# Introduction to Complex Networks Analysis

**Miloš Savić**

**Department of Mathematics and Informatics,**
**Faculty of Sciences,**
**University of Novi Sad, Serbia**

# Complex systems and networks

- System - a set of **interrelated** and **interdependent** components (real or abstract) together forming an integrated whole

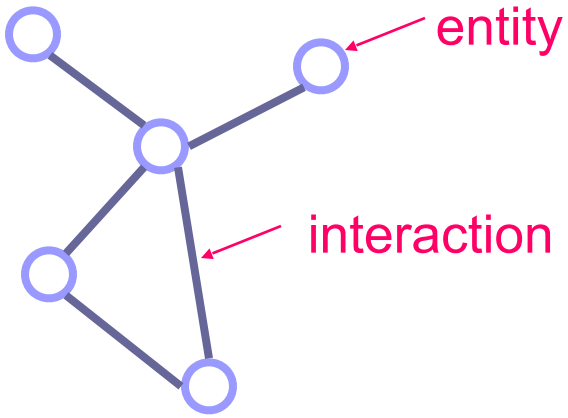
- Complex system – **large number of constituent components that may interact in many different ways**

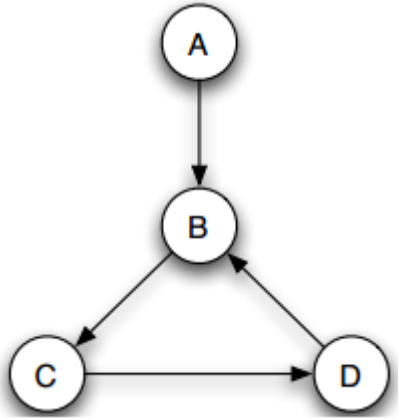
- The behavior of a complex system cannot be inferred from the behavior of its constituent components

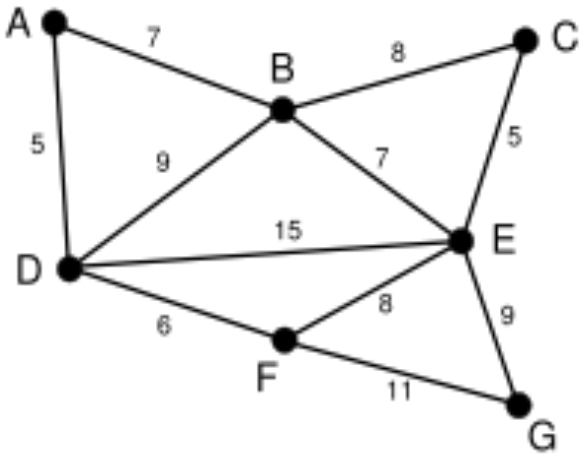
- Network – the most natural way to describe interactions and relations among constituent components of a complex system

# Complex systems and networks

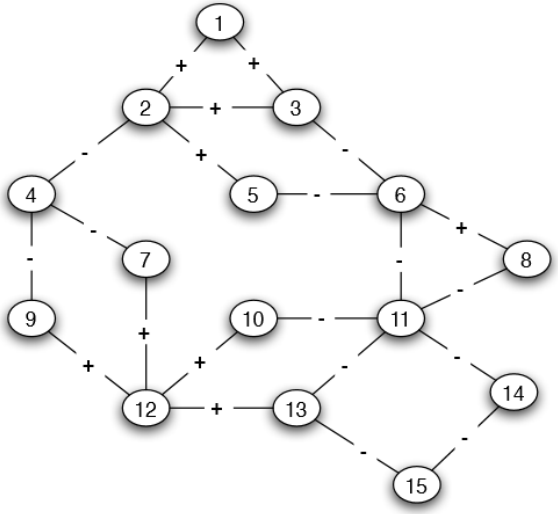
- Network – a *real-world* graph representing interactions and relations among entities within a complex system




| Entities | Interactions | |
|---|---|---|
| vertex | edge, arc | math |
| node | link | computer science |
| site | bond | physics |
| actor | tie, relation | sociology |



# Newman's classification of complex networks

- **Technological networks**
  - networks representing engineered man-made systems
- **Social networks**
  - Interactions and relations among social entities
- **Information networks**
  - Connections between data items
- **Biological networks**
  - Networks representing biological systems and processes

# Technological networks

- Internet
- Power-grids
- Telephone networks
- Transportation networks
  - Airport, railway and road networks
- Software networks
- Networks of electronic circuits
- …

# OO software networks: class collaboration networks

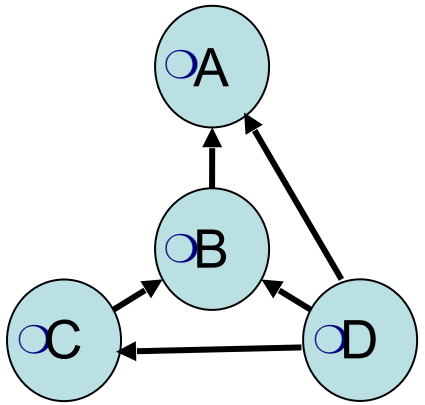
- Nodes: classes and interfaces
- Links: A→B iff
  - A extends, implements or throws B
  - A instantiates objects of B
  - A declares attribute whose type is B
  - B is return type, argument type or type of a local variable in methods of A
  - A calls methods defined in B
- Simplified class diagrams

```
interface A { … }

class B implements A { … }

class C {
    public void methodC() {
        B b = new B();
    }
}

class D extends C implements A {
    public B makeB() {
        return new B();
    }
}
```




In-degree, number of in-coming links, quantifies internal reuse (Fan-in)

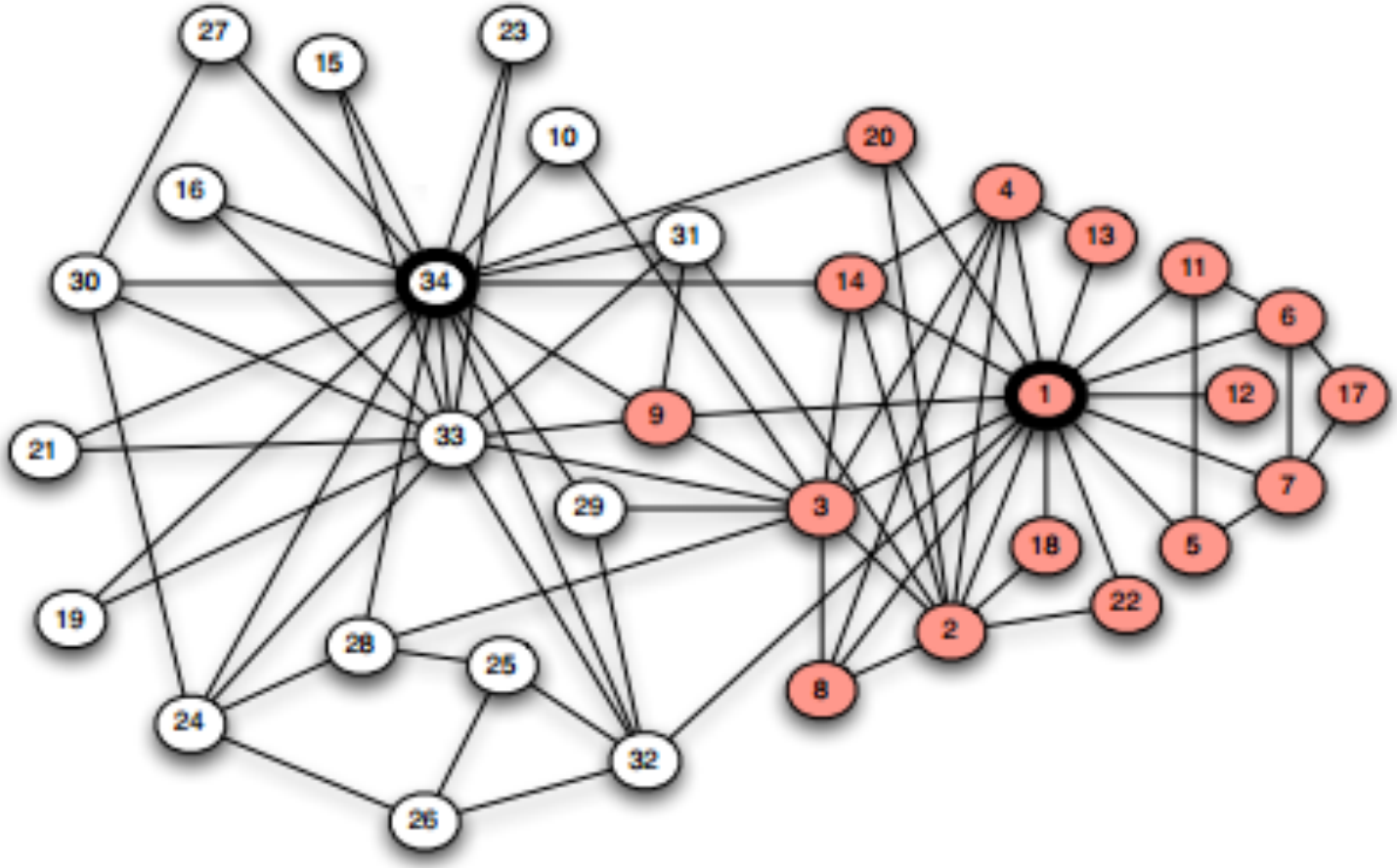
Out-degree, number of out-going links, quantifies internal aggregation (Fan-out)

Total degree (in-degree + out-degree) quantifies class coupling (CK CBO)

# Social networks

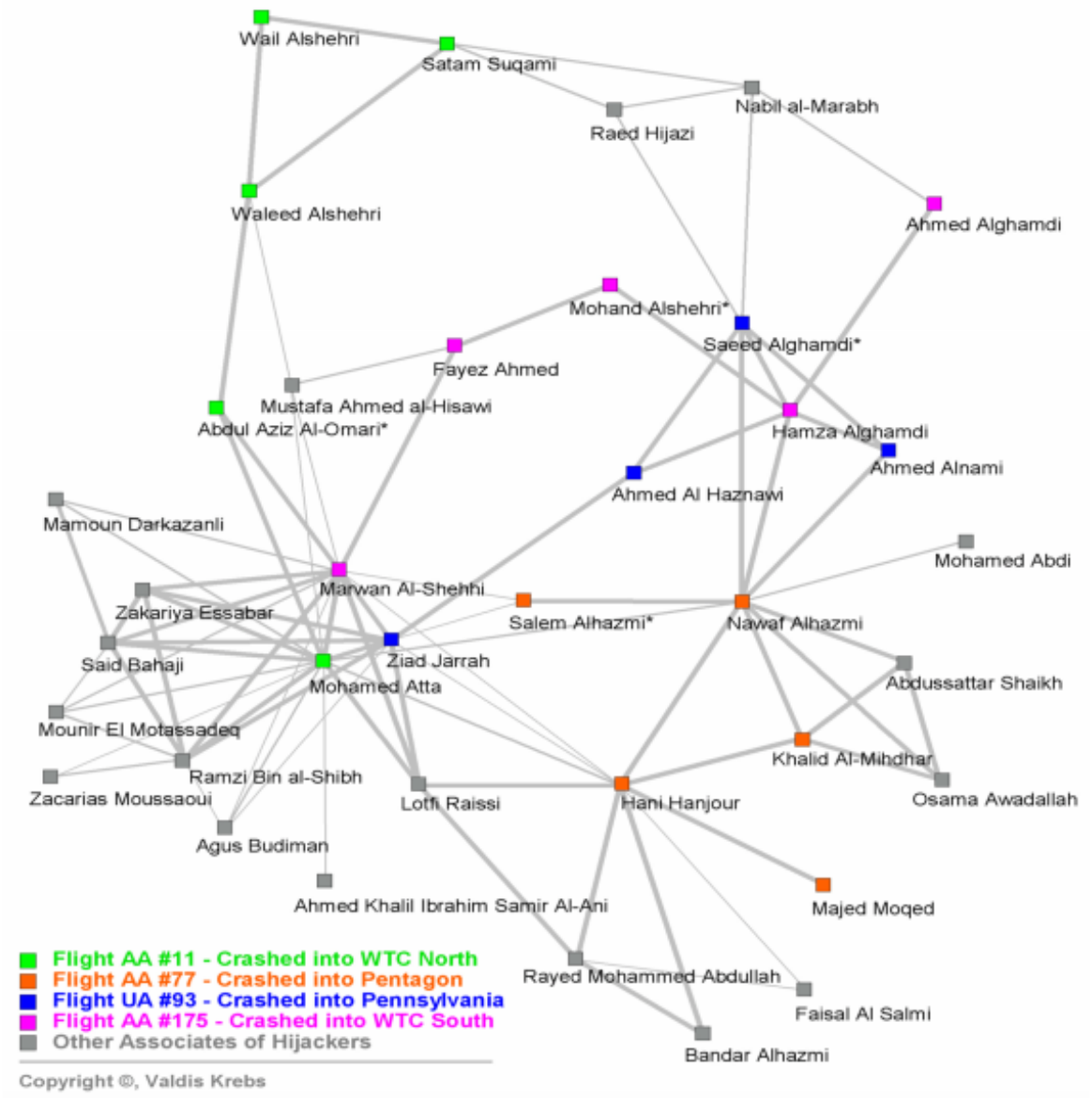
- Facebook, Twitter, MySpace, etc. are not social networks, but social networking sites
  - Platforms/media to establish and maintain social interactions
- **Social network - network-structured data describing interactions among social entities**
- **Social entities**
  - Individuals, social groups, institutions, organizations, companies, political parties, nations
- **Social links**
  - opinions on other individuals (signed social networks)
  - transfers of material resources
  - links denoting collaboration, cooperation and coalition
  - links resulting from behavioral interactions
  - links imposed by formal relations within formally organized social groups
  - links on social networking sites
  - …

# Zachary's karate club




W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**, 452-473 (1977)

# Terroristic organizations




Valdis Krebs. Mapping Networks of Terrorist Cells. *Connections* 24(3): 43-52, 2002

