Social Networks: Theories and Applications

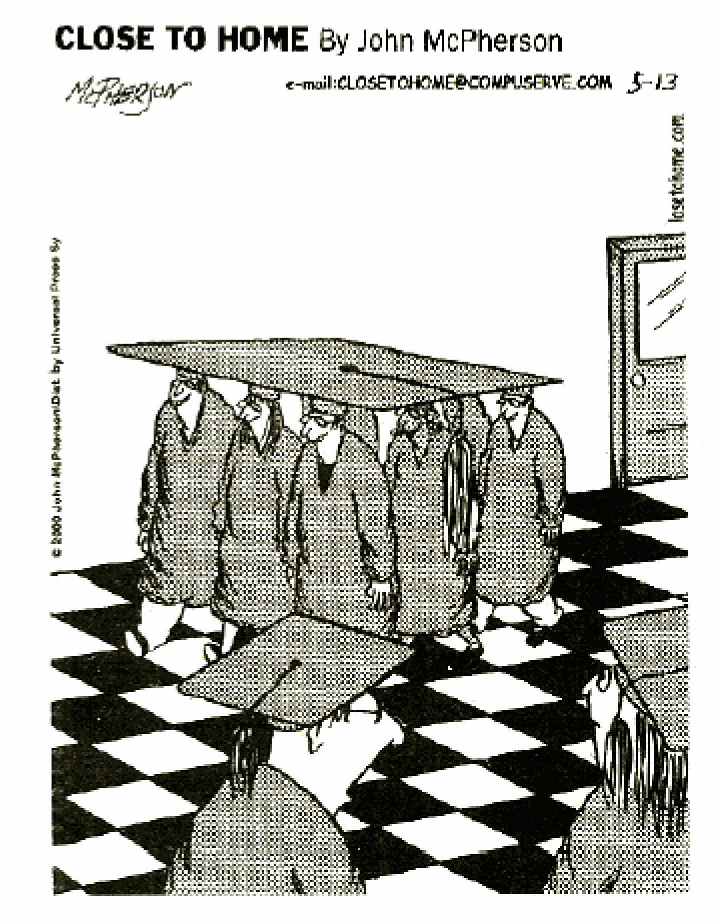
Yves Zenou
Stockholm University

March 16, 2012

Modeling Social Networks: Where we are and where to go

- Some empirical background
- What are the interesting questions?
- Random graph models
  - a few representative examples
  - strengths and weaknesses
- Strategic/Game Theoretic models
  - a few representative examples
  - strengths and weaknesses
- Hybrids and the future

Introduction

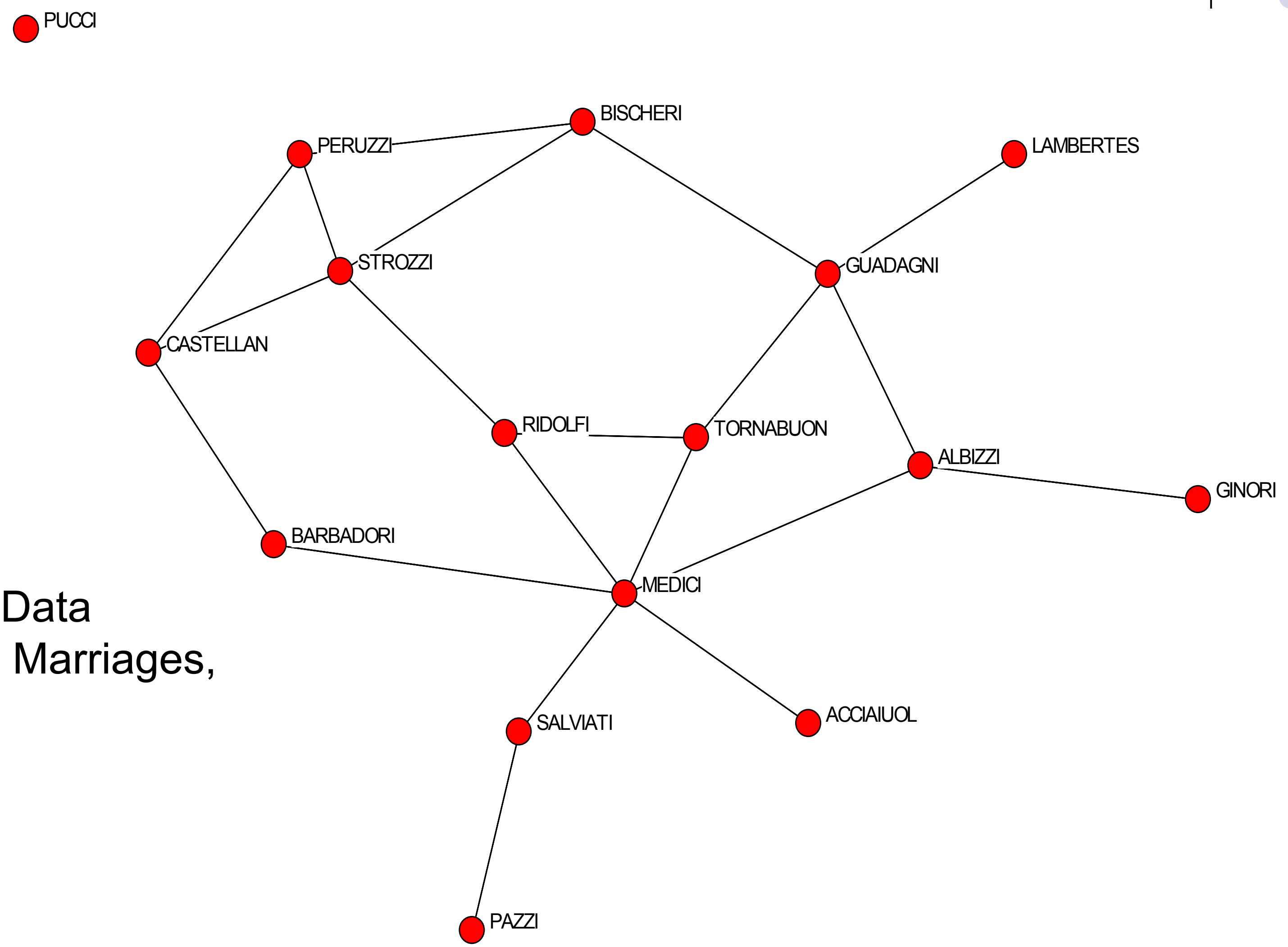


"They always were a tight little clique."

"Miss Wilkins, have some of my close associates in the old-boy network arrange a little surprise party for me."

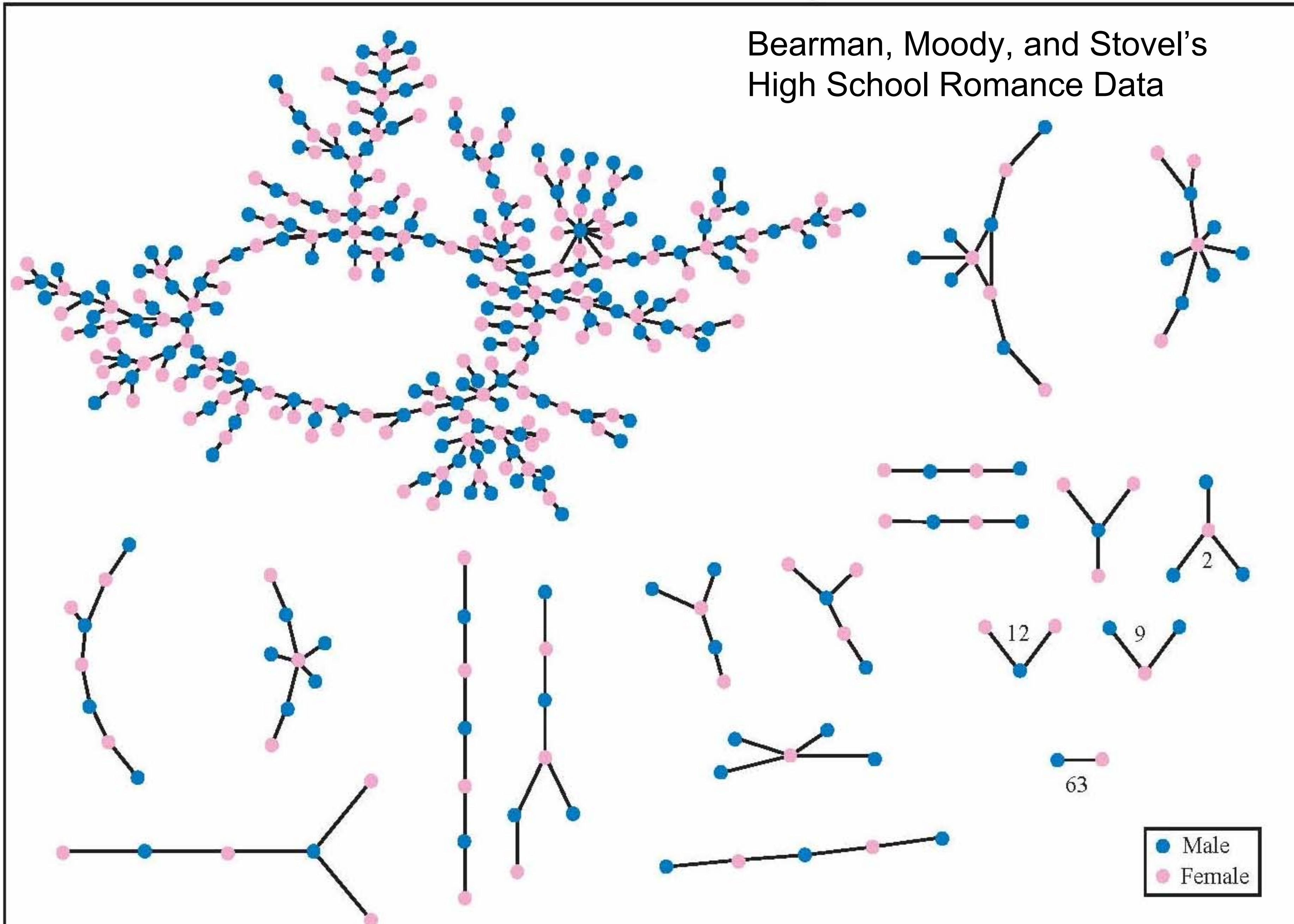
Source: Linton Freeman "See you in the funny pages" *Connections*, 23, 2000, 32-42.

Examples of Social and Economic Networks



Padgett’s Data
Florentine Marriages, 1430’s

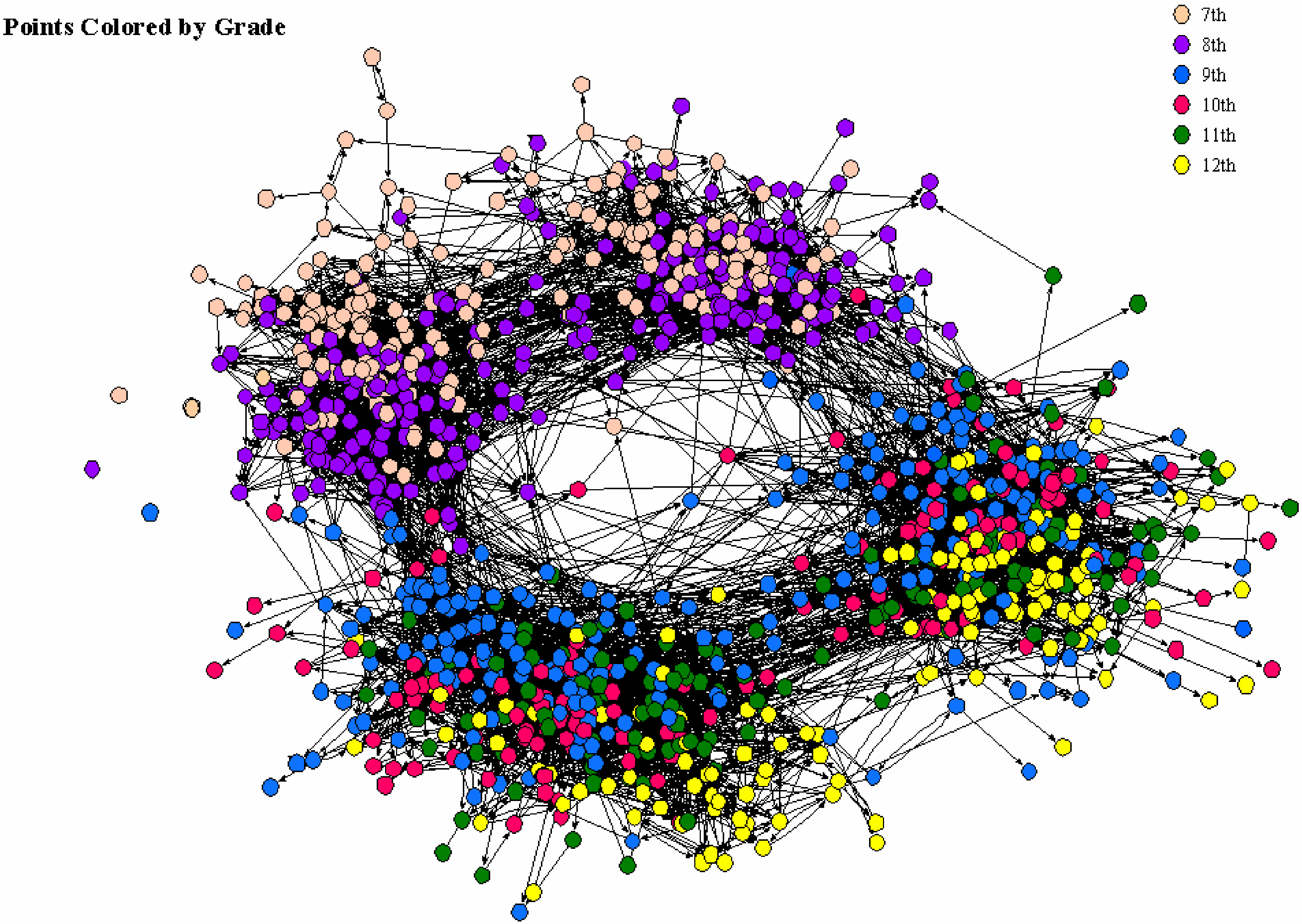
The Structure of Romantic and Sexual Relations at "Jefferson High School"



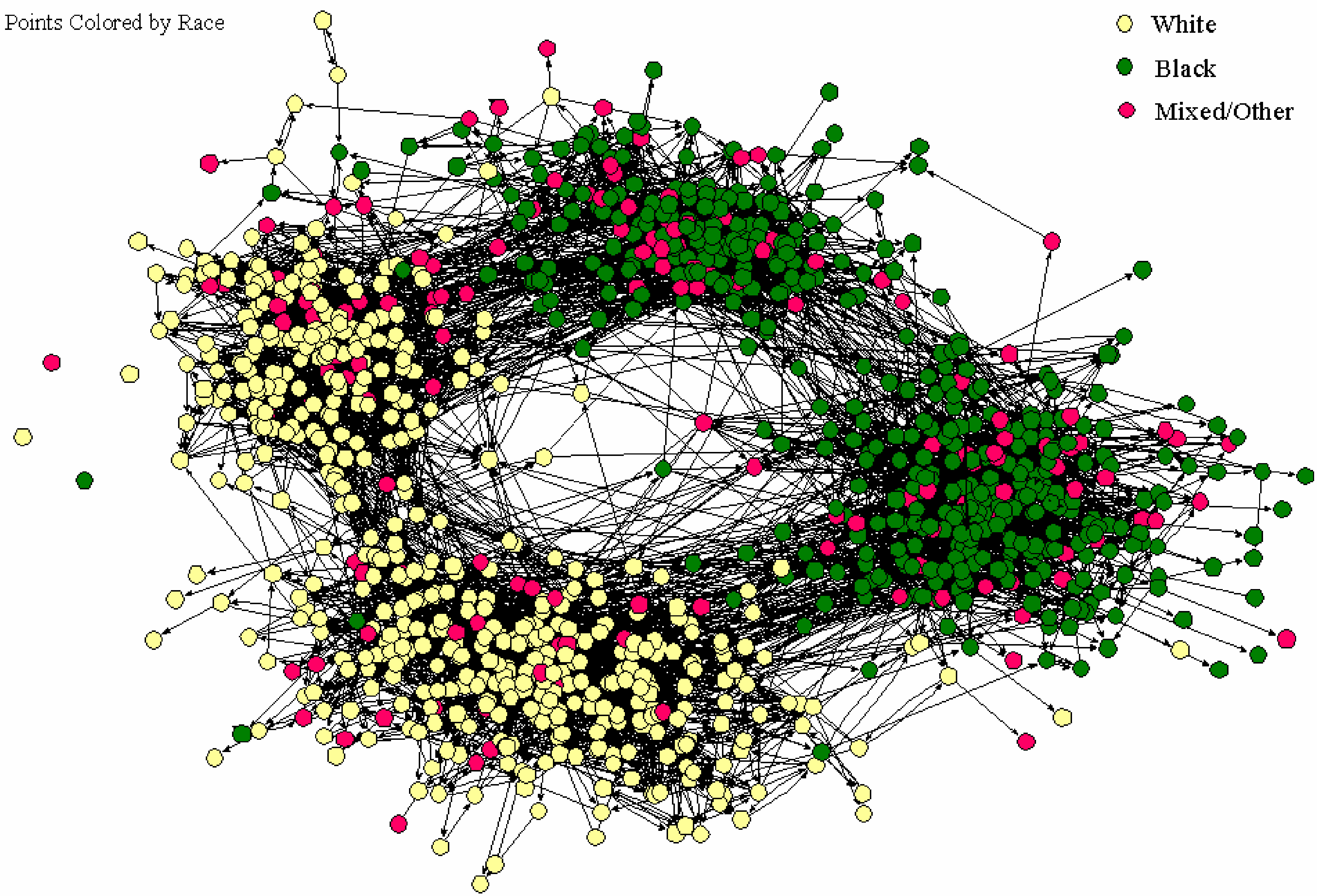
Introduction

High Schools as Networks

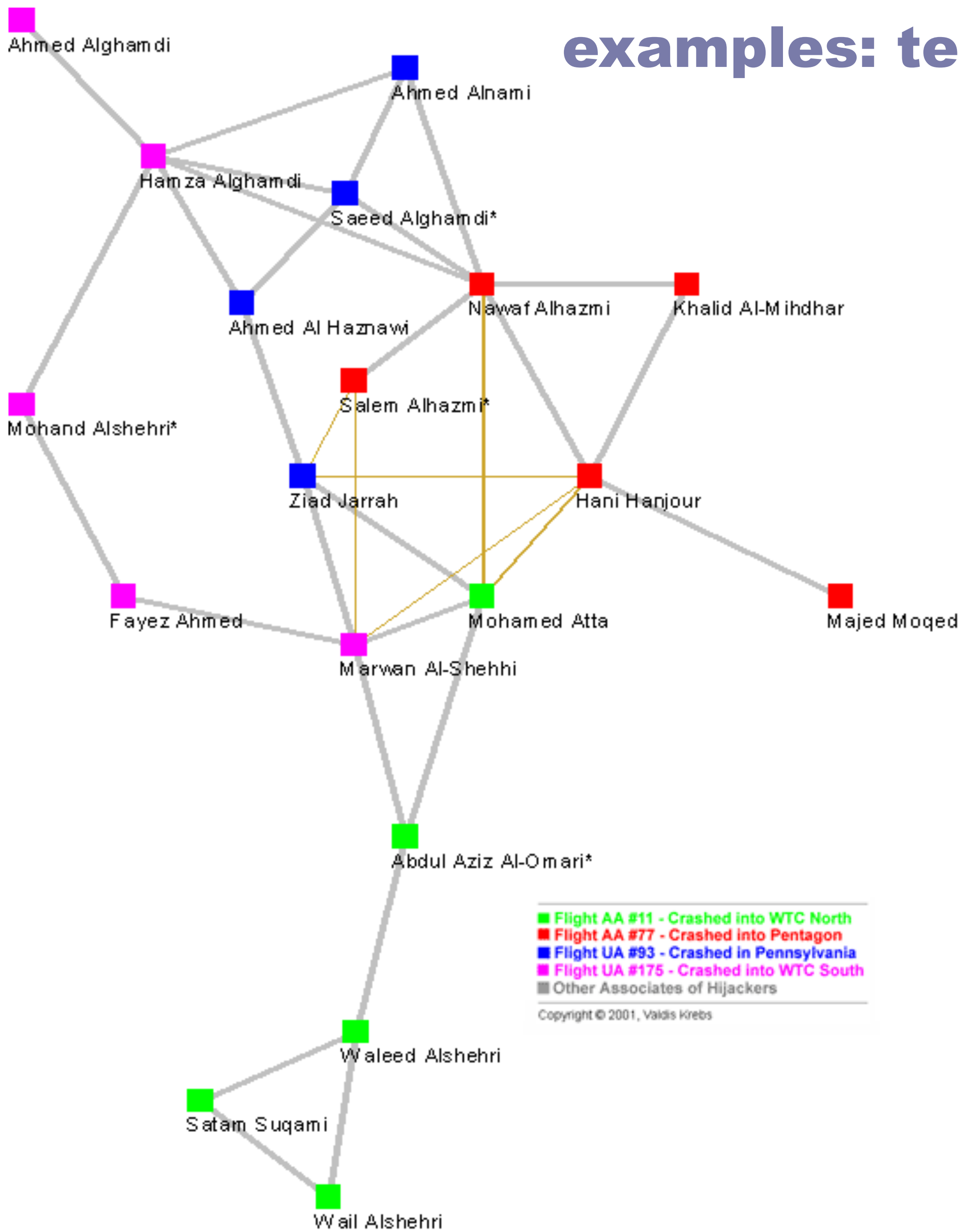
The Social Structure of "Countryside" School District



The Social Structure of "Countryside" School District



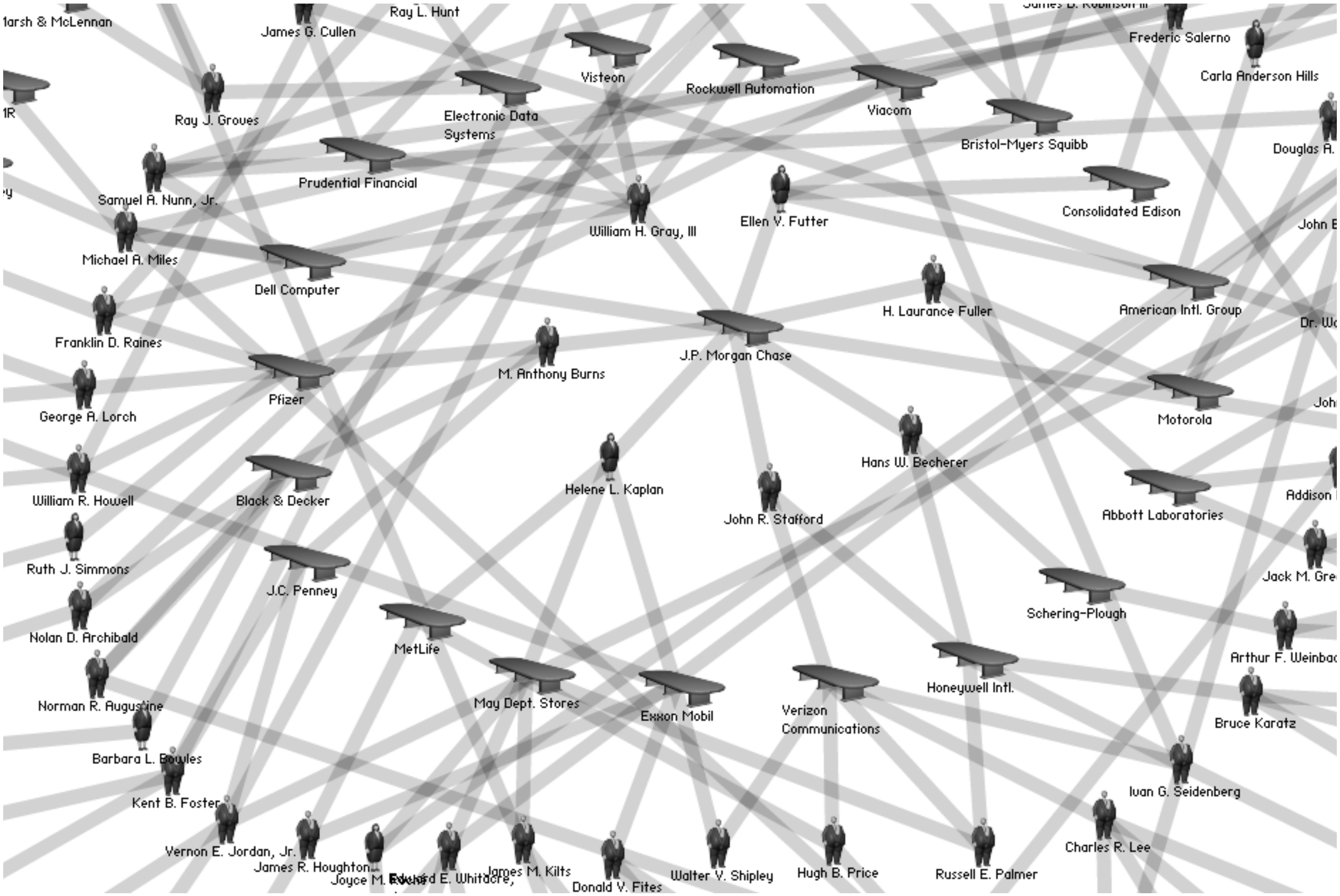
examples: terrorist networks



"Six degrees of Mohammed Atta"

Uncloaking Terrorist Networks, by Valdis Krebs

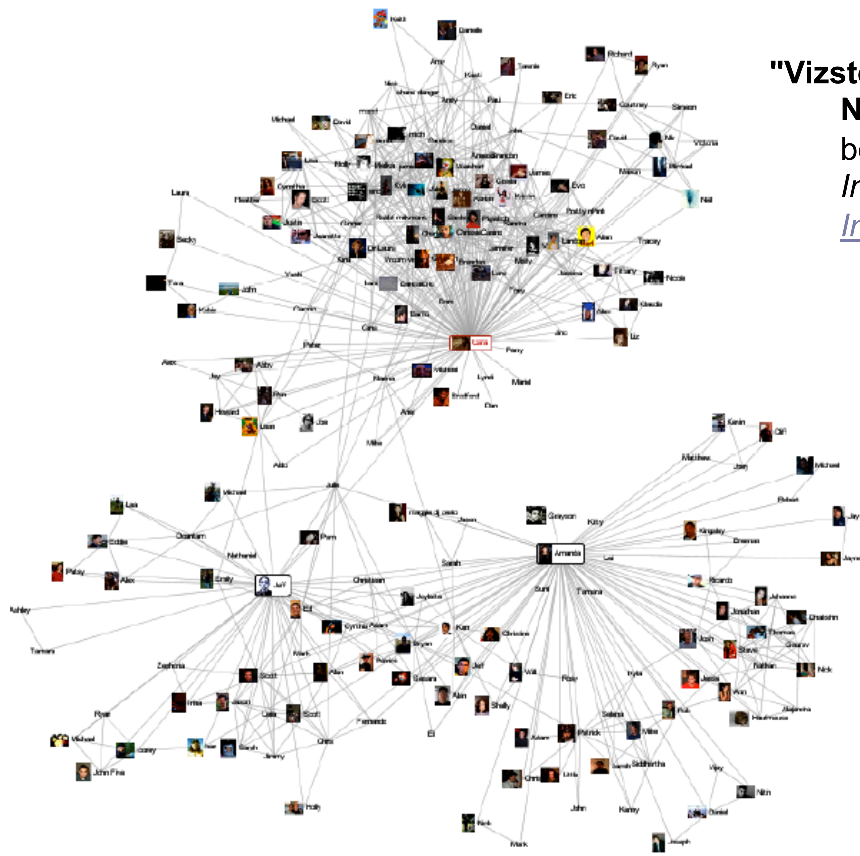
examples: boards of directors



Source: http://theyrule.net

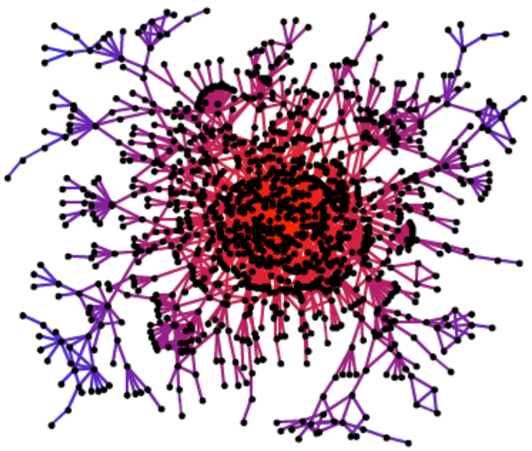
examples: online social networks

- Friendster

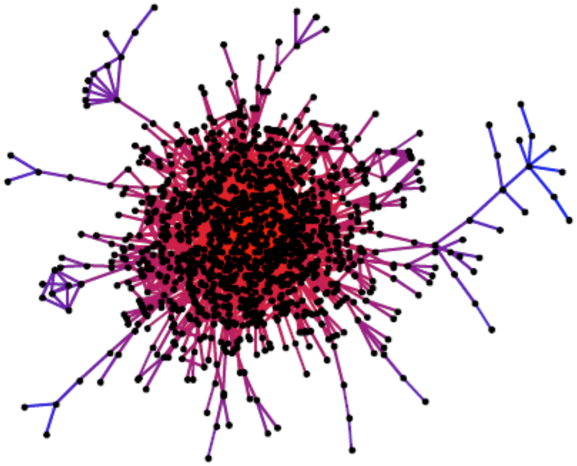


"Vizster: Visualizing Online Social Networks." Jeffrey Heer and danah boyd. *IEEE Symposium on Information Visualization (* *InfoViz 2005*).

examples: Networks of personal homepages



Stanford



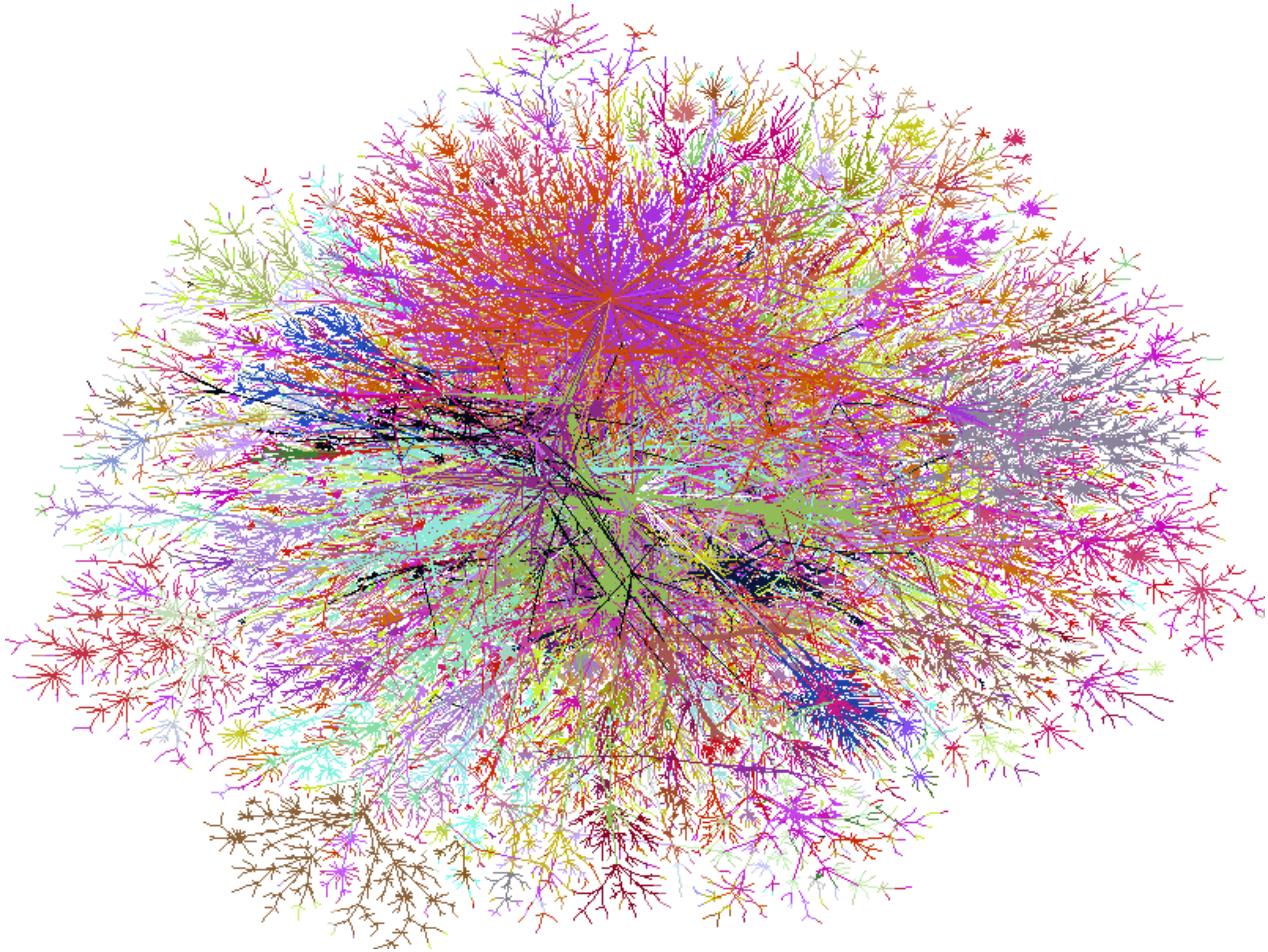
MIT

homophily: what attributes are predictive of friendship?

