the neighbors of node  $i$  form a complete graph, i.e., they all link to each other.

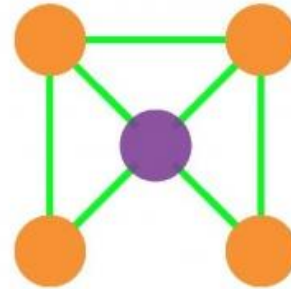
# Clustering Coefficients

## Examples:

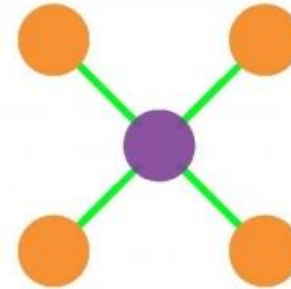
a.



$$C_i = 1$$

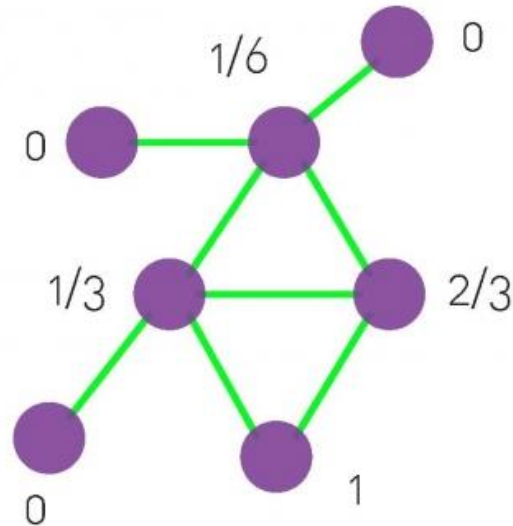


$$C_i = 1/2$$



$$C_i = 0$$

b.



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

$$C_{\Delta} = \frac{3}{8} = 0.375$$

# Clustering Coefficients

## Local Clustering and Redundancy

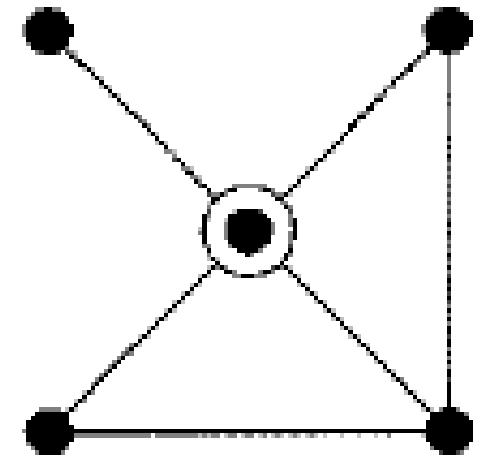
$$C_i = \frac{\text{connected pairs of neighbors of } i}{\text{pairs of neighbors of } i}$$

$$C_{WS} = \frac{1}{n} \sum_{i=1}^n C_i$$

- Redundancy

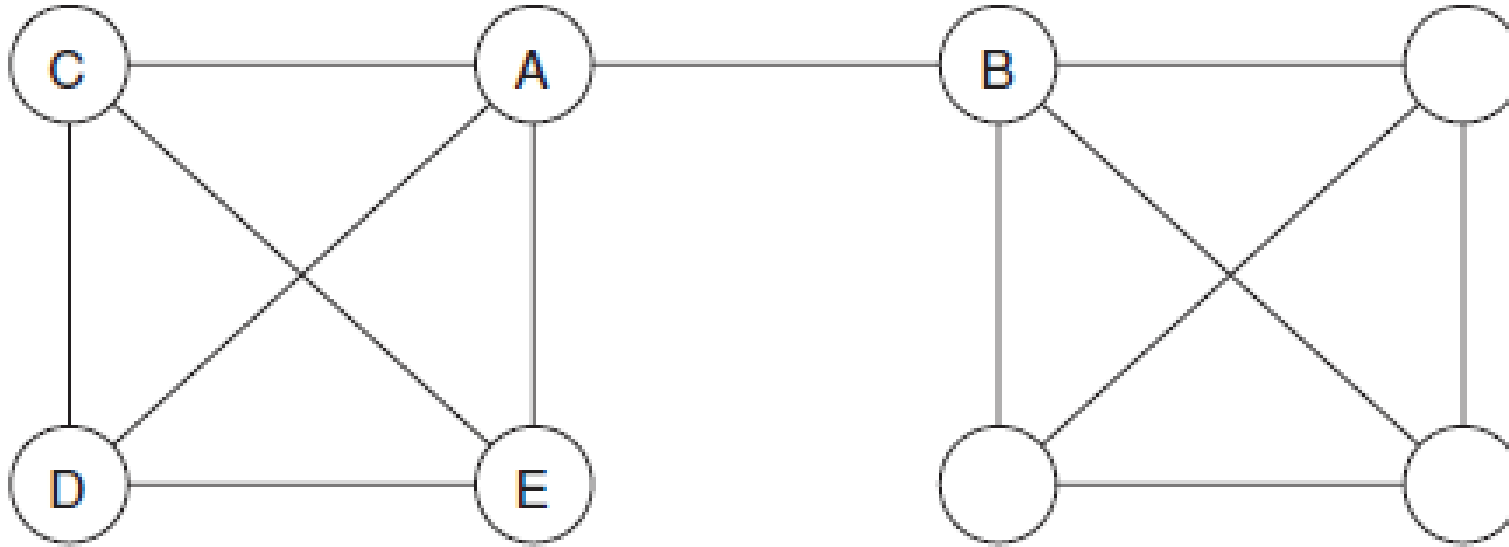
$$C_i = \frac{R_i}{k_i - 1}$$

$$R_i = C_i(k_i - 1)$$



**Redundancy.** The neighbors of the central vertex in this figure have 0, 1, 1, and 2 connections to other neighbors respectively. The redundancy is the mean of these values:  $R_i = \frac{1}{4}(0 + 1 + 1 + 2) = 1$ .