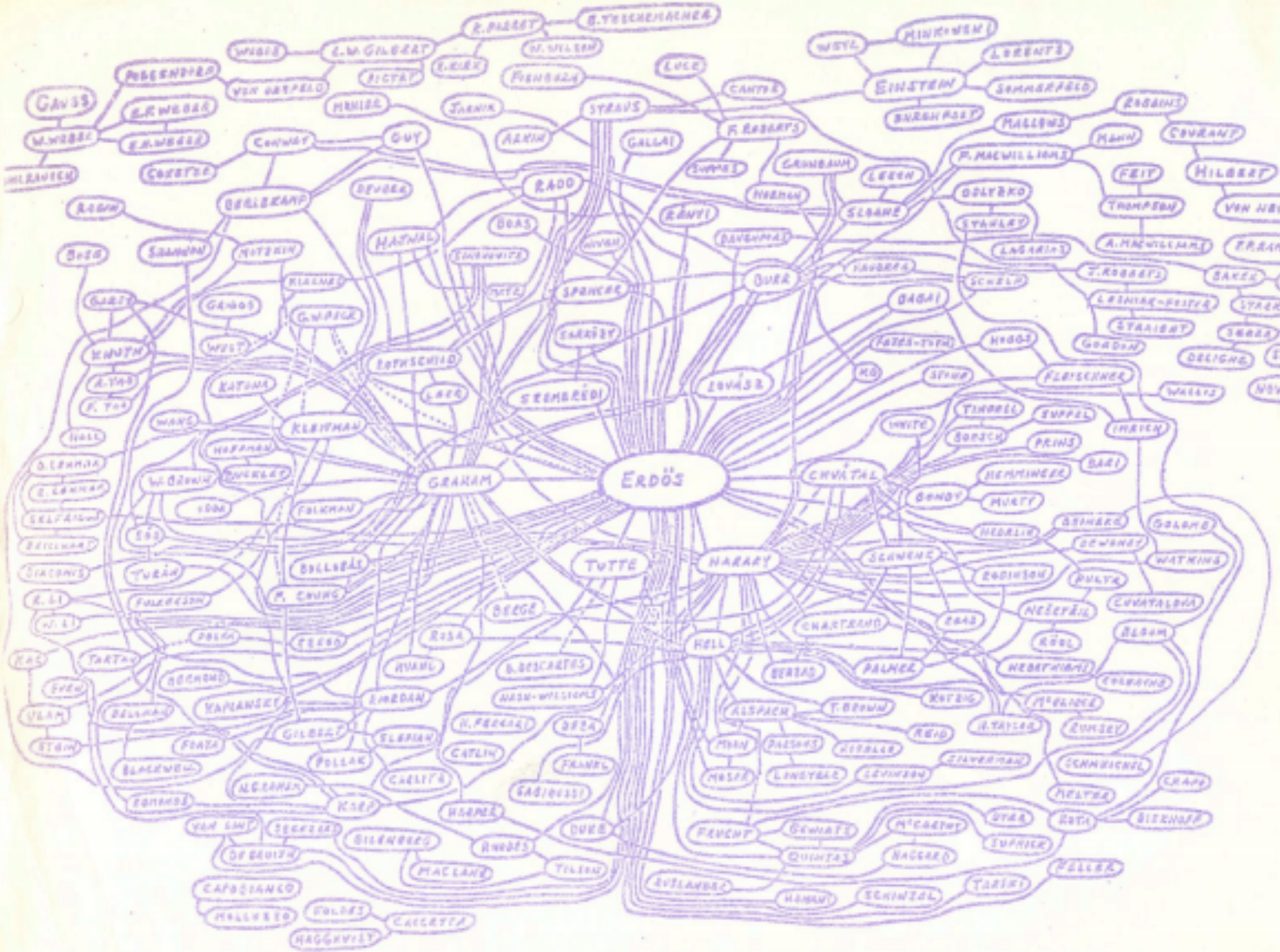
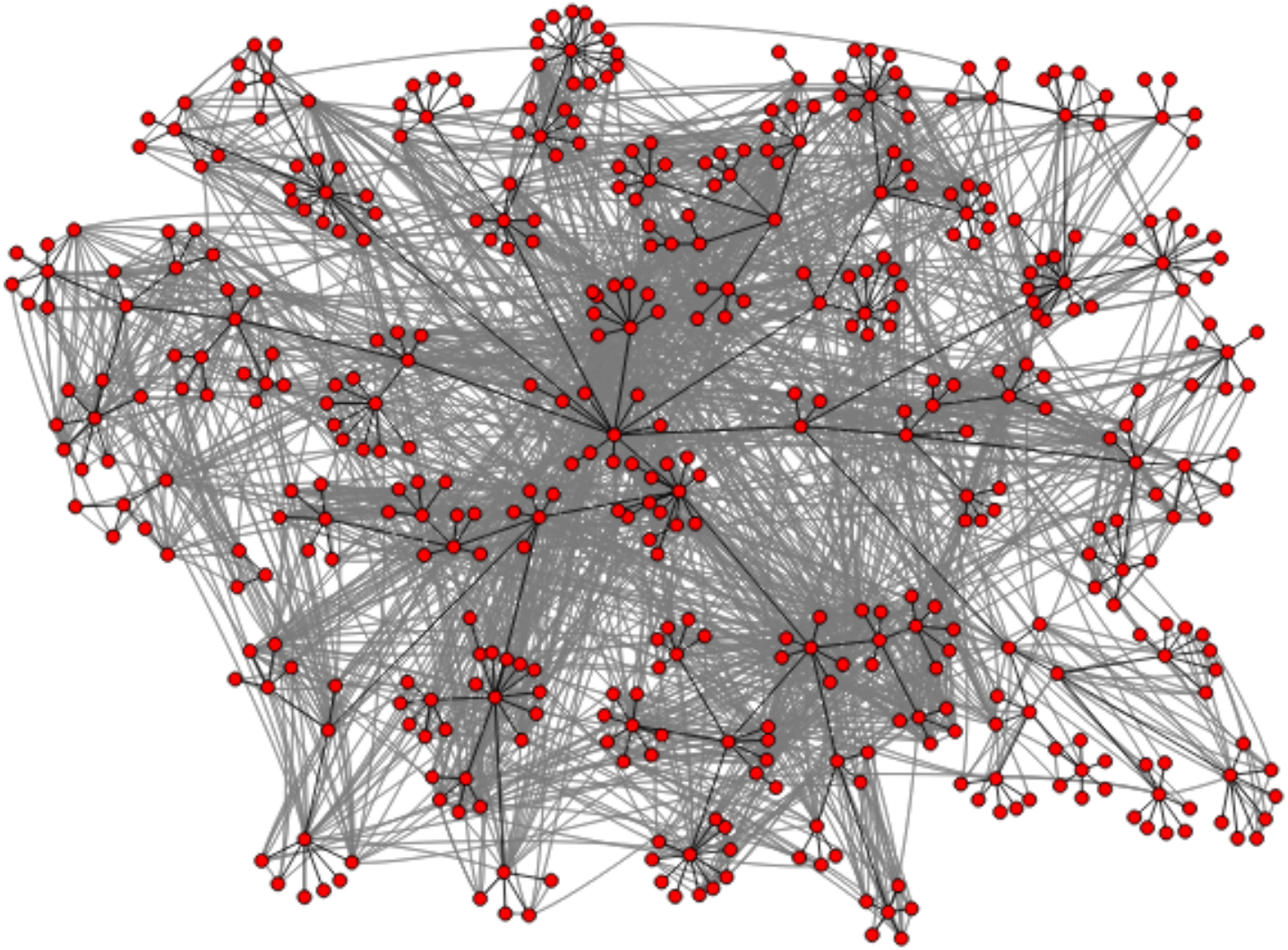



Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

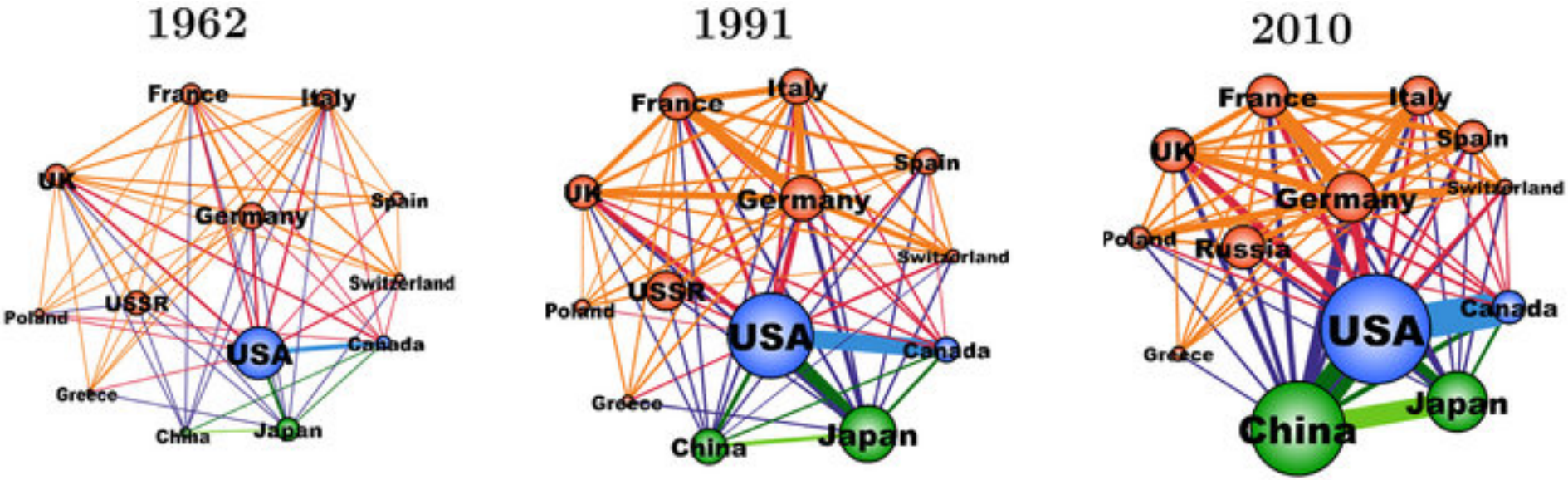
Ron Graham (alias Tom Odda).

# E-mail communication within a company/institution



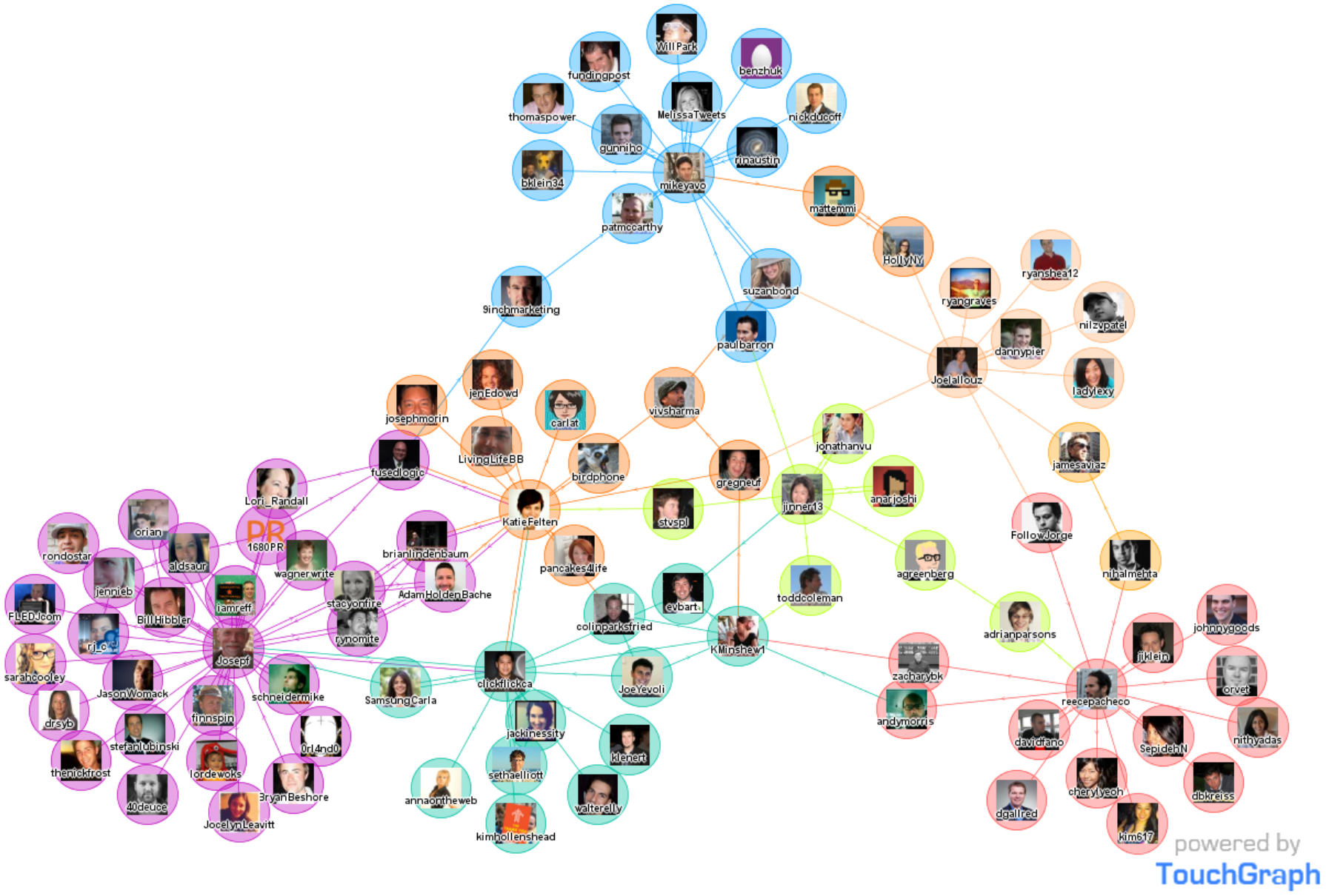
Lada Adamic, Eytan Adar. 2003. How to search a social network?

# Economic networks




Node colors correspond to the continents: orange for Europe, green for Asia, blue for America. The node size corresponds to the GDP of the country. The edge thickness corresponds to the volume of the trade between the countries.