group cohesion

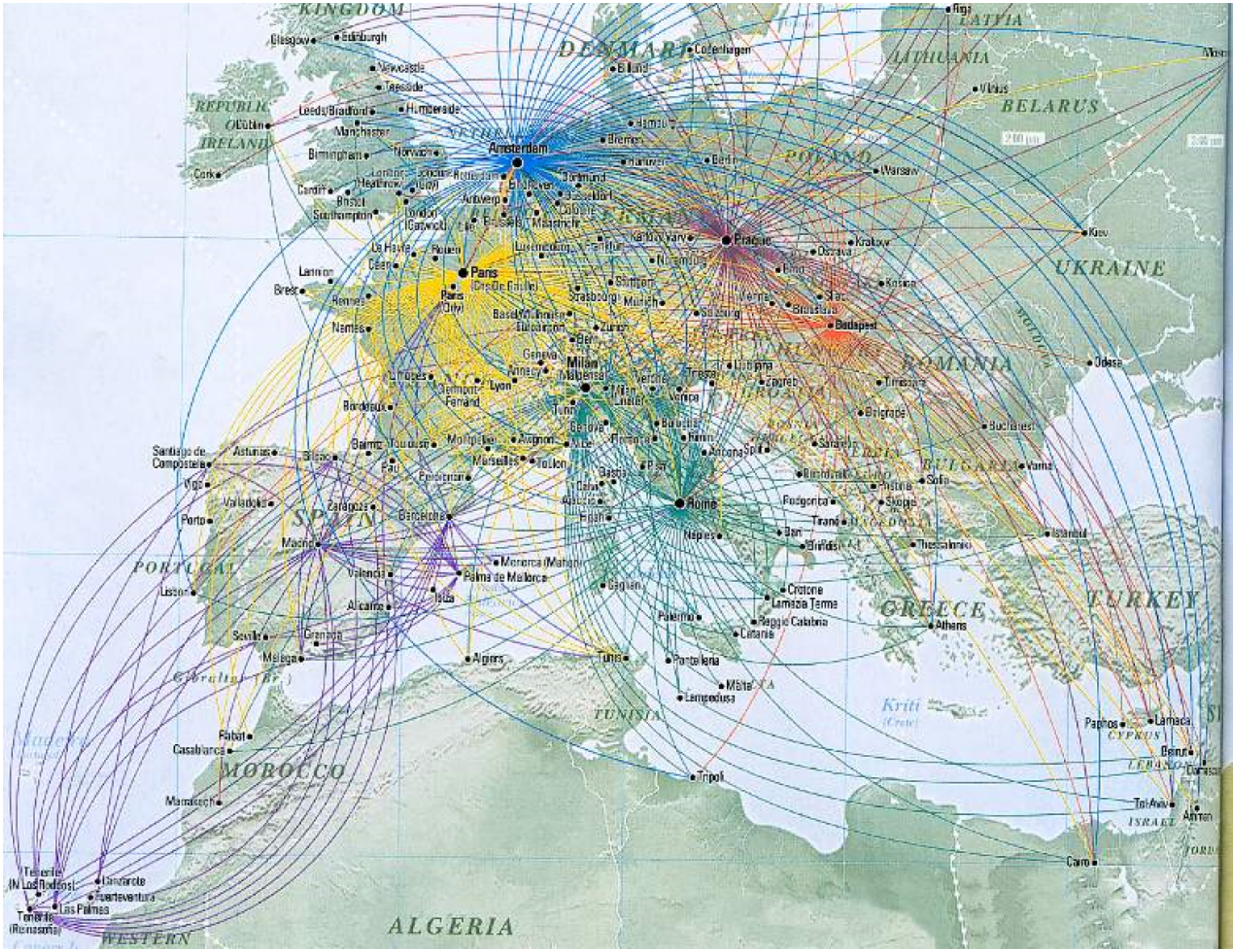
Source: Lada A. Adamic and Eytan Adar, ‘Friends and neighbors on the web’, *Social Networks*, 25(3):211-230, July 2003.

examples: internet



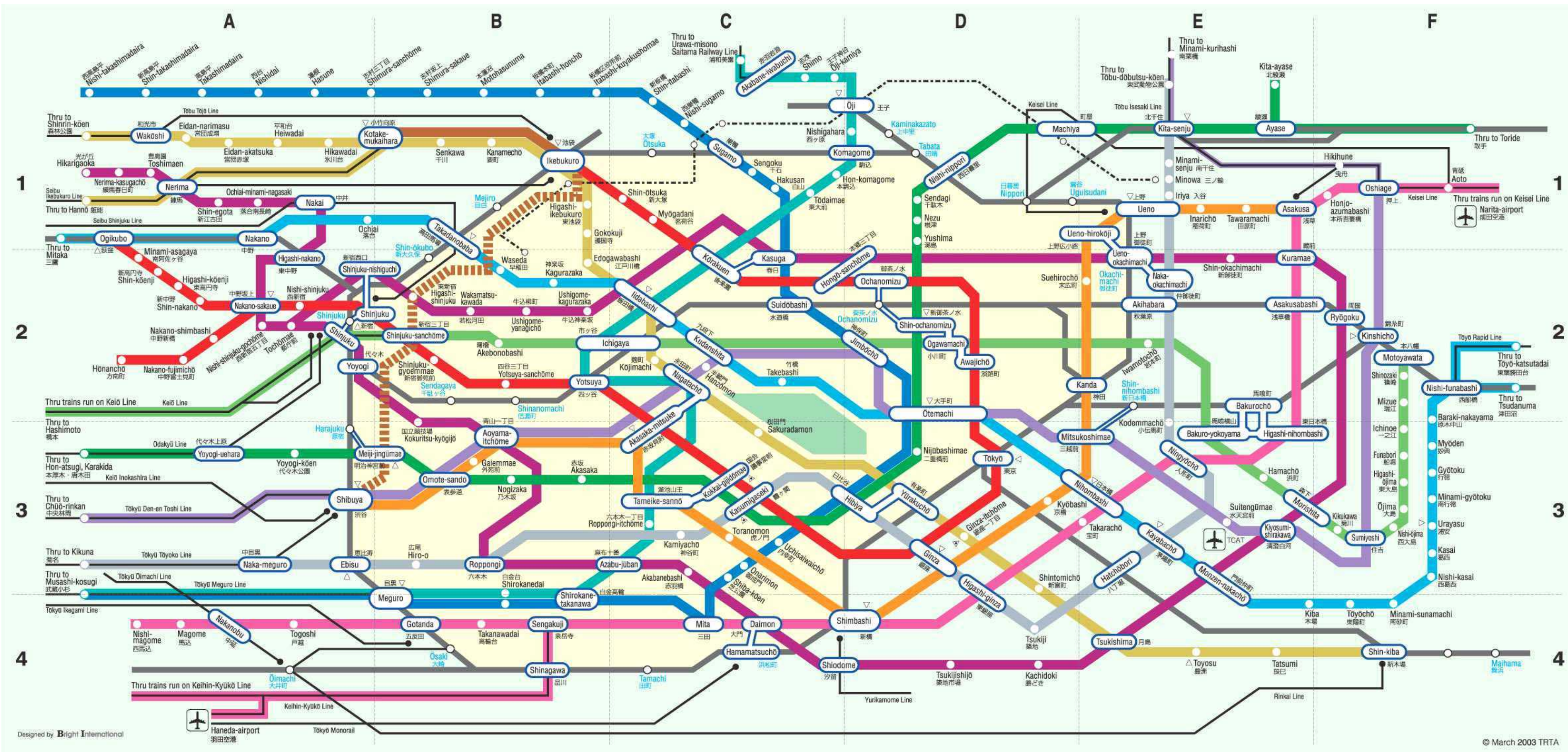
Source: Bill Cheswick http://www.cheswick.com/ches/map/gallery/index.html

examples: airline networks



Source: Northwest Airlines WorldTraveler Magazine

examples: railway networks



Source: TRTA, March 2003 - Tokyo rail map

Two persons belong to the same component if and only if there exists a path between them.

The components can be ordered in terms of their size.

We say that the network has a **giant component** if the largest component constitutes a relatively large part of the population of economists and all other components are small (typically of order $ln[n]$).

Distance 2 is the number of nodes at distance 2 from i.

Evolving network structures

Aggregate statistics

Table: Network statistics for the co-author networks.

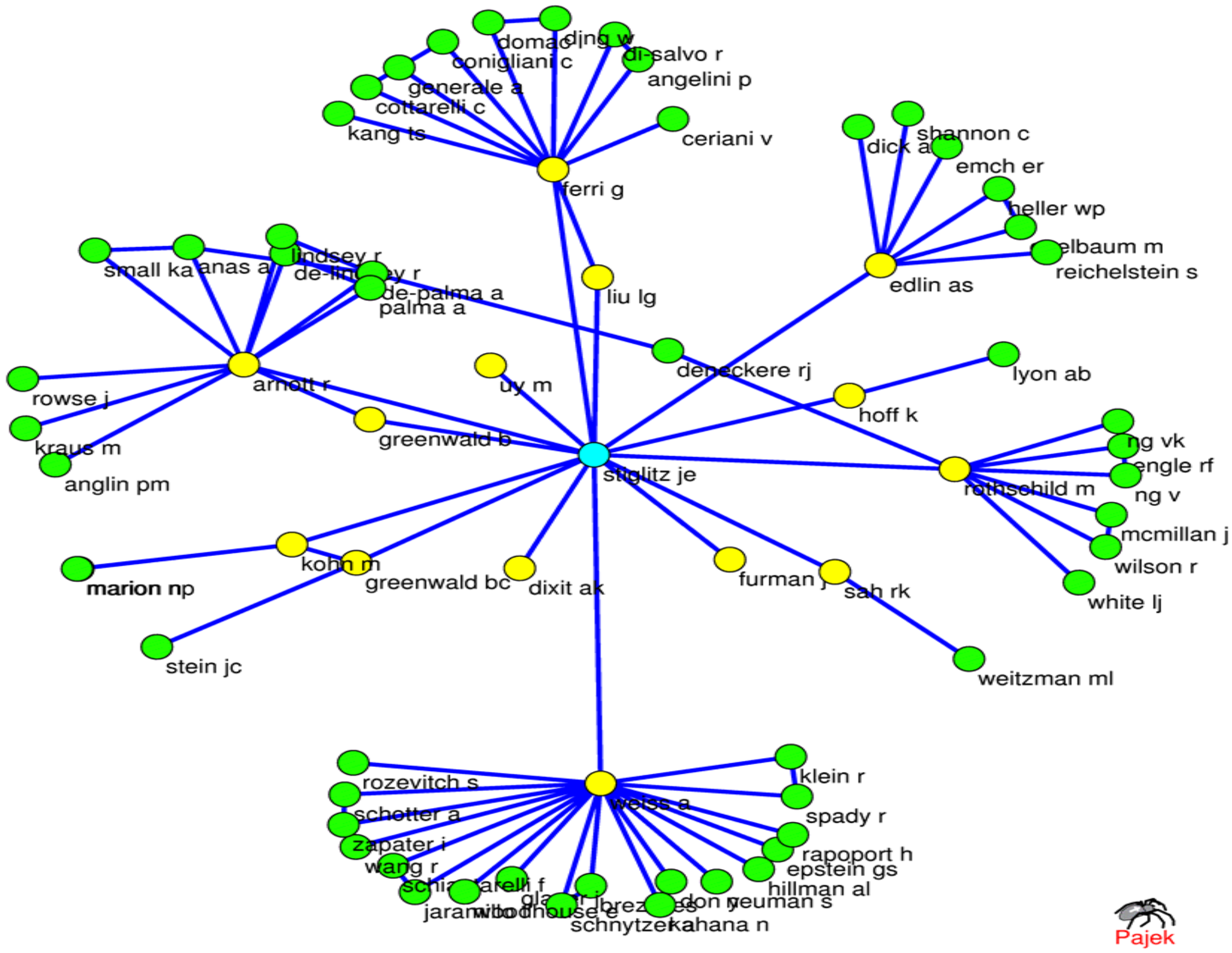
decade	1970s	1980s	1990s
total authors	33770	48608	81217
average degree	.894	1.244	1.672
size of giant component	5253	13808	33027
—as percentage	.156	.284	.407
second largest component	122	30	30
average distance in giant component	12.86	11.07	9.47

Introduction and overview
Hubs and spokes

Table: Most linked economists: 1970's, 1980's and 1990s.

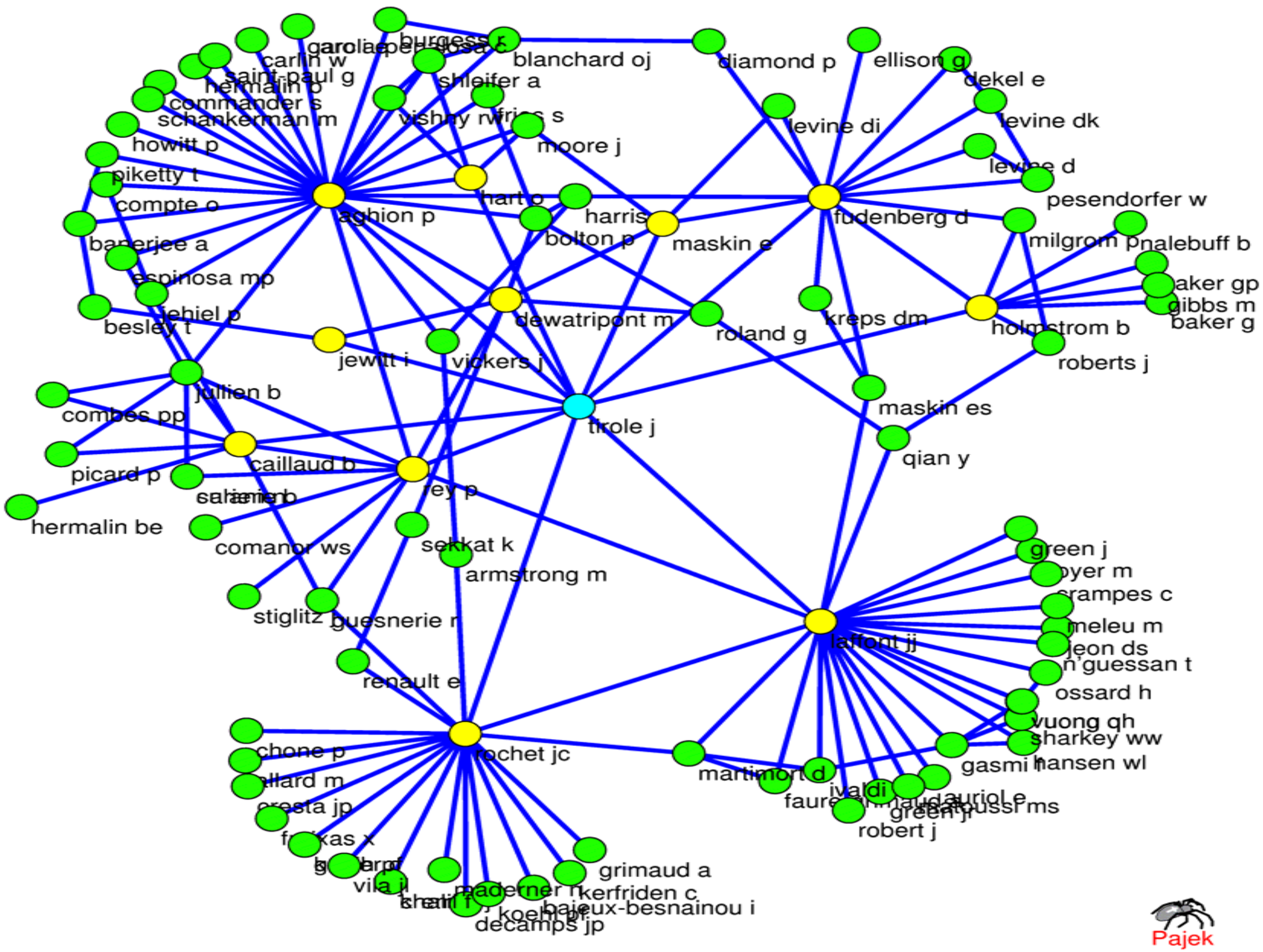
Average Top 100	Papers	% Co-authored	Degree	Dist. 2
1970's	23.87	0.724	11.94	25.67
1980's	28.42	0.827	16.36	49.80
1990's	37.69	0.849	25.31	99.40

Local network of J. Stiglitz in 1990's



Note: Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit.

Local network of J. Tirole in 1990's



Note: Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit.

Evolving social networks

Emerging small world

- **Changes in the network:**
  1. Large growth in the number of authors: from 33770 in the 1970's to 81217 in the 1990's.
  2. Massive growth in giant component: In the 1970's the largest component contained 5,253 nodes (15.6% of all nodes), while in the 1990's it contains 33,027 nodes (40% of all nodes).
  3. Sharp fall in second largest component: from 122 in 1970's to 30 in 1990's.
  4. Sharp decrease in average distance: from 12.86 in 1970's to 9.47 in 1990's.
- Suggests an **emerging small world.** Why has the world become smaller?

Evolving social networks

Emerging small world

- **Changes in the network:**
  1. Large growth in the number of authors: from 33770 in the 1970's to 81217 in the 1990's.
  2. Massive growth in giant component: In the 1970's the largest component contained 5,253 nodes (15.6% of all nodes), while in the 1990's it contains 33,027 nodes (40% of all nodes).
  3. Sharp fall in second largest component: from 122 in 1970's to 30 in 1990's.
  4. Sharp decrease in average distance: from 12.86 in 1970's to 9.47 in 1990's.
- Suggests an **emerging small world.** Why has the world become smaller?

Evolving social networks

Emerging small world

- **Changes in the network:**
  1. Large growth in the number of authors: from 33770 in the 1970's to 81217 in the 1990's.
  2. Massive growth in giant component: In the 1970's the largest component contained 5,253 nodes (15.6% of all nodes), while in the 1990's it contains 33,027 nodes (40% of all nodes).
  3. Sharp fall in second largest component: from 122 in 1970's to 30 in 1990's.
  4. Sharp decrease in average distance: from 12.86 in 1970's to 9.47 in 1990's.
- Suggests an **emerging small world.** Why has the world become smaller?

Evolving social networks

Emerging small world

- **Changes in the network:**
  1. Large growth in the number of authors: from 33770 in the 1970's to 81217 in the 1990's.
  2. Massive growth in giant component: In the 1970's the largest component contained 5,253 nodes (15.6% of all nodes), while in the 1990's it contains 33,027 nodes (40% of all nodes).
  3. Sharp fall in second largest component: from 122 in 1970's to 30 in 1990's.
  4. Sharp decrease in average distance: from 12.86 in 1970's to 9.47 in 1990's.
- Suggests an **emerging small world.** Why has the world become smaller?

Evolving social networks
Emerging small world

- **Changes in the network:**
  1. Large growth in the number of authors: from 33770 in the 1970's to 81217 in the 1990's.
  2. Massive growth in giant component: In the 1970's the largest component contained 5,253 nodes (15.6% of all nodes), while in the 1990's it contains 33,027 nodes (40% of all nodes).
  3. Sharp fall in second largest component: from 122 in 1970's to 30 in 1990's.
  4. Sharp decrease in average distance: from 12.86 in 1970's to 9.47 in 1990's.
- Suggests an **emerging small world.** Why has the world become smaller?

Evolving network structures

Why has the world become smaller?

- A number of key elements of the network have changed: the number of nodes has grown, the distribution is different, and clustering has changed. We would like to assign quantitative significance to the different changes. It is difficult to identify the cause of the changes.
- We use a model due to Jackson-Rogers-Vasquez which combines random and local linking and estimate its its parameters in different periods, 970's, 1980's and 1990's. We then discuss the economics of the changes.
- This is on-going work.

Evolving social networks

The strength of weak ties?

- We may interpret strength of a co-author tie as a function of the number of papers written together. We use our data to define a weighted network and examine the structural validity of the celebrated *strength of weak ties* hypothesis due to Mark Granovetter.
- This part of talk is based on **van der Leij and Goyal (2006).**
- In particular, we would like to know where the strong ties are located – at the center of the network or at the periphery – and how this affects their importance in connecting up the network.

Evolving social networks

The strength of weak ties?

- We may interpret strength of a co-author tie as a function of the number of papers written together. We use our data to define a weighted network and examine the structural validity of the celebrated *strength of weak ties* hypothesis due to Mark Granovetter.
- This part of talk is based on **van der Leij and Goyal (2006).**
- In particular, we would like to know where the strong ties are located – at the center of the network or at the periphery – and how this affects their importance in connecting up the network.

Evolving social networks

The strength of weak ties?

- We may interpret strength of a co-author tie as a function of the number of papers written together. We use our data to define a weighted network and examine the structural validity of the celebrated *strength of weak ties* hypothesis due to Mark Granovetter.
- This part of talk is based on **van der Leij and Goyal (2006).**
- In particular, we would like to know where the strong ties are located – at the center of the network or at the periphery – and how this affects their importance in connecting up the network.

Evolving social networks

The strength of weak ties?

- We may interpret strength of a co-author tie as a function of the number of papers written together. We use our data to define a weighted network and examine the structural validity of the celebrated *strength of weak ties* hypothesis due to Mark Granovetter.
- This part of talk is based on **van der Leij and Goyal (2006).**
- In particular, we would like to know where the strong ties are located – at the center of the network or at the periphery – and how this affects their importance in connecting up the network.

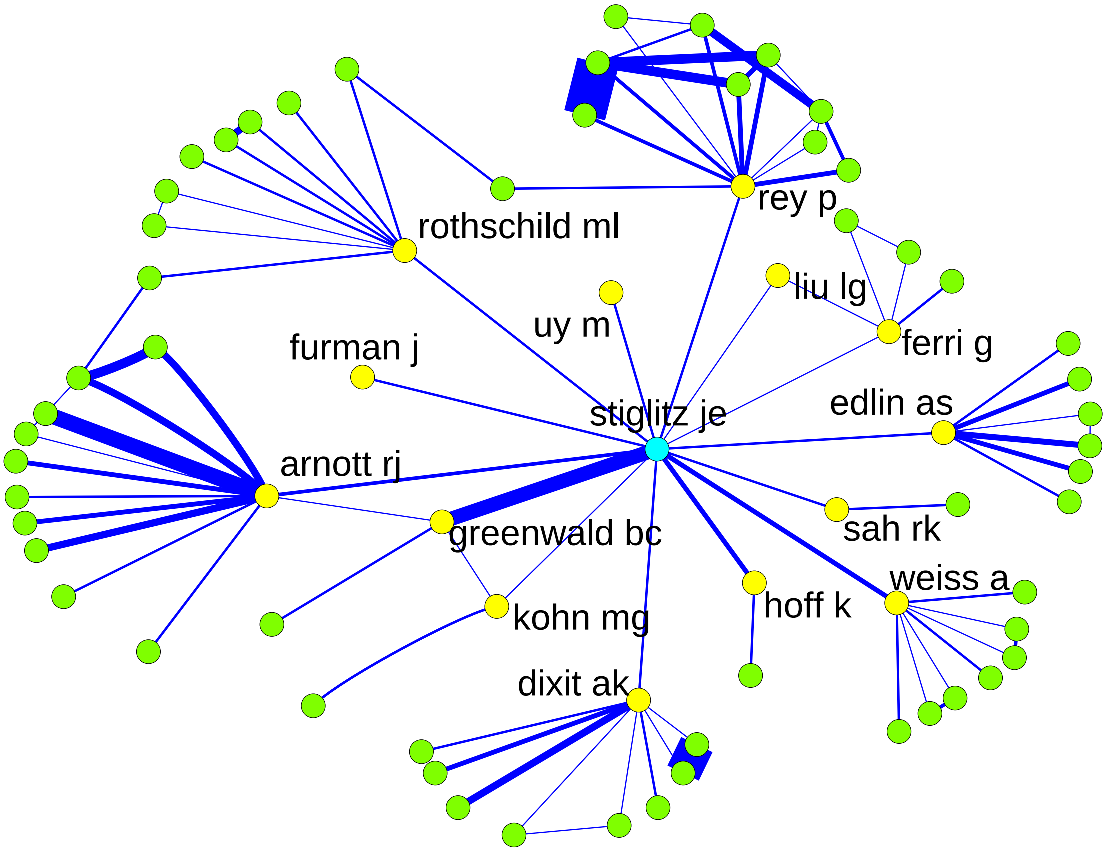


Figure 1: Local network of collaboration of Joseph E. Stiglitz in the 1990s.

Note: The figure shows all authors within distance 2 of J.E. Stiglitz as well as the links between them. The width denotes the strength of a tie. Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit. The figure was created by software program *Pajek*.

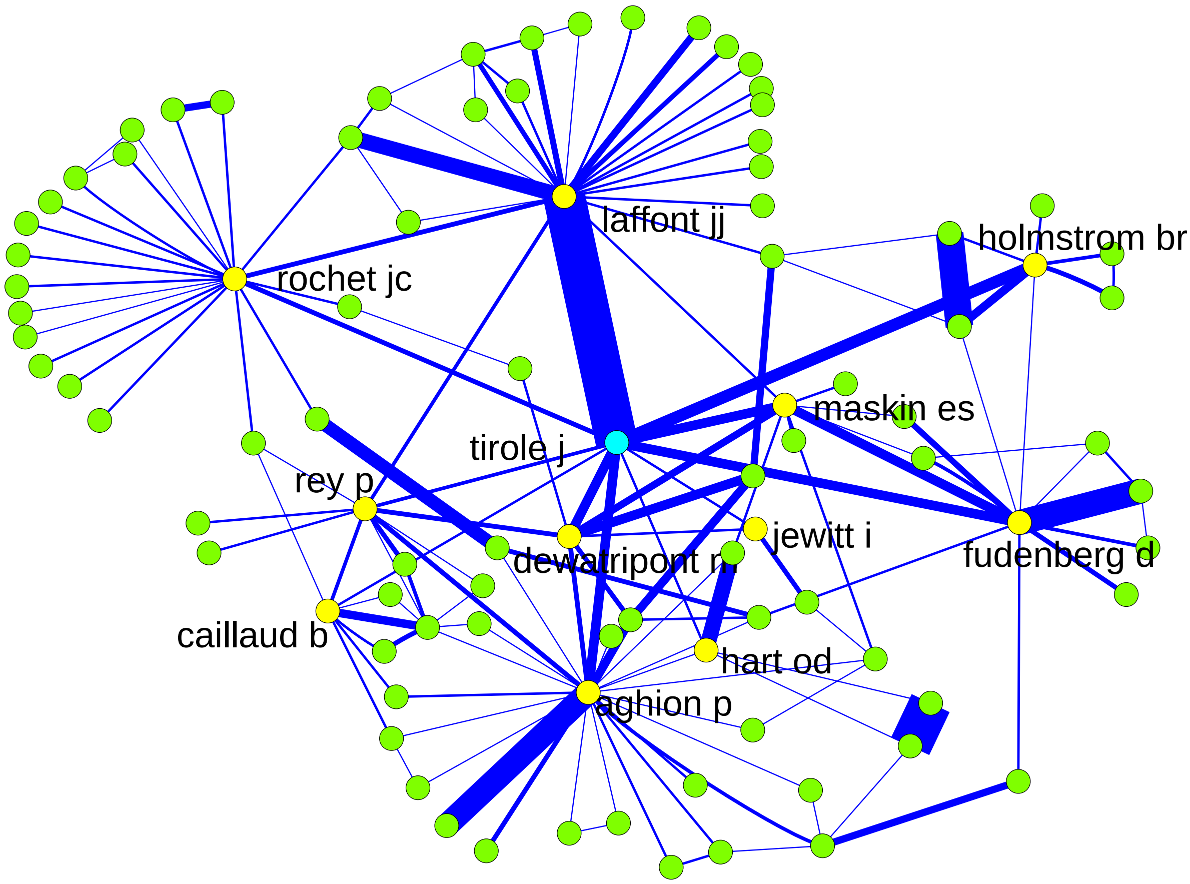


Figure 2: Local network of collaboration of Jean Tirole in the 1990s.

Note: The figure shows all authors within distance 2 of J. Tirole as well as the links between them. The width denotes the strength of a tie. Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit. The figure was created by software program *Pajek*.

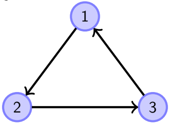
What do we know?

- Networks are prevalent
  - Job contact networks, crime, trade, politics, ...
- Network position and structure matters
  - rich sociology literature
  - Padgett example – Medicis not the wealthiest nor the strongest politically, but the most central
- ``Social'' Networks have special characteristics
  - small worlds, degree distributions...

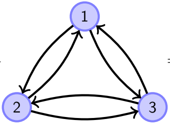
Graphs—1

- We represent a network by a graph (N, g), which consists of a set of nodes $N = \{1, \dots, n\}$ and an $n \times n$ matrix $g = [g_{ij}]_{i,j \in N}$ (referred to as an adjacency matrix), where $g_{ij} \in \{0, 1\}$ represents the availability of an edge from node i to node j.
  - The edge weight $g_{ij} > 0$ can also take on non-binary values, representing the intensity of the interaction, in which case we refer to (N, g) as a weighted graph.
- We refer to a graph as a directed graph (or digraph) if $g_{ij} \neq g_{ji}$ and an undirected graph if $g_{ij} = g_{ji}$ for all $i, j \in N$.

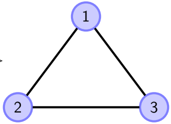
Example 1: $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \Rightarrow$



Example 2: $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \Rightarrow$



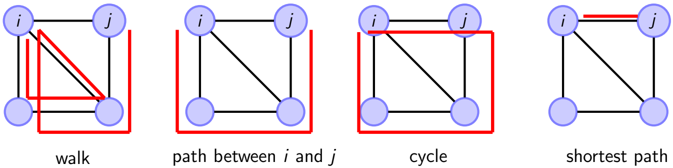
$\Rightarrow$



Walks, Paths, and Cycles—1

- We consider "sequences of edges" to capture indirect interactions.
- For an undirected graph (N, g):
  - A walk is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$.
  - A path between nodes i and j is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$ such that $i_1 = i$ and $i_K = j$, *and each node in the sequence* $i_1, \ldots, i_K$ *is distinct*.
  - A cycle is a path with a final edge to the initial node.
  - A geodesic between nodes i and j is a "shortest path" (i.e., with minimum number of edges) between these nodes.
- A path is a walk where there are no repeated nodes.
- The length of a walk (or a path) is the number of edges on that walk (or path).
- For directed graphs, the same definitions hold with directed edges (in which case we say "a path from node i to node j").

Walks, Paths, and Cycles—2



- *Note:* Under the convention $g_{ii} = 0$, the matrix g^2 tells us number of walks of length 2 between any two nodes:
  - $(g \times g)_{ij} = \sum_k g_{ik} g_{kj}$
  - Similarly, g^k tells us number of walks of length k.

Connectivity and Components

- An undirected graph is connected if every two nodes in the network are connected by some path in the network.
- Components of a graph (or network) are the distinct maximally connected subgraphs.
- A directed graph is
  - connected if the underlying undirected graph is connected (i.e., ignoring the directions of edges).
  - strongly connected if each node can reach every other node by a "directed path".

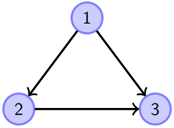
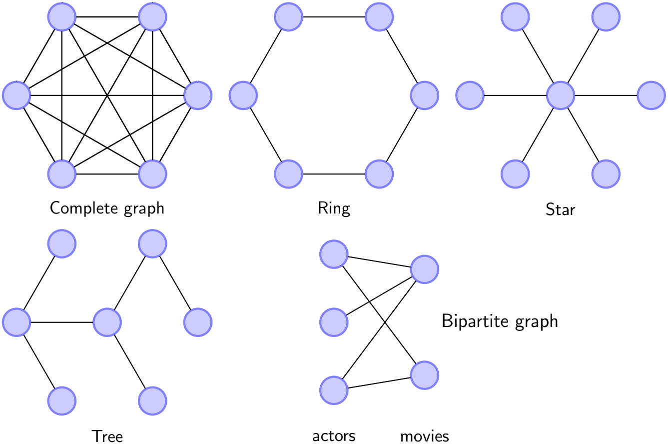


Figure: A directed graph that is connected but not strongly connected

Trees, Stars, Rings, Complete and Bipartite Graphs

- A tree is a connected (undirected) graph with no cycles.
  - A connected graph is a tree if and only if it has $n-1$ edges.
  - In a tree, there is a unique path between any two nodes.



Neighborhood and Degree of a Node

- The neighborhood of node i is the set of nodes that i is connected to.
- For undirected graphs:
  - The degree of node i is the number of edges that involve i (i.e., cardinality of his neighborhood).
- For directed graphs:
  - Node i's in-degree is $\sum_j g_{ji}$.
  - Node i's out-degree is $\sum_j g_{ij}$.

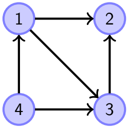


Figure: Node 1 has in-degree 1 and out-degree 2

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] =$

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] = (n-1)p$

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] = (n-1)p$
  - Expected number of edges is $\mathbb{E}[\text{number of edges}] =$

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] = (n-1)p$
  - Expected number of edges is $\mathbb{E}[\text{number of edges}] = \frac{n(n-1)}{2} p$

Betweeness centrality is number of times that a node lies along the shortest path between two others.