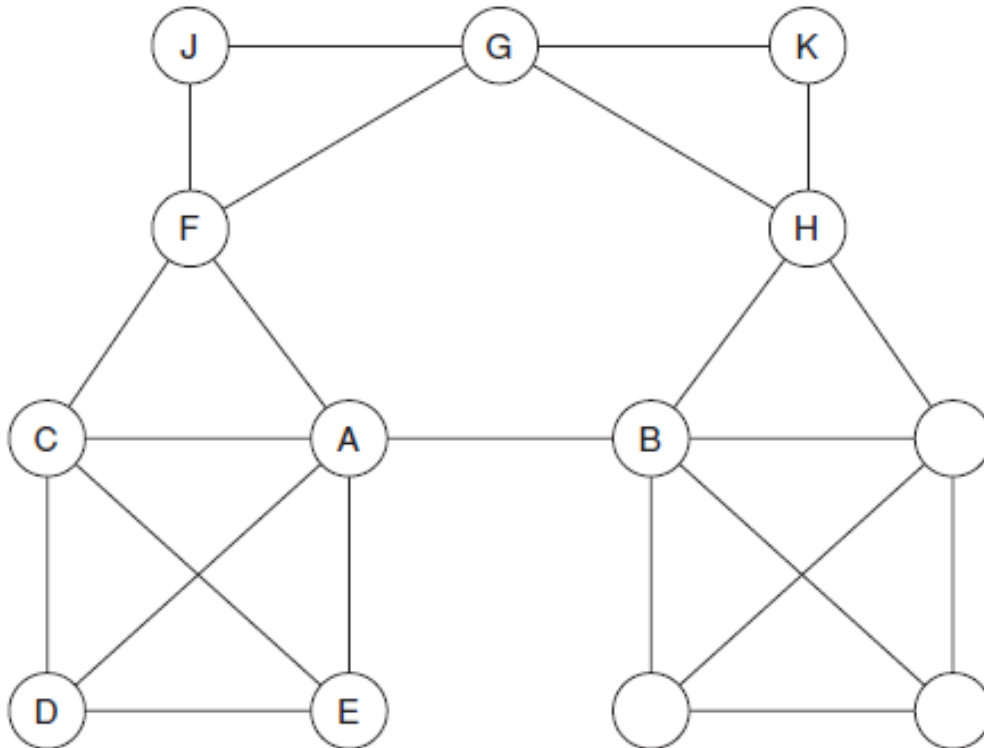
## *Bridges:*



The A-B edge is a **bridge**, meaning that its removal would place A and B in distinct components. Bridges provide nodes with access to parts of the network that are unreachable by other means.

## Bridges: Local bridges

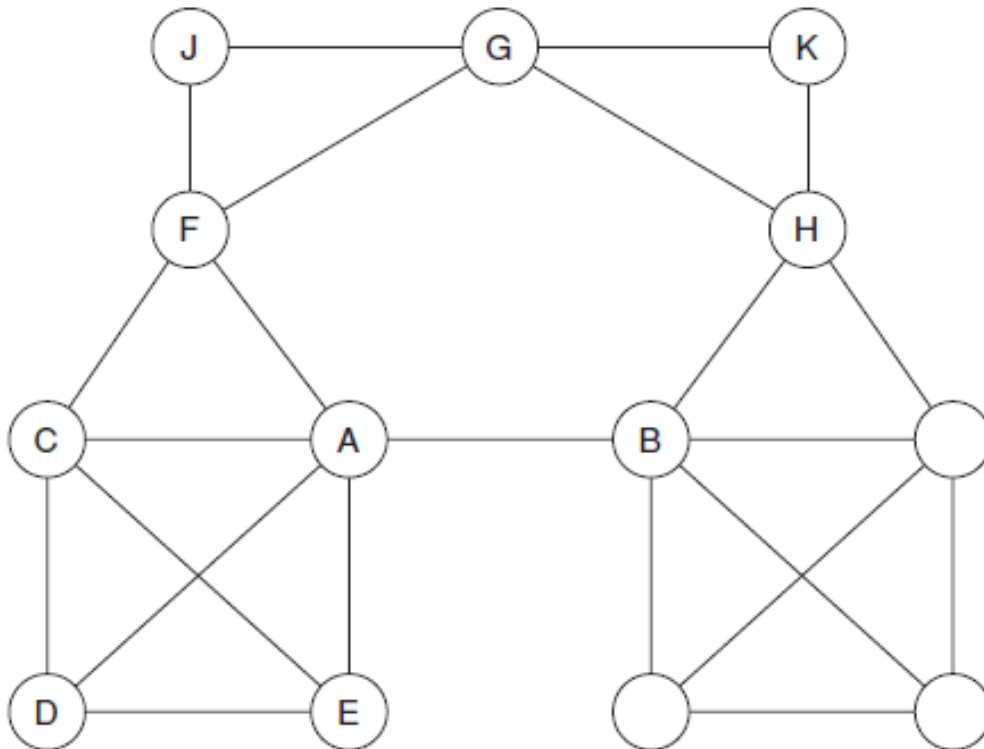
An edge joining two nodes A and B in a graph is a **local bridge** if its endpoints A and B have **no** friends in common – i.e., deleting the edge would increase the distance between A and B to a value strictly more than 2. The **span** of a local bridge is the distance its endpoints would be from each other if the edge were deleted. The definition of a local bridge already makes an implicit connection with triadic closure – an edge is a local bridge precisely when it does not form the side of any triangle in the graph.



The A-B edge is a *local bridge* of **span 4**, since the removal of this edge would increase the distance between A and B to 4.