# Social Webs

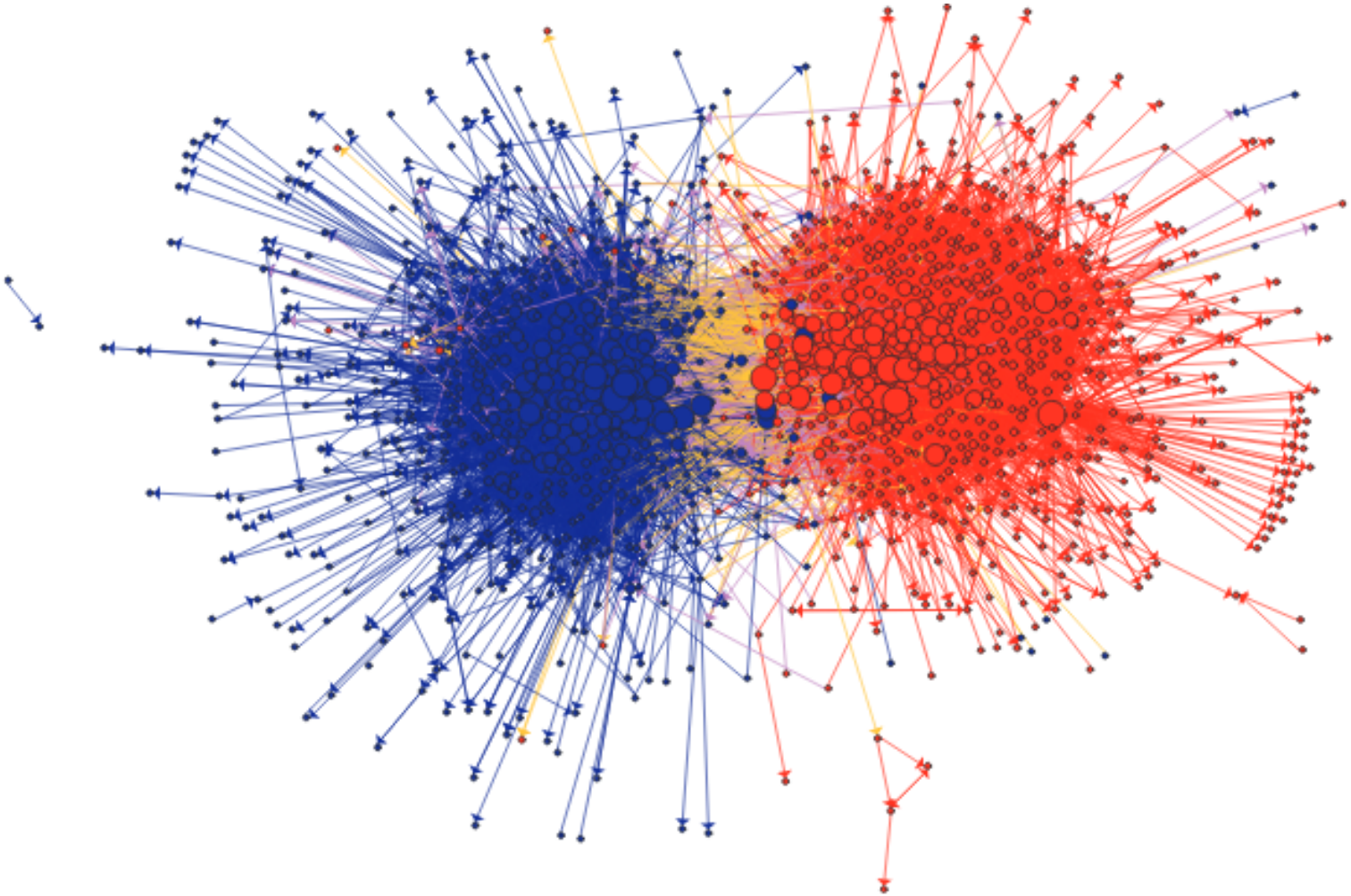


# Information networks

- **WWW network**
  - nodes: WWW pages
  - links: hyperlinks (directed links)
- **Citation networks: references between documents**
  - Scientific papers, patents, legal documents
- **Linguistic networks**
  - **Semantic:** semantic relationships (e.g., synonyms or antonyms) between words or concepts
  - **Structural:** word co-occurrence networks and sentence similarity networks
- **Recommender networks**
  - Bipartite graphs showing preferences of individuals towards some items

- **Tabular datasets can also be transformed to networks**
  - Tabular dataset — a set of data points described by feature vectors
  - k-nearest neighbor networks: A → B if B is among the top k nearest data points to A
  - Eps-radius networks: A and B connected if distance(A, B) < Eps

# Political blogosphere - blog citation network



Lada Adamic, Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog

# Networks are everywhere…

World Wide Web and hyperlink structure

The Internet and router connectivity

Roads, air lines

Power-grids

Collaborations among…

- Movie actors
- Scientists

Sexual interactions

Cellular networks in biology

Food webs in ecology

Phone call networks

Word co-occurrence in text

Neural network connectivity

Protein interaction networks

# Complex network analysis

- Quantitative methods for studying the structure and evolution of complex networks
  - Analysis of direct and indirect connectivity of nodes, identification of connectivity trends and patterns
  - Centrality metrics and algorithms — identification of the most important nodes and links in a network
  - Network comprehension — identification of cohesive subgraphs (clusters/communities), connectivity between and within clusters
  - Identification of evolutionary trends and principles that can explain the evolution of a network at the microscopic, mesoscopic and macroscopic level
  - …

# Frequently observed characteristics of real-world large-scale networks

- Real-world networks are sparse
- Large number of connected components, the existence of a giant connected component
- The existence of hubs — nodes with an extremely large number of links (neighbors)
  - Assortative/disassortative degree mixing patterns
  - Preferential attachment
- Small-world phenomenon in the Watts-Strogatz sense
  - Short distances between nodes
  - Dense ego-networks
- Core-periphery structures in assortative networks
- Community structure
- Densification laws and shrinking diameters

# Real large-scale networks are sparse

- density = the number of links / the maximal number of links the nodes can form

## Most real-world networks are sparse

$$E << E_{max} \quad (\text{or } \bar{k} << N-1)$$

| | | |
|---|---|---|
| WWW (Stanford-Berkeley): | N=319,717 | ⟨k⟩=9.65 |
| Social networks (LinkedIn): | N=6,946,668 | ⟨k⟩=8.87 |
| Communication (MSN IM): | N=242,720,596 | ⟨k⟩=11.1 |
| Coauthorships (DBLP): | N=317,080 | ⟨k⟩=6.62 |
| Internet (AS-Skitter): | N=1,719,037 | ⟨k⟩=14.91 |
| Roads (California): | N=1,957,027 | ⟨k⟩=2.82 |
| Protein (S. Cerevisiae): | N=1,870 | ⟨k⟩=2.39 |

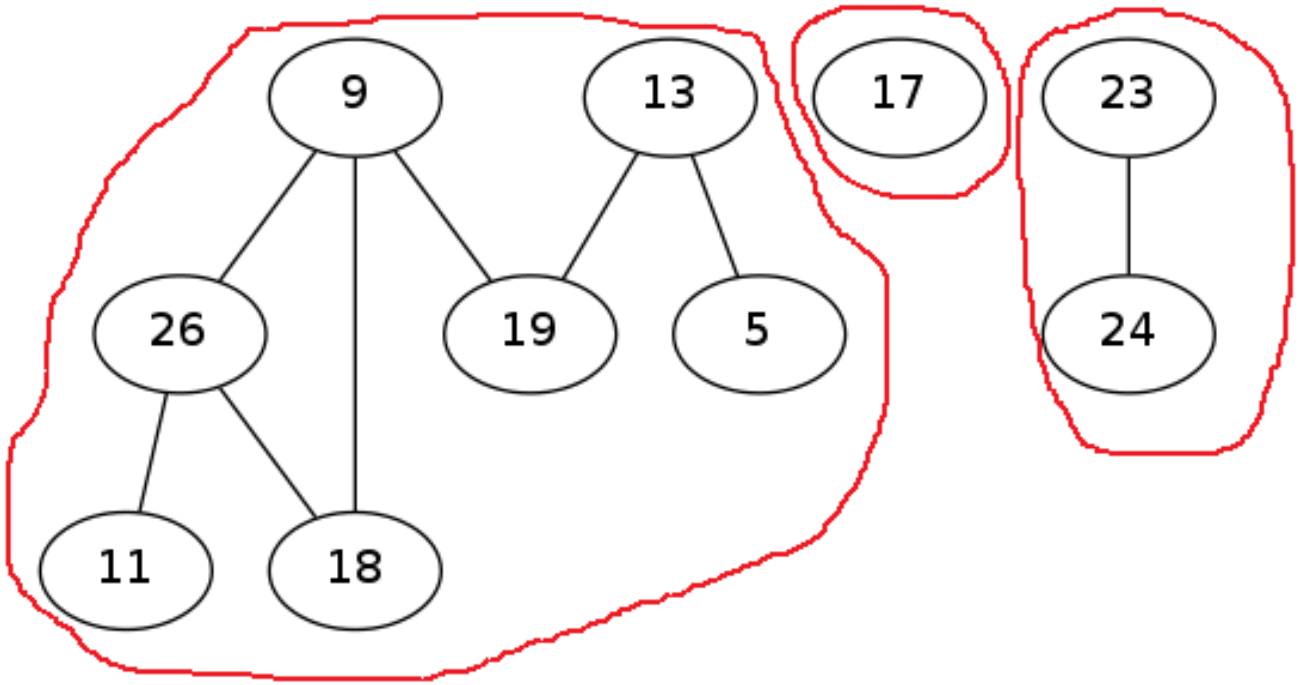
(Source: *Leskovec et al., Internet Mathematics, 2009*)

**Consequence:** Adjacency matrix is filled with zeros!

(**Density ($E/N^2$):** WWW=$1.51\times10^{-5}$, MSN IM = $2.27\times10^{-8}$)

# Connected components in undirected networks

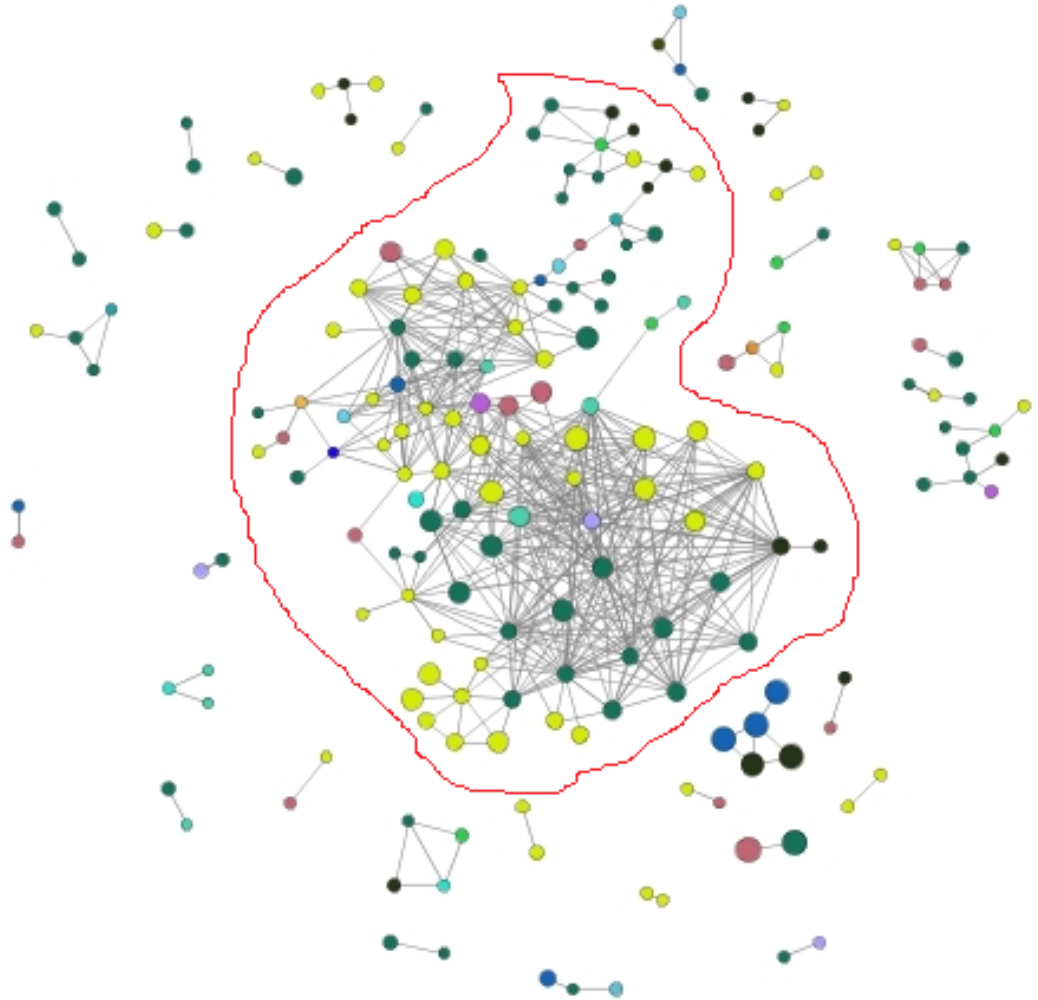
- Connected undirected graph — there is a path between any two nodes
- If a network is not a connected graph then it consists of multiple connected components




- BFS/DFS

# Giant connected component

- A component that encompasses a vast majority of nodes



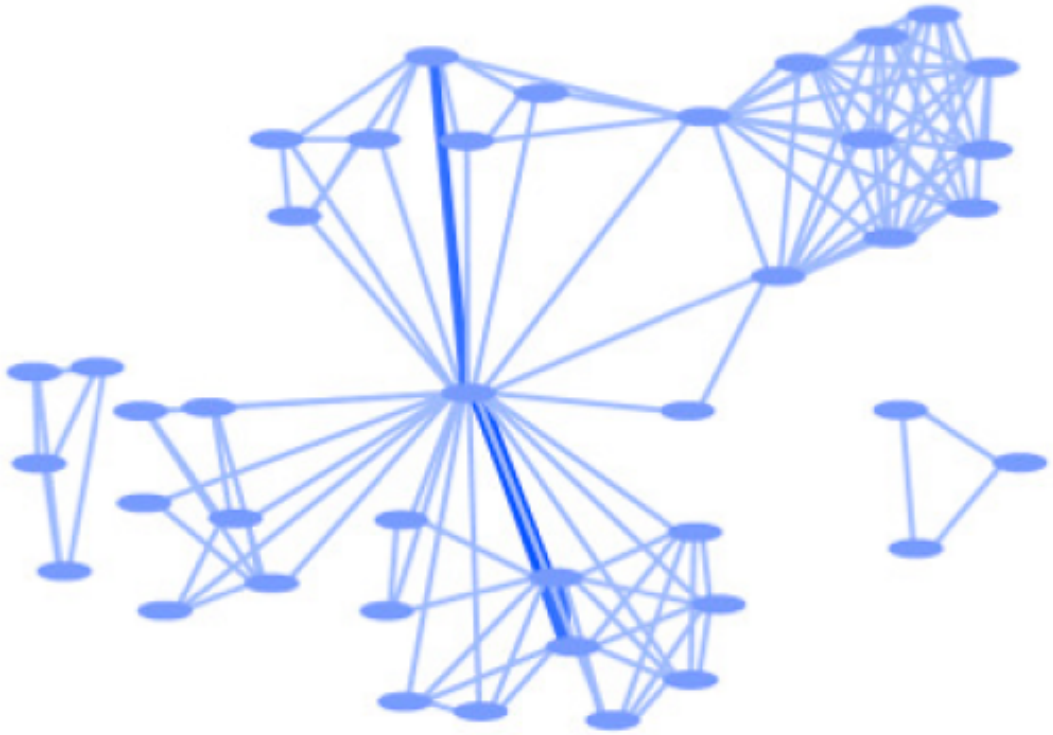


Figure 2.6: The collaboration graph of the biological research center *Structural Genomics of Pathogenic Protozoa (SGPP)* [134], which consists of three distinct connected components. This graph was part of a comparative study of the collaboration patterns graphs of nine research centers supported by NIH's Protein Structure Initiative; SGPP was an intermediate case between centers whose collaboration graph was connected and those for which it was fragmented into many small components.