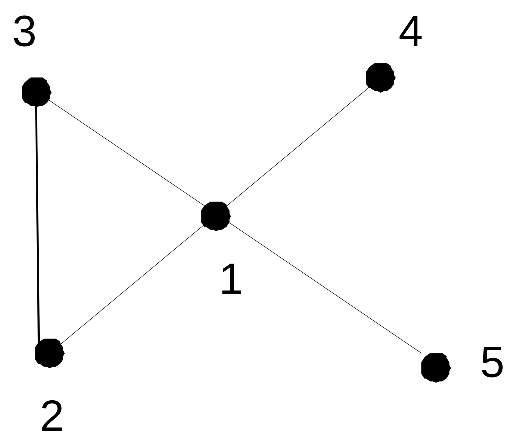
$$B_i = \sum_{j \neq i, k \neq i} \frac{g_{jk}(i)}{g_{jk}}$$

- g_{jk} is the number of geodesic (shortest) paths $d(j,k)$ from j to k,

- $g_{jk}(i)$ is the number of geodesic paths from j to k that pass through i.

Observe that the node that has high betweenness can control the flow of information, acting as a gatekeeper. That node may also serve as a bridge between disparate regions of the network.

Consider the following network:



Adjacency matrix G:

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We have

$$B_1 = \sum_{j \neq i, k \neq i} \frac{g_{jk}(1)}{g_{jk}} = \frac{5}{6} = 0.83333$$

Indeed, there are 6 possible paths that do not include 1 in this network, that is 23, 24, 25, 34 and 45, and only path 23 does not pass through 1.

This means that 83.33% of (shortest) paths in this network pass through 1.

We also have:

$$B_3 = \sum_{j \neq i, k \neq i} \frac{g_{jk}(3)}{g_{jk}} = \frac{0}{6} = 0$$

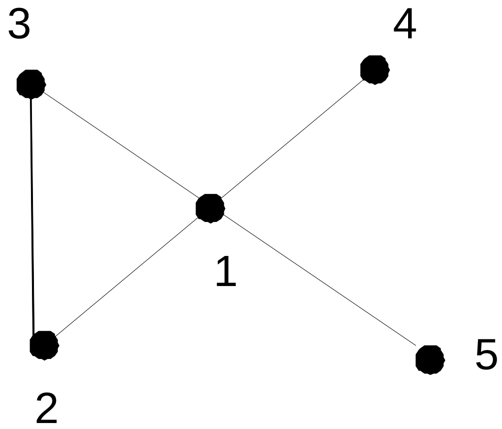
Indeed, there are 6 possible paths that do not include 3 in this network, that is 12, 14, 15, 24, 25 and 45 and none of them passes through 3.

$$B_2 = B_3 = B_4 = B_5 = 0$$

The **closeness centrality** of node i is simply the inverse of the sum of geodesic distances to all other nodes. It is and inverse measure of centrality.

We need to normalize it:

$$CL_i(g) = \frac{n-1}{\sum_{j \neq i} d_{ij}}$$



We have

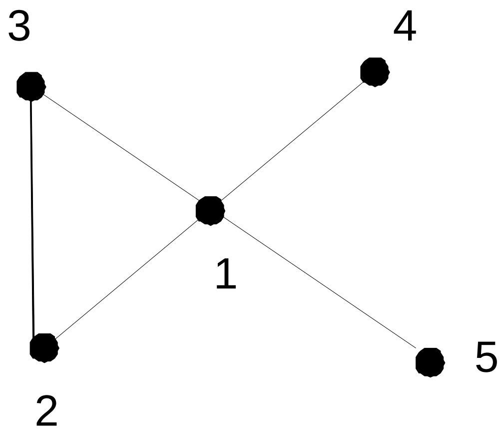
$$CL_1(g) = \frac{4}{1+1+1+1} = 1$$

$$CL_2(g) = CL_3(g) = \frac{4}{1+1+2+2} = \frac{4}{6} = \frac{2}{3} = 0.66667$$

$$CL_4(g) = CL_5(g) = \frac{4}{1+2+2+2} = \frac{4}{7} = \frac{2}{3} = 0.57143$$

The individual clustering for a node i is

$$Cl_i(g) = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered at } i}$$


3
4
1
2
5

The individual clustering for the nodes 2 and 3 are 1, 1, for node 1 is 1/6 and for nodes 4 and 5 is 0, i.e.

$$Cl_2(g) = Cl_3(g) = 1$$

$$Cl_1(g) = \frac{1}{6} = 0.16667$$

$$Cl_4(g) = Cl_5(g) = 0$$

Indeed, for node 1 there are 4 triangles starting at 1 and finishing at 1, i.e. cycles (13-32-21, 14-45-51, 12-25-51, 13-34-41) and 2 triangles that cross 1 (23-34-42 and 25-54-42).

Only one has a clique (13-32-21), i.e. friends of friends are friends, that is 1 is linked to 2 and 1 is linked to 3 implies that 2 and 3 are linked.

If we interpret links in terms of friendships, then, for individual 1, 16.67% of her friends are also friends together.

The average clustering of network g is

$$Cl_i^{avg}(g) = \frac{1}{n} \sum_i Cl_i(g)$$

The overall clustering coeffcient for this network is

$$Cl_i^{avg}(g) = \frac{1}{5} \sum_{i=1}^{5} Cl_i(g) = \frac{3}{8}$$

Eigenvector centrality is a measure of the importance of a node in a network.

It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

Google's PageRank is a variant of the Eigenvector centrality measure.

Another related centrality is Katz-Bonacich centrality.

Let x_i denote the score (i.e. centrality) of the ith node.

Eigenvector centrality: For the ith node, let the centrality score x_i be proportional to the sum of the centrality scores of all nodes which are connected to it. Hence:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{j=n} g_{ij} x_j$$

where $\lambda > 0$ is a constant. In vector/matrix notation:

$$\lambda \mathbf{x} = \mathbf{G} \mathbf{x}$$

Hence we see that $\mathbf{x}$ is an **eigenvector** of the adjacency matrix with eigenvalue λ.

Assuming that we wish the centralities to be non-negative, it can be shown (using the Perron-Frobenius theorem) that λ must be the largest eigenvalue of the adjacency matrix and $\mathbf{x}$ the corresponding eigenvector.

Perron-Frobenius theorem: Since $\mathbf{G}$ of an undirected graph is real and symmetric, all λ_i (eigenvalues) are real. Thus there exists a largest eigenvalue $\lambda^{\mathsf{max}}(\mathbf{G})$ such that $|\lambda_i| \leq \lambda^{\mathsf{max}}$ and there exists an associated nonnegative eigenvector $\mathbf{x}^{\mathsf{max}} \geq 0$ such that $\mathbf{G}\mathbf{x}^{\mathsf{max}} = \lambda^{\mathsf{max}}(\mathbf{G})\mathbf{x}^{\mathsf{max}}$.

The **eigenvector centrality** defined in this way accords each node a centrality that depends both on the number and the quality of its connections:

having a large number of connections still counts for something, but a node with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts.

Eigenvector centrality turns out to be a revealing measure in many situations.

For example, a variant of eigenvector centrality is employed by the well-known Web search engine Google to rank Web pages, and works well in that context.

For example, consider the following star-shaped network with $n = 5$

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Eigenvalues: $2, 0, -2$ so $\lambda^{\mathsf{max}}(\mathbf{G}) = 2$.

Eigenvector corresponding to $\lambda^{\mathsf{max}}(\mathbf{G}) = 2$ is:

$$\mathbf{x}^{\mathsf{max}} = \begin{pmatrix} x_1^{\mathsf{max}} \\ x_2^{\mathsf{max}} \\ x_3^{\mathsf{max}} \\ x_4^{\mathsf{max}} \\ x_5^{\mathsf{max}} \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

The Bonacich network centrality The kth power $\mathbf{G}^k = \mathbf{G}^{(k \ \overset{times}{\cdots})}\mathbf{G}$ of the adjacency matrix $\mathbf{G}$ keeps track of indirect connections in $\mathbf{g}$.

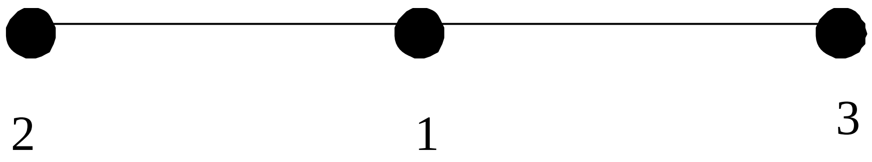
The coefficient $g_{ij}^{[k]}$ in the (i, j) cell of $\mathbf{G}^k$ gives the number of paths of length k in $\mathbf{g}$ between i and j.

Definition 0.1 *Given a vector* $\mathbf{u} \in \mathbb{R}_+^n$*, and* $\phi \geq 0$ *a small enough scalar, we define the vector of Bonacich centralities of parameter* ϕ *in the network* $\mathbf{g}$ *as:*

$$\mathbf{b_u}(\mathbf{g}, \phi) = (\mathbf{I} - \phi\mathbf{G})^{-1}\mathbf{u} = \sum_{p=0}^{+\infty} \phi^p \mathbf{G}^p \mathbf{u}.$$

- $\mathbf{A}$ Adjacency matrix of network $\mathbf{G}$

- Example: Network g with three individuals (star)



- Adjacency matrix

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

For $k \geq 1$

$$\mathbf{A}^{2k} = \begin{bmatrix} 2^k & 0 & 0 \\ 0 & 2^{k-1} & 2^{k-1} \\ 0 & 2^{k-1} & 2^{k-1} \end{bmatrix} \qquad \text{and} \qquad \mathbf{A}^{2k+1} = \begin{bmatrix} 0 & 2^k & 2^k \\ 2^k & 0 & 0 \\ 2^k & 0 & 0 \end{bmatrix}$$

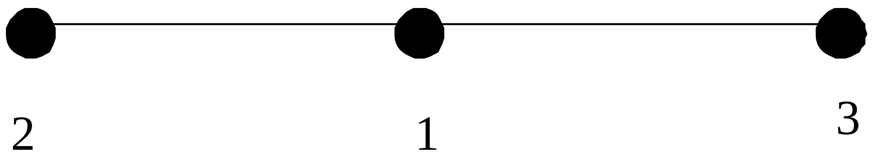
For example

$$\mathbf{A}^3 = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix}$$

$\mathbf{A}^3$: two paths of length three between 1 and 2: $12 \rightarrow 21 \rightarrow 12$ and $12 \rightarrow 23 \rightarrow 32$.

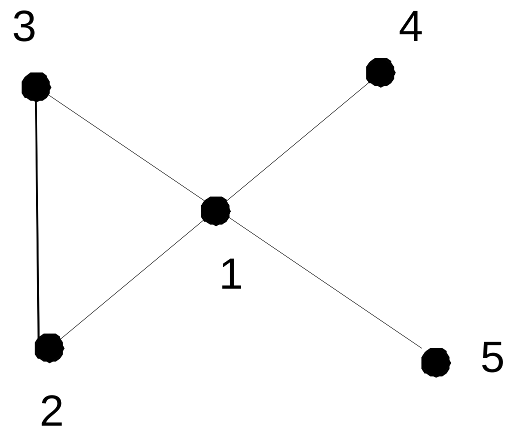
But no path of length three from i to i.

- Example: Network g with three individuals (star)



When λ is small enough (i.e. $\lambda < 1/2^{0.5}$ since $2^{0.5}$ largest eigenvalue), then the vector of Bonacich network centralities is:

$$\mathbf{b}(G,\lambda) = \begin{bmatrix} b_1(G,\lambda) \\ b_2(G,\lambda) \\ b_2(G,\lambda) \end{bmatrix} = \frac{1}{\left(1-2\lambda^2\right)} \begin{bmatrix} 1+2\lambda \\ 1+\lambda \\ 1+\lambda \end{bmatrix}$$


3
4
1
5
2

Determine **Katz-Bonacich centrality** by accounting for only path of length 1 and 2.

Define $m_{ij}(\mathbf{g}, a) = \sum_{k=0}^{+\infty} \phi^k g_{ij}^{[k]}$. This counts the number of paths in g starting from i and ending at j, where paths of length k are weighted by ϕ^k. Observe that $g_{ii}^{[k]} = 0$ but not necessarily $g_{ii}^{[3]}$.

Bonacich index is (when j can be equal to i):

$$\begin{aligned} b_i(\mathbf{g}, \phi) &= m_{ii}(\mathbf{g}, \phi) + \sum_{j \neq i} m_{ij}(\mathbf{g}, \phi) \\ &= \sum_{k=0}^{+\infty} \phi^k g_{ii}^{[k]} + \sum_{k=0, j \neq i}^{+\infty} \phi^k g_{ij}^{[k]} \end{aligned}$$

If ϕ is small enough (less than 1 over the largest eigenvalue of $\mathbf{G}$), then

$$\mathbf{b}(\mathbf{g}, \phi) = [\mathbf{I} - \phi \mathbf{G}]^{-1} \mathbf{1}$$

We have

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

which implies that

$$\mathbf{G1} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

We have:

$$\mathbf{G}^2 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}^2 = \begin{pmatrix} 4 & 1 & 1.0 & 0 & 0 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and thus

$$\mathbf{G}^2\mathbf{1} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}^2 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 6 \\ 4 \\ 4 \end{pmatrix}$$

Thus, individual 1's Bonacich centrality (up to length 2) is (using the first row of $\mathbf{G1}$ and $\mathbf{G}^2\mathbf{1}$):

$$b_1(\mathbf{g}, \phi) = 4\phi + 6\phi^2$$

whereas individual 3's Bonacich centrality (up to length 2) is (using the third row of $\mathbf{G1}$ and $\mathbf{G}^2\mathbf{1}$):

$$b_3(\mathbf{g}, \phi) = 2\phi + 6\phi^2$$

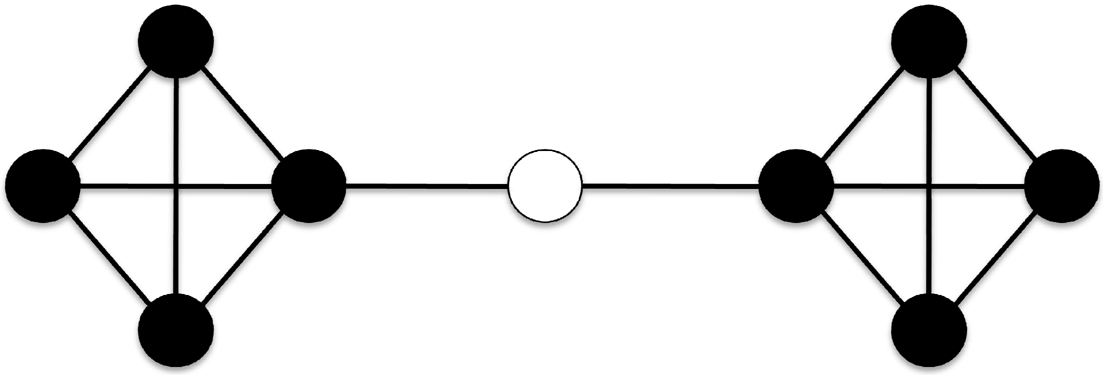
It is easy verified that $b_1(\mathbf{g}, \phi) > b_3(\mathbf{g}, \phi)$.

Different *centrality measures* to capture the prominence of actors inside a network.

Degree centrality: counts the number of connections an agent has.

Bonacich centrality: gives to any individual a particular numerical value for each of his/her direct connection. Then, give a smaller value to any connection at distance two and an even smaller value to any connection at distance three; etc. When adding up all these values, we end up with a new numerical value that is now capturing both direct and indirect connections of any order.

Betweenness centrality: calculates the relative number of indirect connections (or shortest paths) in which the actor into consideration is involved in with respect to the total number of paths in the network.



Agent in the middle: **Lowest degree** centrality, **highest betweenness** centrality.

Bonacich centrality: Depends on the value of the discount factor.

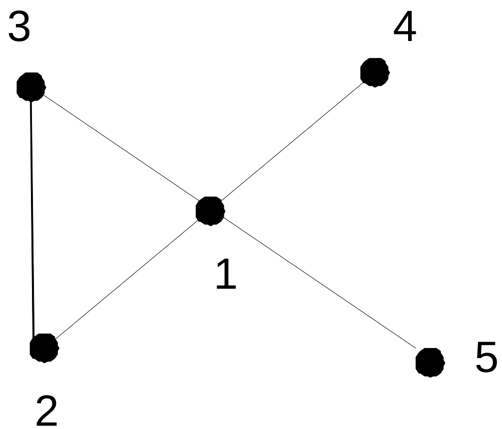
For small discount factors (i.e. indirect links give less benefits), this agent is the less central one while for high levels of discount (i.e. direct links are weighted less), this agent is the most central.

Diameter and **average path length** of a network g:

$$DIAM_g = \max_{i,j} d(i,j)$$

where $d(i,j)$ denotes the length of the shortest path (or geodesic) between node i and j.

$$APL_g = \frac{\sum_{i \geq j} d(i,j)}{n(n-1)/2}$$


3
4
1
5
2

$$DIAM_g = \max_{i,j} d(i,j) = 2$$

where $d(i,j)$ denotes the length of the shortest path (or geodesic) between node i and j.

$$APL_g = \frac{\sum_{i \geq j} d(i,j)}{n(n-1)/2} = \frac{1+1+1+1+1+2+2+2+2+2}{5 \times 4/2} = 1.5$$

Degree Distributions

- The degree distribution, $P(d)$, of a network is a description of relative frequencies of nodes that have different degrees d.
  - For a given graph: $P(d)$ is a histogram, i.e., $P(d)$ is the fraction of nodes with degree d.
  - For a random graph model: $P(d)$ is a probability distribution.
- Two types of degree distributions:
  - $P(d) \leq c\, e^{-\alpha d}$, for some $\alpha > 0$ and $c > 0$: The tail of the distribution **falls off faster than an exponential**, i.e., large degrees are unlikely.
  - $P(d) = c\, d^{-\gamma}$, for some $\gamma > 0$ and $c > 0$: Power-law distribution: The tail of the distribution is **fat**, i.e., there tend to be many more nodes with very large degrees.
    - Appear in a wide variety of settings including networks describing incomes, city populations, WWW, and the Internet
    - Also known as a scale-free distribution: a distribution that is unchanged (within a multiplicative factor) under a rescaling of the variable
    - Appear linear on a $\log - \log$ plot
- What is the degree distribution of the Erdös-Renyi model?

Consider a set of n nodes, every pair of nodes being connected with probability p and then independently consider each possible link.

Degree distribution $P(d)$ of the Erdos-Renyi model

$$P(\text{deg} = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$$

Networks in Labor Markets

- Myers and Shultz (1951)- textile workers:
  - 62% first job from contact
  - 23% by direct application
  - 15% by agency, ads, etc.
- Rees and Shultz (1970) – Chicago market:
  - Typist 37.3%
  - Accountant 23.5%
  - Material handler 73.8%
  - Janitor 65.5%, Electrician 57.4%…
- Granovetter (1974), Corcoran et al. (1980), Topa (2001), Ioannides and Loury (2004) ...

Facts about the use of networks in labor markets

- Granovetter (1973) –use of contacts across job types
  - 44 percent technical jobs
  - 56 percent professional jobs
  - 65 percent managerial jobs
- Corcoran *et al.* (1980) –contacts across gender and race
  - 52 percent white males; 47.1 percent white females
  - 58.5 percent black males; 43 percent black females

Search Channel (%)	U.S.	France	Italy	Spain
Social Contacts	52.0	33.1	24.1	39.4
Direct Applications	–	25.9	33.2	35.0
Newspapers	9.4	9.8	3.9	5.3
Employment Agencies	5.8	9.8	7.4	7.4
Other	32.8	21.4	31.4	12.9

Other Settings

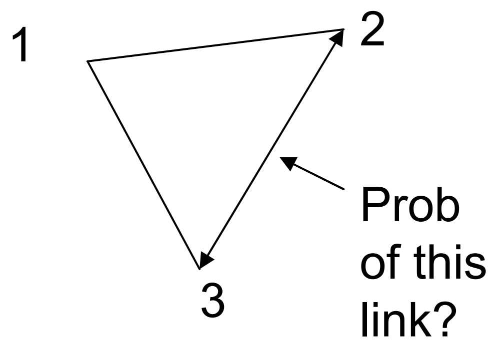
- Networks and social interactions in crime:
  - Reiss (1980, 1988) - 2/3 of criminals commit crimes with others
  - Glaeser, Sacerdote and Scheinkman (1996) - social interaction important in petty crime, among youths, and in areas with less intact households
- Networks and Markets
  - Uzzi (1996) - relation specific knowledge critical in garment industry
  - Weisbuch, Kirman, Herreiner (2000) – repeated interactions in Marseille fish markets
- Social Insurance
  - Fafchamps and Lund (2000) – risk-sharing in rural Phillipines
  - De Weerdt (200
- Sociology literature – interlocking directorates, aids transmission, language, ...

Stylized Facts: Small diameter

- Milgram (1967) letter experiments
  - median 5 for the 25% that made it
- Actors in same movie (Kevin Bacon Oracle)
  - Watts and Strogatz (1998) – mean 3.7
- Co-Authorship studies
  - Grossman (1999) Math mean 7.6, max 27,
  - Newman (2001) Physics mean 5.9, max 20
  - Goyal et al (2004) Economics mean 9.5, max 29
- WWW
  - Adamic, Pitkow (1999) – mean 3.1 (85.4% possible of 50M pages)

High Clustering Coefficients - distinguishes ``social’’ networks

- Watts and Strogatz (1998)
  - .79 for movie acting
- Newman (2001) co-authorship
  - .496 CS, .43 physics, .15 math, .07 biomed
- Adamic (1999)
  - .11 for web links (versus .0002 for random graph of same size and avg degree)



Distribution of links per node: Power Laws

- Plot of log(frequency) versus log(degree) is ``approximately'' linear in upper tail

- prob(degree) = c degree^{-a}
  - log[prob(degree)] = log[c] – a log[degree]

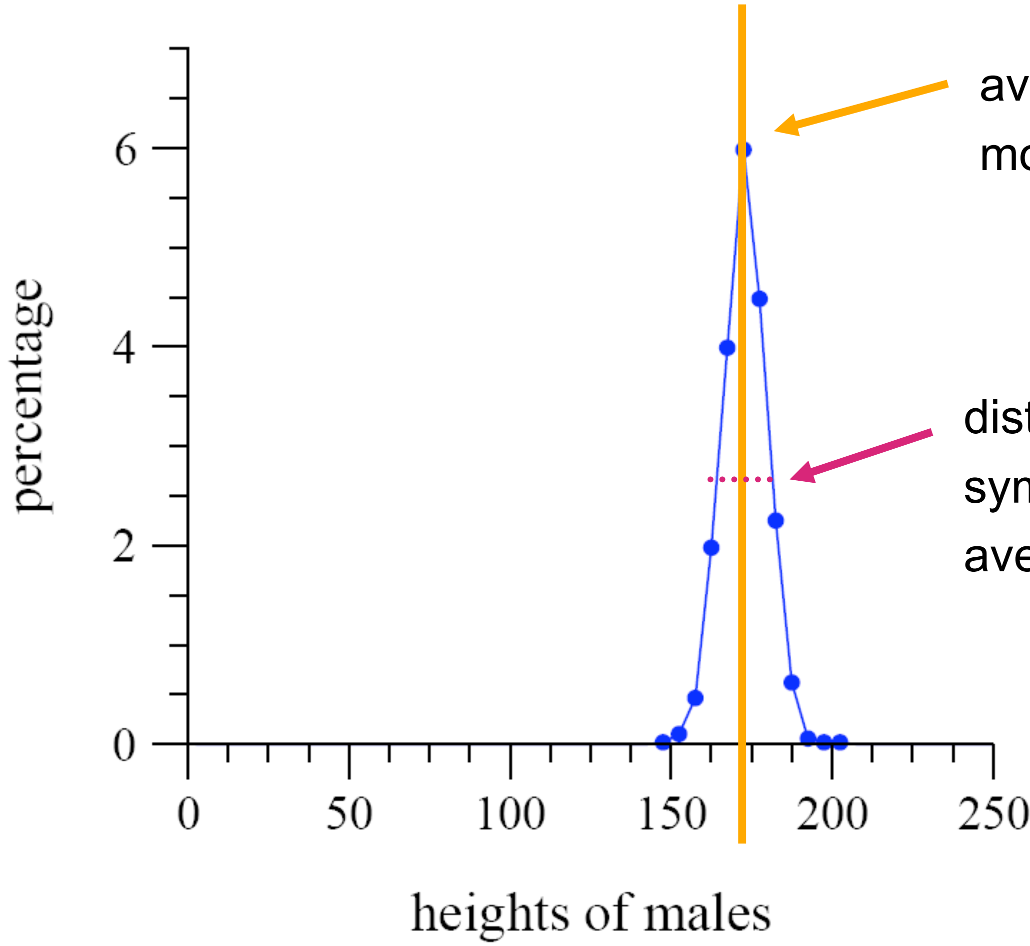
- Fat tails compared to random network

- Related to other settings: Pareto (1896), Yule (1925), Zipf (1949), Simon (1955),

What is a heavy tailed-distribution?

- Right skew
  - normal distribution (not heavy tailed)
    - e.g. heights of human males: centered around 180cm (5’11’’)
  - Zipf’s or power-law distribution (heavy tailed)
    - e.g. city population sizes: NYC 8 million, but many, many small towns
- High ratio of max to min
  - human heights
    - tallest man: 272cm (8’11”), shortest man: (1’10”) *ratio: 4.8* from the Guinness Book of world records
  - city sizes
    - NYC: pop. 8 million, Duffield, Virginia pop. 52, *ratio: 150,000*

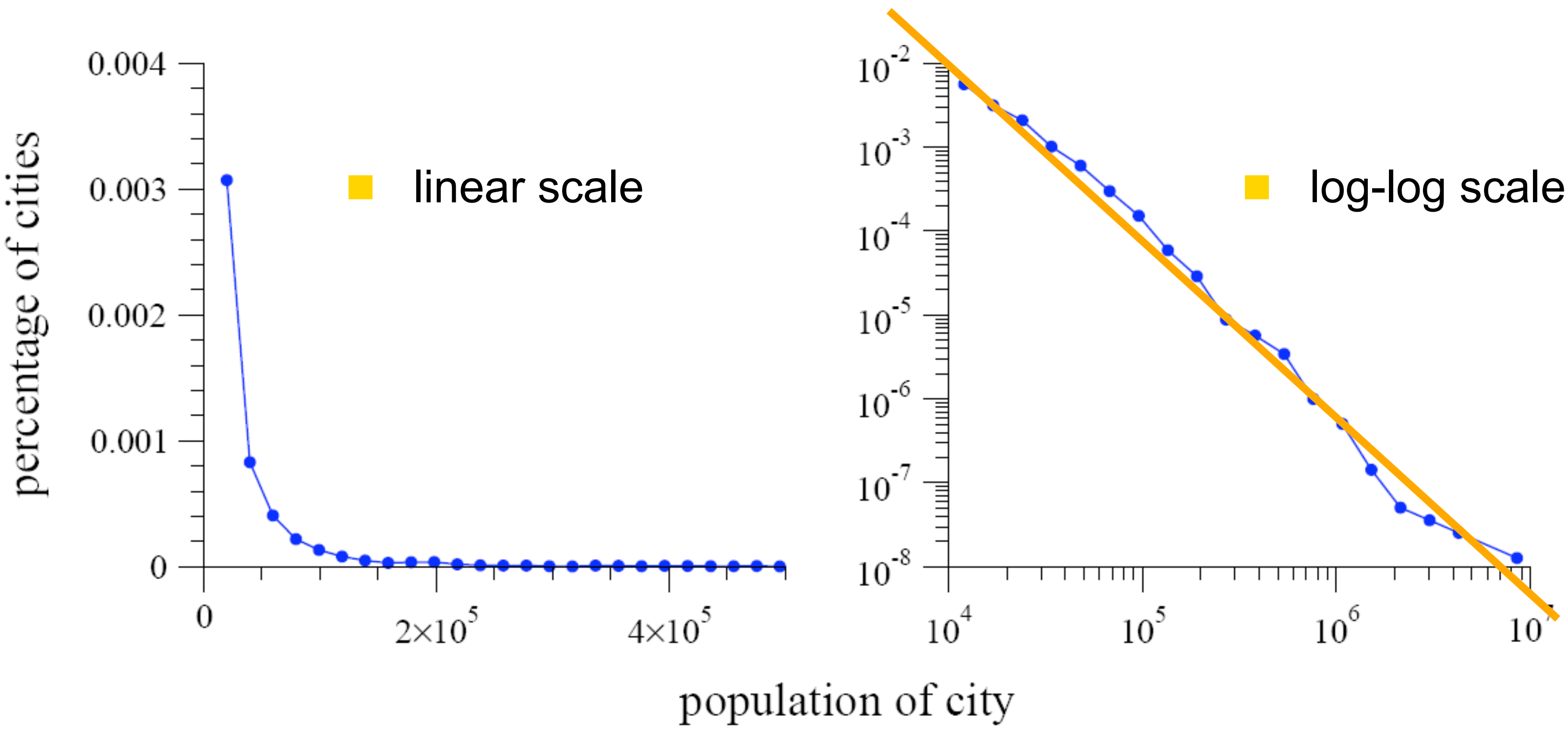
Normal (also called Gaussian) distribution of human heights



average value close to most typical

distribution close to symmetric around average value

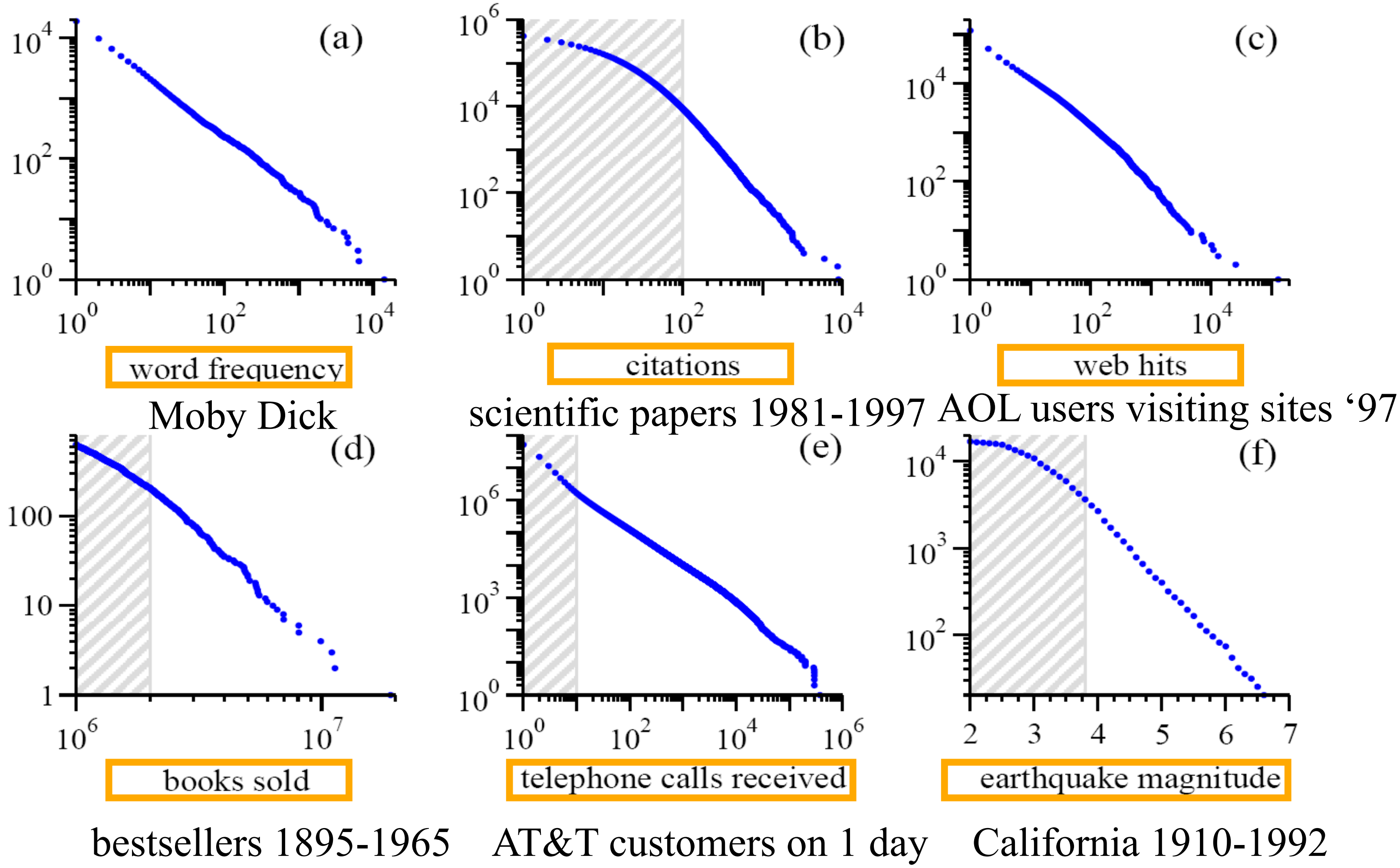
Power-law distribution



- high skew (asymmetry)
- straight line on a log-log plot

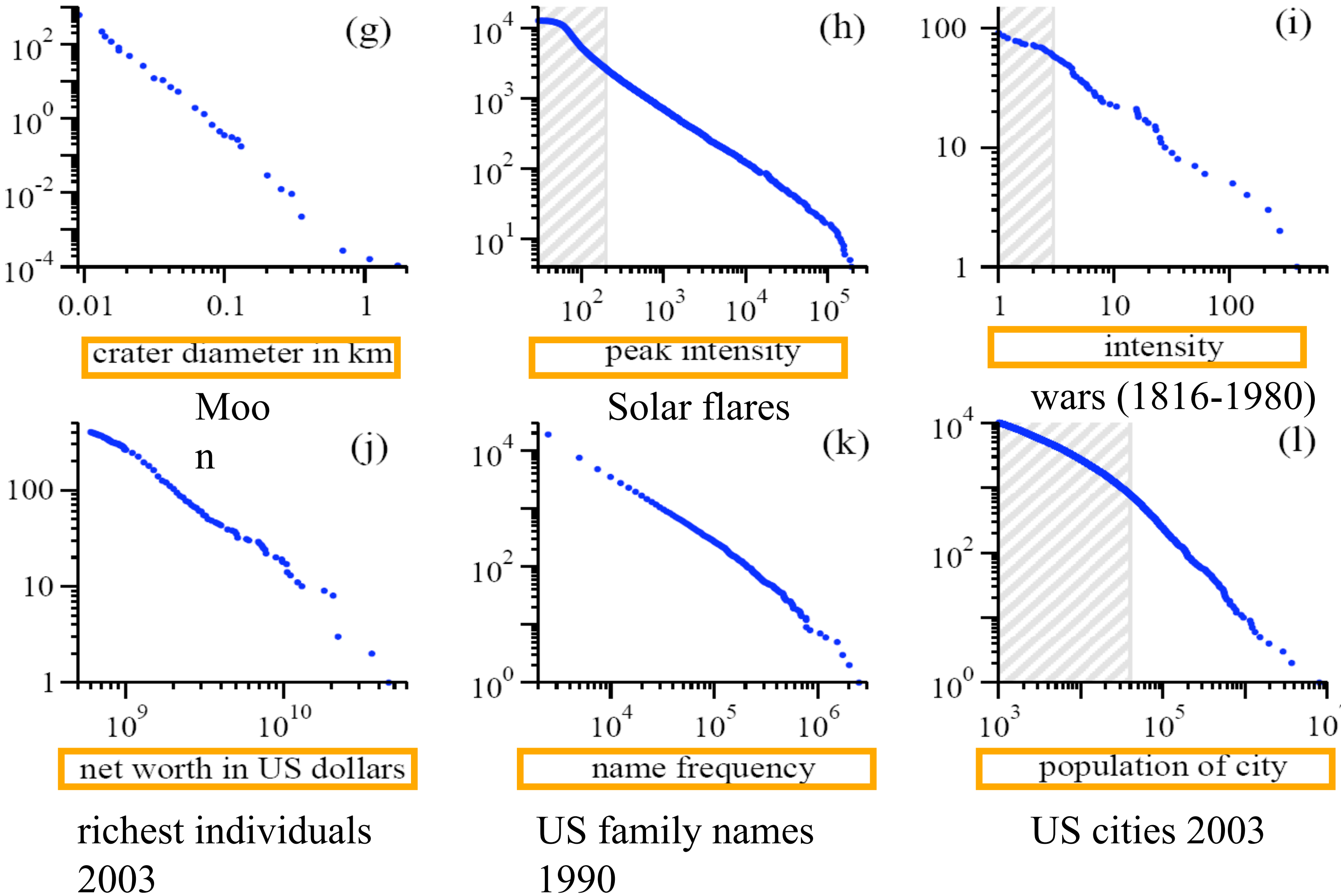
Power laws are seemingly everywhere

note: these are cumulative distributions, more about this in a bit…



Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

Yet more power laws



Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

Power law distribution

- Straight line on a log-log plot

$$\ln(p(x)) = c - \alpha \ln(x)$$

- Exponentiate both sides to get that *p(x)*, the probability of observing an item of size 'x' is given by