## Bridges: Strong and Weak Ties

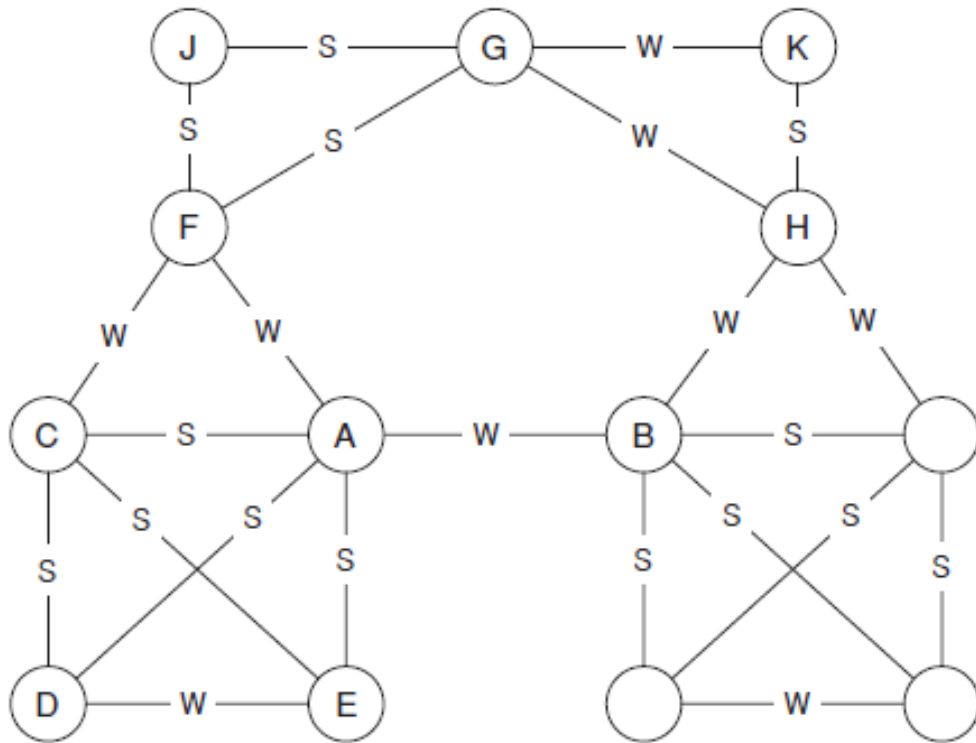
Granovetter (1960) found that many people learned information leading to their current jobs through personal contacts. More strikingly, these personal contacts were often described by interview subjects as acquaintances rather than close friends. His observation suggests that if a node like A is going to get truly new information – the kind that leads to a new job – it might come from a friend, B connected by a local bridge. The closely-knit groups to which you belong, although they are filled with people eager to help, are also filled with people who know roughly the same things that you do.





## *Bridges: Strong and Weak Ties*

In general, links can have a wide range of possible strengths, but for conceptual simplicity – we will categorize all links in the social network as belonging to one of two types: strong ties (the stronger links, corresponding to friends) and weak ties (the weaker links, corresponding to acquaintances).

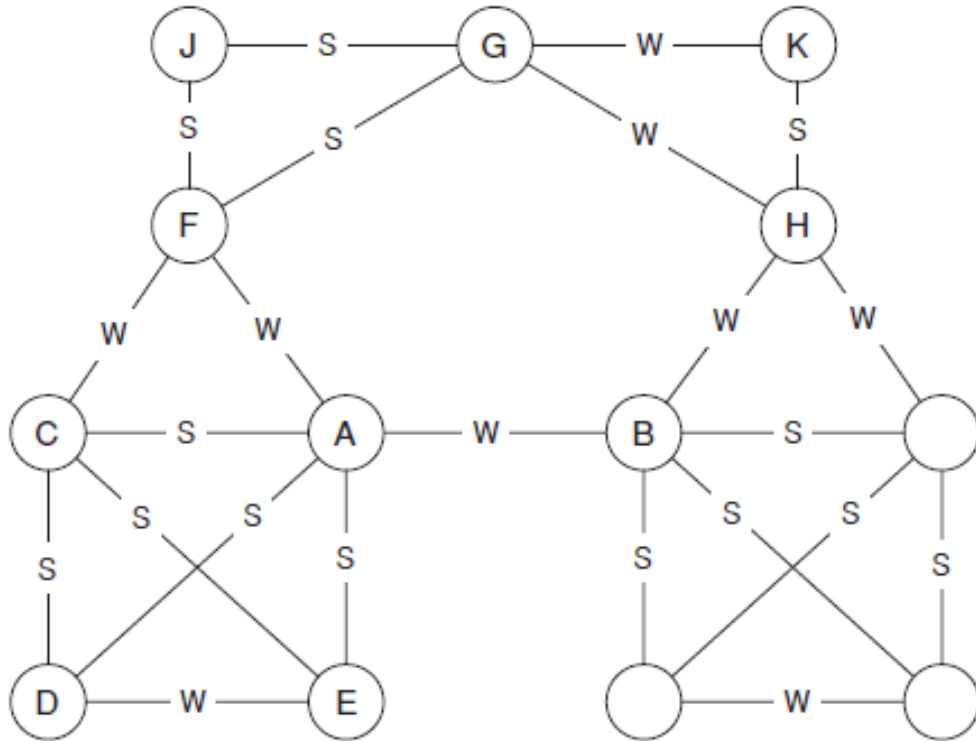


Each edge of the social network from left figure is labeled as either a strong tie ( $S$ ) or a weak tie ( $W$ ) to indicate the strength of the relationship. The labeling in the figure satisfies the Strong Triadic Closure property at each node: if the node has strong ties to two neighbors, then these neighbors must have at least a weak tie between them.

## Bridges: Strong and Weak Ties

**Strong Triadic Closure Property:** If a node A has edges to nodes B and C, then the B-C edge is especially likely to form if A's edges to B and C are both strong ties.

Granovetter's observation implies – A node A violates the Strong Triadic Closure property if it has strong ties to two other nodes B and C, and there is no edge at all (either a strong or weak tie) between B and C. We say that a node A satisfies the Strong Triadic Closure property if it does not violate it.