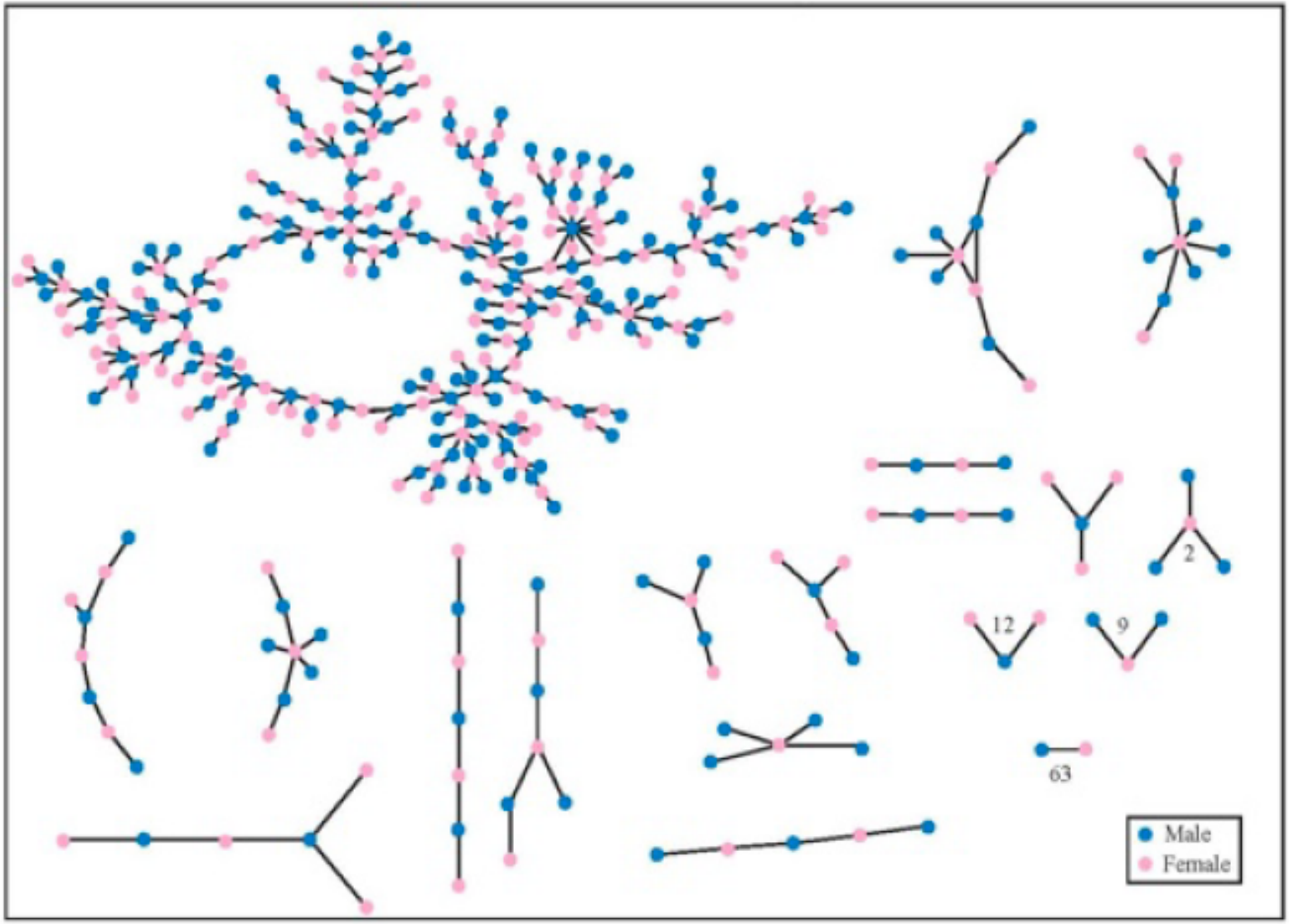
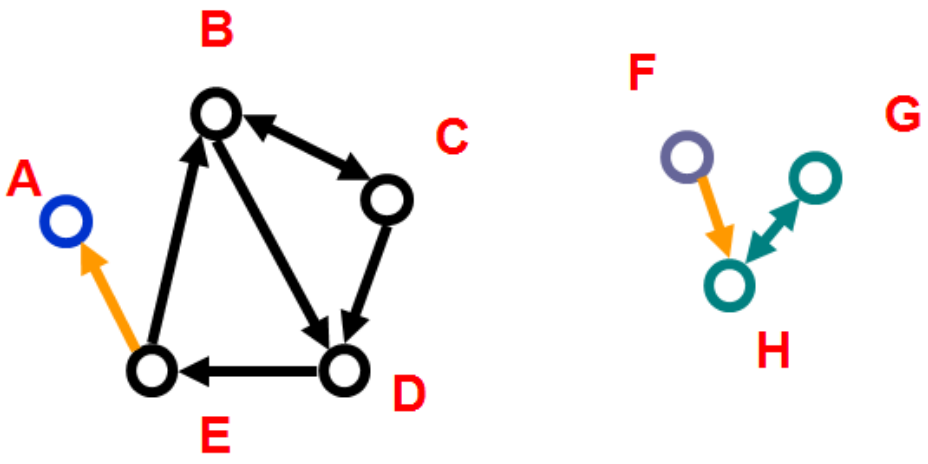



Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].

**The fact that this network contains such a large component is significant when one thinks about the spread of sexually transmitted diseases, a focus of the researchers performing the study**

# Components in directed networks

- **Weakly connected components**
  - connected components in the undirected projection of a directed network
- **Strongly connected components**
  - for every two nodes A and B
    - there is a directed path from A to B, and
    - a directed path from B to A




Weakly connected components:

{A, B, C, D, E}

{F, G, H}

Strongly connected components:

{B, C, D, E}

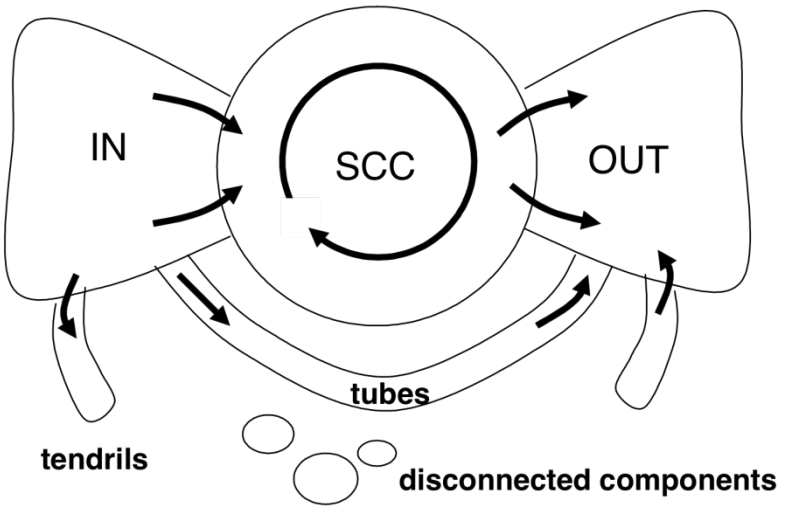
{G, H}

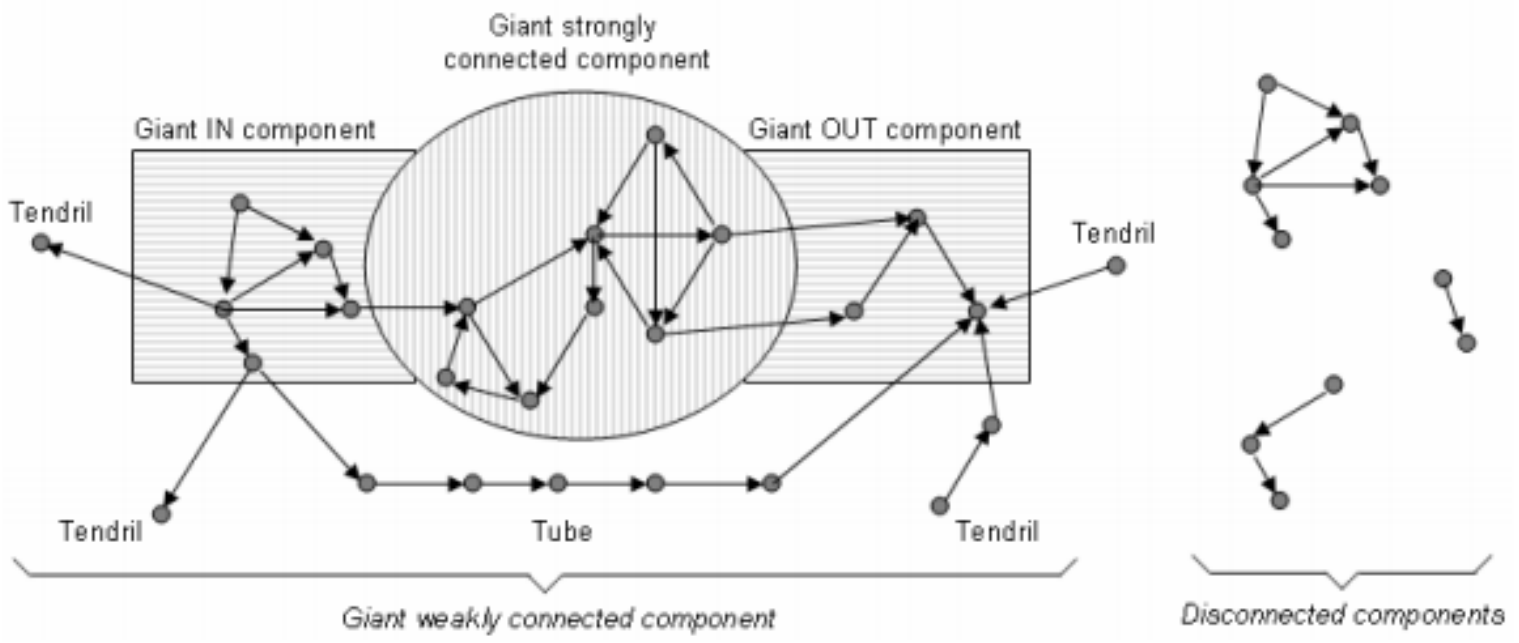
# Page rank

- Metric proposed by Google to rank web pages
- **Metric of node centrality for directed networks**
- Metric related to random walks
- Random walker (surfer) initially positioned at a random page in the WWW network
  - With probability β goes to a random out-neighbor of the current node
  - **Teleportation rule: with probability 1 - β goes to a randomly chosen node**
- PageRank(x) increases with the number of times random walker visits x
- **Why teleportation rule?**

# Bow-tie structure of the Web

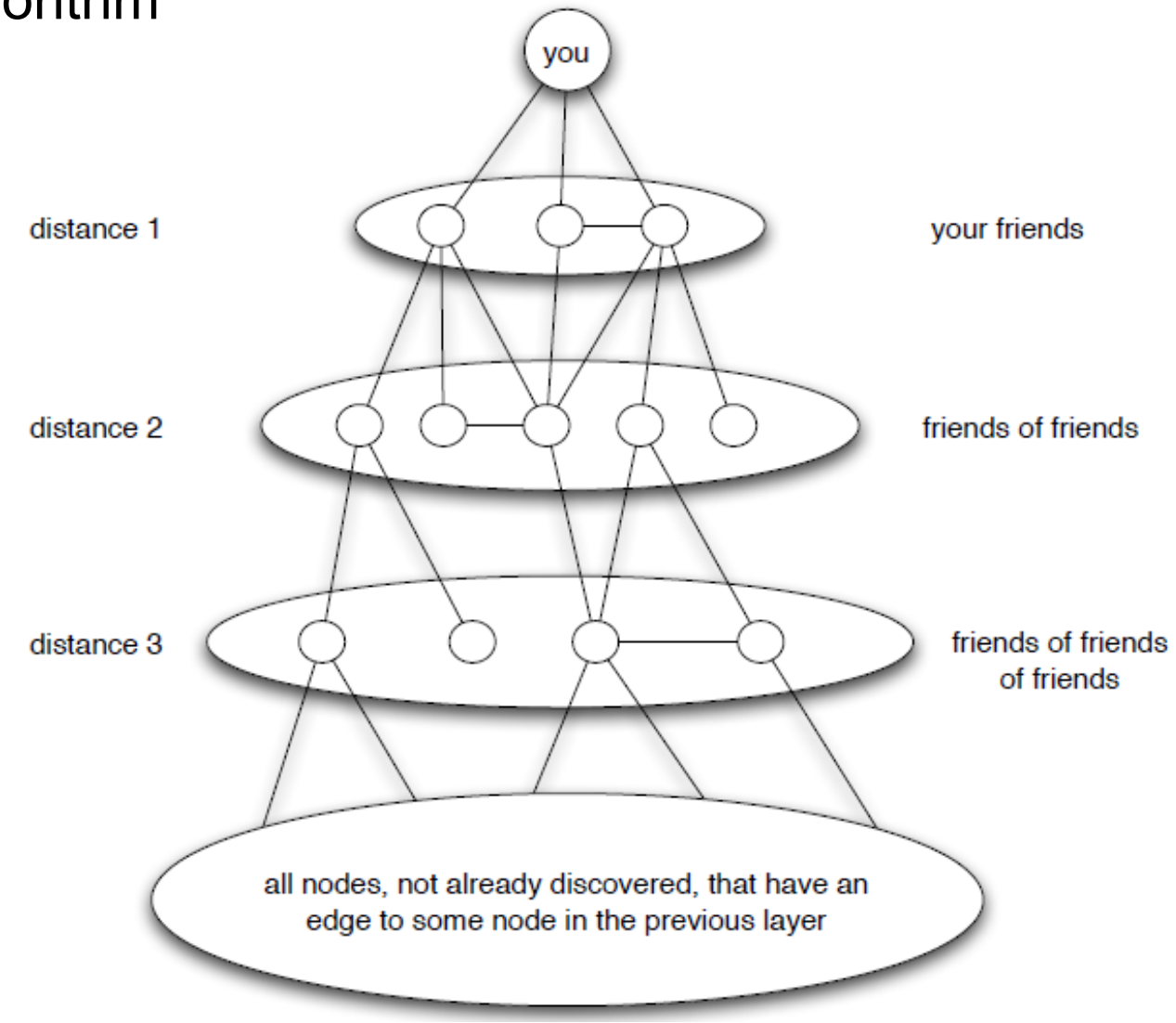
- Broder et al. 1999
- 200 million web pages
- 1.5 bilion links
- The network has a GWCC
- SCC – 27.5%
- IN, OUT – 21.5%
- Tendrils, tubes – 21.5%
- Out of GWCC – 8%



# Node distance

- Distance(A, B) = the length of the shortest path between A and B
- BFS algorithm