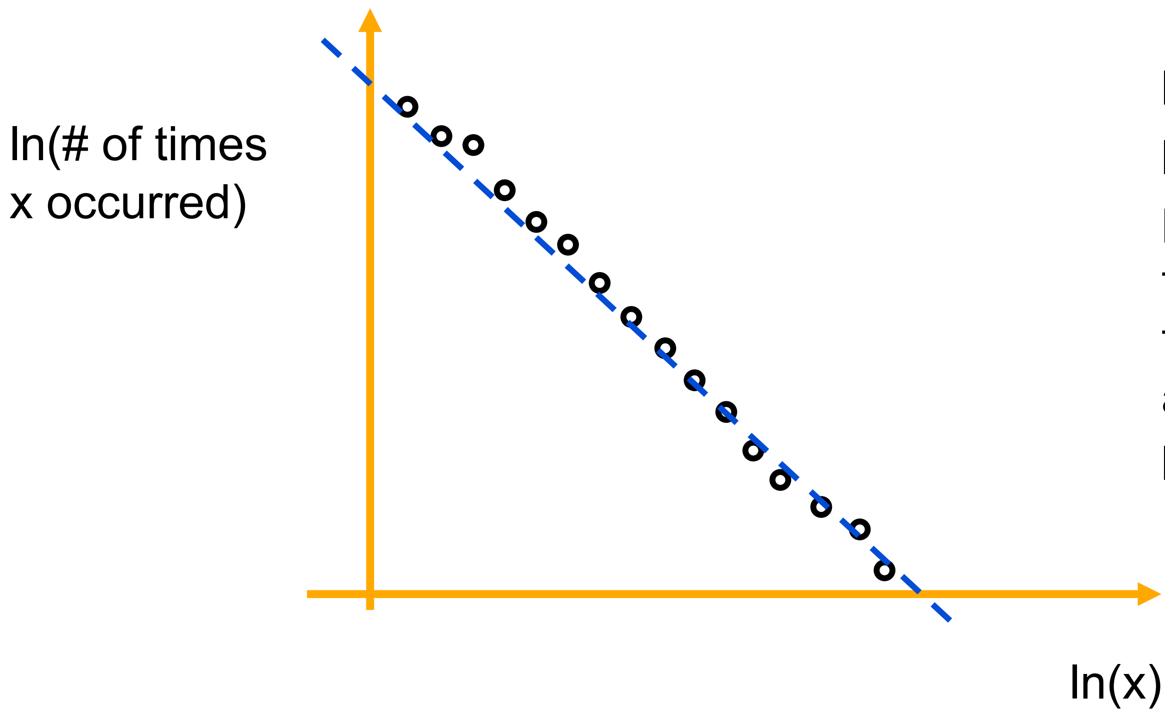
$$p(x) = Cx^{-\alpha}$$

normalization constant (probabilities over all *x* must sum to 1)

power law exponent α

Fitting power-law distributions

- Most common and not very accurate method:
  - Bin the different values of x and create a frequency histogram



ln(*x*) is the natural logarithm of x,

but any other base of the logarithm will give the same exponent of a because

$\log_{10}(x) = \ln(x)/\ln(10)$

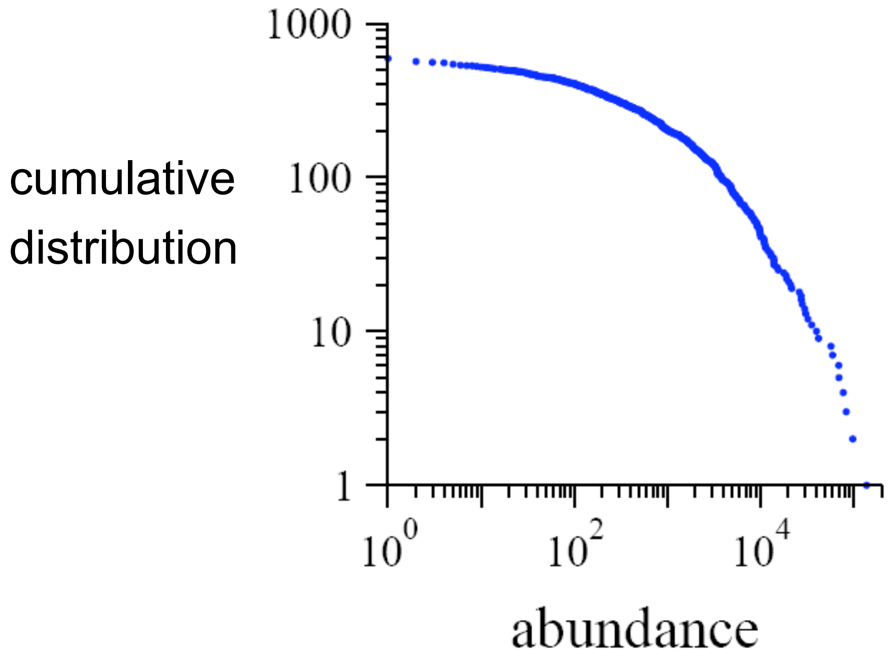
x can represent various quantities, the indegree of a node, the magnitude of an earthquake, the frequency of a word in text

Many real world networks are power law

	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
internet	2.5
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

Hey, not everything is a power law

- number of sightings of 591 bird species in the North American Bird survey in 2003.



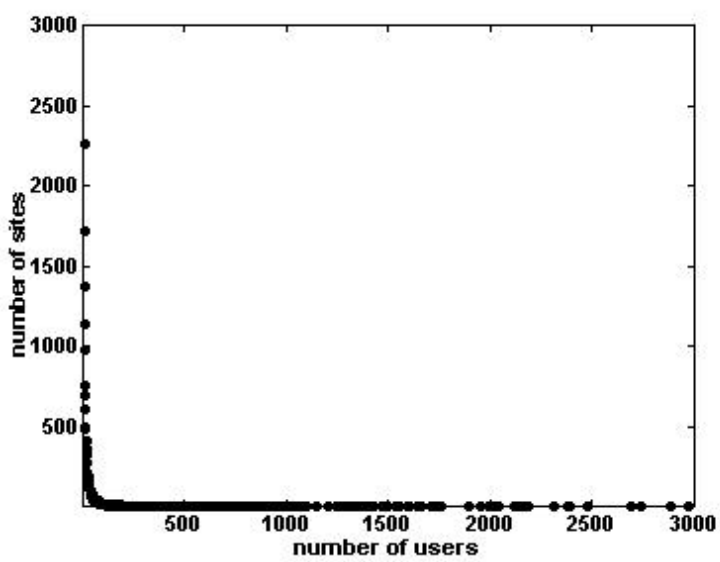
- another example:
  - size of wildfires (in acres)

Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

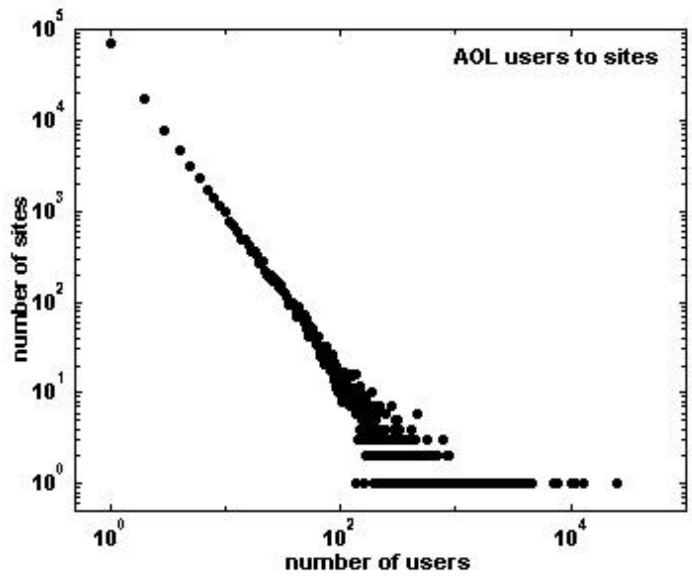
Not every network is power law distributed

- reciprocal, frequent email communication
- power grid
- Roget's thesaurus
- company directors…

Example on a real data set: number of AOL visitors to different websites back in 1997



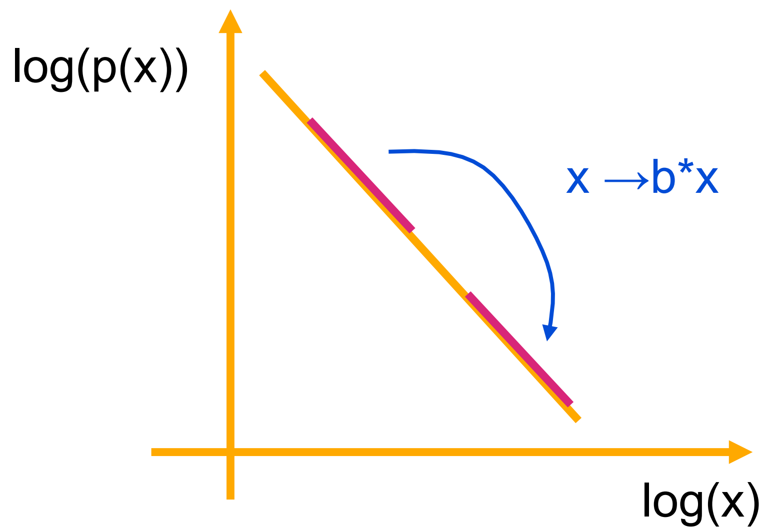
simple binning on a linear scale



simple binning on a log-log scale

What does it mean to be scale free?

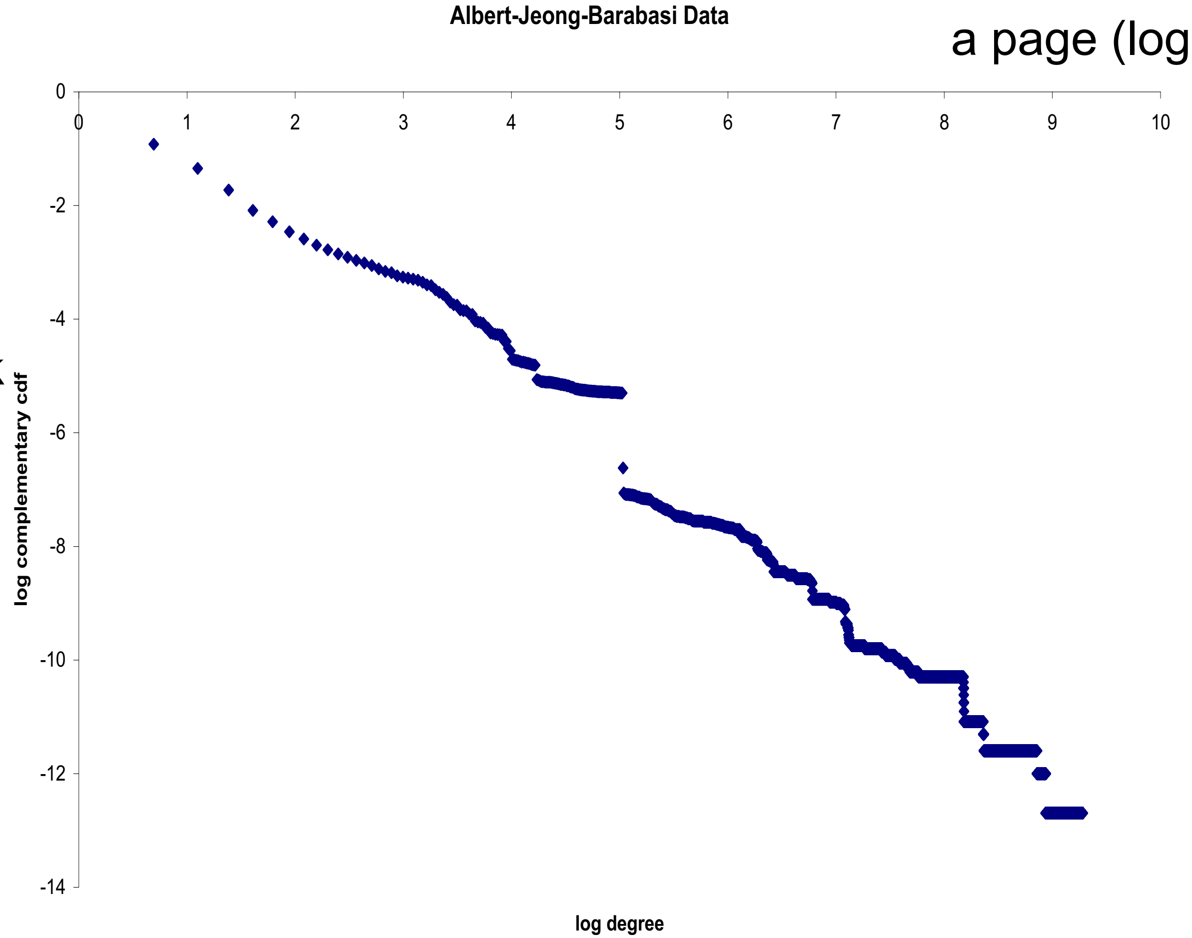
- A power law looks the same no mater what scale we look at it on (2 to 50 or 200 to 5000)
- Only true of a power-law distribution!
- $p(bx) = g(b)\ p(x)$ – shape of the distribution is unchanged except for a multiplicative constant
- $p(bx) = (bx)^{-\alpha} = b^{-\alpha}\ x^{-\alpha}$



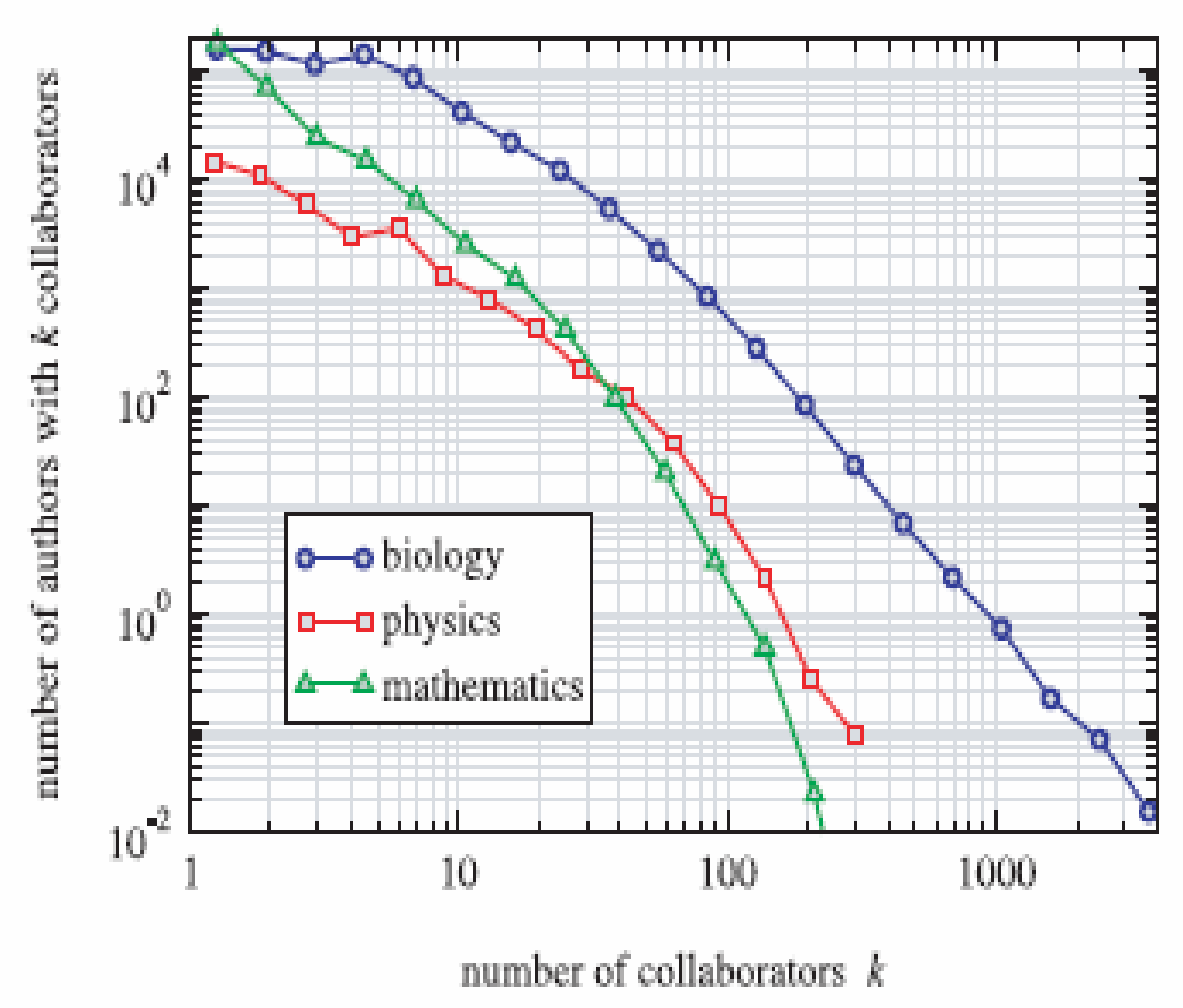
Degree – ND www Albert, Jeong, Barabasi (1999)

number of links to a page (log scale)

fraction of pages with more than k links (log)



Co-Authorship Data, Newman and Grossman



Liljeros et al. (*Nature* 2001): 1996 Swedish survey of sexual behaviour.

Evidence that the distribution of sexual partners follows a power law distribution

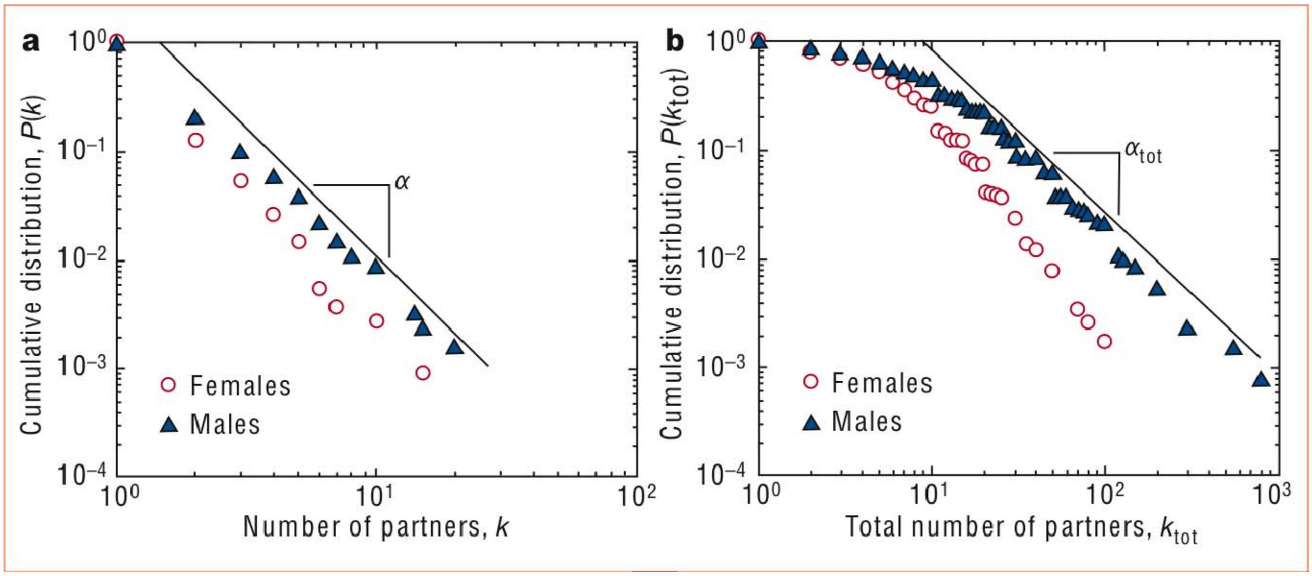


Figure 2 Scale-free distribution of the number of sexual partners for females and males. **a,** Distribution of number of partners, k, in the previous 12 months. Note the larger average number of partners for male respondents: this difference may be due to 'measurement bias' — social expectations may lead males to inflate their reported number of sexual partners. Note that the distributions are both linear, indicating scale-free power-law behaviour. Moreover, the two curves are roughly parallel, indicating similar scaling exponents. For females, $\alpha = 2.54 \pm 0.2$ in the range $k > 4$, and for males, $\alpha = 2.31 \pm 0.2$ in the range $k > 5$. **b,** Distribution of the total number of partners k_{tot} over respondents' entire lifetimes. For females, $\alpha_{tot} = 2.1 \pm 0.3$ in the range $k_{tot} > 20$, and for males, $\alpha_{tot} = 1.6 \pm 0.3$ in the range $20 < k_{tot} < 400$. Estimates for females and males agree within statistical uncertainty.

Three Key Questions:

- How does network structure affect interaction and behavior?

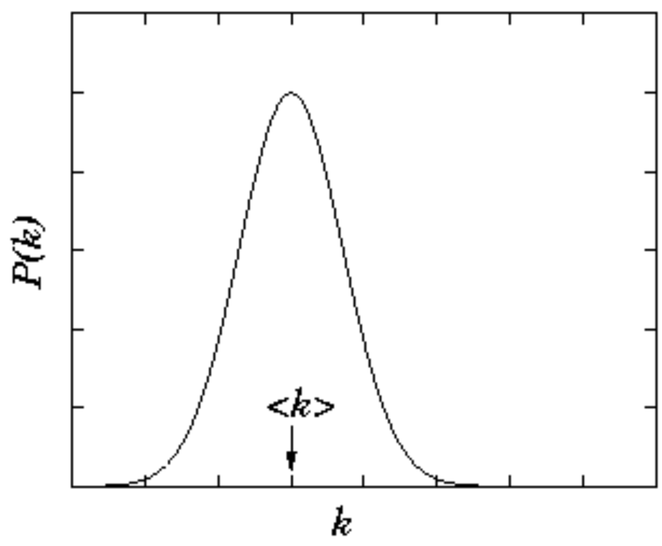
- Which networks form?
  - Game theoretic reasoning
  - dynamic random models

- When do efficient networks form?
  - Intervention - design incentives?

modeling networks: random networks

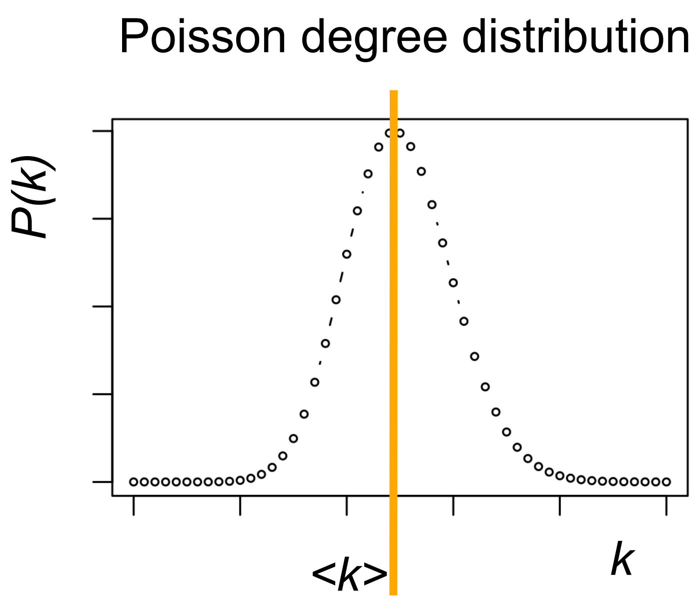
- Nodes connected at random
- Number of edges incident on each node is Poisson distributed

Poisson distribution



Simplest random network

- Erdos-Renyi random graph: each pair of nodes is equally likely to be connected, with probability p.
- p = 2*E/N/(N-1)
- Poisson degree distribution is narrowly distributed around <k> = p*(N-1)



Random graph model

- The degree distribution is given by
  - coinflips to see how many people you'll be connected to, one coin flip per each of the (n-1) other nodes
  - probability *p,* of connecting

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

$$p_k = \frac{z^k e^{-z}}{k!}$$

Binomial

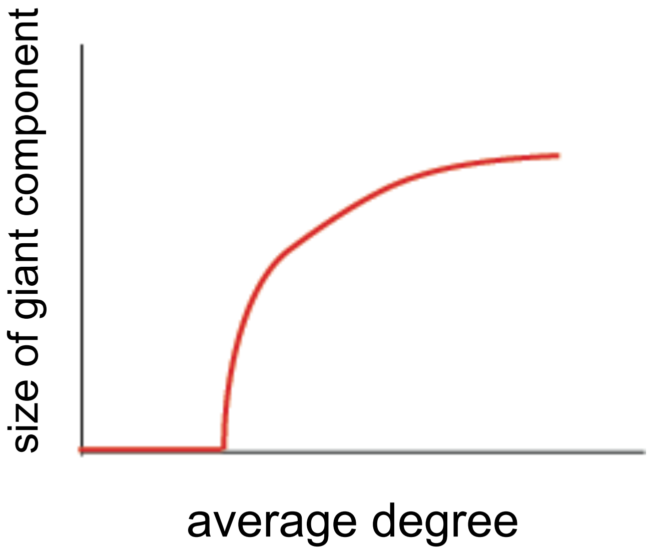
↓ limit *p* small

Poisson

↓ limit large *n*

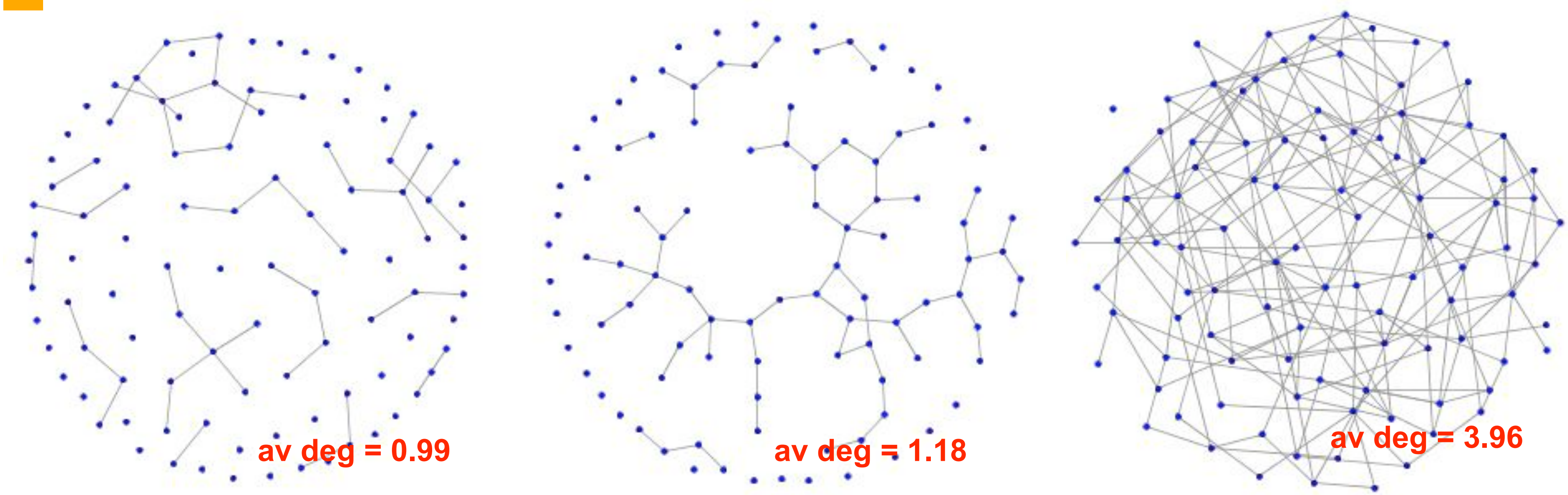
Normal

Percolation threshold in Erdos-Renyi Graphs

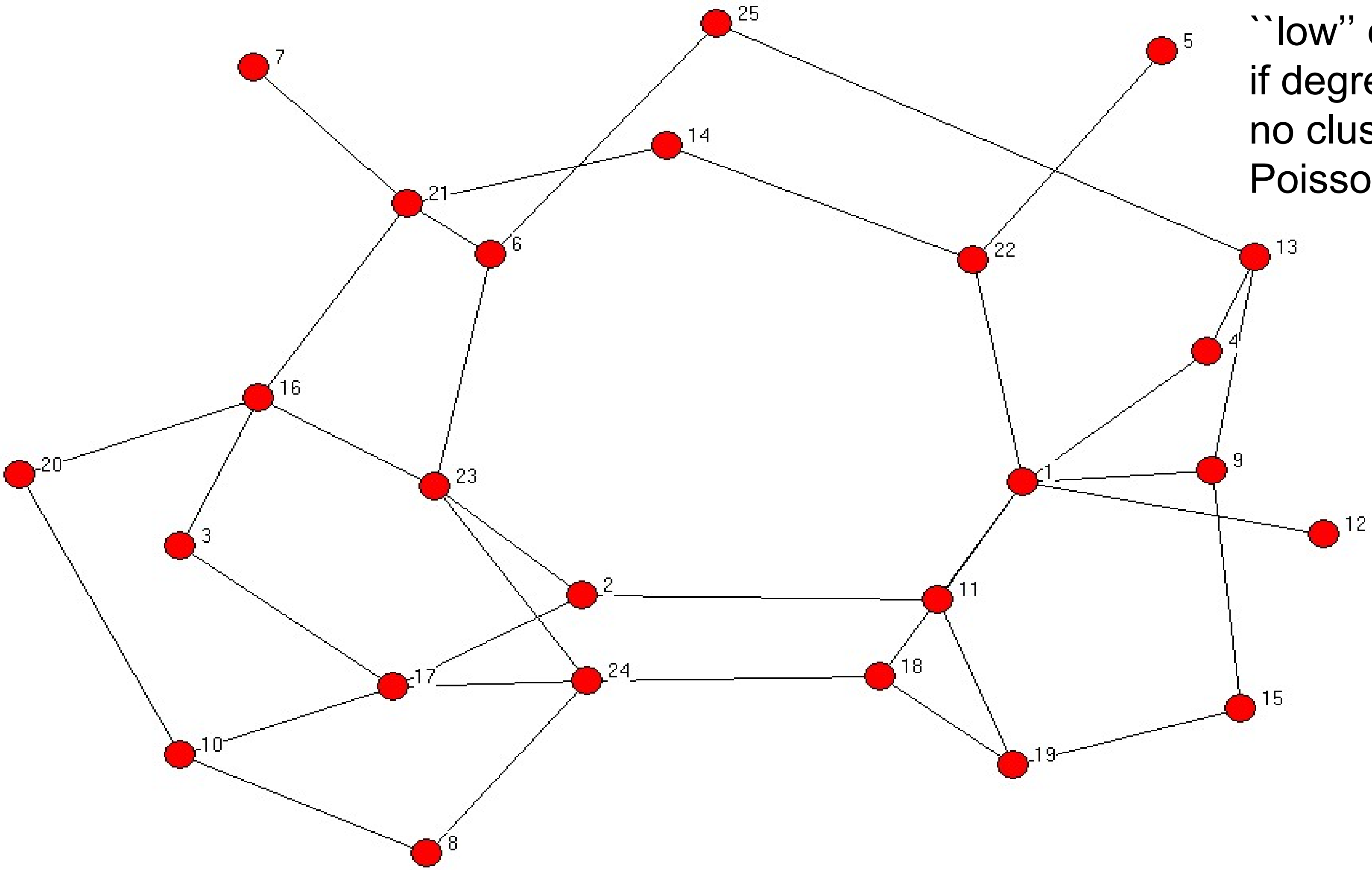


Percolation theshold: how many edges need to be added before the giant component appears?