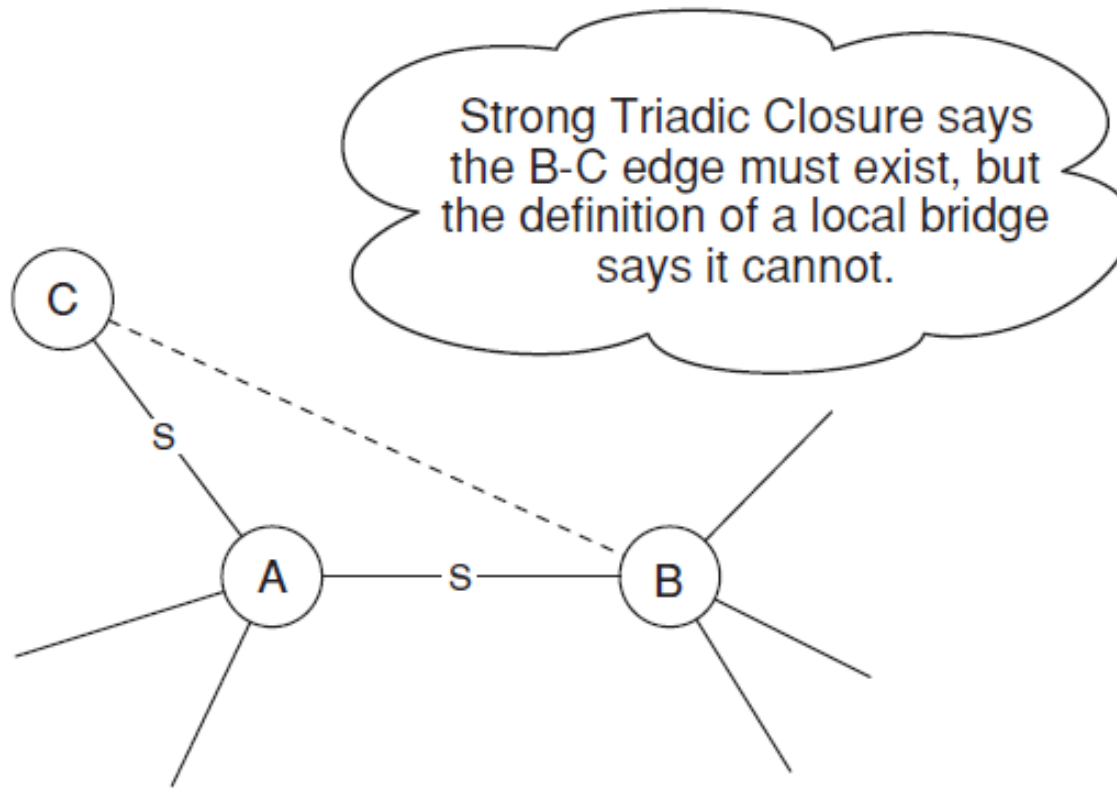


In the figure, no node violates the Strong Triadic Closure property; hence, all nodes satisfy the property. *If* the A-F edge were to be a *strong* tie rather than a *weak* tie, then nodes A and F would both *violate* the Strong Triadic Closure property: node A would now have strong ties to nodes E and F without there being an E-F edge, and node F would have strong ties to both A and G without the presence of an A-G edge. Node H satisfies the Strong Triadic Closure property since it only has a strong tie to one other node.