# Small-world networks

**The average node distance is much smaller than the number of nodes in the network (d ~ log(N))**

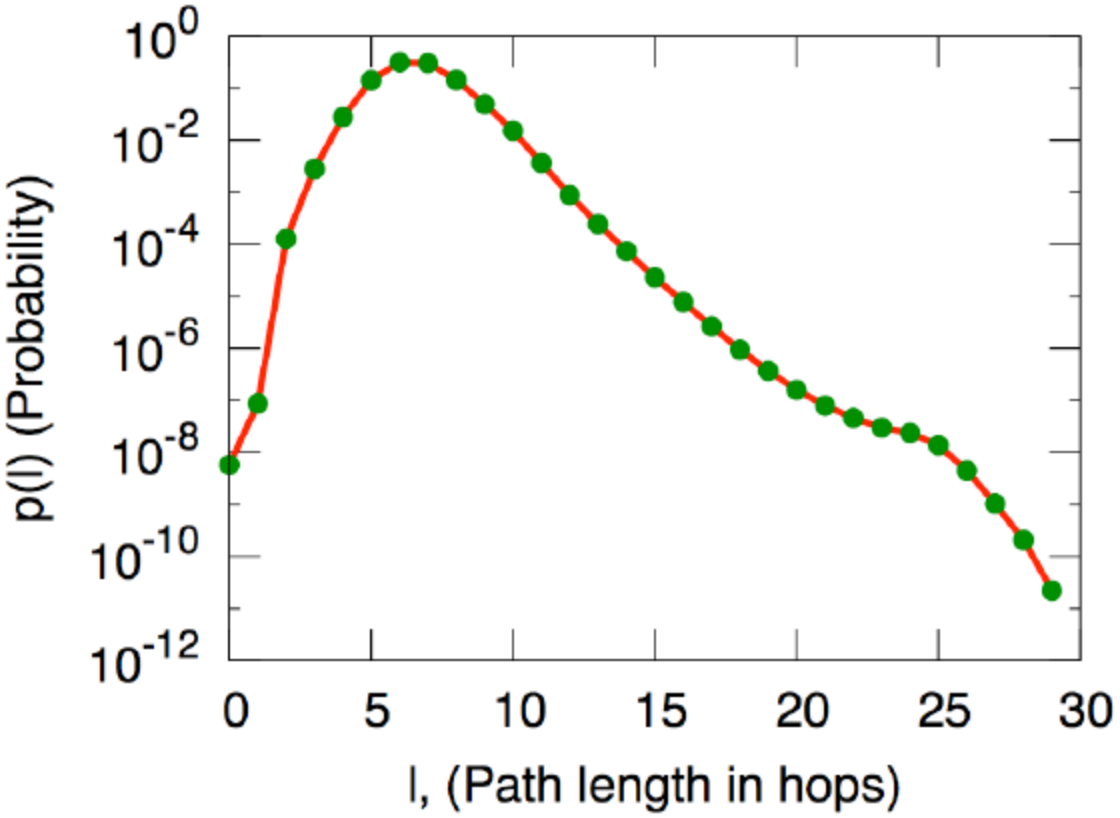
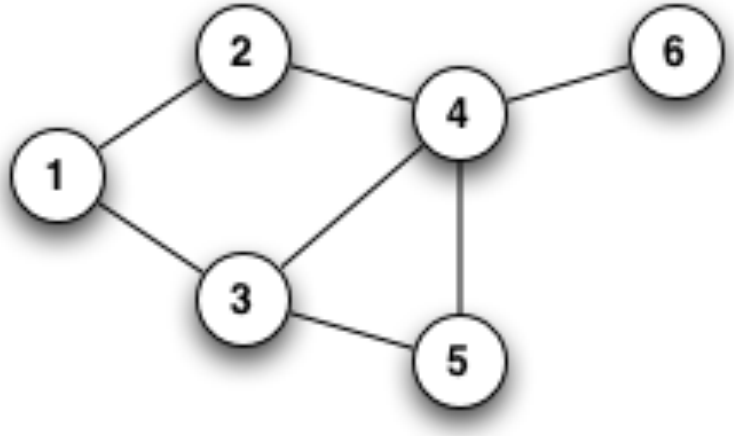



Figure 2.11: The distribution of distances in the graph of all active Microsoft Instant Messenger user accounts, with an edge joining two users if they communicated at least once during a month-long observation period [273].

# Node degree

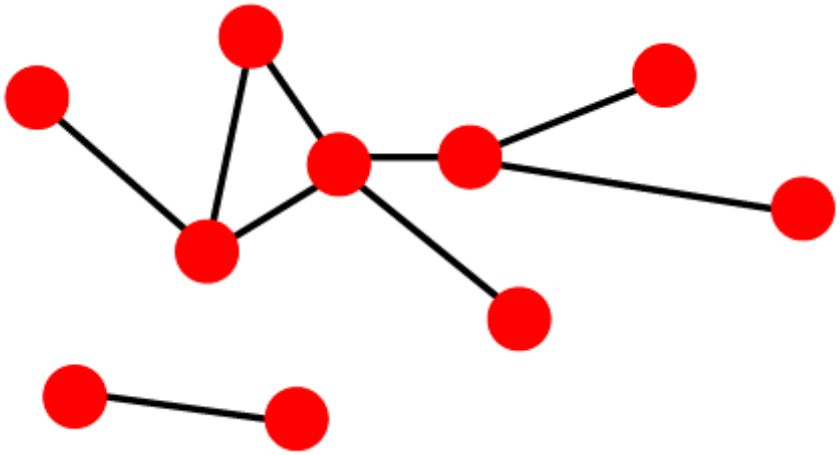
- degree(x) = the number of links incident with x
  = the number of x's neighbors
- the most basic metrics to assess importance of nodes
  - e.g. in social networks: degree is a metric of social capital
    higher number of contacts —> broader possibilities to spread ideas/opinions/interests and influence others
- Directed networks: in-degree and out-degree
- Isolated nodes and hubs

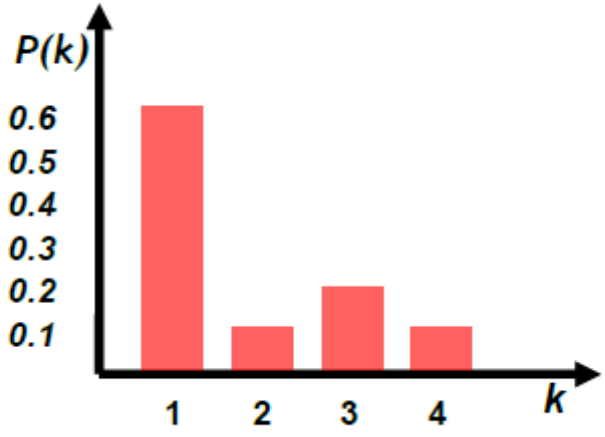



| Node | Degree |
|---|---|
| 1 | 2 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 2 |
| 6 | 1 |

# Degree distribution

- Summarizes the connectivity of the nodes in a network
  - X — randomly chosen node
  - $k_X$ — the degree of X
  - $P(k) = P\{k_X = k\}$
  - $CCD(k) = P\{k_X \geq k\}$

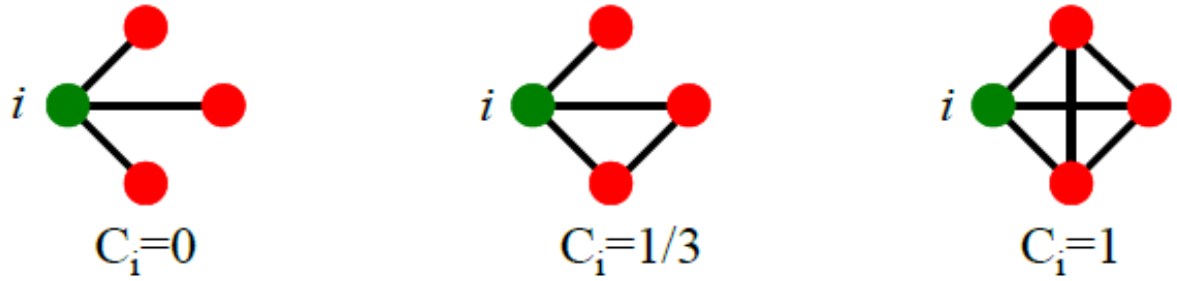




# Transitivity of links

- **Clustering coefficient:**
  - What portion of $i$'s neighbors are connected?
  - Node $i$ with degree $k_i$
  - $C_i \in [0,1]$
  - $C_i = \dfrac{2e_i}{k_i(k_i-1)}$

    where $e_i$ is the number of edges between the neighbors of node $i$

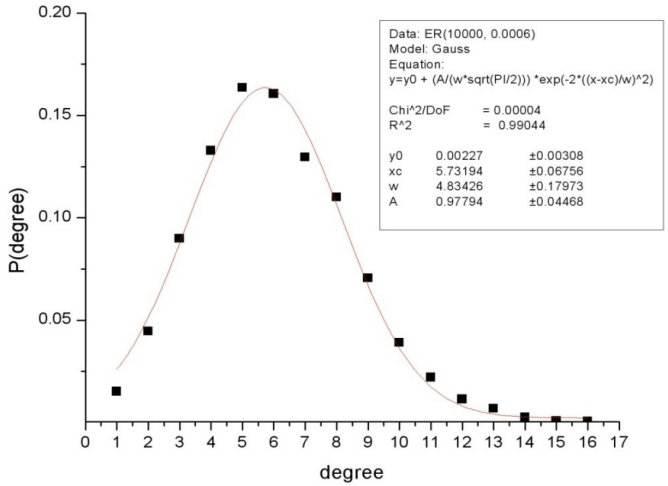



$C_i$=0 $C_i$=1/3 $C_i$=1

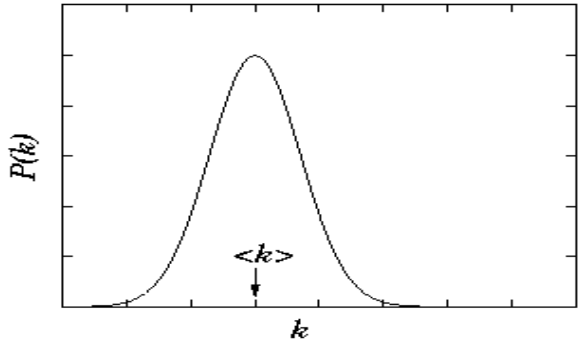
- **Average Clustering Coefficient:** $C = \dfrac{1}{N}\sum_{i}^{N} C_i$

# Erdos-Renyi random graphs

- ER random graph, ER(N, p): N nodes, every pair of nodes connected with probability p
- Poisson degree distribution
  - no hubs
- Small-world graphs
  - d ~ log(N)
- C = p
- pN > 1 → giant connected component emerges

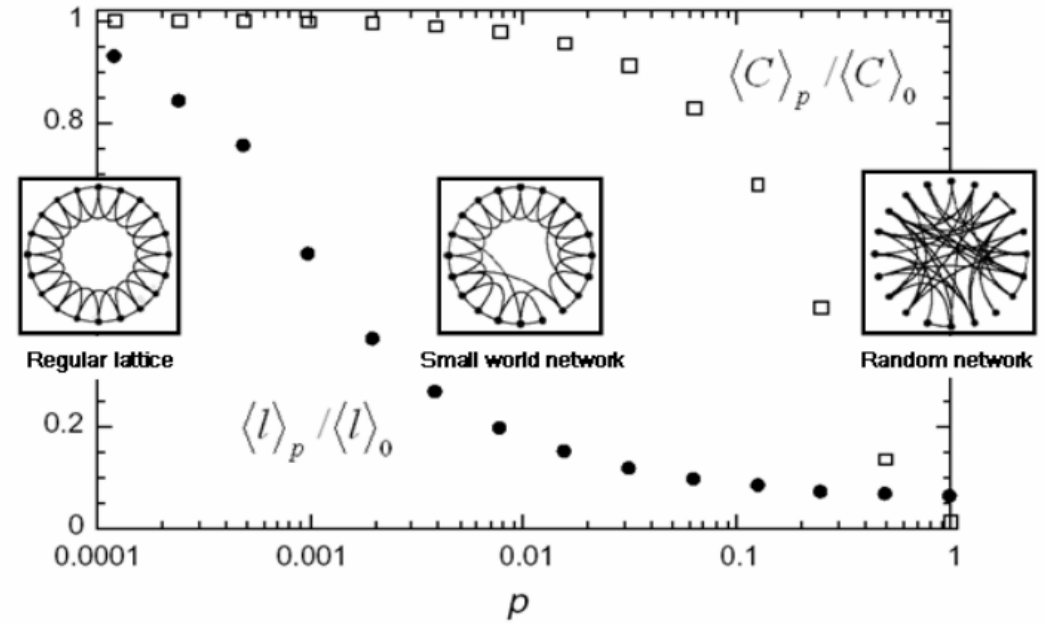



Poisson degree distr.



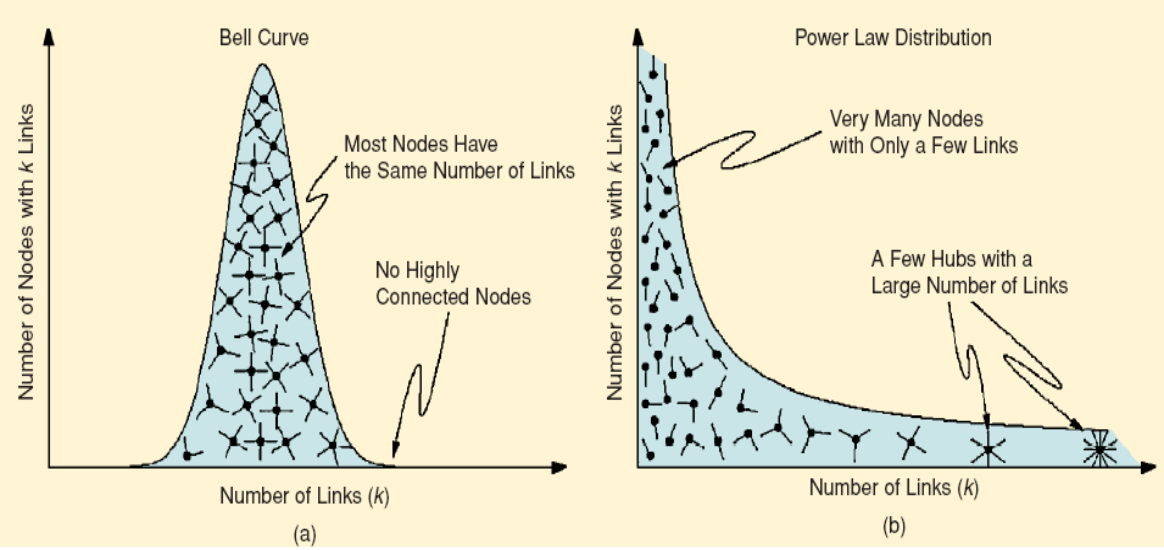
# 1999, Watts & Strogatz

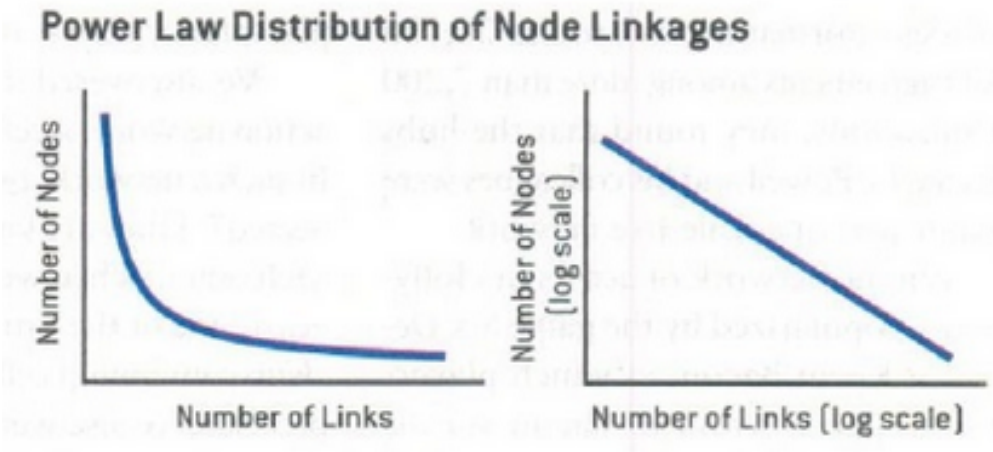
- Watts and Strogatz analyzed three real-world networks
  - collaboration network of Hollywood movie actors
  - power-grid network of Western USA
  - neural network of C. Elegans (a worm)
- Empirical findings
  - short distances between nodes (the small-world property)
  - **clustering coefficients drastically larger than clustering coefficients of comparable random graphs**