As the average degree increases to z = 1, a giant component suddenly appears



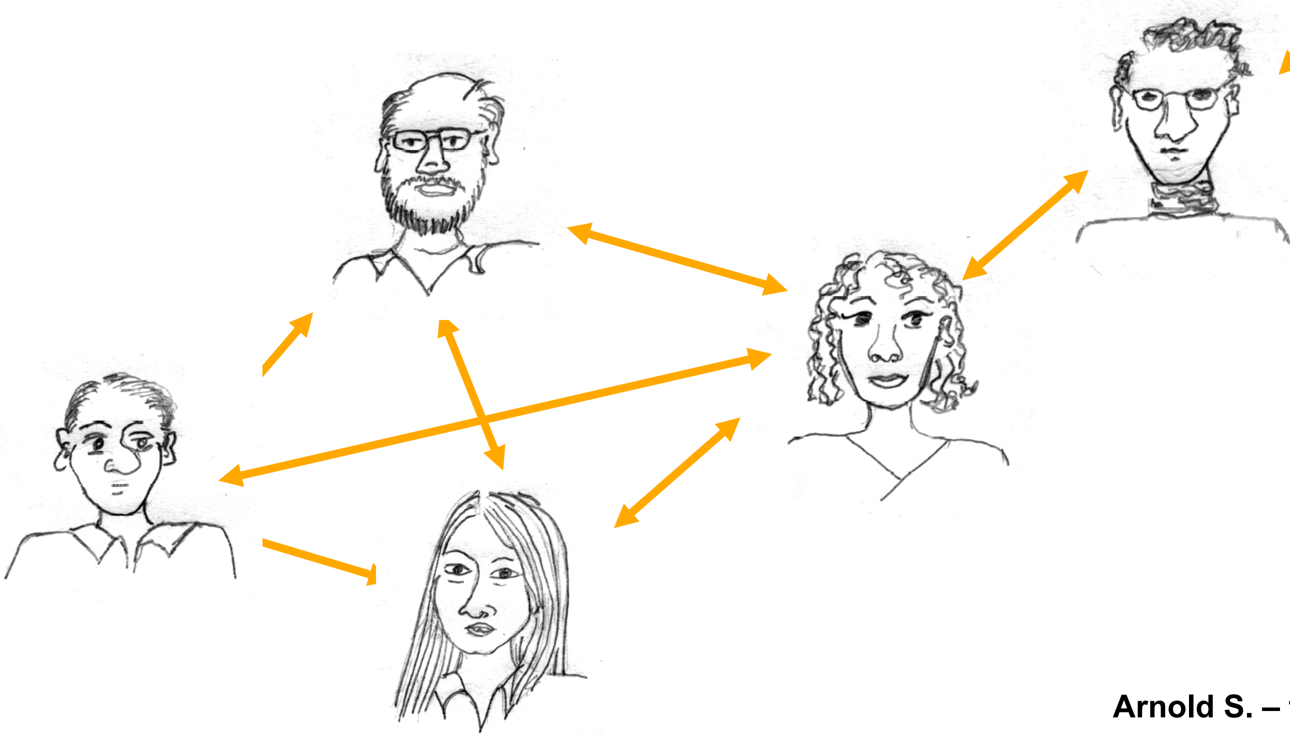
Random Graphs: Bernoulli (Erdos and Renyi (1960))



``low'' diameter if degree is high, no clustering, Poisson degree

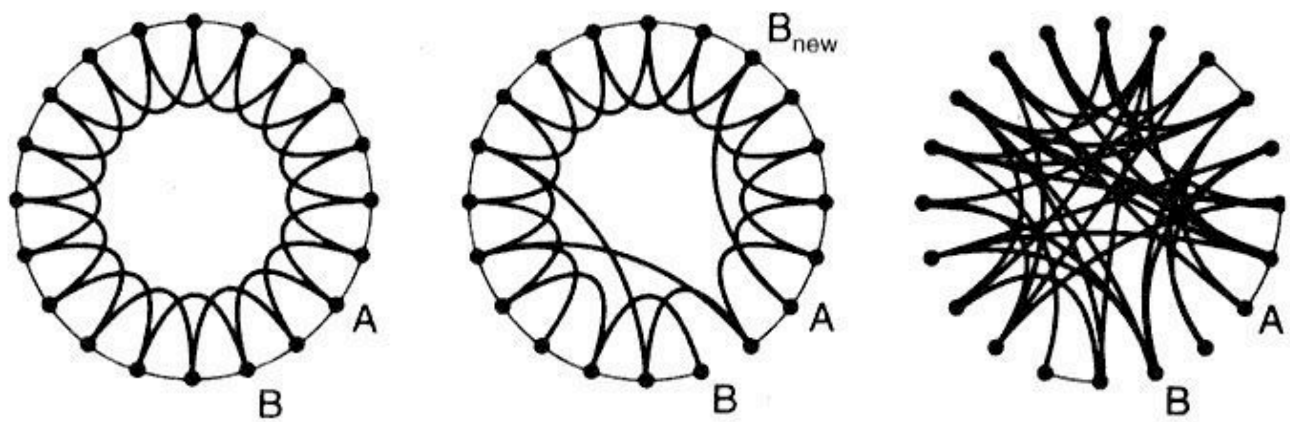
modeling networks: small worlds

- Small worlds
  - a friend of a friend is also frequently a friend
  - but only six hops separate any two people in the world



Small world models

- Duncan Watts and Steven Strogatz
  - a few random links in an otherwise structured graph make the network a small world: the average shortest path is short



regular lattice:
my friend's friend is always my friend

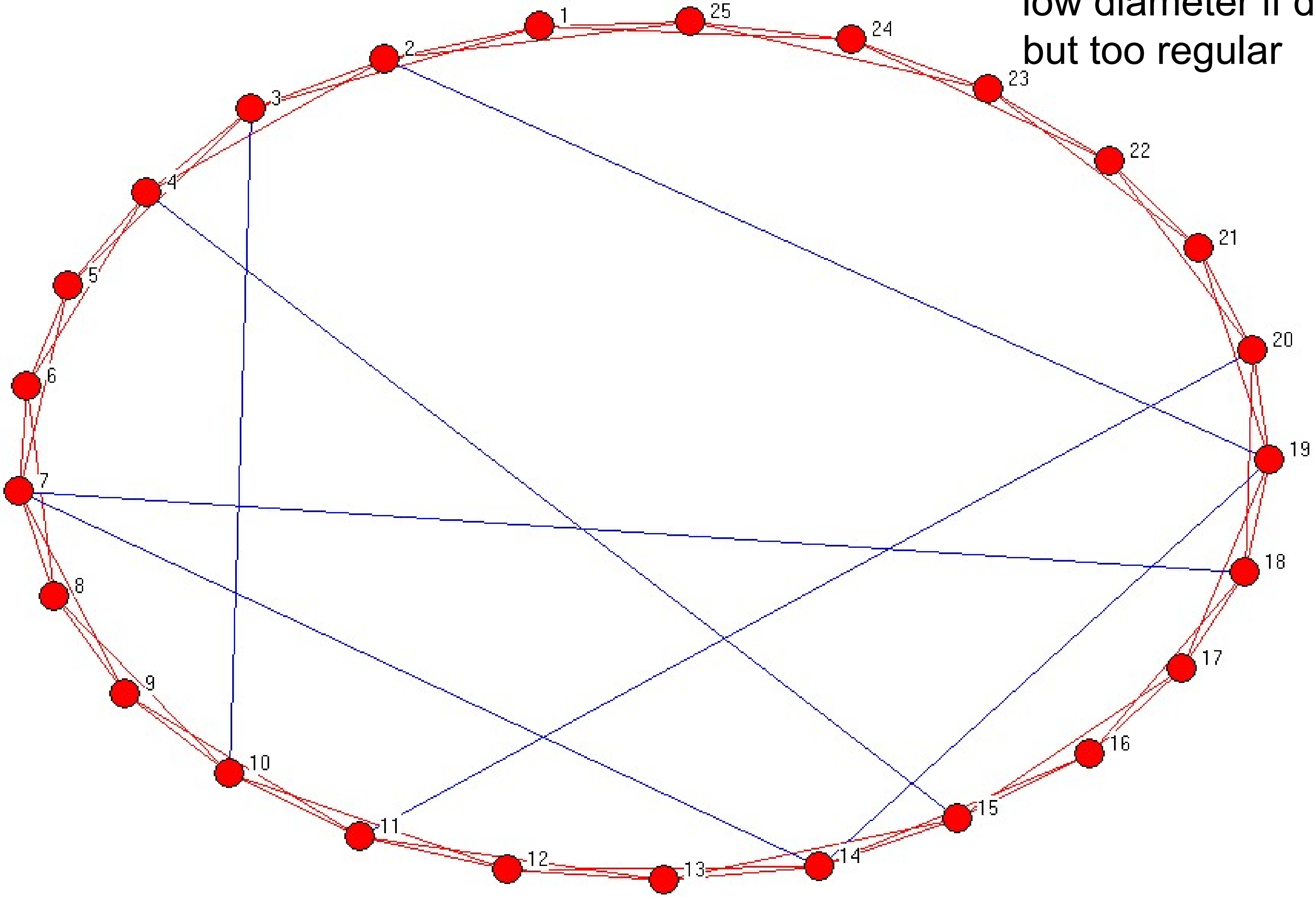
small world:
mostly structured with a few random connections

random graph:
all connections random

Source: Watts, D.J., Strogatz, S.H.(1998) Collective dynamics of 'small-world' networks. Nature 393:440-442.

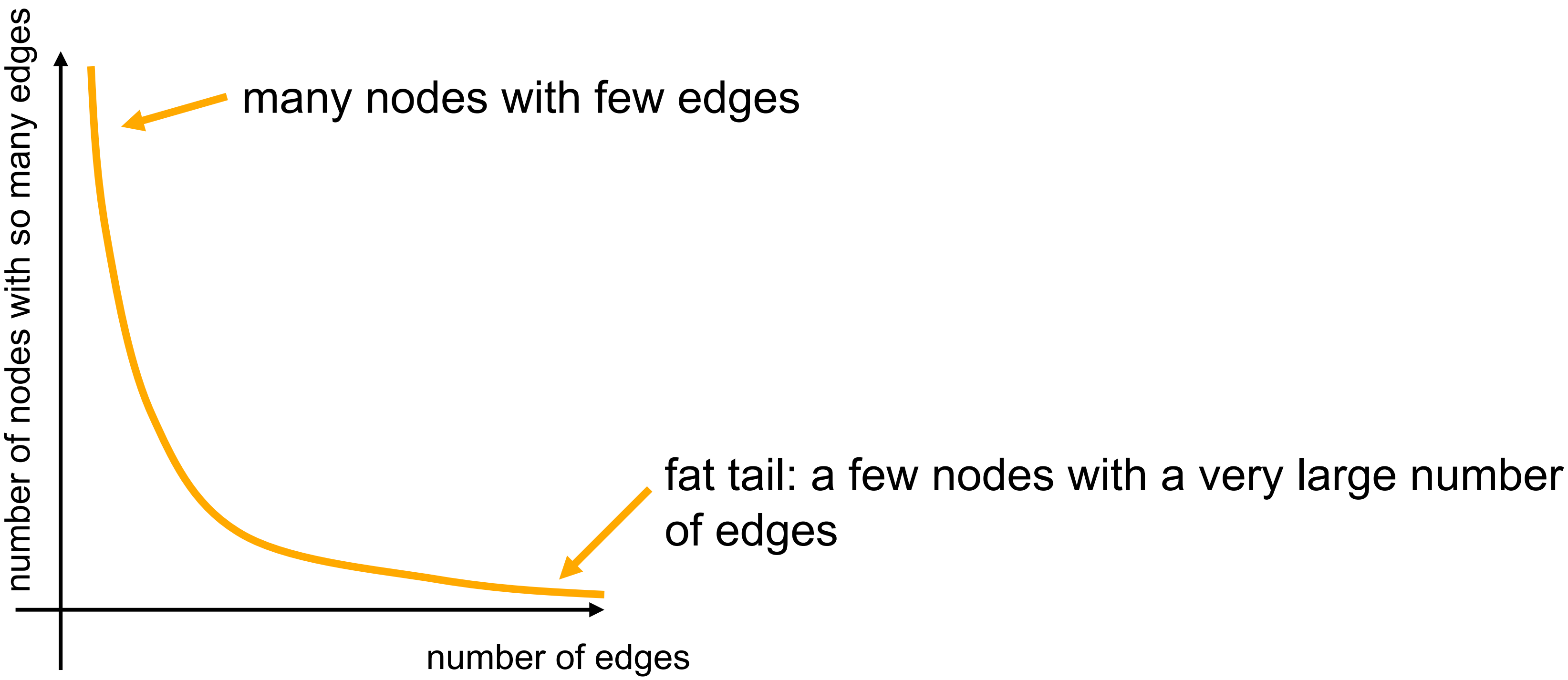
Rewired lattice (Watts and Strogatz (1999))

high clustering
low diameter if degree is high
but too regular



modeling networks: power law networks

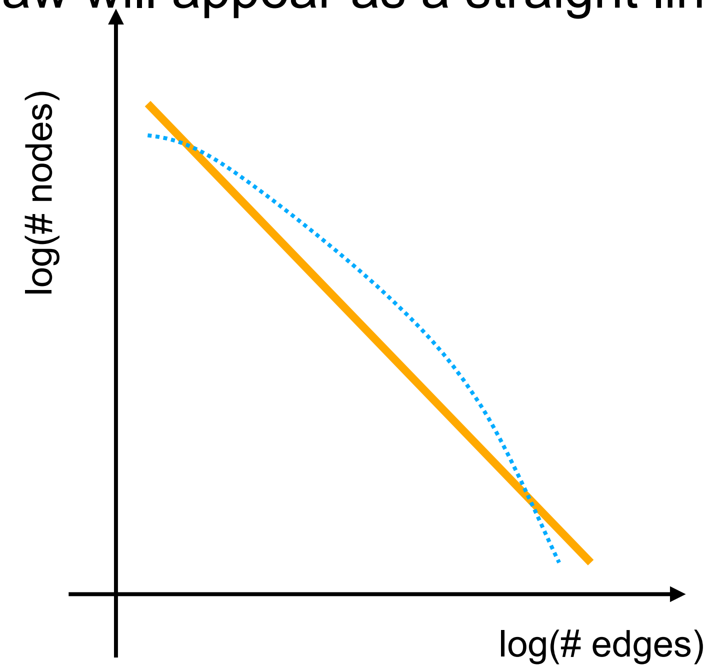
- Many real world networks contain hubs: highly connected nodes
- Usually the distribution of edges is extremely skewed



no “typical” number of edges

But is it really a power-law?

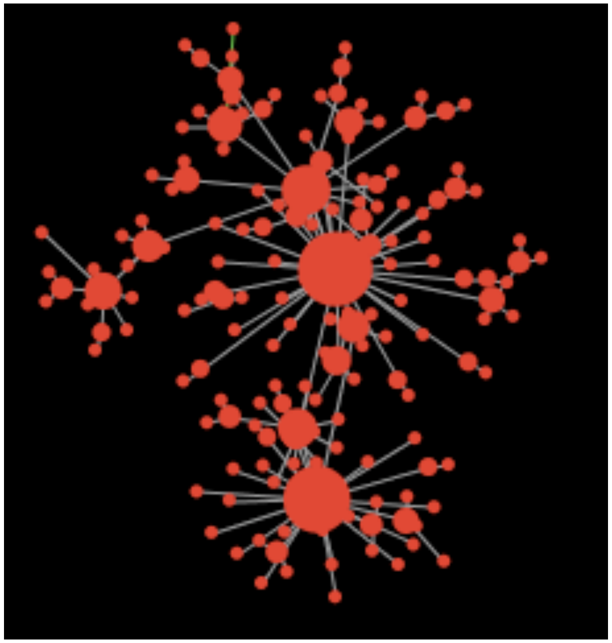
- A power-law will appear as a straight line on a log-log plot:



- A deviation from a straight line could indicate a different distribution:
  - exponential
  - lognormal

network growth & resulting structure

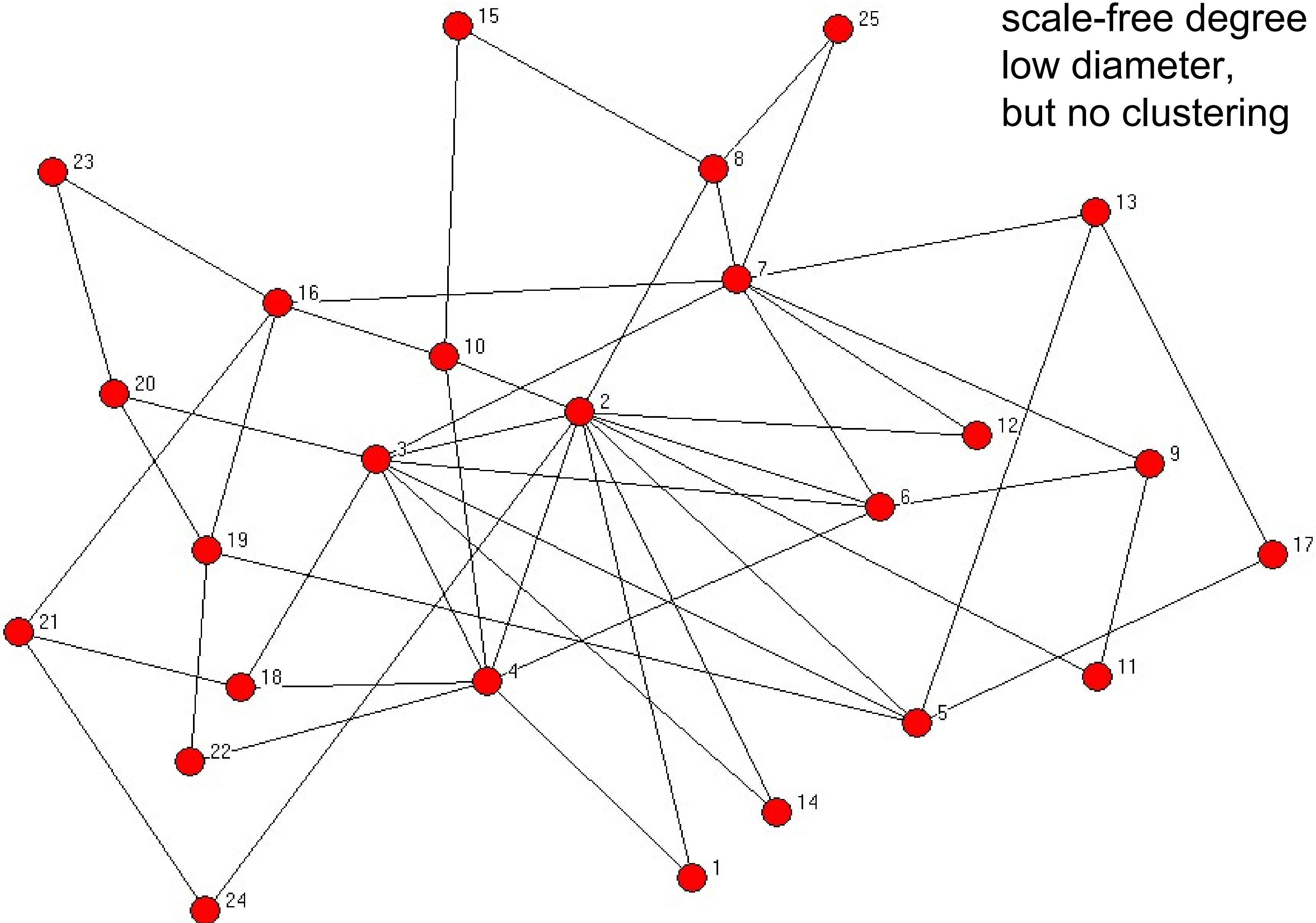
- random attachment: new node picks any existing node to attach to
- preferential attachment: new node picks from existing nodes according to their degrees



http://projects.si.umich.edu/netlearn/NetLogo4/RAndPrefAttachment.html

Preferential Attachment (Barabasi and Albert (2001))

scale-free degree distribution
low diameter,
but no clustering



Jackson and Rogers (AER 2007)

▲ Nodes are players

▲ Indexed by date of birth $t = \{1, 2, 3, ...\}$

▲ Find m_r other nodes at random

▲ Search their neighborhoods to find m_s more nodes.

think of entering at a random web page and following its links

▲ Attach to a given node if net utility is positive

random utility or

increasing in node's degree

Degree Distribution

Expected increase in the in-degree of a node i

prob found at random

prob found through search

$$p\,(\,m_r/t\;+\;d_i\;[m_s/(t\,m)])$$

prob linked to given found

number of neighbors

prob my neighbor is entry point

m – average links/node,   r – ratio random/search

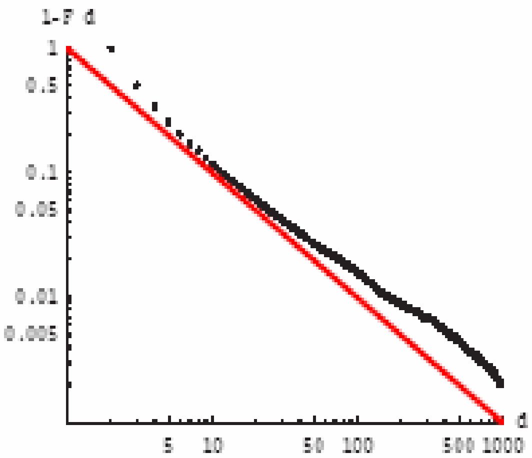
Proposition (Mean field)

The degree distribution of the mean field approximation to the process has a degree distribution having complementary cdf of

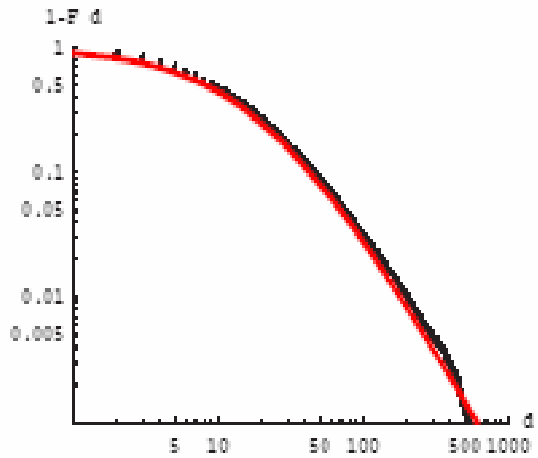
$$F(d) = 1 - (rm)^{1+r} (d + rm)^{-(1+r)}$$

Clustering is bounded away from 0 and decreasing in r

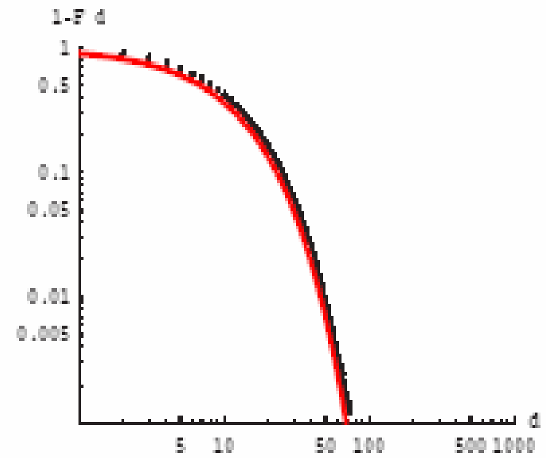
Varying the relative Random and Search probabilities



r=0



r=1



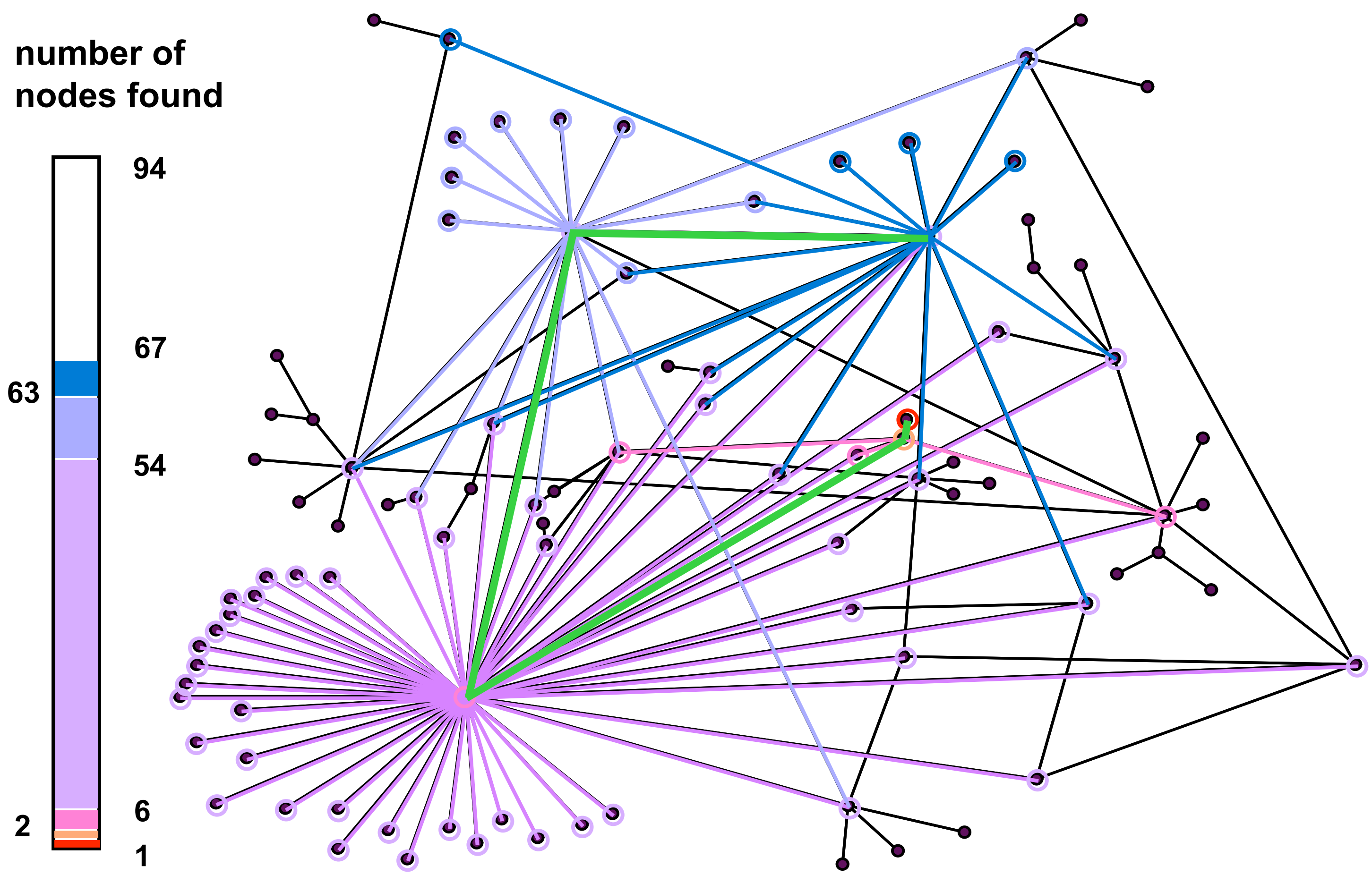
r= ∞

What implications does this have?

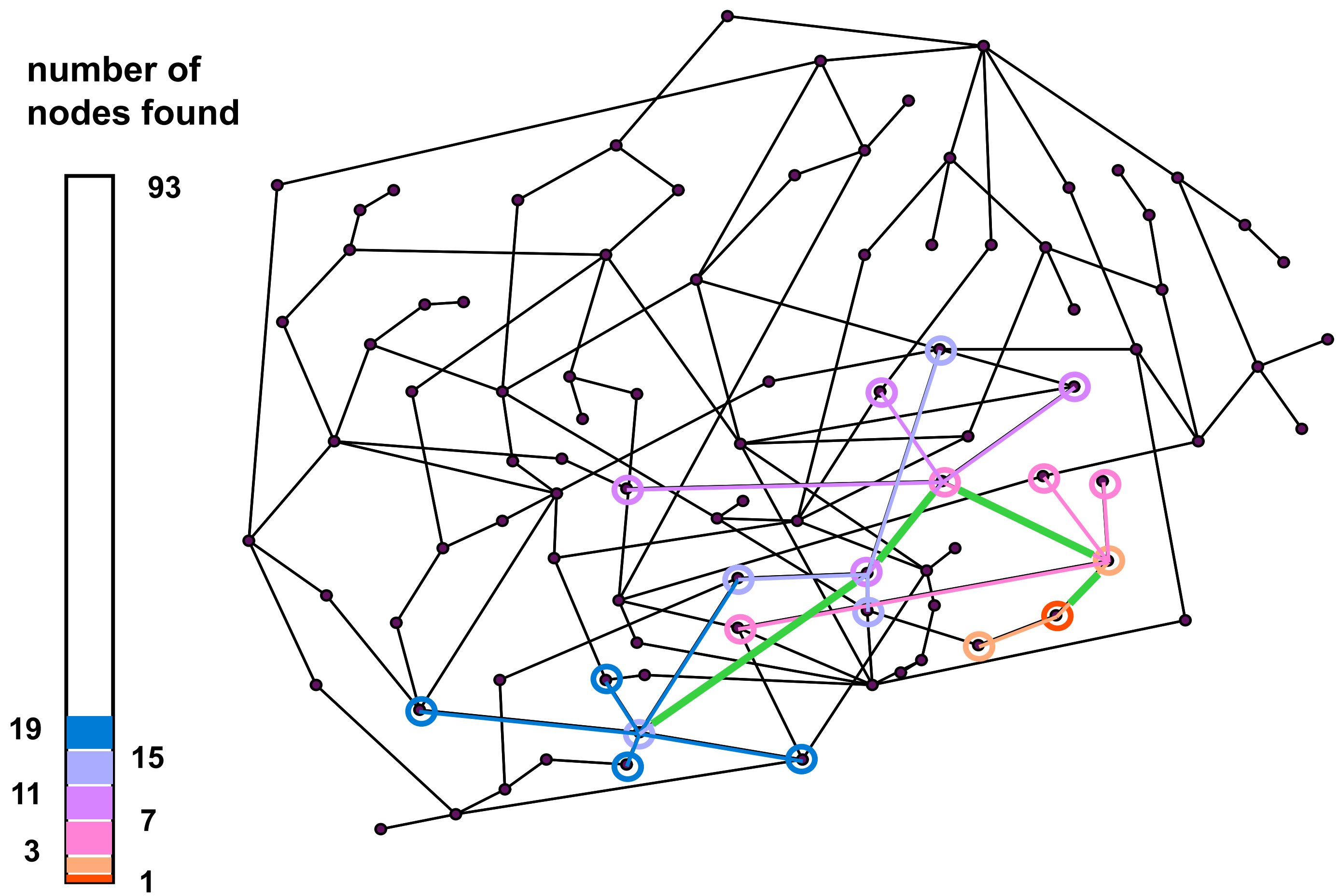
- Robustness
- Search
- Spread of disease
- Opinion formation
- Spread of computer viruses
- Gossip

How do we search?