# Bridges: Strong and Weak Ties

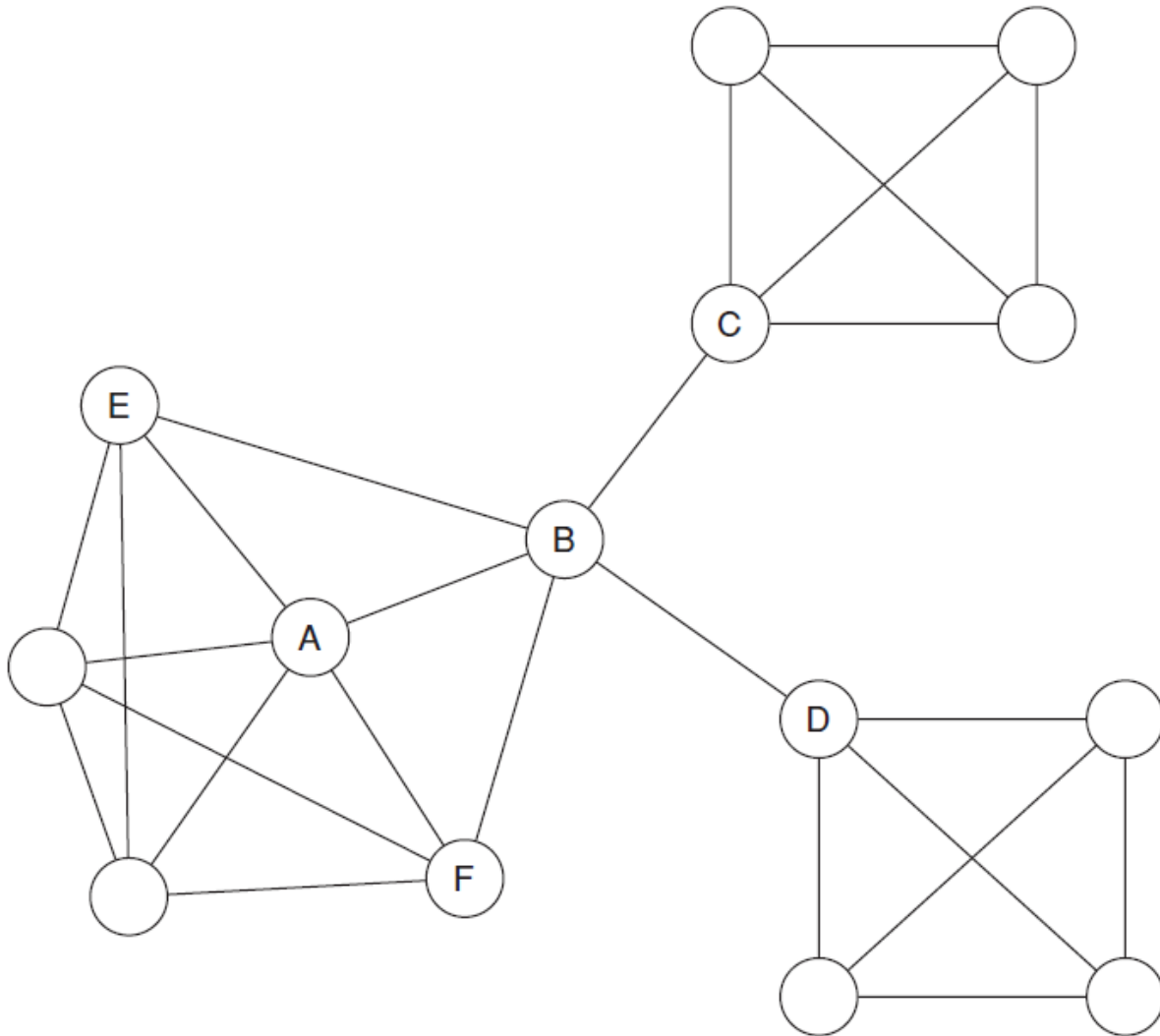
## ***Local Bridge and Weak Ties:***

If a node A in a network satisfies the Strong Triadic Closure property and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.



If a node satisfies Strong Triadic Closure and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie. The figure illustrates the reason why: if the A-B edge is a strong tie, then there must also be an edge between B and C, meaning that the A-B edge cannot be a local bridge.

# Bridges: Embeddedness, Structure Holes, Social Capital



**Embeddedness** of an edge in a network to be the number of common neighbors shared by the two endpoints. The A-B edge has an embeddedness of 2, because A and B have the two common neighbors E and F.

Nodes with multiple local bridges, spans a **structural hole** in the organization – the “empty space” in the network between two sets of nodes that do not otherwise interact closely.

In the literature that **social capital** stands for the ability of actors to secure benefits by virtue of membership in social networks or other social structures

# *Reciprocity*

*How likely is it that the node you point to will point to you as well?*

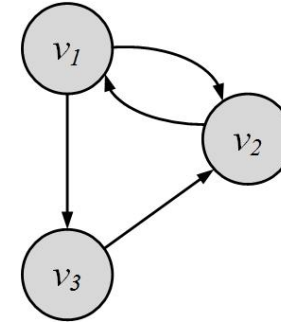
If there is a directed edge from vertex  $i$  to vertex  $j$  in a directed network and there is also an edge from  $j$  to  $i$  then the edge from  $i$  to  $j$  is ***reciprocated***. (Likewise,  $j$  to  $i$  is also ***reciprocated***.) Pairs of edges like this are also called ***co-links***. The reciprocity  $r$  is defined as

$$r = \frac{1}{m} \sum_{ij}^n A_{ij} A_{ji} = \frac{1}{m} \text{Tr} A^2$$

# Reciprocity

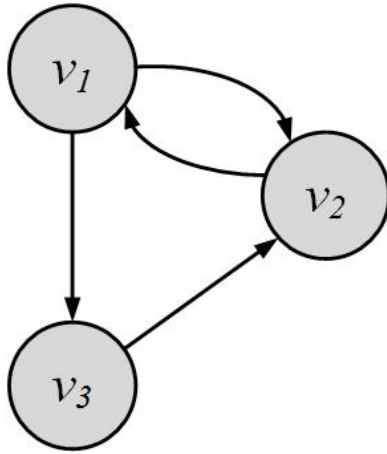
*If you become my friend, I'll be yours*

- **Reciprocity** is a simplified version of transitivity
  - It considers closed loops of length 2
- If node  $v$  is connected to node  $u$ ,
  - $u$  by connecting to  $v$ , exhibits **reciprocity**



$$r = \frac{1}{m} \sum_{ij}^n A_{ij} A_{ji} = \frac{1}{m} \text{Tr} A^2$$

## *Reciprocity: Example*



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



Reciprocal nodes:  $v_1, v_2$

$$r = \frac{1}{m} \text{Tr} A^2 = \frac{1}{4} \text{Tr} \left( \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \right) = \frac{2}{4} = \frac{1}{2}$$

# *Signed Edges and Structural balance*

- Friends / Enemies
- Friend of friend  $\rightarrow$  ?
- Enemy of my enemy  $\rightarrow$  ?
- **Structural balance:** only loops of even number of “negative links”
- Structurally balanced  $\rightarrow$  partitioned into groups where internal links are positive and between group links are negative