# 1999, Barabasi & Albert

- Empirical analysis of a network encompassing 800 million WWW pages
- **Power-law degree distributions, hubs, preferential attachment**



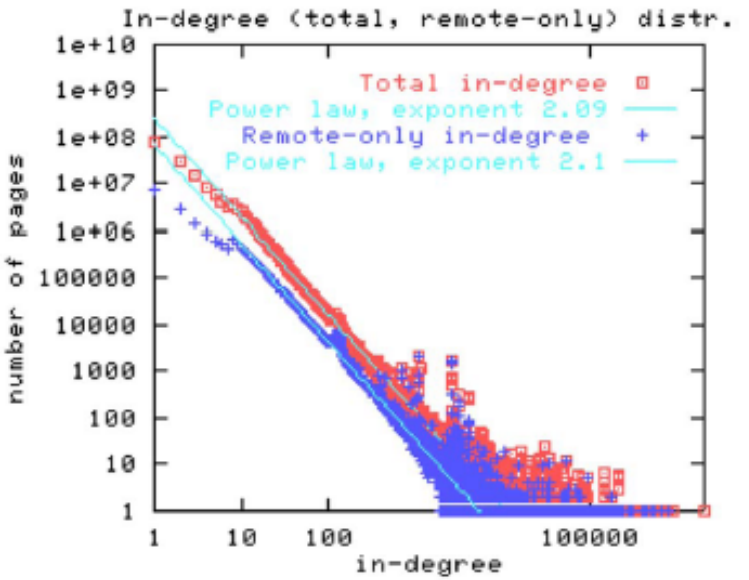



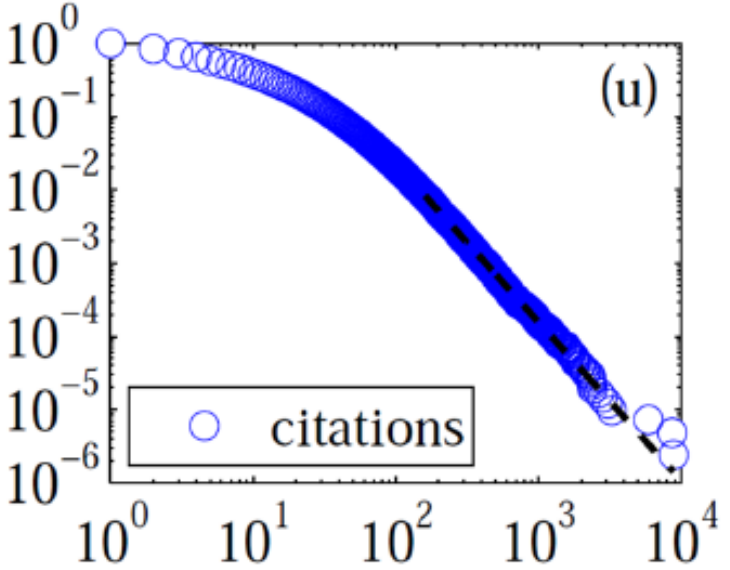

$$P(k) \sim k^{-\gamma} \rightarrow$$

$$\log P(k) \sim -\gamma \log k$$

# Power-law degree exponent is typically $2 < \alpha < 3$

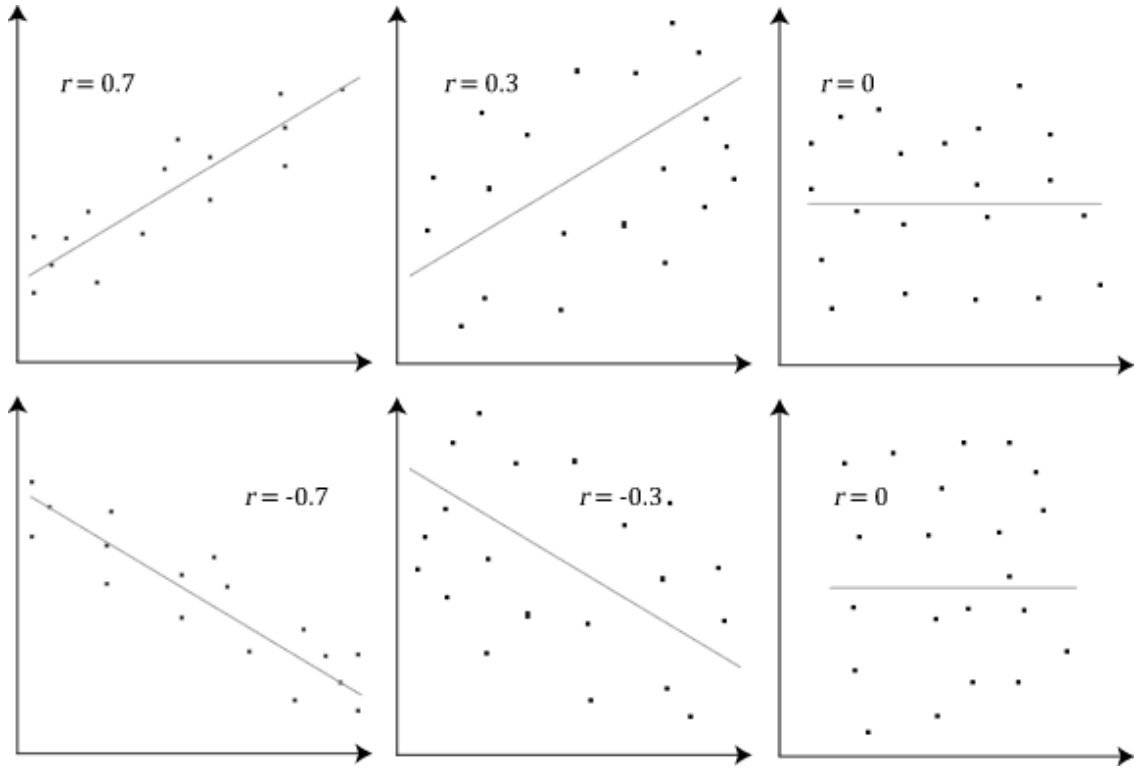
- Web graph:
  - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$ [Broder et al. 00]
- Autonomous systems:
  - $\alpha = 2.4$ [Faloutsos$^3$, 99]
- Actor-collaborations:
  - $\alpha = 2.3$ [Barabasi-Albert 00]
- Citations to papers:
  - $\alpha \approx 3$ [Redner 98]
- Online social networks:
  - $\alpha \approx 2$ [Leskovec et al. 07]



# Mining connectivity trends

- M — some metric applicable to nodes (e.g. node degree)
- Project the network to a 2D plane
  - One link — one point
  - Link l = (A, B) — point (M(A), M(B))
  - Pearson/Spearman correlation coefficient for points on the plot



# Node degree mixing patterns

**M. E. J. Newman – Assortative mixing in networks, 2002**

Social networks — assortative mixing patterns

hubs tend to be directly connected to other hubs

homophily in social networks

Biological and technological networks — disassortative mixing patterns

hubs tend to avoid other hubs

| 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 |
| food web[i] | 92 | −0.276 |

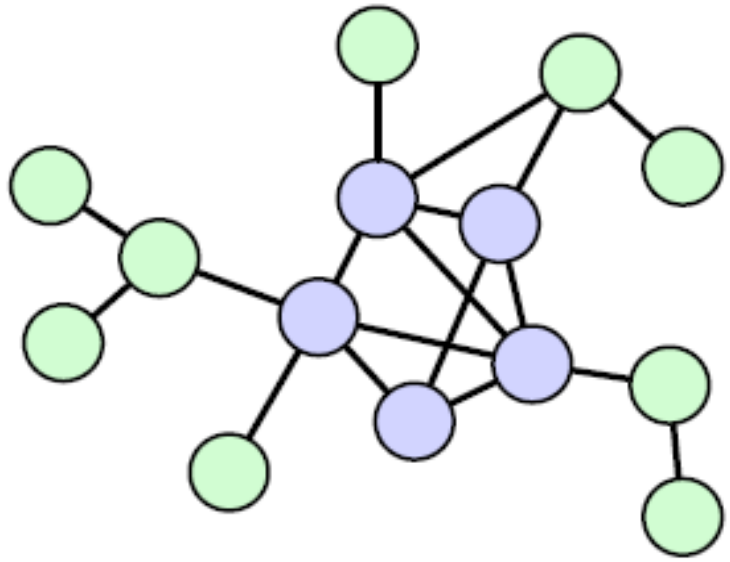
# Core-periphery structure

- Assortative networks with localized hubs
- **k-core —** maximal sub-graph S containing nodes whose degree is higher than or equal to k in S

```
void identifyCore(int k) {
      while network contains a node whose degree is < k:
            remove nodes whose degree is < k
      remaining nodes constitute k-core
}
```

**localized hubs:**

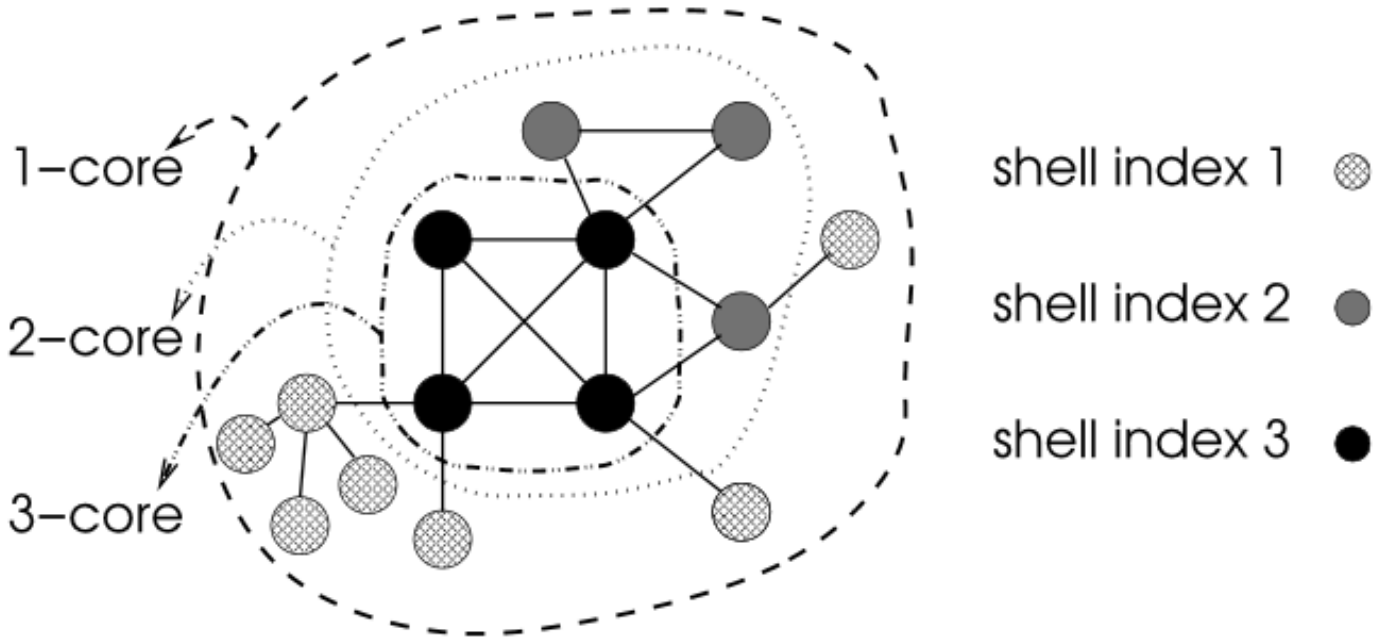
a k-core for a large k is
a connected graph or
has a giant connected component



core-periphery

# k-core decomposition

- k-cores are nested
- shell-index(x) = k — x belongs to k-core, but not to (k+1)-core
- Hubs with
  - high shell-index: hubs connected to other hubs
  - low shell-index: hubs connected to low-degree nodes



# Centrality metrics

- Metrics to rank and identify the most important nodes/links in the network

- Fundamental node centrality metrics originate from social network analysis
  - Betweenness centrality
  - Closeness centrality
  - Eigenvector centrality

- Information retrieval
  - centrality metrics for directed graphs inspired by eigenvector centrality
  - Page rank and HITS hub and authority scores

# Betweenness centrality

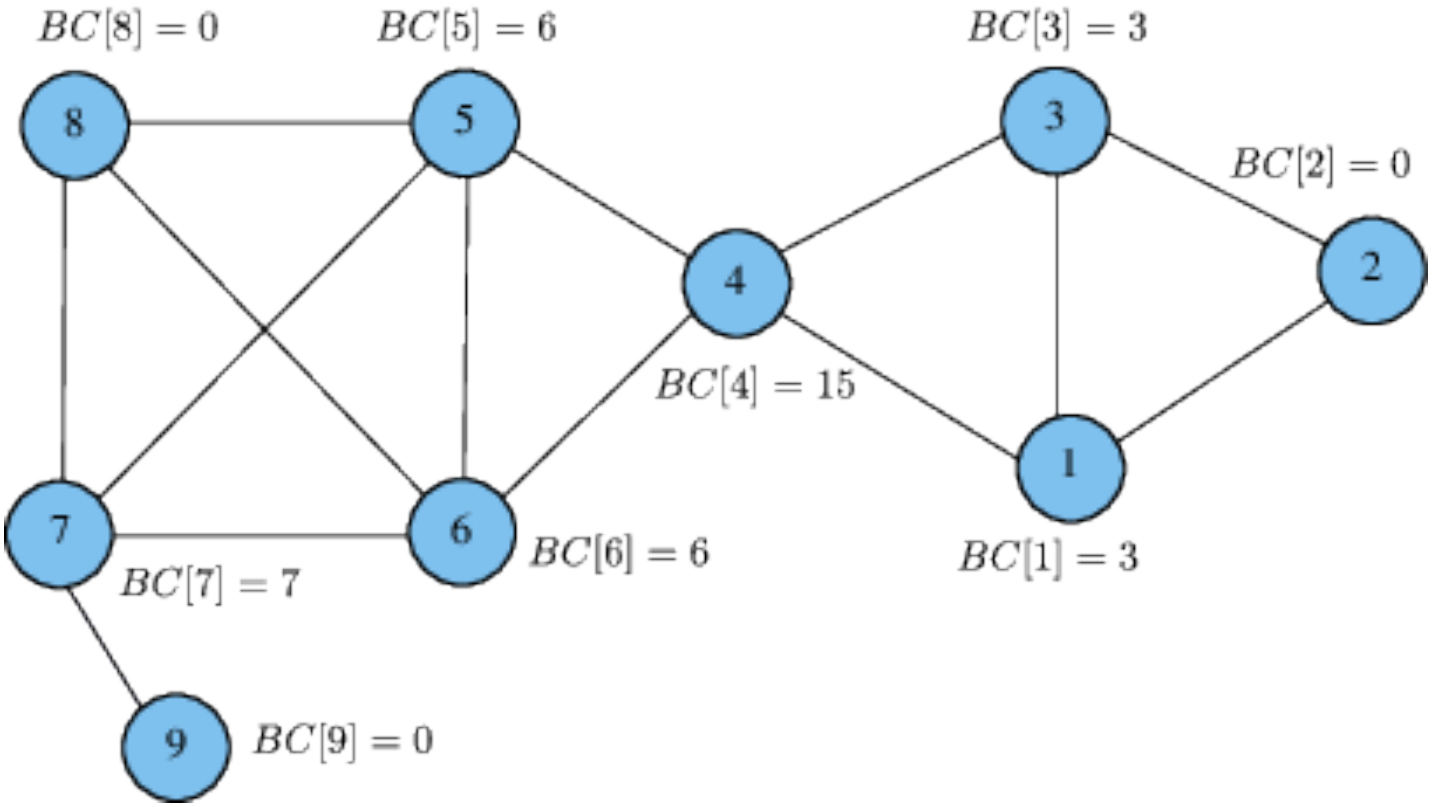
- A node is important if it is located on a large number of shortest paths between other nodes
  - Such node is in a position to control, maintain and influence information flow through the network

**Definition 2.38 (Betweenness centrality).** The betweenness centrality of a node $z$ in a graph $G$, denoted by $C_b(z)$, is the extent to which $z$ is located on the shortest paths between two arbitrary nodes different than $z$:

$$C_b(z) = \sum_{x,y \in V, x \neq y \neq z} \frac{\sigma(x,y,z)}{\sigma(x,y)} \tag{2.10}$$

where $\sigma(x,y)$ is the total number of shortest paths between $x$ and $y$, and $\sigma(x,y,z)$ is the total number of shortest paths between $x$ and $y$ passing through $z$.