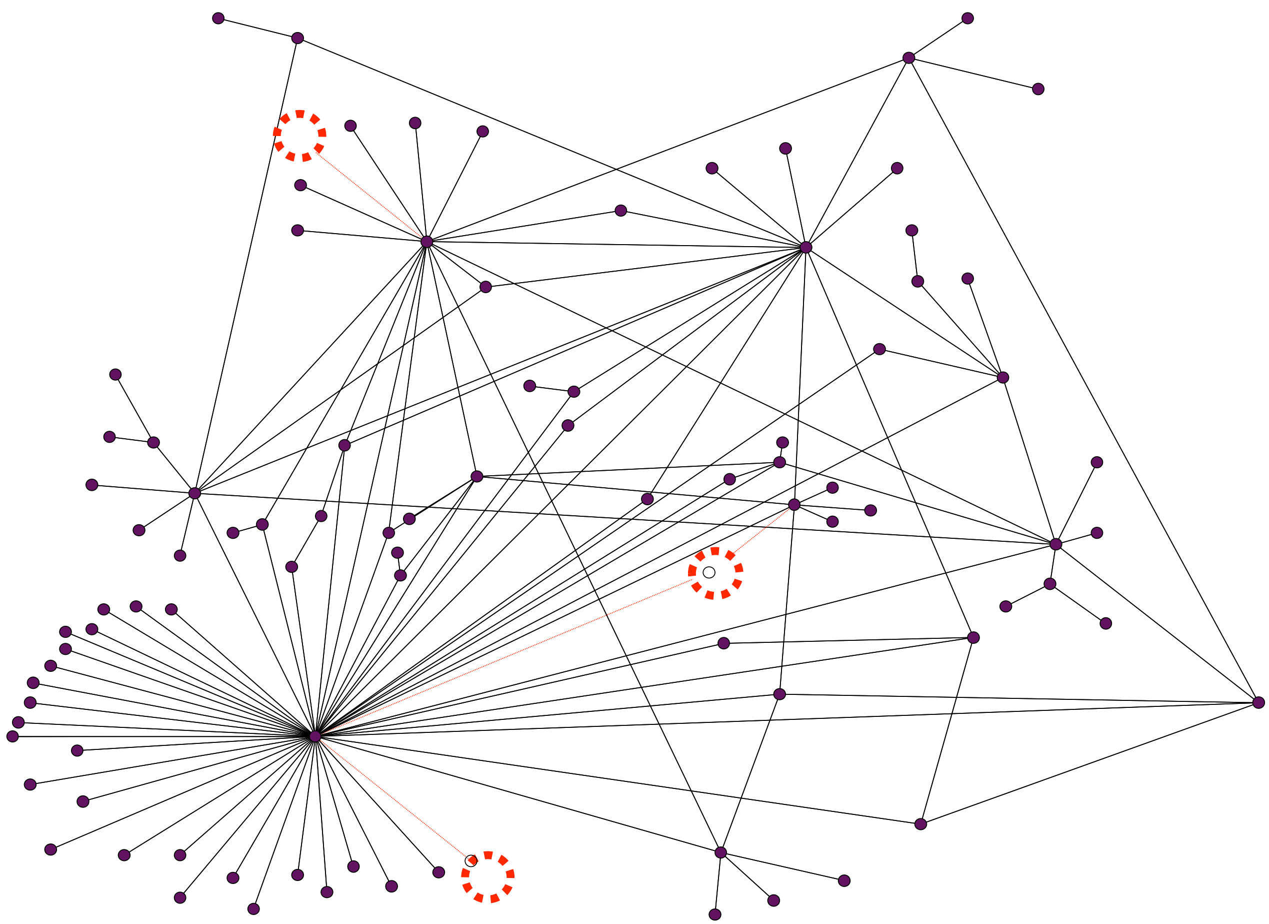
power-law graph



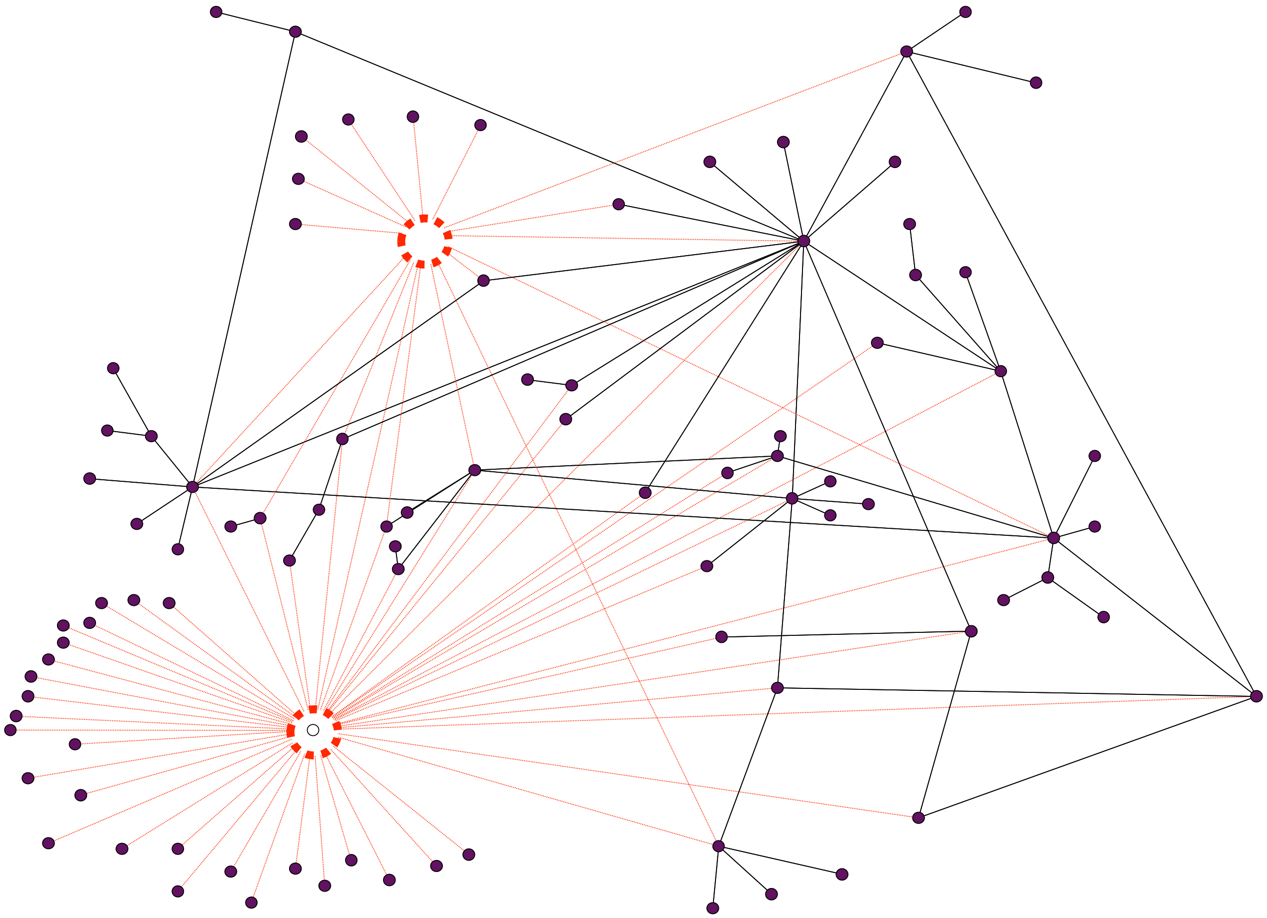
Poisson graph



Power-law networks are robust to random breakdown

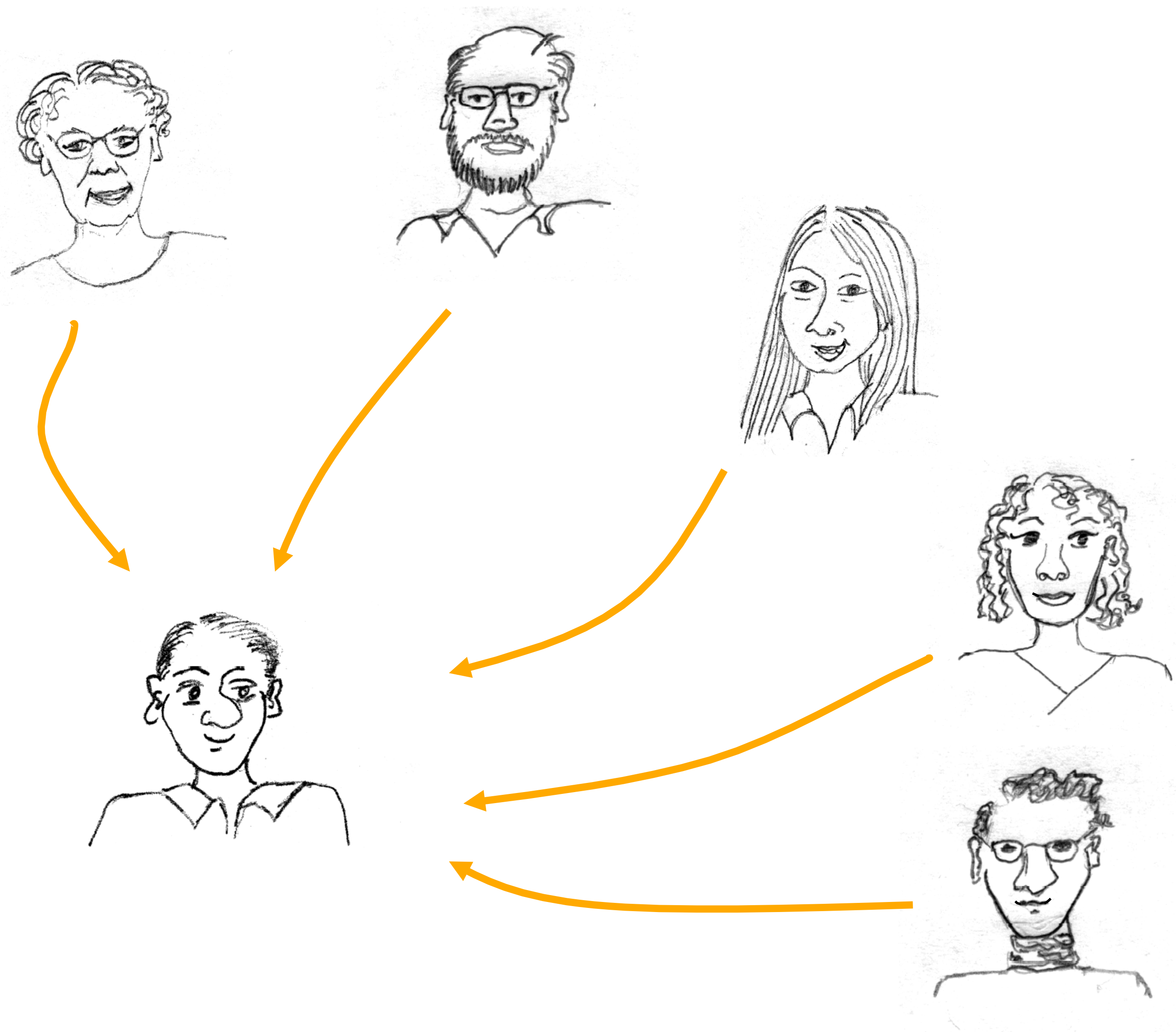


But are especially vulnerable to targeted attack

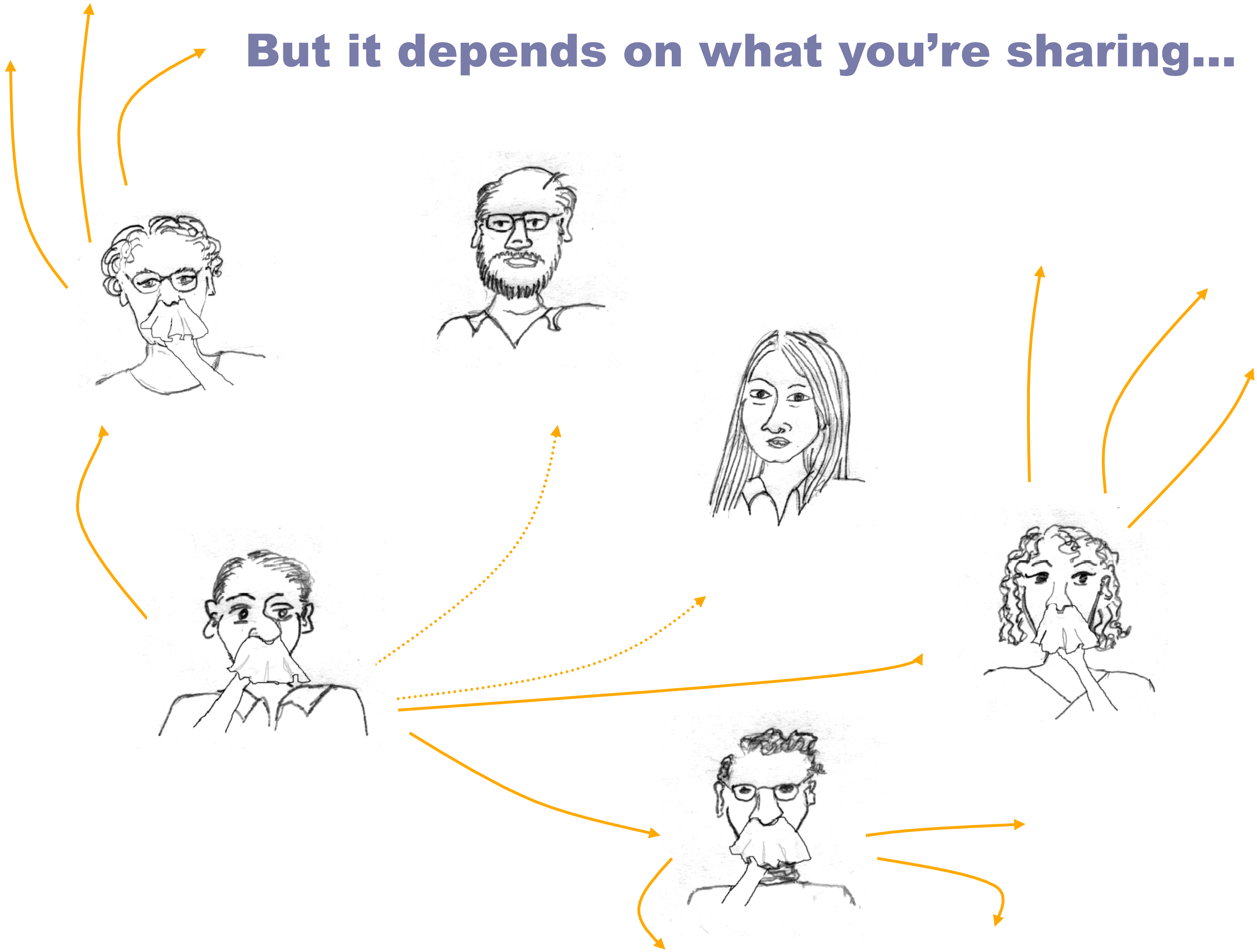


- Targeting and removing hubs can quickly break up the network

In social networks, it's nice to be a hub



But it depends on what you're sharing...

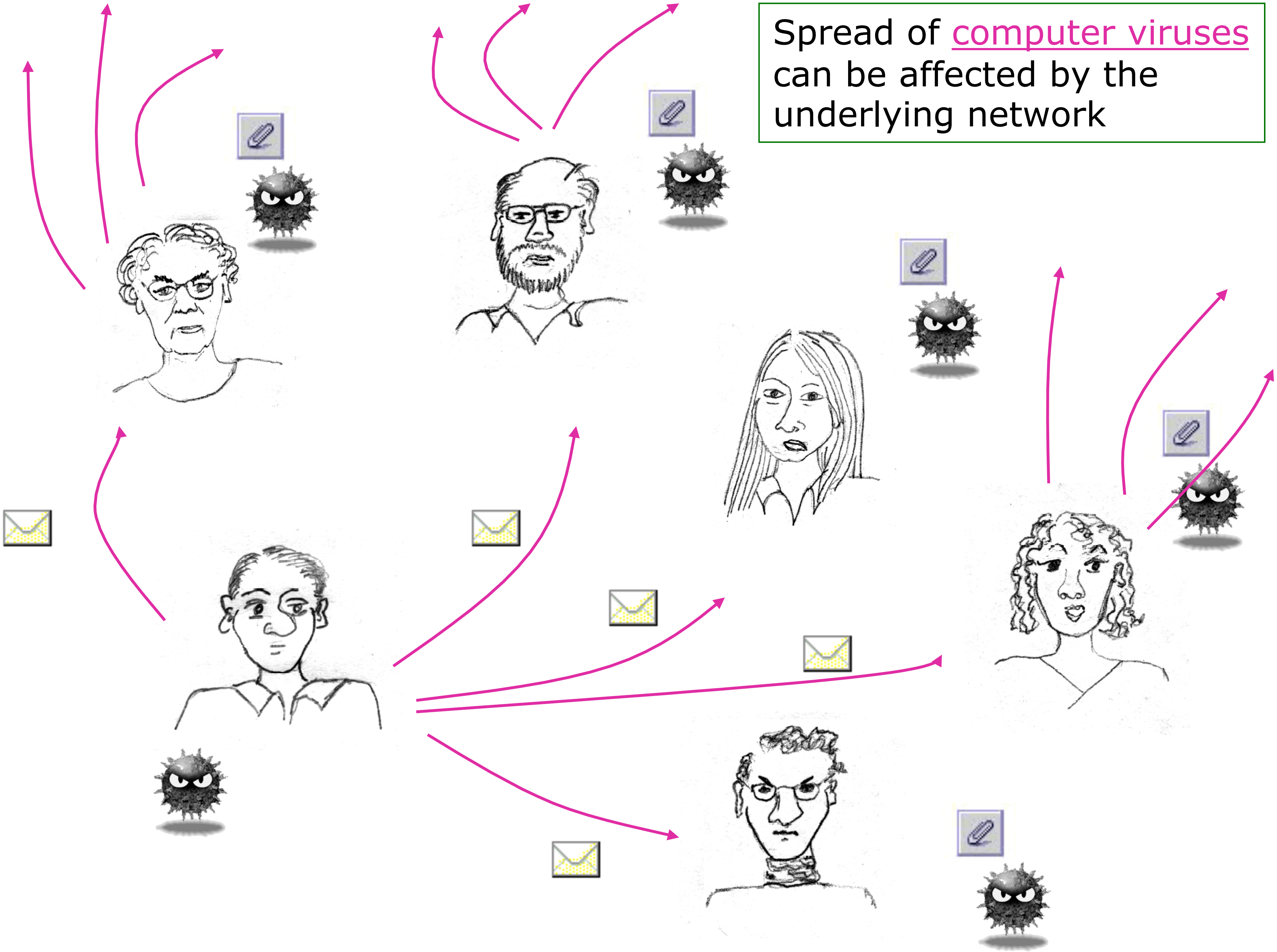


The role of hubs in epidemics

- In a power-law network, a virus can persist no matter how low its infectiousness

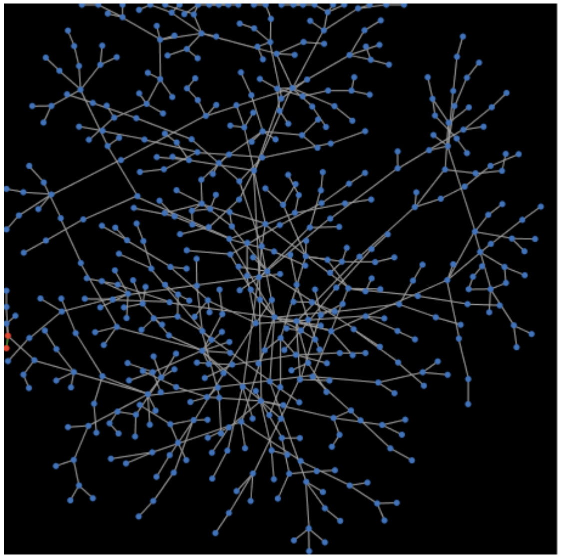
- Many real world networks do exhibit power-laws:
  - needle sharing
  - sexual contacts
  - email networks

Spread of computer viruses can be affected by the underlying network

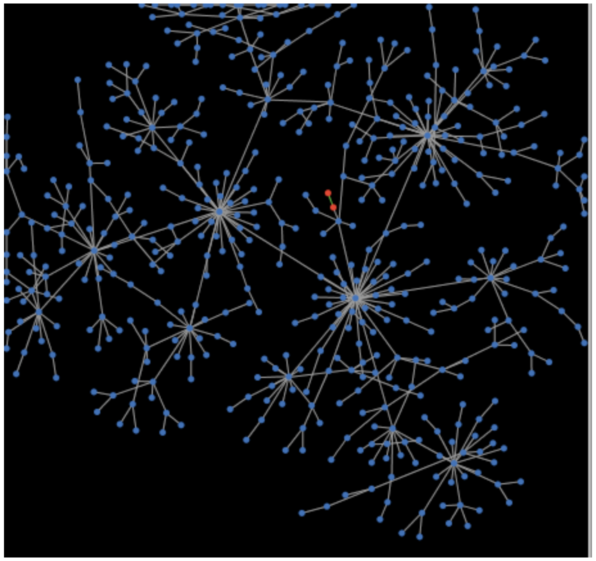


SI models & network structure

- Will random or preferential attachment lead to faster diffusion?



random growth



preferential growth

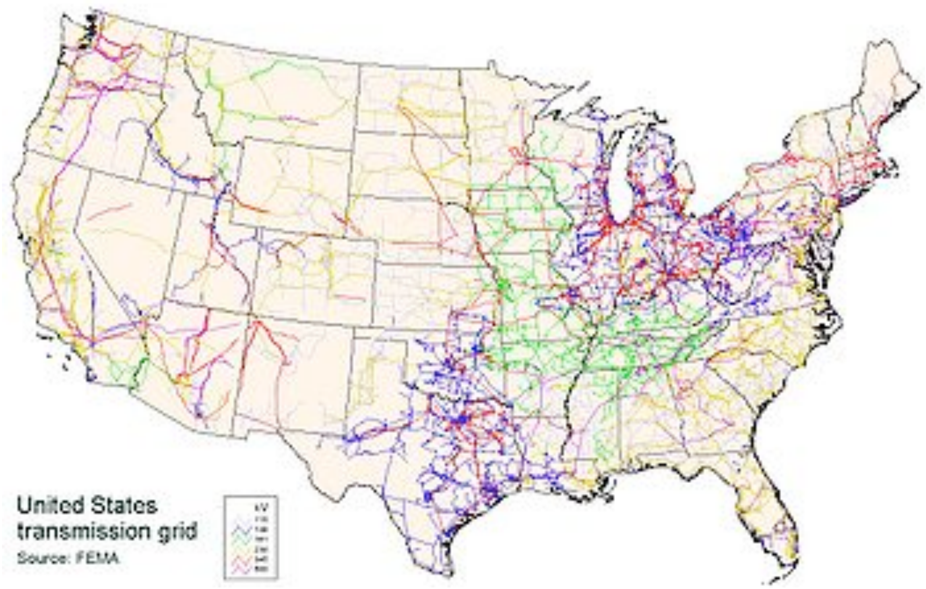
http://projects.si.umich.edu/netlearn/NetLogo4/BADiffusion.html

resilience:
power grids and cascading failures

- Vast system of electricity generation, transmission & distribution is essentially
  a single network

- Power flows through all paths from source to sink (flow calculations are important for other networks, even social ones)

- All AC lines within an interconnect must be in sync

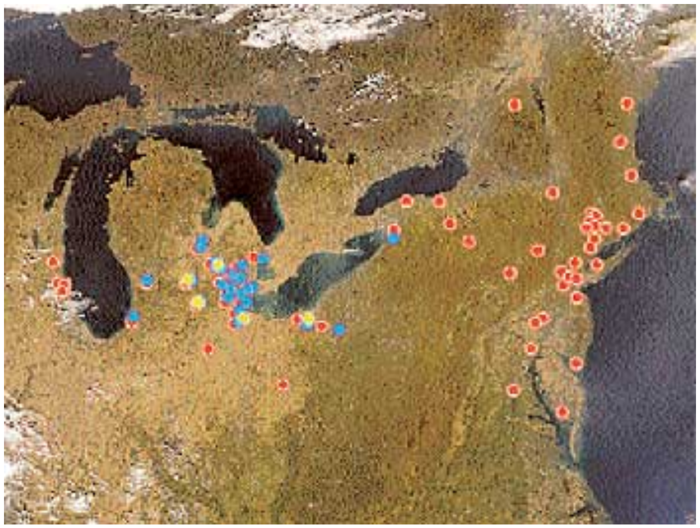


- If frequency varies too much (as line approaches capacity), a circuit breaker takes the generator out of the system

- Larger flows are sent to neighboring parts of the grid – triggering a cascading failure

Source: .wikipedia.org/wiki/File:UnitedStatesPowerGrid.jpg

Cascading failures



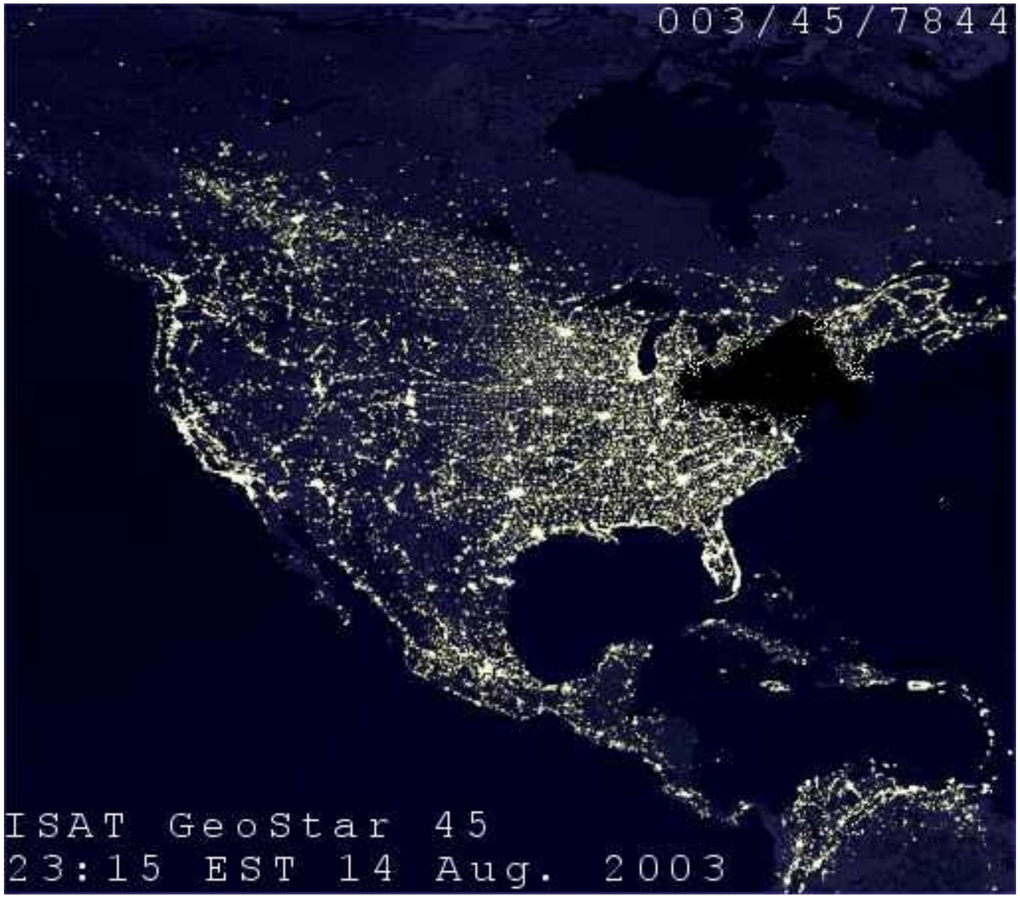
- **1:58 p.m.** The Eastlake, Ohio, First Energy generating plant shuts down (maintenance problems).
- **3:06 p.m.** A First Energy 345-kV transmission line fails south of Cleveland, Ohio.
- **3:17 p.m**. Voltage dips temporarily on the Ohio portion of the grid. Controllers take no action, but power shifted by the first failure onto another power line causes it to sag into a tree at 3:32 p.m., bringing it offline as well. While Mid West ISO and First Energy controllers try to understand the failures, they fail to inform system controllers in nearby states.
- **3:41 and 3:46 p.m.** Two breakers connecting First Energy's grid with American Electric Power are tripped.
- **4:05 p.m.** A sustained power surge on some Ohio lines signals more trouble building.
- **4:09:02 p.m.** Voltage sags deeply as Ohio draws 2 GW of power from Michigan.
- **4:10:34 p.m.** Many transmission lines trip out, first in Michigan and then in Ohio, blocking the eastward flow of power. Generators go down, creating a huge power deficit. In seconds, power surges out of the East, tripping East coast generators to protect them.

Source: Eric J. Lerner, "What's wrong with the electric grid?" http://www.aip.org/tip/INPHFA/vol-9/iss-5/p8.html

(dis) information cascades

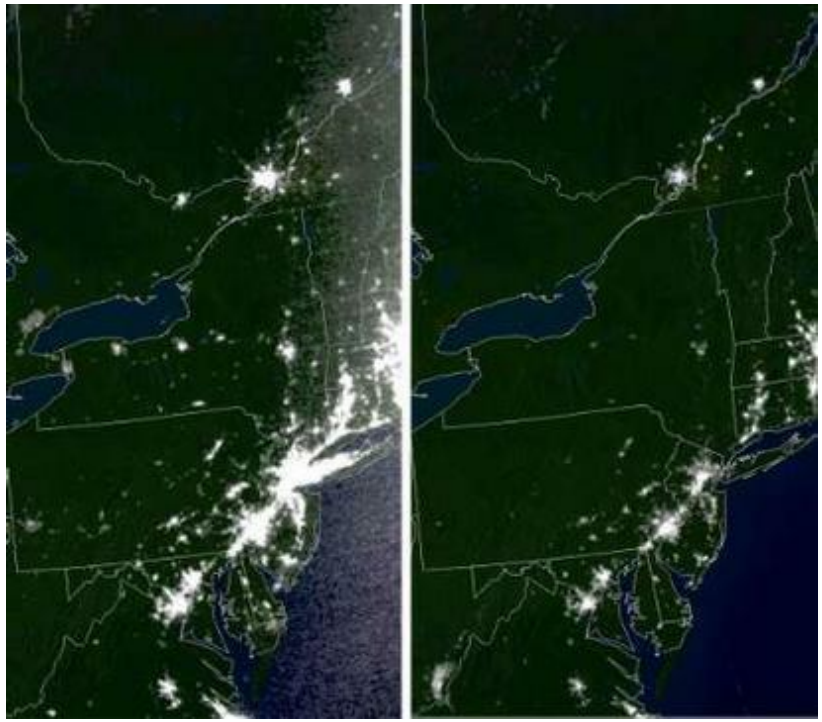
- Rumor spreading
- Urban legends
- Word of mouth (movies, products)

- Web is self-correcting:
  - Satellite image hoax is first passed around, then exposed, hoax fact is blogged about, then written up on urbanlegends.about.com



Source: undetermined

Actual satellite images of the effect of the blackout



20 hours prior to blackout

7 hours after blackout

Source: NOAA, U.S. Government

Advantages of Random Graph Models

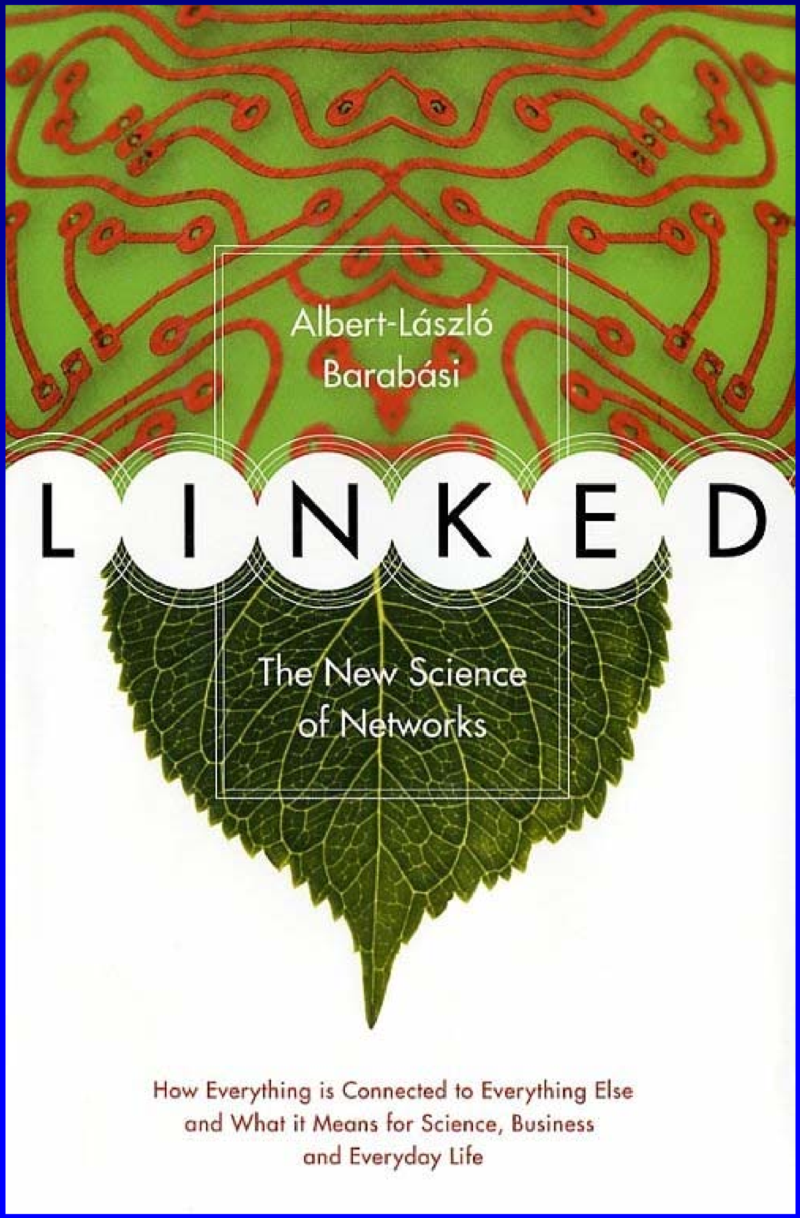
- Generate large networks with well identified properties
- Mimic real networks (at least in some characteristics)
- Tie a specific property to a specific process

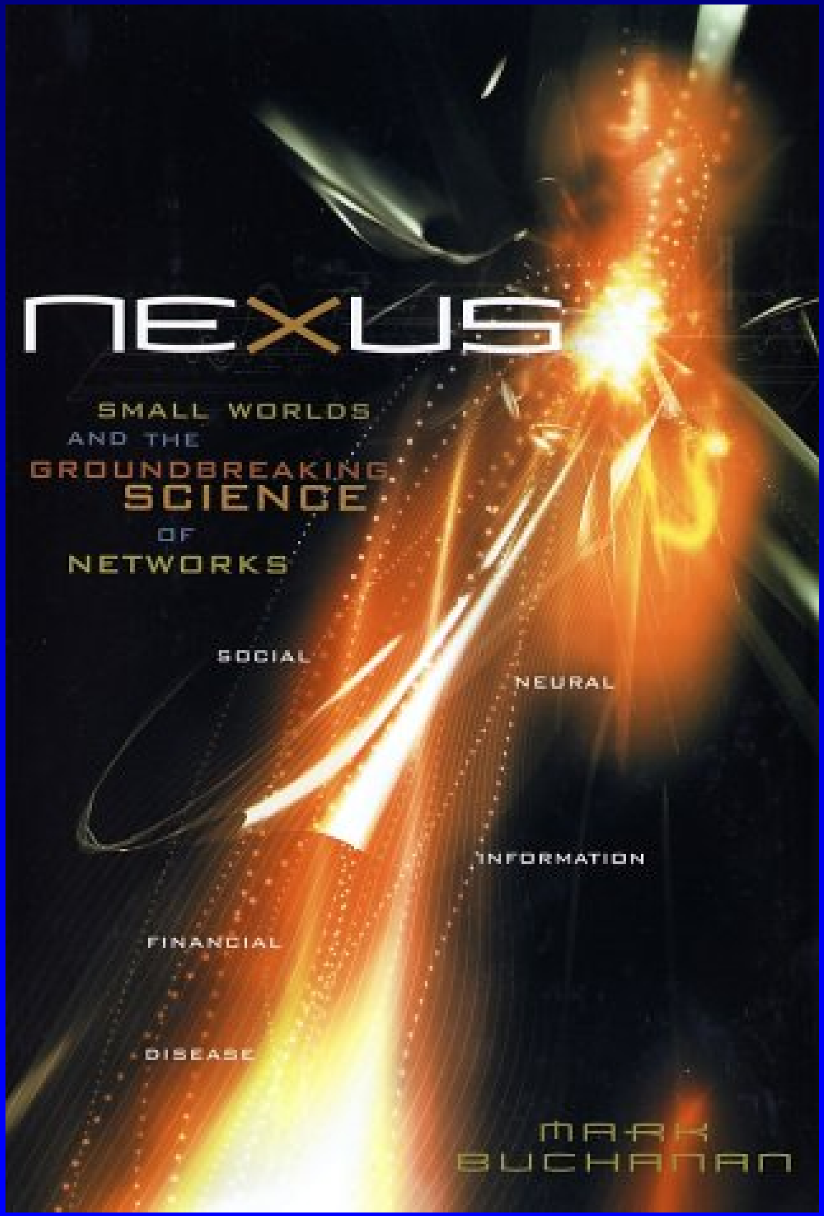
What's Missing From Random Graph Models?