# *Social Balance Theory*

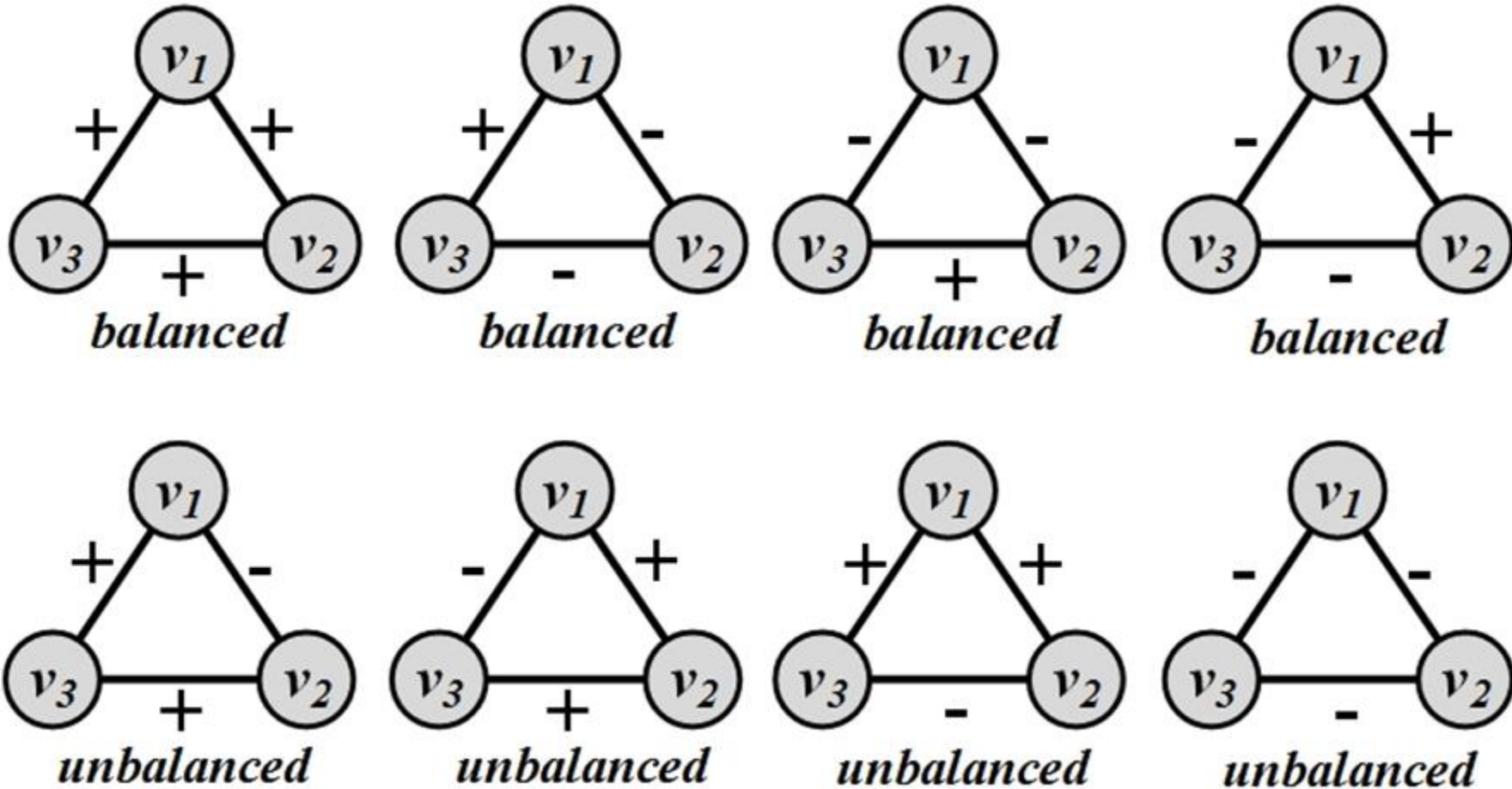
- Consistency in friend/foe relationships among individuals
- Informally, friend/foe relationships are consistent when

*The friend of my friend is my friend,  
The friend of my enemy is my enemy,  
The enemy of my enemy is my friend,  
The enemy of my friend is my enemy.*

- In the network
  - Positive edges demonstrate friendships ( $w_{ij} = 1$ )
  - Negative edges demonstrate being enemies ( $w_{ij} = -1$ )
- Triangle of nodes  $i, j$ , and  $k$ , is balanced, if and only if
  - $\omega_{ij}$  denotes the value of the edge between nodes  $i$  and  $j$

$$\omega_{ij}\omega_{jk}\omega_{ki} \geq 0$$

## *Social Balance Theory: Possible Combinations*

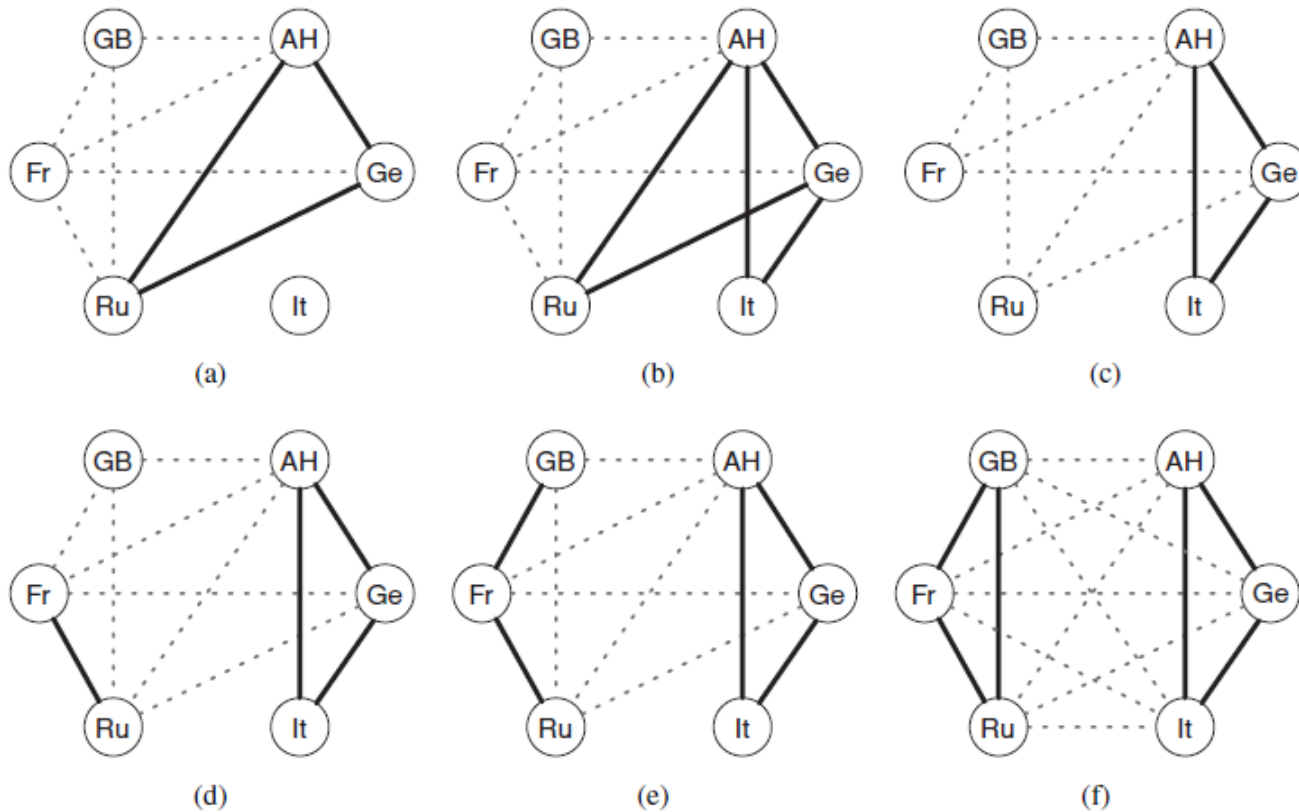


For any *cycle* if the multiplication of edge values become positive, then the cycle is socially balanced.