# Betweenness centrality




**Low-degree nodes may have a high betweenness**

nodes connecting cohesive sub-graphs (clusters)

**High-degree nodes may have a low betweenness**

hubs in large and dense clusters

# Closeness centrality

- A node is important if it is in proximity to a large number of other nodes
  - Spreading/diffusion processes: information originating at nodes having a high closeness centrality quickly propagate through the network

**Definition 2.41 (Closeness centrality).** The closeness centrality of a node $z$ in a graph $G$, denoted by $C_c(z)$, is inversely proportional to the total distance between $z$ and all other nodes in $G$:

$$C_c(z) = \frac{1}{\sum_{i \in V \setminus \{z\}} d_{zi}} \tag{2.17}$$

# Eigenvector centrality

- Recursively defined centrality: a node is important if it is directly connected to other important nodes

**Definition 2.44 (Eigenvector centrality).** The eigenvector centrality of a node $z$ in a graph $G$, denoted by $C_e(z)$, is proportional to the sum of eigenvector centralities of its neighbors:

$$C_e(z) = \frac{1}{\lambda} \sum_{i \in N(z)} C_e(i) \tag{2.21}$$

where $\lambda$ is a constant and $N(z)$ denotes the set of nodes directly connected to $z$, i.e. $N(z) = \{w : \{w,z\} \in E\}$.

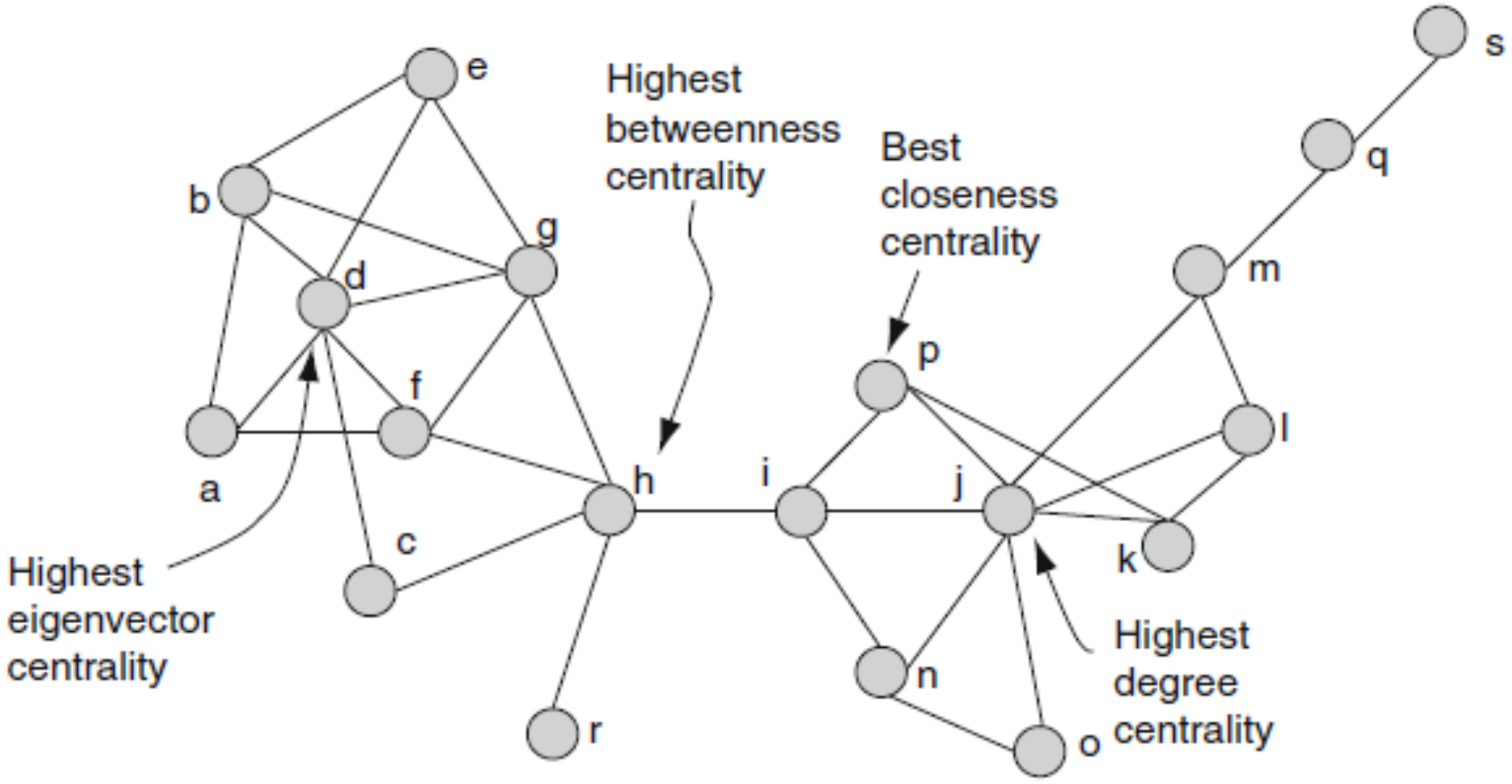
**input** : a graph $G = (V, E)$
$M$ – a number indicating the maximal number of iterations
$\varepsilon$ – a real value indicating the desired precision
**output:** $C_e$ – a vector containing eigenvector centralities of nodes in $G$

$N = |V|$
$C_e = [1, 1, 1, \ldots, 1]_N$
$C_e^n = [0, 0, 0, \ldots, 0]_N$
$\Delta = \varepsilon$
$i = M$

**while** $\Delta \geq \varepsilon \wedge i > 0$ **do**
    $i = i - 1$
    **foreach** $v \in V$ **do**
        // determine the new value of eigenvector centrality of $v$ according to
        // the current eigenvector centralities of its neighbors
        $C_e^n[v] = 0$
        **foreach** $w \in V : \{v, w\} \in E$ **do**
            $C_e^n[v] = C_e^n[v] + C_e[w]$
        **end**
    **end**
    // normalize $C_e^n$ to an unit vector
    $C_e^n = C_e^n / \|C_e^n\|$
    // compute the $L_1$ distance between $C_e^n$ and $C_e$
    $\Delta = 0$
    **foreach** $v \in V$ **do**
        $\Delta = \Delta + |C_e^n[v] - C_e[v]|$
    **end**
    // copy $C_e^n$ to $C_e$ for the next iteration
    $C_e = C_e^n$
**end**

# Centrality metrics

- Moderate to strong correlations between different centrality metrics in real-world complex networks
  → Node rankings by different centrality metrics are similar, but not identical



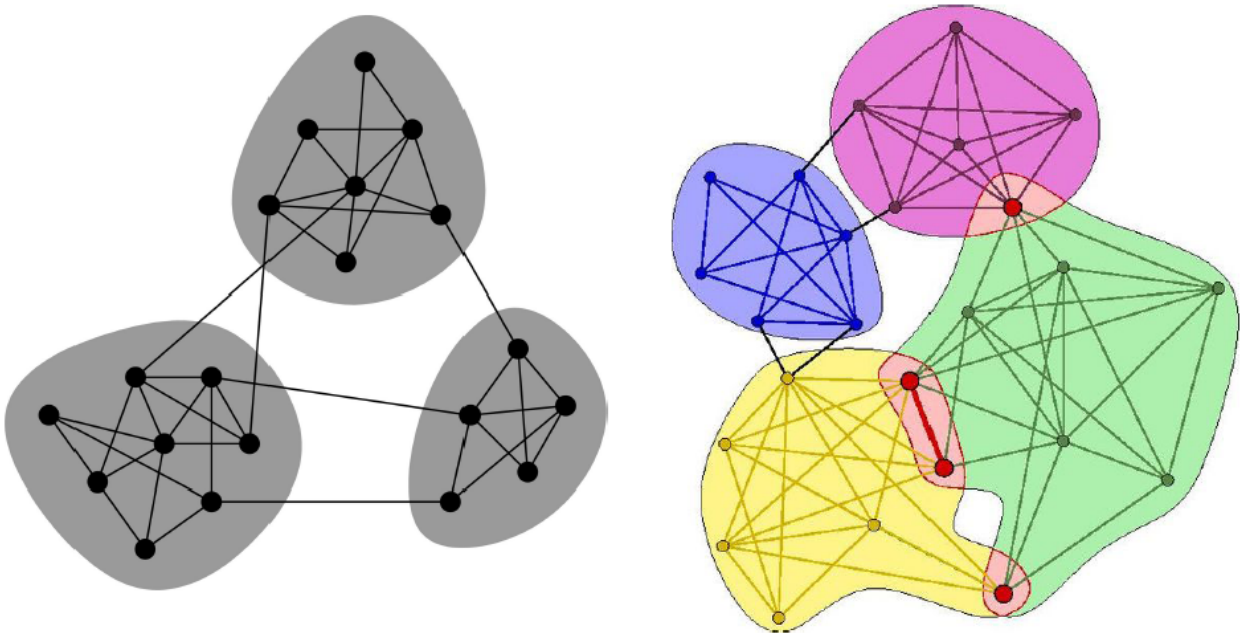
Betweenness centrality

Closeness centrality

Eigenvector centrality

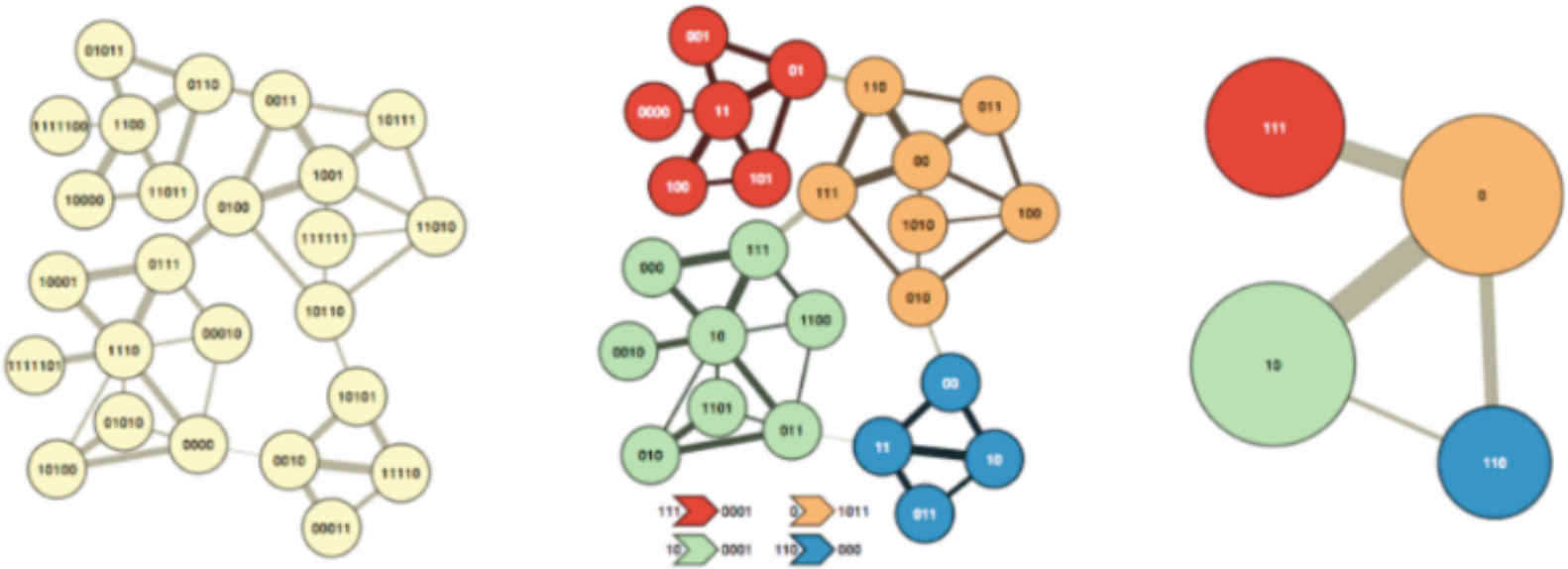
# Community structure

- Community (module, node cluster)
  - a subgraph that is more densely/strongly internally connected than with the rest of the network
- Automatic identification of communities — community detection algorithms
- Overlapping and non-overlapping community partitions