- The ``Why''?
  - Why this process? (lattice, preferential attach...)
- Implications of network structure: economic and social context or relevance?
  - welfare and how can it be improved...
- Careful Empirical Analysis
  - ``Scale-Free'' may not be
  - No fitting of models to data (models aren't rich enough to fit across applications)

Introduction

Scientific Importance of Networks:





Economic/Game Theoretic Models

- Welfare analysis – agents get utility from networks
  - $u_i(g)$
  - Efficient Networks: argmax $\sum u_i(g)$

- Decision making agents form links and/or choose actions

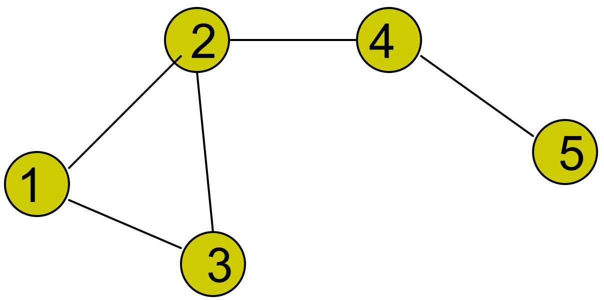
Example: Connections Model

Jackson and Wolinsky (1996):

- benefit from a friend is δ
- benefit from a friend of a friend is δ^2,...
- cost of a link is c

$u_2 = 3\delta + \delta^2 - 3c$

$u_1 = 2\delta + \delta^2 + \delta^3 - 2c$



$u_5 = \delta + \delta^2 + 2\delta^3 - c$

- Pairwise Stable networks
  - $u_i(g) \geq u_i(g\text{-}ij)$ for each i and ij in g
  - $u_i(g+ij) \geq u_i(g)$ implies $u_j(g+ij) \geq u_j(g)$ for each ij not in g

Efficient Networks

- low cost: $c < \delta - \delta^2$
  - complete network is efficient
- medium cost: $\delta - \delta^2 < c < \delta + (n-2)\delta^2/2$
  - star network is efficient
    - minimal number of links to connect
    - connection at length 2 is more valuable than at 1 ($\delta - c < \delta^2$)
- high cost: $\delta + (n-2)\delta^2/2 < c$
  - empty network is efficient

Pairwise Stable Networks:

- low cost: $c < \delta - \delta^2$
  - complete network is pairwise stable (and efficient)

- medium/low cost: $\delta - \delta^2 < c < \delta$
  - star network is pairwise stable (and efficient)
  - others are also pairwise stable

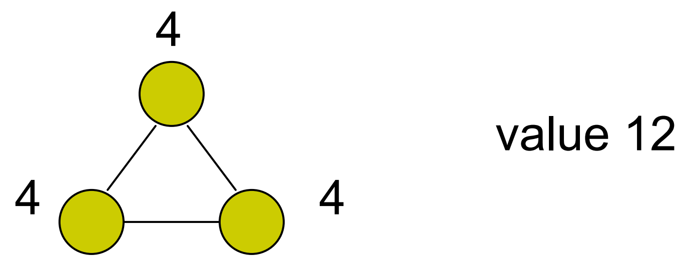
- medium/high cost: $\delta < c < \delta + (n-2)\delta^2/2$
  - star network is not pairwise stable (no loose ends)
  - nonempty pairwise stable networks are over-connected and may include too few agents

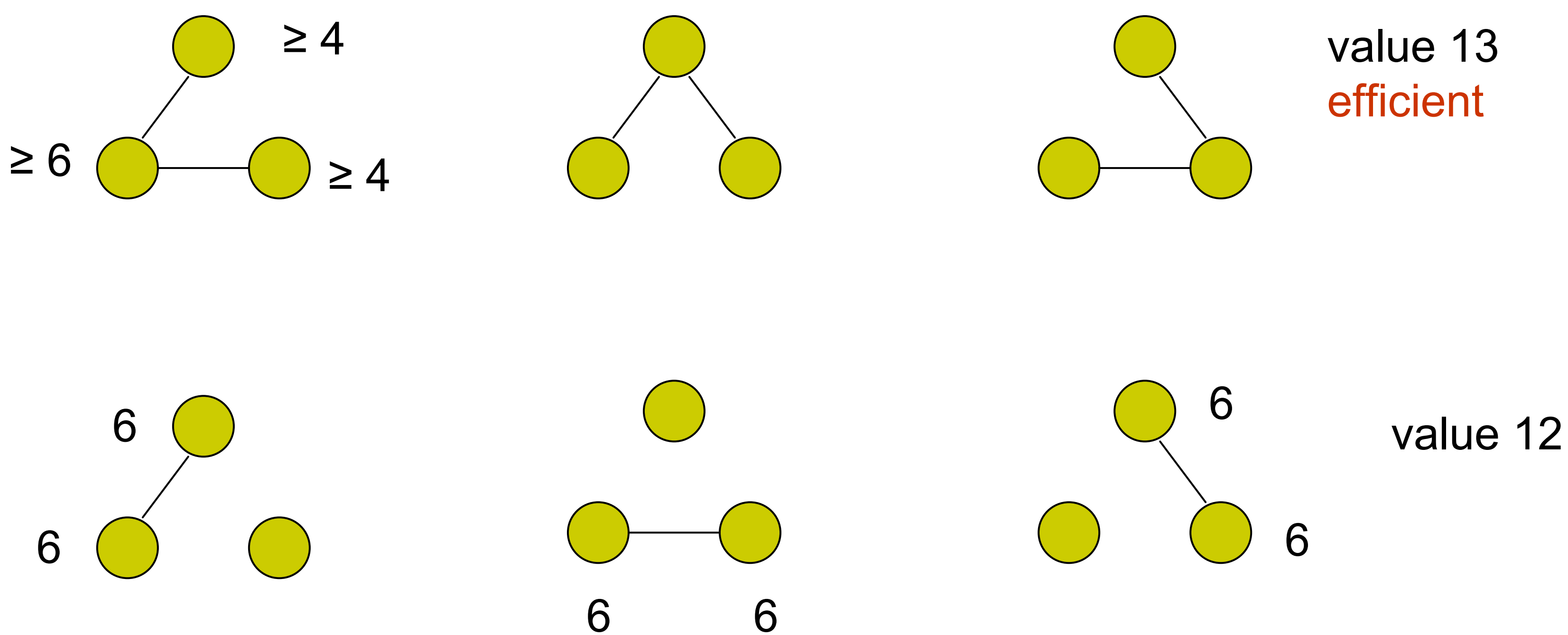
- high cost: $\delta + (n-2)\delta^2/2 < c$
  - empty network is pairwise stable (and efficient)

Transfers cannot always help

anonymity: same transfers to identical players

balance: no transfers outside of component





Rich literature on such issues

- loosen anonymity (Dutta-Mutuswami (1997))
- directed networks (Bala-Goyal (2000), Dutta-Jackson (2000),...)
- bargaining when forming links (Currarini-Morelli(2000), Slikker-van den Nouweland (2000), Mutuswami-Winter(2002), Bloch-Jackson (2004))
- dynamic models (Aumann-Myerson (1988), Watts (2001), Jackson-Watts (2002ab), Goyal-Vega-Redondo (2004), Feri (2004), Lopez-Pintado (2004),...)
- farsighted models (Page-Wooders-Kamat (2003), Dutta-Ghosal-Ray (2003), Deroian (2003),...)
- allocating value (Myerson (1977), Meessen (1988), Borm-Owen-Tijs (1992), van den Nouweland (1993), Qin (1996), Jackson-Wolinsky (1996), Slikker (2000), Jackson (2005)...)
- modeling stability (Dutta-Mutuswami (1997), Jackson-van den Nouweland (2000), Gilles-Sarangi (2003ab), Calvo-Armengol and Ikilic (2004),...)
- experiments (Callander-Plott (2001), Corbae-Duffy (2001), Pantz-Zeigelmeyer (2003), Charness-Corominas-Bosch-Frechette (2001), Falk-Kosfeld (2003), ...)

Models of Networks in Context

- crime networks (Glaeser-Sacerdote-Scheinkman (1996), Ballester, Calvo, Zenou (2003),...)
- markets (Kirman (1997), Tesfatsion (1997), Weisbach-Kirman-Herreiner (2000), Kranton-Minehart (2002), Corominas-Bosch (2005), Wang-Watts (2002), Galeotti (2005),Kakade et al (2005)...)
- labor networks (Boorman (1975), Montgomery (1991, 1994), Calvo (2000), Arrow-Borzekowski (2002), Calvo-Jackson (2004,2005), Cahuc-Fontaine (2004), Currie...)
- insurance (Fafchamps-Lund (2000), DeWeerdt (2002), Bloch-Genicot-Ray (2004),...
- IO (Bloch (2001), Goyal-Moraga (2001), Goyal-Joshi (2001), Belleflamme-Bloch (2002),Billard-Bravard (2002), ...)
- international trade (Casella-Rauch (2001), Furusawa-Konishi (2003),
- public goods (Bramoulle-Kranton (2004)
- airlines (Starr-Stinchcombe (1992), Hendricks-Piccione-Tan (1995))
- network externalities in goods (Katz-Shapiro (1985), Economides (1989, 1991) , Sharkey (1991)...)
- organization structure (Radner (), Radner-van Zandt (), Demange (2004)...)
- learning (Bala-Goyal (1998), Morris (2000), DeMarzo-Vayanos-Zweibel (2003), Gale-Kariv (2003), Choi-Gale-Kariv (2004),...)

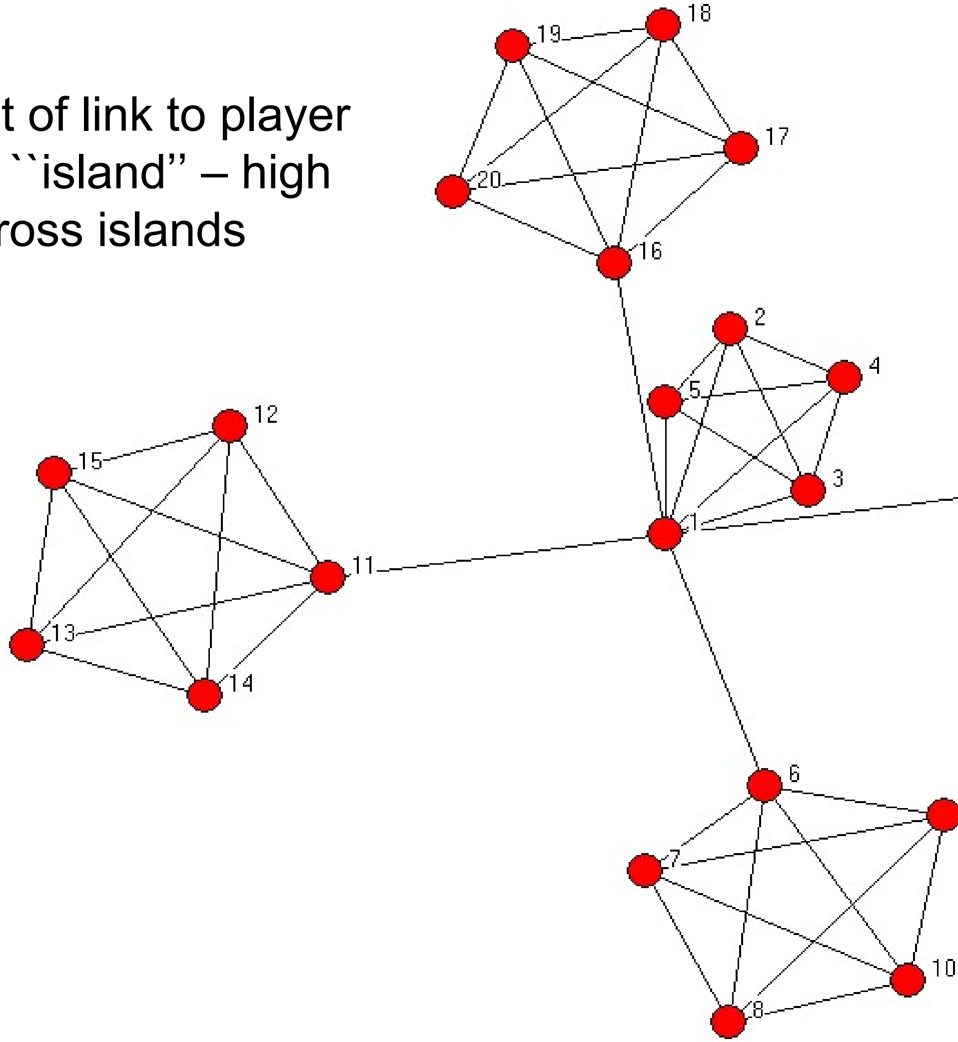
Can economic models match observables?

- Small worlds related to costs/benefits
  - low costs to local links – high clustering
  - high value to distant connections – low diameter

Geographic Connections (Johnson-Gilles (2000), Carayol-Roux (2003), Galeotti-Goyal-Kamphorst (2004), Jackson-Rogers (2004))

low cost of link to player on own ``island'' – high cost across islands

high clustering,
low diameter,
but regular degree



Friendship Formation, Oppositional Identity, and Segregation

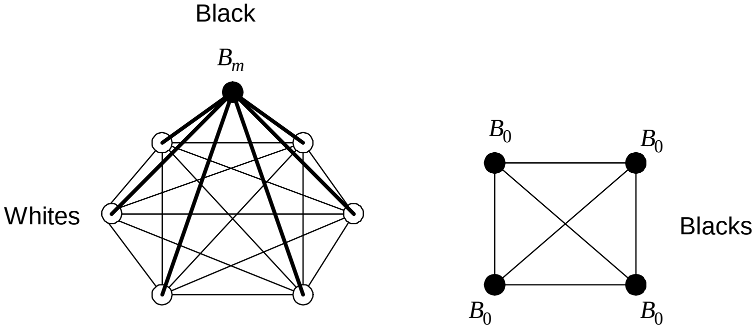
Joan de Marti
Universitat Pompeu Fabra

Yves Zenou
Stockholm University and IFN

November 2008

Low Within-Community Costs

Consider the following network



If

$$C > \frac{\left[\delta + \left(n^B - 1\right)\delta^2 - c\right]\left(n^W + 1\right)}{n^W - 1}$$

then this network, where not fully intraconnected communities prevail and where one black is assimilated and has an oppositional identity while all other blacks are separated, is pairwise stable.

Advantages of an economic approach

- Payoffs allow for a welfare analysis
  - Identify tradeoffs – incentives versus efficiency

- Tie the nature of externalities to network formation...

- Put network structures in context

- Account for (and *explain*) some observables

What's missing from Game theoretic models?

- Stark network structures emerge
  - need more heterogeneity

- over-emphasize choice versus chance determinants for *large* applications?

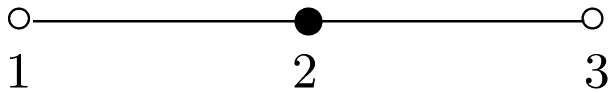
- more on network structure and outcomes

Games on Networks

Calvó-Armengol and Jackson (AER 2004)

Labor Market Networks

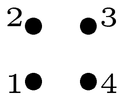
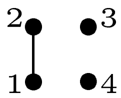
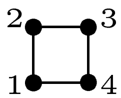
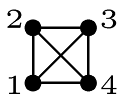
Do friends of friends help or hurt?



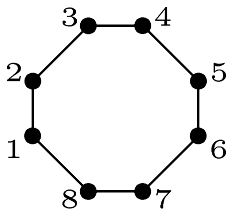
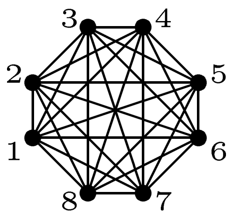
S_{1t} and S_{3t} are negatively correlated. But in the long run agent 1 can benefit from 3's presence.

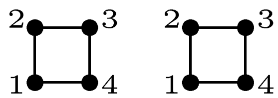
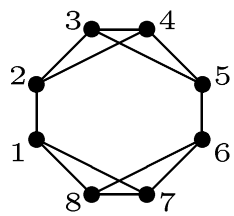
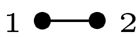
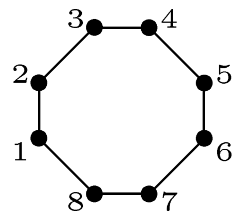
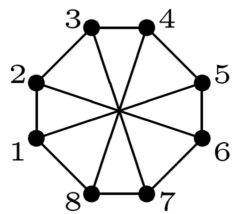
We sort out the short run competition from the longer run benefits by dividing p and b by some T.

Unskilled labor: $n = 4$**,** $a_i = .100$ **and** $b_i = .015$

g	$\text{Prob}(S_1 = 0)$	$\text{Corr}(S_1, S_2)$	$\text{Corr}(S_1, S_3)$
	.132	–	–
	.083	.041	–
	.063	.025	.019
	.050	.025	.025

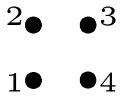
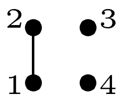
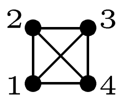
Unskilled labor: $n = 8$, $a_i = .100$ **and** $b_i = .015$

g	$\text{Prob}(S_1 = 0)$	$\text{Corr}(S_1, S_2)$	$\text{Corr}(S_1, S_3)$	$\text{Corr}(S_1, S_4)$
	.060	.023	.003	.001
	.030	.014	.014	.014

Network	Jobless	Network	Jobless	Network	Jobless
	13.2 %		6.3 %		4.9 %
	8.3 %		6.0 %		4.8 %

Duration dependence

Prob($S_1 = 0$) conditional on a string of at least so many periods of unemployment:

g	1 period	2 periods	10 periods	limit
	.901	.901	.901	.901
	.824	.825	.830	.901
	.695	.700	.722	.901

Ballester, Calvó-Armengol and Zenou

Peer effects in Networks

Econometrica 2006

Utility function for all $i = 1, ..., n$.

$$u_i = \underbrace{\alpha x_i - \frac{1}{2}(\beta - \gamma) x_i^2}_{\text{Own Concavity}} - \underbrace{\gamma \sum_{j=1}^{n} x_i x_j}_{\text{Global Substituability}} + \underbrace{\lambda \sum_{j=1}^{n} g_{ij} x_i x_j}_{\text{Local Complementarity}}$$

The network Bonacich centrality To each network $\mathbf{g}$, we associate its adjacency matrix $\mathbf{G} = [g_{ij}]$.

Symmetric zero-diagonal square matrix that keeps track of the direct connections in $\mathbf{g}$.

The kth power $\mathbf{G}^k = \mathbf{G}^{(k \overset{}{\underset{\cdots}{times}})}\mathbf{G}$ of the adjacency matrix $\mathbf{G}$ keeps track of indirect connections in $\mathbf{g}$.

The coefficient $g_{ij}^{[k]}$ in the (i, j) cell of $\mathbf{G}^k$ gives the number of paths of length k in $\mathbf{g}$ between i and j.

The Bonacich centrality of node i is $b_i(\mathbf{g}, a) = \sum_{j=1}^{n} m_{ij}(\mathbf{g}, a)$, and counts the *total* number of paths in $\mathbf{g}$ starting from i.

It is the sum of all loops starting from i and ending at i, and all outer paths that connect i to every other player $j \neq i$:

$$b_i(\mathbf{g}, a) = \underbrace{m_{ii}(\mathbf{g}, a)}_{self-loops} + \sum_{j \neq i} \underbrace{m_{ij}(\mathbf{g}, a)}_{out-paths}.$$

Note that, by definition, $m_{ii}(\mathbf{g}, a) \geq 1$, and thus $b_i(\mathbf{g}, a) \geq 1$.

Nash equilibrium and Bonacich centrality

The largest eigenvalue $\mu_1(\mathbf{G})$ of $\mathbf{G}$, also called index of $\mathbf{g}$.

Theorem. $[\beta \mathbf{I} - \lambda \mathbf{G}]^{-1}$ is well-defined and non-negative iff $\beta > \lambda \mu_1(\mathbf{G})$. Then, $\mathbf{\Sigma}$ has a unique Nash equilibrium, interior:

$$\mathbf{x}^* (\mathbf{\Sigma}) = \frac{\alpha}{\beta + \gamma b(\mathbf{g}, \lambda/\beta)} \mathbf{b}(\mathbf{g}, \lambda/\beta)$$

Equilibrium outcomes are proportional to node centrality.

Best-reply functions

$$BR_i(\mathbf{y}_{-i}) = \phi \sum_{j=1}^{n} g_{ij} y_j + \sum_{m=1}^{M} \beta_m x_i^m - pf + \eta_k + \varepsilon_i$$

$$\underbrace{Alice}_{y_A \uparrow \Delta} \quad \rightarrow \quad \underbrace{Bob}_{y_B \uparrow \phi \Delta} \quad \rightarrow \quad \underbrace{Charlie}_{y_C \uparrow \phi^2 \Delta}$$

- Direct complementarities induce indirect complementarities of all possible order.
- There is a discount of distance ϕ^{distance}.
- This means that ϕ cannot be too large.

Network structure and peer effects

Standard *peer effects*, or intragroup externalities, are homogeneous across group members, an *average* influence.

Here, instead, the intragroup externality varies across group members with their network location:

$$x_i^* = \frac{b_i(\mathbf{g}, \lambda/\beta)}{b(\mathbf{g}, \lambda/\beta)} x^*.$$

The Bonacich network centrality captures the *variance* in peer effects.

Comparative statics

Theorem. Let $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$ symmetric s.t. $(\alpha, \beta, \gamma, \lambda) = (\alpha', \beta', \gamma', \lambda')$ and $\mathbf{G} \leq \mathbf{G}'$. If $\beta > \lambda \mu_1(\mathbf{G}')$, then $x^*(\mathbf{\Sigma}') > x^*(\mathbf{\Sigma})$.

The denser the pattern of local complementarities, the higher the aggregate activity level.

In words, the denser the pattern of local complementarities, the higher the aggregate outcome, as players can rip more complementarities in $\mathbf{g}'$ than in $\mathbf{g}$.

The geometric intuition for this result is clear. Recall that $b(\mathbf{g}, \lambda^*)$ counts the total number of weighted paths in $\mathbf{g}$. This is obviously an increasing function in $\mathbf{g}$ (for the inclusion ordering), as more links imply more such paths.

Network-based policies

The planner can fine-tune the exogenous payoff parameters. This only shifts the distribution of individual outcomes.

The planner can also manipulate the network geometry: the distribution of individual outcomes is now completely modified.

We focus on a particular policy, removing one node, and we identify the *optimal target* in the network, the *key player*.

The planner's problem: finding the Achille's heel

When i is removed, all cross effects with i vanish. We get $\mathbf{\Sigma^{-i}}$.

The impact on aggregate equilibrium outcome is twofold: less players (*direct* effect), a different matrix (*indirect*).

The planner's problem is:

$$\max_i \{x^*(\mathbf{\Sigma}) - x^*(\mathbf{\Sigma}^{-i})\} \quad \Leftrightarrow \quad \min_i \{b(\mathbf{g}^{-i}, \lambda/\beta)\}.$$

The network inter-centrality measure

The equilibrium Bonacich centrality fails to internalize all the network externalities, but the planner needs to internalize them all!

The inter-centrality keeps track of all the cross-contributions in Bonacich centralities across agents:

$$c_i(\mathbf{g}, a) = \frac{b_i^2(\mathbf{g}, a)}{m_{ii}(\mathbf{g}, a)}.$$

The Bonacich centrality of player i counts the number of paths in $\mathbf{g}$ that stem from i.

The intercentrality counts the total number of such paths that hit i;

It is the sum of i's Bonacich centrality and i's contribution to every other player's Bonacich centrality.

Holding $b_i(\mathbf{g},a)$ fixed, $c_i(\mathbf{g},a)$ decreases with the proportion of i's Bonacich centrality due to self-loops $m_{ii}(\mathbf{g},a)/b_i(\mathbf{g},a)$.

The key player

Theorem. If $\beta > \lambda\mu_1(\mathbf{G})$, the key player i^* is such that $c_{i^*}(\mathbf{g}, \lambda/\beta) \geq c_i(\mathbf{g}, \lambda/\beta)$, for all i.

The key player is the one with highest inter-centrality.

In fact, $c_i(\mathbf{g}, a) = b_i(\mathbf{g}, a) + \sum_{j \neq i} [b_j(\mathbf{g}, a) - b_j(\mathbf{g}^{-i}, a)]$.

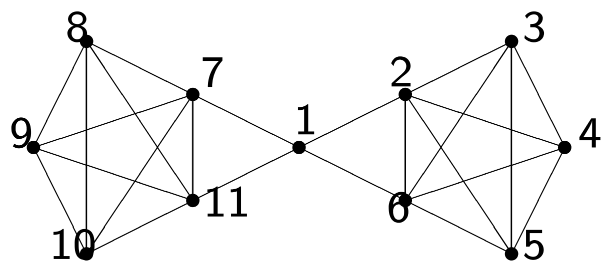
Is the key player always the more active criminal?

Holding $b_i(g, \phi)$ fixed, the intercentrality $d_i(g, \phi)$ of player i *decreases* with the proportion $m_{ii}(g, \phi)/b_i(g, \phi)$ of i's Bonacich centrality due to self-loops, and *increases* with the fraction of i's centrality amenable to out-walks.

Not always true.

Consider this network g with eleven criminals.

Figure 1: A bridge network



We distinguish three different types of equivalent actors in this network, which are the following:

Type	Criminals
1	1
2	2, 6, 7 and 11
3	3, 4, 5, 8, 9 and 10

Role of location in the network

Criminals are ex identical: $\boldsymbol{\alpha} = \mathbf{1}$

$$\mathbf{b_1}\left(g,\phi\right) = \left(\mathbf{I} - \phi\mathbf{G}\right)^{-1}\mathbf{1}$$

$y_i^* = b_{1_i}\left(g,\phi\right)$ and $d_{i^*}(g,\phi) = b_1(g,\phi) - b_1^{[-i]}(g,\phi)$.

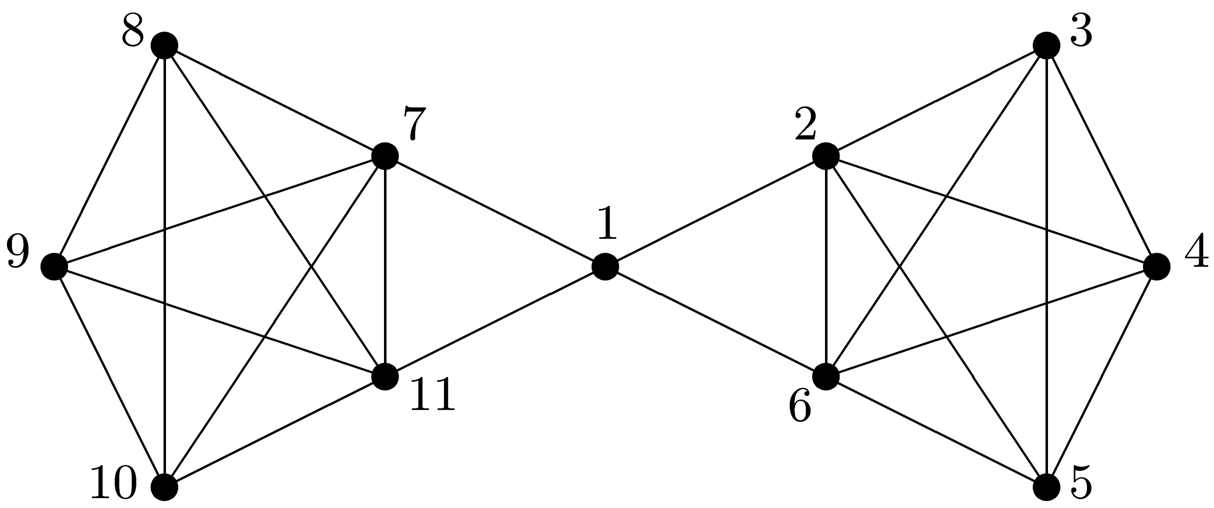
Take $\phi = 0.2$.

Table 1a: Key player versus Bonacich centrality in a bridge network

Player Type	1	2	3
$y_i = b_i$	8.33	9.17*	7.78
d_i	41.67*	40.33	32.67

From individual key player to key group

The key group maximizes a group inter-centrality measure $c_S(\mathbf{g}, a)$.



S	θ_S
$\{2,7\}^*$	67.22
$\{2,8\}$	64.01
$\{3,8\}$	59.39
$\{1,2\}$	56.66
$\{2,6\}$	50.41
$\{2,3\}$	46.96
$\{3,4\}$	42.15

Social multiplier

Peer effects

- Do social interactions matter for aggregate outcomes?
- One answer: peer effects amplify elasticities through a social multiplier.
- Peers can affect individual behavior in many domains.
  - Crime, student outcomes, workplace outcomes, investment, consumption, political views, technology adoption, group participation, etc.
- If these interactions feed back into each other, we get two implications:
  1. Small environmental changes have amplified effects through altered peer behavior.
  2. Excess variance in outcomes across groups.

Model of peer effects

- Economy partitioned into non-overlapping groups; group of agent i, denoted $G(i)$, has N members.
- Action of i determined as

$$A_i = \alpha X_i + \beta \bar{X}_{-i} + \gamma \bar{A}_{-i} + \epsilon_i.$$

  where X_i is observed and ϵ_i unobserved individual characteristics.
- Two types of peer effects:
  - β measures “contextual effect” that peer characteristics affect my action.
  - γ is “endogenous effect” of friend action.
- Example: higher GPA because my peer works hard (endogenous effect), or because she is smart (contextual effect).

Econometric problems

Suppose we regress GPA on roommate's GPA and characteristics.