# Social Balance Theory: Theorem and Applications

**The Balance Theorem:** If a labeled complete graph is balanced, then either all pairs of nodes are friends, or else the nodes can be divided into two groups,  $X$  and  $Y$ , such that each pair of people in  $X$  likes each other, each pair of people in  $Y$  likes each other, and everyone in  $X$  is the enemy of everyone in  $Y$ .

## An Application: International Relations



**The evolution of alliances in Europe, 1872–1907**  
(the abbreviations GB, Fr, Ru, It, Ge, and AH stand for Great Britain, France, Russia, Italy, Germany, and Austria-Hungary, respectively):

- (a) Three Emperors' League, 1872–1881;
- (b) Triple Alliance, 1882;
- (c) German–Russian Lapse, 1890;
- (d) French–Russian Alliance, 1891–1904;
- (e) Entente Cordiale, 1904;
- (f) British Russian Alliance, 1907.

Solid dark edges indicate friendship while dotted edges indicate enmity. Note how the network slides into a balanced labeling – and into World War I.

# *Homophily and Assortative Mixing*

People have a strong tendency to associate with others whom they perceive as being similar to themselves in some way. This tendency is called *homophily* or *assortative mixing*. In a similar way, one also encounters *disassortative mixing*, the tendency for people to associate with others who are *unlike* them.

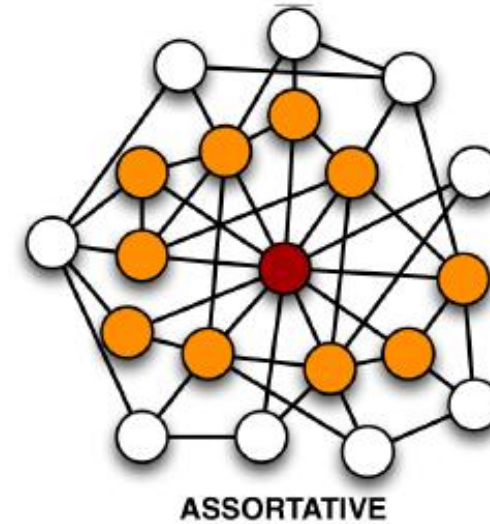
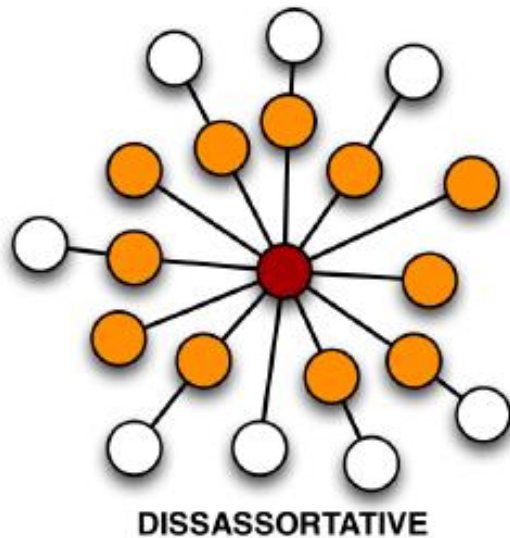
- **Assortativity:** Tendency to be linked with nodes that are similar in some way
  - *Humans*: age, race, nationality, language, income, education level, etc.
  - *Citations*: similar fields than others
  - *Web-pages*: Language
- **Disassortativity:** Tendency to be linked with nodes that are different in some way
  - Network providers: End users vs. other providers
- ***Assortative mixing*** can be based on
  - Enumerative characteristic
  - Scalar characteristic
  - Degree
  - .....

# Assortativity

Assortativity – The average degree of neighbors

Two point degree correlation

A qualitative analysis of the 2-vertex degree correlation can be given by assortativity



In assortative networks high degree nodes tend to connect more to high degree nodes and in disassortative networks they are more likely to be connected to low degree nodes.



## *Assortativity: An Example*

- The friendship network in a US high school in 1994
- Colors represent races,
  - **White**: whites
  - **Grey**: blacks
  - **Light Grey**: hispanics
  - **Black**: others
- High assortativity between individuals of the same race



## *Assortativity Significance*

- **Assortativity significance**

- The difference between measured assortativity and expected assortativity
- The higher this difference, the more significant the assortativity observed

- **Example**

- In a school, half the population is white and the other half is Hispanic.
- We expected 50% of the connections to be between members of different races.
- If all connections are between members of different races, then we have a significant finding

## Modularity ( $Q$ )

- Extend to which a node is connected to a like in network
  - + if there are more edges between nodes of the same type than expected value
  - - otherwise

$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

$\delta(c_i, c_j)$  is 1 if  $c_i$  and  $c_j$  are of same type, and 0 otherwise, and the quantity  $A_{ij} - \frac{k_i k_j}{2m}$  ( $= B_{ij}$ ) is often called the modularity matrix.

$$Q = \sum_r (e_{rr} - a_r^2)$$

$e_{rr}$  is the fraction of edges that join same type of vertices, where  $e_{rs} = \frac{1}{2m} \sum_{ij} A_{ij} \delta(c_i, r) \delta(c_j, s)$