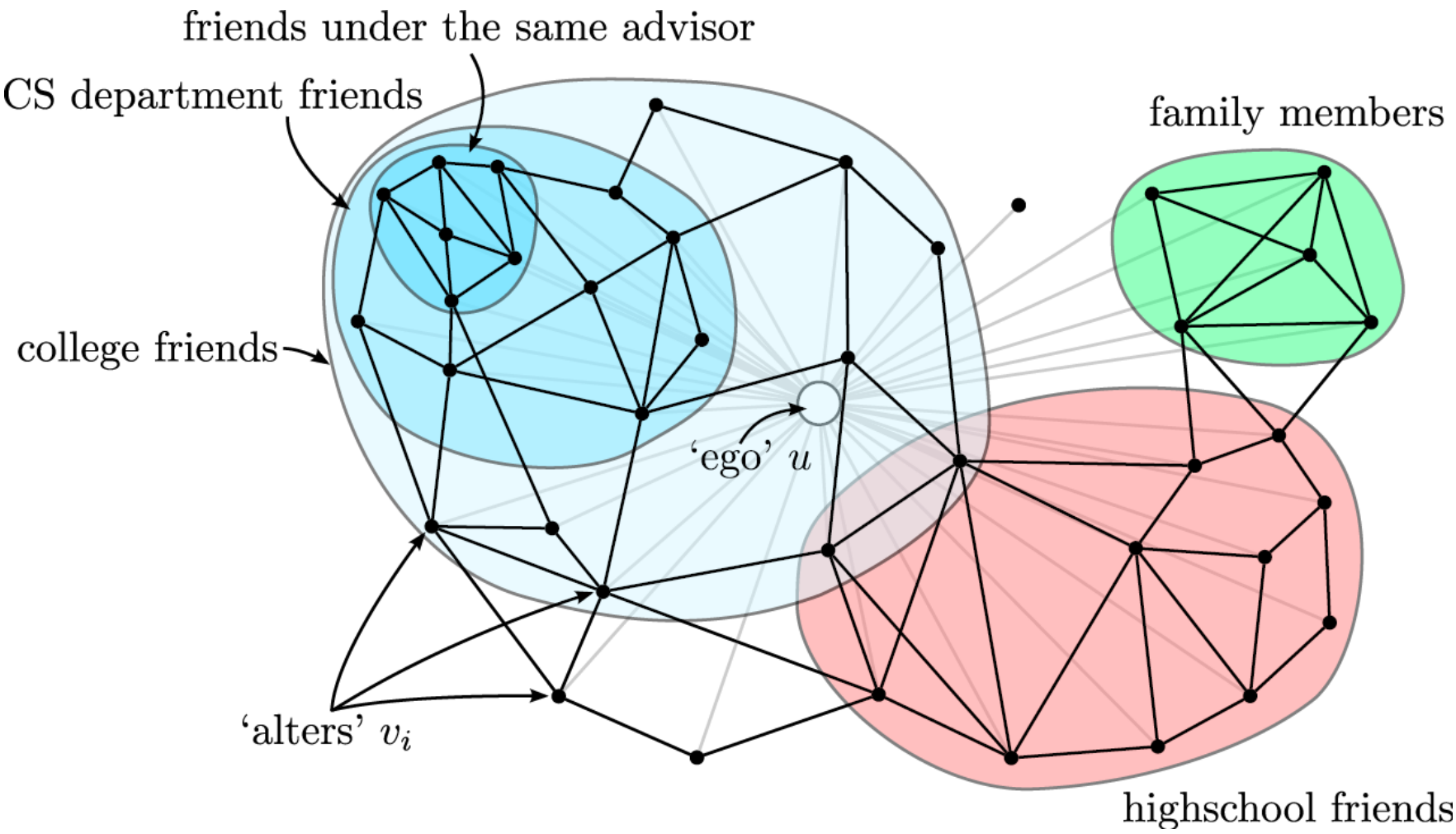


# Network comprehension

- Block model — a network of communities

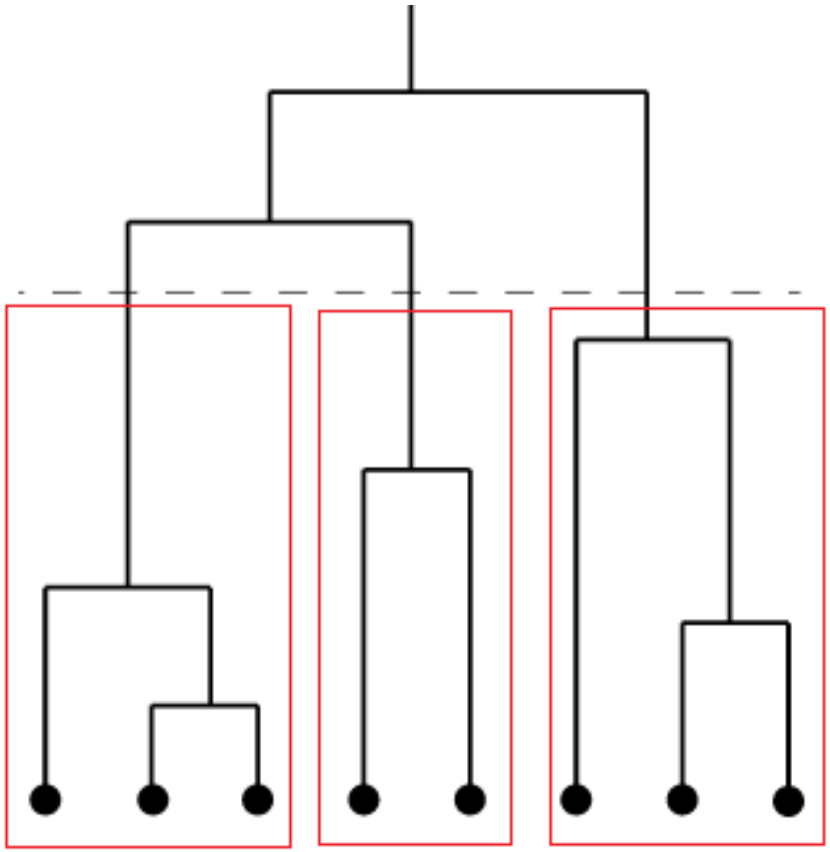


# Nested (hierarchical) communities

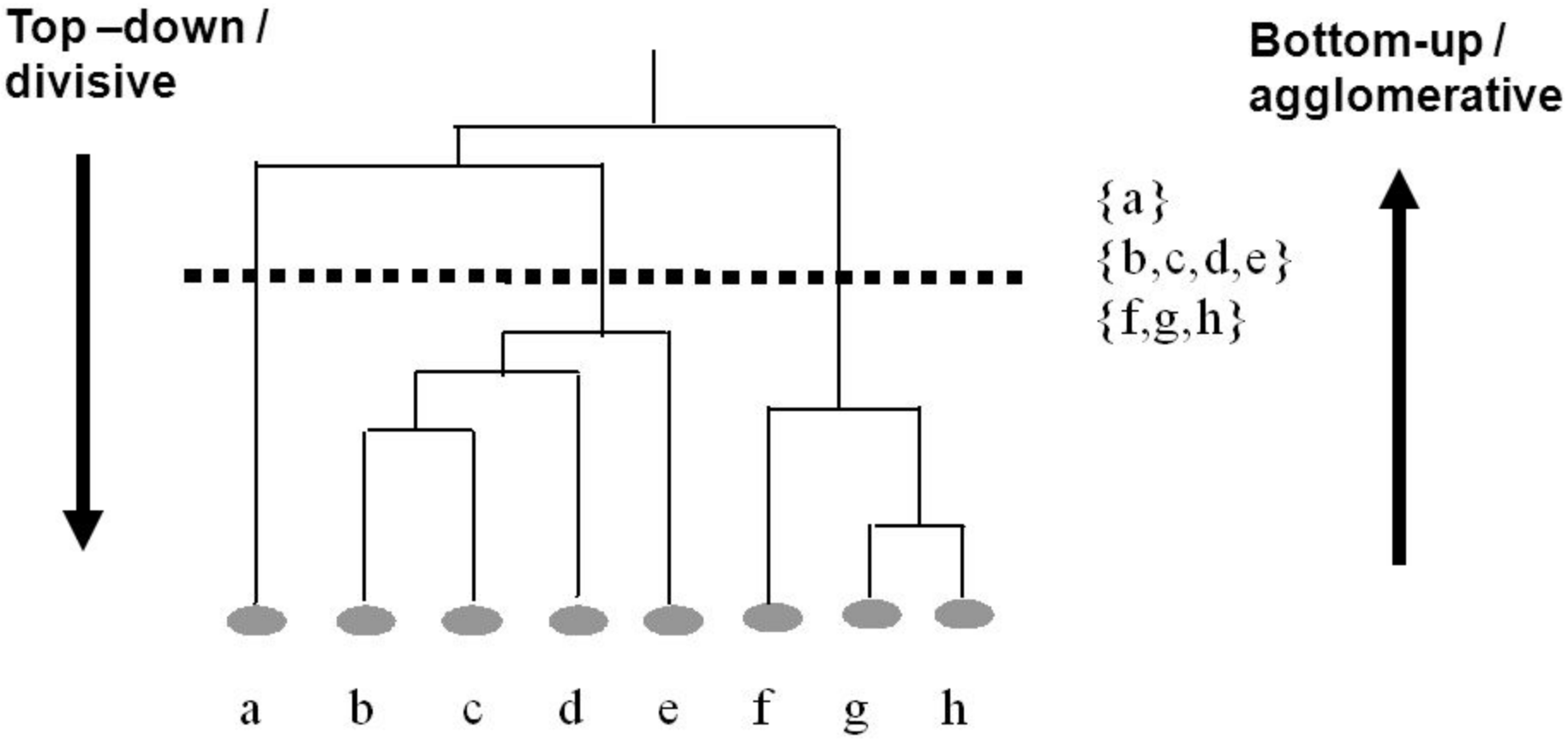


# Nested communities

- Nested communities can be represented by a dendrogram
  - Tree whose leafs are network nodes
  - Nodes in a subtree of the dendrogram form one community
- One dendrogram cut — one non-overlapping community partition



# Hierarchical clustering



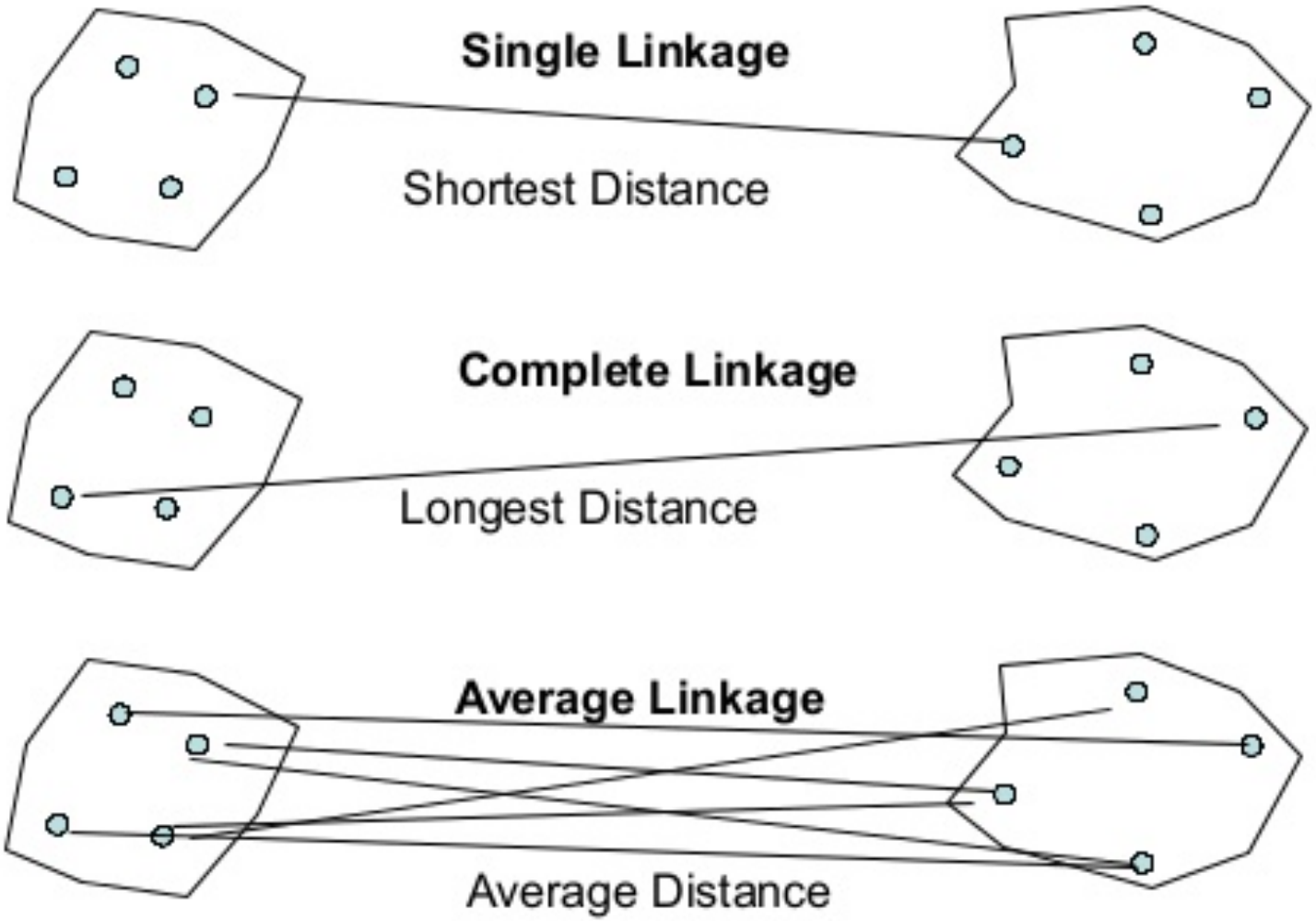
# Hierarchical agglomerative clustering

- Start with the partitioning in which nodes form singleton communities (one node one community), i.e. form leafs of the dendrogram

- Find **two most similar communities** A and B and join them into a community C
  - Form a new node C in the dendrogram such that Parent(A) = Parent(B) = C

- We need a function quantifying similarity/distance between communities
- Such function can be derived from a function quantifying similarity/distance between nodes

# Node similarity/distance —> Cluster similarity/distance



# Node similarity/distance

- The length of the shortest path between two nodes
- Similarity based on random walks: the probability that a random walker reaches X from Y in k random walk steps
- The number of common neighbors $|\Gamma(x) \cap \Gamma(y)|$
  - **Facebook:** mutual friends
- Jaccard coefficient $\frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$
  - The number of common neighbors normalized to avoid inherent bias towards pairs of hubs (high-degree nodes)
- **Other metrics:** Adamic-Adar, Katz, personalized PageRank, cosine similarity, SimRank (recursively defined similarity)

- Santo Fortunato, 2009, “Community detection in graphs”
  - traditional clustering and graph partitioning methods (e.g., K-means, Kernighan-Lin, etc.)
  - Divisive algorithms
    - Repeatedly remove links that are likely to be inter-communitarian links to form the dendrogram
    - Measures indicating inter-communitarian links: edge betweenness centrality, edge clustering coefficient, edge information centrality
  - Modularity-based algorithms
    - heuristics to maximize the modularity measure
      - X — a subgraph in the network
      - Q(X) = the number of links in X - the expected number of links in X under some null random network model
  - Dynamic algorithms
    - Discovering communities by dynamical processes running on the network (e.g. label propagation)
  - Method-based on statistical inference
    - fitting stochastic block models

# Literature

- **Books**
  - M. E. Newman. Network: An Introduction, 2010
  - D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World, 2010
  - Brandes and Erlebach (Eds). Network Analysis: Methodological Foundations, 2005
  - Aggarwal (Ed). Social Network Data Analytics, 2011

- **Articles**
  - M. E. Newman. The structure and function of complex networks
  - Boccaletti et al. Complex networks: Structure and dynamics
  - S. Fortunato. Community detection in graphs
  - Liben-Nowell and Kleinberg. The link prediction problem for social networks