$a_r$  is the fraction of ends of edges attached to vertex type  $r$ , where  $a_r = \frac{1}{2m} \sum_i k_i \delta(c_i, r)$



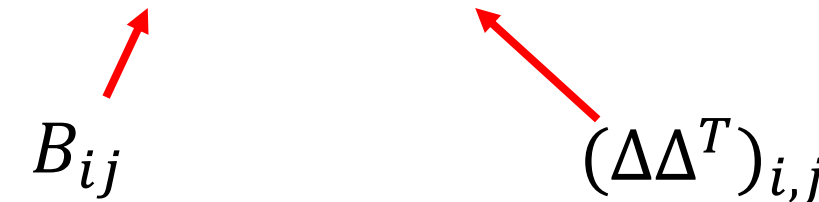
## *Modularity: Matrix Form*

The Modularity matrix takes the form

$$B = A - k k^T / 2m$$

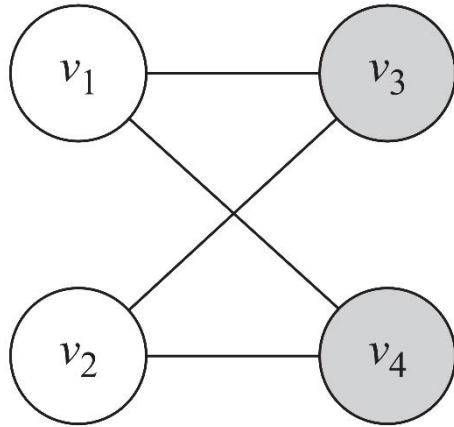
where  $k \in \mathbb{R}^{n \times 1}$  is the degree vector

Modularity can be reformulated as

$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) = \frac{1}{2m} \text{Tr}(B \Delta \Delta^T) = \frac{1}{2m} \text{Tr}(\Delta^T B \Delta)$$


$B_{ij}$   $(\Delta \Delta^T)_{i,j}$

## Modularity: Example



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \Delta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, k = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, m = 4$$

$$B = A - kk^T / 2m = \begin{bmatrix} -0.5 & -0.5 & 0.5 & 0.5 \\ -0.5 & -0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}$$

$$Q = \frac{1}{2m} \text{Tr}(\Delta^T B \Delta) = -0.5$$

The number of edges between nodes of the **same color** is less than the expected number of edges between them

## *Assortative Mixing by enumerative characteristics*

*Modularity* is almost always less than 1, hence we can normalize it with the  $Q_{max}$  value

$$r = \frac{Q}{Q_{max}} = \frac{\sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)}{2m - \sum_{ij} \left( \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)}$$

## *Assortative Mixing by Scalar Characteristics*

$$r = \frac{\sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) x_i \cdot x_j}{\sum_{ij} \left( k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) x_i \cdot x_j}$$

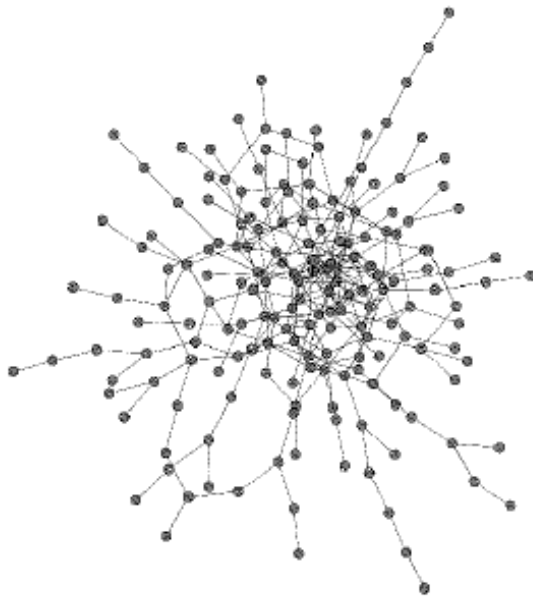
- $r=1$ , perfectly assortative
- $r=-1$ , perfectly disassortative
- $r=0$ , non-assortative

➤ Usually node degree is used as scale

$$r = \frac{\sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) k_i \cdot k_j}{\sum_{ij} \left( k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) k_i \cdot k_j}$$

# Assortative Mixing by Degree

A special case of assortative mixing according to a scalar quantity, and one of particular interest, is that of *mixing by degree*. In a network that shows *assortative mixing by degree*, high-degree vertices will be preferentially connected to other high-degree vertices, and the low to low. In particular, in an assortative network, where the high-degree nodes tend to stick together, one expects to get a clump or core of such high-degree nodes in the network surrounded by a less dense periphery of nodes with lower-degree. This *core/periphery* structure is a common feature of social networks. If a network is disassortatively mixed by degree then high-degree vertices tend to be connected to low-degree ones, creating star-like features in the network that are often readily visible.



Assortative  
Network



Disassortative  
Network

## *Assortativity Coefficient of Various Networks*

| Network                       | $n$       | $r$                    |
|-------------------------------|-----------|------------------------|
| Physics coauthorship (a)      | 52 909    | 0.363                  |
| Biology coauthorship (a)      | 1 520 251 | 0.127                  |
| Mathematics coauthorship (b)  | 253 339   | 0.120                  |
| Film actor collaborations (c) | 449 913   | 0.208                  |
| Company directors (d)         | 7 673     | 0.276                  |
| Internet (e)                  | 10 697    | −0.189                 |
| World-Wide Web (f)            | 269 504   | −0.065                 |
| Protein interactions (g)      | 2 115     | −0.156                 |
| Neural network (h)            | 307       | −0.163                 |
| Marine food web (i)           | 134       | −0.247                 |
| Freshwater food web (j)       | 92        | −0.276                 |
| Random graph (u)              |           | 0                      |
| Callaway <i>et al.</i> (v)    |           | $\delta/(1 + 2\delta)$ |
| Barabási and Albert (w)       |           | 0                      |