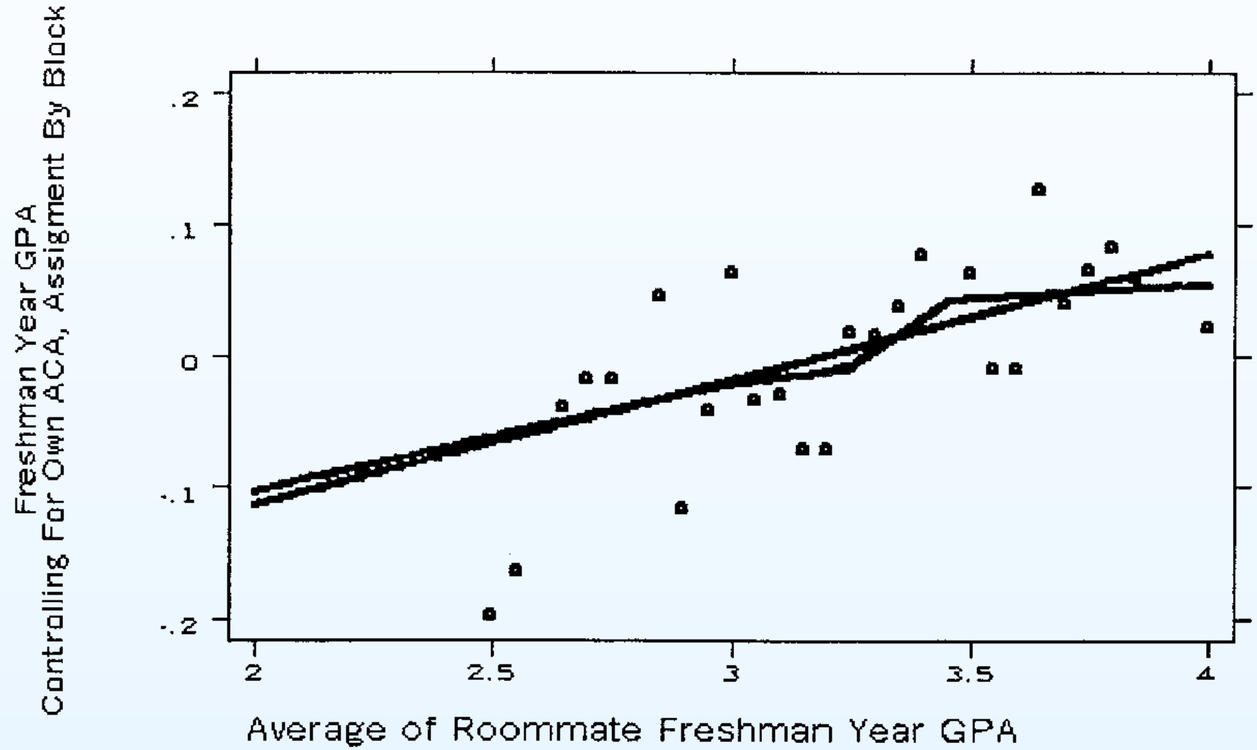
$$A_i = \alpha X_i + \beta \bar{X}_{-i} + \gamma \bar{A}_{-i} + \epsilon_i.$$

Three problems in identifying coefficients in this regression:

1. Reflection problem: group members' actions determined jointly in equilibrium, hence ϵ_i is correlated with $\bar{A}_{-i}$.
2. Selection: if agents select similar peers, ϵ_i will be correlated with both $\bar{X}_{-i}$ and $\bar{A}_{-i}$.
3. Common shocks: if peer group exposed to same shock, ϵ_i will be correlated with ϵ_{-i} and hence A_{-i}.

Example: Dartmouth data.

Regression of own GPA on roommate GPA



- Reflection: slope coefficient cannot be given causal interpretation.
- Selection: no, because roommates randomized.
- Common shocks: if a drummer lives above room.

Toy model

- Focus only on endogenous peer effects. Action of i is

$$A_i = \theta_i + \frac{\gamma}{N-1} \sum_{j \in G(i), j \neq i} A_j$$

where $\theta_i = \sum_k \beta_k X_k^i + \epsilon_i$ represents exogenous forces.

- Average across group members yields

$$\bar{A} = \frac{1}{1-\gamma} \bar{\theta}$$

- Using this we can derive:

$$A_i = \left(1 + \frac{\gamma^2}{(1-\gamma)(N-1+\gamma)}\right) \theta_i + \frac{\gamma}{N-1+\gamma} \sum_{j \in G(i), j \neq i} \theta_j .$$

Social multiplier

- Suppose that (1) X independent across people, (2) ϵ orthogonal to X.
- Social multiplier is ratio of “aggregate coefficients” to “individual coefficients” in regression of A on X:

$$\frac{N-1+\gamma}{(1-\gamma)(N-1)+\gamma}.$$

- Approaches $1/(1-\gamma)$ as N grows large.
- Two predictions:
  1. Higher coefficients in regressions at higher level of aggregation.
  2. “Excess variance” in $\bar{A}$ across groups.
- In reality, X and ϵ will often be correlated at group level.

Beer in high school and joining fraternity in college

TABLE 1. SOCIAL MULTIPLIERS IN FRATERNITY PARTICIPATION
DARTMOUTH ROOMMATE DATA: EFFECT OF BACKGROUND CHARACTERISTICS ON PARTICIPATION IN FRATERNITIES AT THE INDIVIDUAL LEVEL AND THREE LEVELS OF AGGREGATION

	(1) Member of fraternity or sorority	(2) Room average level membership	(3) Floor average membership	(4) Dorm average membership
Drank beer in high school	0.1040	0.0984	0.1454	0.2320
	(0.0258)	(0.0399)	(0.0812)	(0.1930)
Constant	0.0482	0.1980	0.7993	2.2277
	(0.1455)	(0.2266)	(0.4594)	(1.1421)
R-squared	0.04	0.05	0.03	0.08
Observations	1579	700	197	57
Average group size	1	2.3	8.0	28

- Random assignment addresses identification problems.
- Evidence supports a multiplier in fraternity membership.
  - Inverting formula, obtain $\gamma = .38$ at the floor and $\gamma = .56$ at the dorm level.

Lessons and next steps

- Social interactions can amplify economic forces.
- But there are econometric problems. Hence **randomized evaluation** important for network studies!
  - Growing evidence on peer effects in crime, technology adoption, 401K membership, incentives at work.
  - Less direct evidence about the social multiplier.
- Peer effects likely operate through friends.
  - Should incorporate network structure in model/evidence.
- For some peer effects, economic mechanism provides more structure.
  - Information transmission: we can use models of how agents learn.

Calvó-Armengol, Patacchini and Zenou

Peer Effects and Social Networks in Education

Review of Economic Studies 2009

Empirics: Test predictions of our peer-effect model

Dataset of friendship networks in the United States from the National Longitudinal Survey of Adolescent Health (AddHealth).

Role of network location in education.

Empirical issues: endogenous network formation, unobserved individual, school and network heterogeneity.

Richness of the information provided by the AddHealth data

Use of both within and between network variations

11,491 pupils distributed over 181 networks.

Direct estimation of the model:

A one-standard deviation increase in the Katz-Bonacich index translates into roughly 7 percent of a standard deviation in education outcome.

Influence of peers on education outcomes

Standard approach: instrumental variables (e.g. Evans et al., 1992)

or a natural experiment (e.g. Angrist and Lavy, 1999; Sacerdote, 2001; Zimmerman, 2003)

Nearly no studies that have adopted a more structural approach to test a specific peer effect model in education

(Glaeser et al. (1996) peer effecs in criminal behavior).

Here:

Stress the role of the structure of social networks in explaining individual behavior.

Build a theoretical model of peer effects

Direct empirical test of our model on the network structure of peer effects

We characterize the exact conditions on the geometry of the peer network, so that the model is fully identified

Utility function

$$u_i(\mathbf{y}^0, \mathbf{z}; \mathbf{g}) = \theta_i \, y_i^0 - \frac{1}{2}\left(y_i^0\right)^2 + \mu g_i z_i - \frac{1}{2} z_i^2 + \phi \sum_{j=1}^{n} g_{ij} z_i z_j .$$

$$\theta_i(\mathbf{x}) = \underbrace{\sum_{m=1}^{M} x_i^m}_{i\text{'s observable heterogeneity}} + \underbrace{\frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} g_{ij} \, x_j^m}_{i\text{'s friends (average) observable heterogeneity}}$$

x_i^m set of M variables accounting for observable differences in individual, neighborhood and school characteristics of individual i.

Equilibrium behavior Nash equilibrium

$$\max_{y_i^0, z_i} u_i(\mathbf{y}^0, \mathbf{z}; \mathbf{g}) = \theta_i\, y_i^0 - \frac{1}{2}\left(y_i^0\right)^2 + \mu g_i z_i - \frac{1}{2} z_i^2 + \phi \sum_{j=1}^{n} g_{ij} z_i z_j .$$

Best reply function for each $i = 1, ..., n$:

$$y_i^{0*}(\mathbf{x}) = \theta_i(\mathbf{x}) = \sum_{m=1}^{M} x_i^m + \frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} g_{ij}\, x_j^m$$

$$z_i^*(\mathbf{g}) = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j$$

Individual outcome (educational achievement) is the sum of these two different efforts:

$$y_i^*(\mathbf{x}, \mathbf{g}) = \underbrace{y_i^{0*}(\mathbf{x})}_{\text{idiosyncratic}} + \underbrace{z_i^*(\mathbf{g})}_{\text{peer effect}} .$$

We can decompose additively individual behavior into an exogenous part and an endogenous peer effect component that depends on the individual under consideration:

Denote by $\omega(\mathbf{g})$ the largest eigenvalue of the adjacency matrix $\mathbf{G} = \left[g_{ij}\right]$ of the network.

Proposition 3.1 *Suppose that $\phi\omega(\mathbf{g}) < 1$. Then, the individual equilibrium outcome is uniquely defined and given by:*

$$y_i^* (\mathbf{x}, \mathbf{g}) = \theta_i (\mathbf{x}) + \frac{\mu}{\phi} b_i (\mathbf{g}, \phi) . \tag{3}$$

Rewriting (3) as

$$y_i^* (\mathbf{x}, \mathbf{g}) = \left(1 + \frac{\mu}{\phi} \frac{b_i (\mathbf{g}, \phi)}{\theta_i^0 (\mathbf{x})} \right) y_i^{0*} (\mathbf{x}) ,$$

Peer influence acts as a multiplier on the behavior of the isolated individual.

Alternative formulation of the model Our utility function: each individual i chooses two different effort levels, y_i^0 and z_i.

Assume utility function with only one type of effort z_i:

$$u_i(\mathbf{z};\mathbf{g}) = \mu g_i z_i - \frac{1}{2} z_i^2 + \phi \sum_{j=1}^{n} g_{ij} z_i z_j. \tag{4}$$

Best reply function for each individual i

$$z_i^*(\mathbf{g}) = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j$$

When $\phi\omega(\mathbf{g}) < 1$

$$z_i^*(\mathbf{g}) = \frac{\mu}{\phi} b_i(\mathbf{g}, \phi)$$

Educational achievement of each individual i:

$$y_i^*(\mathbf{x}, \mathbf{g}) = \theta_i(\mathbf{x}) + z_i^*(\mathbf{g})$$

$\theta_i(x)$ contextual effects

4 Data

Unique database on friendship networks from the National Longitudinal Survey of Adolescent Health (AddHealth).

Adolescents' behavior in the United States: Data on students in grades 7-12 from a nationally representative sample of roughly 130 private and public schools in years 1994-95.

Respondents' demographic and behavioral characteristics, education, family background and friendship.

Friendship networks

Friendship information is based upon actual friends nominations.

Pupils were asked to identify their best friends from a school roster (up to five males and five females).

The limit in the number of nominations is not binding.

Less than 1% of the students in our sample show a list of ten best friends.

Less than 3% a list of five males and roughly 4% name five females.

On average, they declare to have 5.48 friends with a small dispersion around this mean value (the standard deviation is equal to 1.29).

The corresponding figures for male- and female-friends are 2.78 (with standard deviation equal to 1.85) and 3.76 (with standard deviation equal to 1.04).

A link exists between two friends if at least one of the two individuals has identified the other as his/her best friend (undirected networks)

We also consider directed networks: 14% of relationships are not reciprocal.

For each school, we obtain all the networks of (best) friends.

Education achievements

Grade achieved by each student in mathematics, history and social studies and science, ranging from D or lower to A, the highest grade (recoded 1 to 4).

School performance index

Final sample: 11,964 pupils distributed over 199 networks.

Descriptive statistics

Table 1: Descriptive statistics

(11,964 pupils; 199 networks)

	Mean	St. Dev.	Min	Max
Female	0.41	0.35	0	1
Black or African American	0.17	0.31	0	1
Other races	0.12	0.15	0	1
Age	15.29	1.85	10	19
Religion practice	3.11	1.01	1	4
Health status	3.01	1.77	0	4
School attendance	3.28	1.86	1	6
Student grade	9.27	3.11	7	12
Organized social participation	0.62	0.22	0	1
Motivation in education	2.23	0.88	1	4
Relationship with teachers	0.12	0.34	0	1
Social exclusion	2.26	1.81	1	5
School attachment	2.59	1.76	1	5
Parental care	0.69	0.34	0	1
Household size	3.52	1.71	1	6
Two married parent family	0.41	0.57	0	1
Single parent family	0.23	0.44	0	1
Public assistance	0.12	0.16	0	1
Mother working	0.65	0.47	0	1

	Mean	St. Dev.	Min	Max
Parental education	3.69	2.06	1	5
Parent age	40.12	13.88	33	75
Parent occupation manager	0.11	0.13	0	1
Parent occupation professional or technical	0.09	0.21	0	1
Parent occupation office or sales worker	0.26	0.29	0	1
Parent occupation manual	0.21	0.32	0	1
Parent occupation military or security	0.09	0.12	0	1
Parent occupation farm or fishery	0.04	0.09	0	1
Parent occupation retired	0.06	0.09	0	1
Parent occupation other	0.11	0.16	0	1

Table 1: Descriptive statistics (continued)

	Mean	St. Dev.	Min	Max
Neighborhood quality	2.99	2.02	1	4
Residential building quality	2.95	1.85	1	4
Neighborhood safety	0.51	0.57	0	1
Residential area suburban	0.32	0.38	0	1
Residential area urban - residential only	0.18	0.21	0	1
Residential area commercial properties - retail	0.12	0.15	0	1
Residential area commercial properties - industrial	0.13	0.18	0	1
Residential area type other	0.19	0.25	0	1
Friend attachment	0.49	0.54	0	1
Friend involvement	1.88	1.56	0	3
Friend contacts	0.89	0.12	0	1
Physical development	3.14	2.55	1	5
Self esteem	3.93	1.33	1	6

4.1 Descriptive evidence

Mean and the standard deviation of network size are 60.42 and 24.48, respectively.

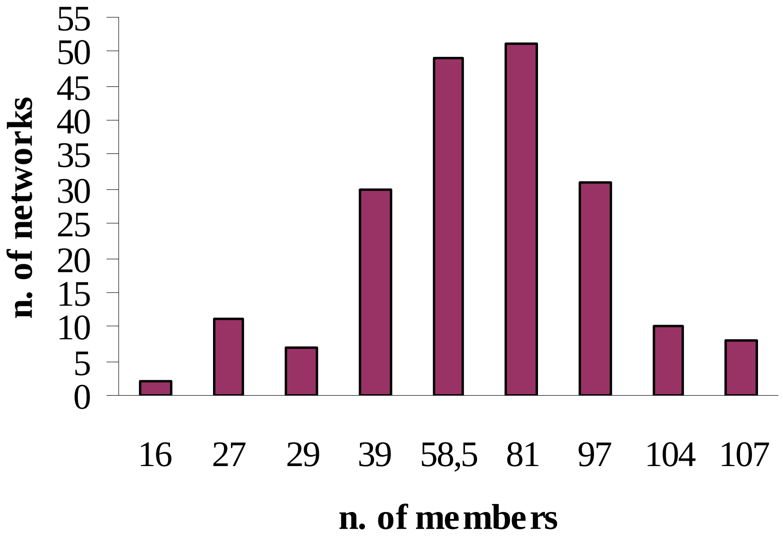


Figure 2. The empirical distribution of adolescent networks

Smallest network in our sample.

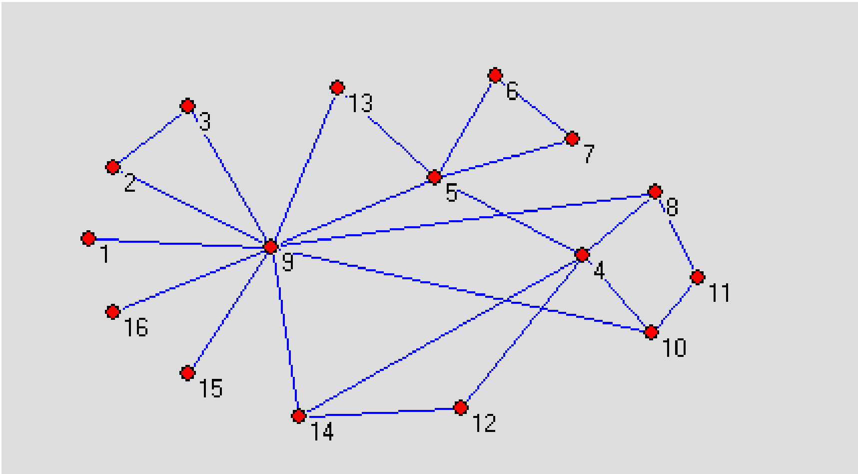


Figure 3. Smallest network of adolescents ($n = 16$)

The largest network in our sample is almost seven times bigger and has 107 members.

5 Empirical strategy

Proposition 3.1: Actual empirical relationship between $b_i(\mathbf{g}, \phi)$ and the observed effort level y_i^*.

Maximally connected components:

Two agents in a network are either directly linked or indirectly linked through a sequence of intermediate agents (connectedness).

Two agents in different networks cannot be connected through any such sequence (maximality).

$\mathbf{g}_\kappa$ network encompassing n_κ different individuals.

K such networks in the economy. $\sum_{\kappa=1}^{K} n_\kappa = n$.

Theory:

$$y_i(\mathbf{x},\mathbf{g}) = \underbrace{\sum_{m=1}^{M} x_i^m + \frac{1}{g_i}\sum_{m=1}^{M}\sum_{j=1}^{n} g_{ij}\, x_j^m}_{\text{idiosyncratic}} + \underbrace{z_i(\mathbf{x},\mathbf{g})}_{\text{peer effect}}.$$

$$z_i = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j,\ i = 1, ..., n$$

Empirical counterpart

$$\begin{aligned} y_{i,\kappa} &= \sum_{m=1}^{M} \beta_m x_{i,\kappa}^m + \frac{1}{g_{i,\kappa}} \sum_{m=1}^{M}\sum_{j=1}^{n_\kappa} \gamma_m g_{ij,\kappa} x_{j,\kappa}^m + \eta_\kappa + \varepsilon_{i,\kappa}, \qquad (5) \\ \varepsilon_{i,\kappa} &= \mu g_{i,\kappa} + \phi \sum_{j=1}^{n_\kappa} g_{ij,\kappa} \varepsilon_{j,\kappa} + \upsilon_{i,\kappa},\ \ i = 1, ..., n;\ \ \kappa = 1, ..., K, \end{aligned}$$

η_κ is an (unobserved) network-specific component (constant over individuals in the same network)

$\varepsilon_{i,\kappa}$ residual of individual i's level of activity in the network g that is *not* accounted for neither by individual heterogeneity and contextual effects nor by (unobserved) network-specific components.

$\sum_{j=1}^{n_\kappa} g_{ij,\kappa}\varepsilon_{j,\kappa}$ is the spatial lag term and ϕ is the spatial autoregressive parameter.

Matrix notation:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{DGX}\boldsymbol{\gamma} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \\
\boldsymbol{\varepsilon} &= \mu\mathbf{G1} + \phi\mathbf{G}\boldsymbol{\varepsilon} + \boldsymbol{\nu},
\end{aligned}$$

Our model is a variation of the Anselin (1988) spatial error model.

Using the Maximum Likelihood approach (see, e.g. Anselin, 1988), we estimate jointly $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}$, $\widehat{\phi}$, $\widehat{\mu}$.

These values: measure the relative importance of

individual characteristics, $\widehat{\beta}_1, ..., \widehat{\beta}_m$ (e.g. parental education, school and neighborhood quality),

contextual effects, $\widehat{\gamma}_1, ..., \widehat{\gamma}_m$ (e.g. average parental education of each individual's best friends, etc.),

the individual Katz-Bonacich centrality index, $\widehat{\phi}$ and $\widehat{\mu}$,

in shaping individuals' behavior (equation (3) in Proposition 3.1).

5.1 Identification of peer effects

Assessment of the effects of peer pressure on individual behavior, i.e. the identification of endogenous social effects

Typically characterized by econometric issues

First problem: The endogenous sorting of individuals into groups

Second problem: The reflection problem (Manski, 1993).

6 Empirical results

6.1 The Katz-Bonacich network centrality index

Model (5) is estimated using the Maximum Likelihood approach.

	ML (with network fixed effects)
Number of best friends (μ)	0.0314**
	(0.0149)
Peer effects (ϕ)	0.5667***
	(0.1433)
Individual socio-demographic variables	yes
Family background variables	yes
Protective factors	yes
Residential neighborhood variables	yes
Contextual effects	yes
School fixed effects	yes
$R^2 = 0.8987$	

Notes:
- Number of observations: 2,079,871 (11,491pupils, 181networks)
- Regressions are weighted to population proportions
- Standard errors in parentheses.
Coefficients marked with one (two) [three] asterisks
are significant at 10 (5) [1] percent level

Estimated μ and ϕ are both positive and highly statistically significant.

Calculate the Katz-Bonacich measure by fixing the value of ϕ at the point estimate $\widehat{\phi}$.

Derived Katz-Bonacich measures range from 0.32 to 3.48, with an average of 1.65 and a standard deviation of 2.79.

The estimated impact of this variable on education outcomes that is predicted by the theory, i.e. $\widehat{\mu}/\widehat{\phi}$ is statistically significant.

A one-standard deviation increase in the Katz-Bonacich index translates into roughly 7 percent of a standard deviation in education outcome

This effect is about 17 percent for parental education.

7.1 An alternative measure of network unit centralities

Two dimensions of centrality, connectivity and betweenness.

Bonacich centrality is an index of connectivity since it counts the number of *any* path stemming from a given node, not just optimal paths.

Degree centrality The individual-level *degree centrality* is simply each individual's number of direct friends:

$$\delta_i(\mathbf{g}_\kappa) = g_i = \sum_{j=1}^{n} g_{ij}$$

To compare networks of different sizes, this measure is normalized to be in an interval from 0 to 1, where values 0 and 1 indicate the smallest and the highest possible centrality.

$$\delta_i^*(\mathbf{g}_\kappa) = \frac{g_i}{n_\kappa - 1} = \frac{\sum_{j=1}^{n} g_{ij}}{n_\kappa - 1}$$

In our data, the normalized degree centrality index $\delta_i^*(\mathbf{g}_\kappa)$ has a mean equal to 0,35 and a standard deviation equal to 0,18.

Closeness centrality The standard measure of centrality according to *closeness* of individual i is given by:

$$c_i(\mathbf{g}_\kappa) = \frac{1}{\sum_j d_{ij,\kappa}}$$

where d_{ij} is the geodesic distance (length of the shortest path) between individuals i and j.

As a result, the closeness centrality of individual i is the inverse of the sum of geodesic distances from i to the $n-1$ other individuals (i.e. the reciprocal of its "farness").

Compared to degree centrality, the closeness measure takes into account not only direct connections among individuals but also indirect connections.

Compared to the Katz-Bonacich centrality, the closeness measure assumes a weight of one to each indirect connection.

Relative closeness centrality measure as:

$$c_i^*(\mathbf{g}_\kappa) = \frac{n_\kappa - 1}{\sum_j d_{ij,\kappa}}$$

where $n_\kappa - 1$ is the maximum possible distance between two individuals in network κ. This measure takes value between 0 and 1.

The mean and the standard deviation in our data of this normalized index are 0.49 and 0.27.

Betweenness centrality measure of agent i in a network component $\mathbf{g}_\kappa$

$$f_i(\mathbf{g}_\kappa) = \sum_{j<l} \frac{\#\text{ of shortest paths between } j \text{ and } l \text{ through } i \text{ in } \mathbf{g}_\kappa}{\#\text{ of shortest paths between } j \text{ and } l \text{ in } \mathbf{g}_\kappa}$$

where j and l denote two given agents in $\mathbf{g}_\kappa$.

For undirected networks, a normalized version of this measure is:

$$f_i^*(\mathbf{g}_\kappa) = \frac{f_i(\mathbf{g}_\kappa)}{(n_\kappa - 1)(n_\kappa - 2)/2},$$

where n_κ is the size of the network $\mathbf{g}_\kappa$.

Betweenness is a parameter-free network measure.

In our data, the normalized betweenness measure f_i^* has a mean equal to 0.45 and a standard deviation equal to 0.51.

Table 4. Explanatory power of unit centrality measures

Dependent variable: school performance index

	OLS	OLS	OLS
Degree centrality	0.2508*	-	-
	(0.1475)		
Closeness centrality	-	0.2892	-
		(0.2599)	
Beetweenness centrality	-	-	0.0621
			(0.0698)
Individual socio-demographic variables	yes	yes	yes
Family background variables	yes	yes	yes
Protective factors	yes	yes	yes
Residential neighborhood variables	yes	yes	yes
Contextual effects	yes	yes	yes
School fixed effects	yes	yes	yes
R^2	0.7958	0.8202	0.8001

Notes:

- Number of observations: 2,079,871 (11,491 pupils, 181 networks)
- Control variables are those listed in Appendix 3.
- Network fixed-effects OLS estimators are reported. They are within-group estimates where individuals are grouped by networks
- Regressions are weighted to population proportions

Out of the three measures, only degree centrality shows a statistically slighly significant impact (i.e., at the 10% significance level).

When such an effect is translated in terms of standard deviations, its impact on educational outcomes is not even one third of the effect exerted by the Katz-Bonacich centrality index (roughly 2.1 percent versus 7 percent).

Two main explanations

1) Bonacich centrality is not an arbitrary network measure but results from a positive analysis that maps network topology to equilibrium behavior.

2) Betweenness centrality is a parameter-free network index. It only depends on the network geometry.

A Dynamic Model of Network Formation with Strategic Interactions

Michael D. König
ETH Zurich,

Claudio J. Tessone
ETH Zurich,

Yves Zenou
Stockholm University and IFN

November 2010

Structure of real-life networks: characterized by low diameter (small world), high clustering, and “scale-free” degree distributions.

Need to analyze *how* and *why* networks form,

the impact of network structure on agents’ outcomes,

the evolution of networks over time.

This is the aim of this paper

Network formation: Two literatures

Random network formation: mainly dynamic and links are formed randomly

(Erdös and Rényi, PM 1959, Albert and Barabási , RMP 2002; Jackson and Rogers, AER 2004).

Strategic network formation: mainly static (game theory, pairwise stability)

(Myerson, 1991; Jackson and Wolinsky JET 1996; Bala and Goyal, Econometrica 2000)

Random approach: insight into *how* networks form (i.e. matches the characteristics of real-life networks)

Strategic approach: *Why* networks form

Games on networks: take the network as given and study how the network structure impacts on outcomes and individual decisions.

(Ballester, Calvó-Armengol, Zenou, Econometrica 2006, Bramoulle and Kranton, JET 2008; Galeotti, Goyal, Jackson, Vega-Redondo, Yariv, Restud 2009)

Here: introduce strategic interactions in a non-random dynamic network formation game.

Two-stage game at each period of time:

First stage: agents play their equilibrium contributions proportional to their Bonacich centrality (as in Ballester et al, 2006)

Second stage: a randomly chosen agent can update her linking strategy by creating a new link or deleting an existing link as a local best response to the current network.

Result:

At each period of time, the network generated by this dynamic formation process is a *nested split graph*.

Nested split graphs have a very nice and simple structure that make them very tractable to work with.

Here: A complex dynamic network formation model can be characterized by a simple structure in terms of networks it generates.

Result:

There exists a unique stationary network, which is a nested split graph.

We show under which conditions these networks emerge and that there exists a sharp transition between hierarchical and flat network structures.

Result:

Our model is able to match the characteristics of most real-life networks.

We show that the stationary networks emerging in our link formation process are characterized by *short path length* with *high clustering* (small worlds), *exponential degree distributions* with *power law tails* and *negative degree-clustering correlation*.

Without capacity constraints in the number of links an agent can maintain: networks are *dissortative*.

With capacity constraints: networks are *assortative*.

A Network Game with Linear Quadratic Payoffs

- We consider a network game where payoffs are interdependent (Ballester et al., 2006)

- $\mathbf{A}$ is the symmetric $n \times n$ adjacency matrix of network $\mathbf{G}$.

- Each agent $i = 1, .., n$ selects an effort level $x_i \geq 0$ and obtains a payoff of ($\lambda > 0$):

$$u_i(x_1, ..., x_n) = \underbrace{x_i - \frac{1}{2}x_i^2}_{\text{own concavity}} + \underbrace{\lambda \sum a_{ij} x_i x_j}_{\text{local complementarities}}$$

Nash Equilibrium

Theorem:[3]

- An interior Nash equilibrium in pure strategies $x^* \in \mathbb{R}^n_+$ is such that for all $i = 1, ..., n$ we have that (FOC) $\frac{\partial u_i(\mathbf{x}^*)}{\partial x_i} = 0$. From this one can show that the Nash equilibrium is given by

$$x_i^* = b_i(G, \lambda)$$

  where $b_i(G, \lambda)$ is the Bonacich network centrality of parameter $\lambda < 1/\lambda_{\mathrm{PF}}(G)$ and $\lambda_{\mathrm{PF}}(G)$ the largest real eigenvalue in the network G.
- Equilibrium payoffs are given by

$$u_i(\mathbf{x}^*, G) = \frac{1}{2}(x_i^*)^2 = \frac{1}{2} b_i^2(G, \lambda).$$

[3] C. Ballester, A. Calvo-Armengol and Y. Zenou, *Who's Who in Networks. Wanted: The Key Player*, Econometrica 74(5):1403–1417, 2006

Formal definition of best responses of an agent in network $G(t)$

Definition 0.1 *Consider the current network* $G(t, L(t))$ *with agents* $N = \{1, ..., n\}$ *and links* $L(t)$*. Let* $G(t)+ij$ *be the graph obtained from* $G(t)$ *by the addition of the edge* $ij \notin G(t)$ *between agents* $i \in N$ *and* $j \in N$*. Further, let* $\boldsymbol{\pi}^*(G(t)) = (\pi_1^*(G(t)), ..., \pi_n^*(G(t)))$ *denote the profile of Nash equilibrium payoffs of the agents in* $G(t)$ *following from the payoff function with parameter* $\lambda < 1/\lambda_{PF}(G(t))$*.*

Then agent j *is a best response of agent* i *if* $\pi_i^*(G(t)+ij) \geq \pi_i^*(G(t)+ik)$ *for all* $j, k \in N \setminus (\mathcal{N}_i \cup \{i\})$*. Agent* j *may not be unique. The set of agent* i*'s best responses is denoted by* $BR_i(G(t))$*.*

Definition 0.2 *We define the network formation process* $(G(t))_{t=0}^{\infty}$, $G(t) = (N, L(t))$, *as a sequence of networks* $G(0), G(1), G(2), \ldots$ *in which at every step* $t = 0, 1, 2, \ldots\}$, *an agent* $i \in N$ *is uniformly selected at random. Then one of the following two events occurs:*

a. *With probability* $\alpha \in (0, 1)$ *agent* i *initiates a link to a best response agent* $j \in BR_i(G(t))$. *Then the link* ij *is created if* $i \in BR_j(G(t))$ *is a best response of* j, *given the current network* $G(t)$. *If* $BR_i(G(t))$ *is not unique, then* i *selects randomly one agent in* $BR_i(G(t))$.

b. *With probability* $1-\alpha$ *the link* $ij \in G(t)$ *is removed such that* $\pi_i^*(G(t) - ij) \leq \pi_i^*(G(t) - ik)$ *for all* $j, k \in \mathcal{N}_i$. *If agent* i *does not have any link then nothing happens.*

Stationary Networks: Characterization

The network formation process $(G(t))_{t \in T}$ is a Markov process. Denote by Ω the state space of $(G(t))_{t=0}^{\infty}$. Ω consists of all possible unlabeled nested split graphs on n nodes. It can be shown that the number of possible states is $|\Omega| = 2^{n-1}$. Thus the number of states is finite and the transition between states can be represented by a transition matrix $\mathbf{P}$.

$(G(t))_{t=0}^{\infty}$ is a Markov chain. Indeed, the network $G(t+1) \in \Omega$ is obtained from $G(t)$ by removing or adding a link to $G(t)$. Thus, the probability of obtaining $G(t+1)$ depends only on $G(t)$ and not on the previous networks for $t' < t$.

The number of possible networks $G(t)$ is finite for any discrete time $t \geq 0$ and the transition probabilities from a network $G(t)$ to $G(t+1)$ do not depend on t. Therefore, $(G(t))_{t=0}^{\infty}$ is a finite state, discrete, time-homogeneous Markov process.

Transition matrix

$$\mathbf{P}_{ij} = P(G(t+1) = G_j \mid G(t) = G_i) \text{ for any } G_i, G_j \in \Omega$$

Proposition 0.2 *The network formation process* $(G(t))_{t=0}^{\infty}$ *induces an ergodic Markov chain on the state space* Ω *with a unique stationary distribution* $\boldsymbol{\mu}$*. In particular, the state space* Ω *is finite and consists of all possible unlabeled nested split graphs on* n *nodes, where the number of possible states is given by* $|\Omega| = 2^{n-1}$*.*

Uniqueness: This Markov chain is irreducible and aperiodic.

There exists a unique stationary distribution μ satisfying: $\mu\,\mathbf{P} = \mu$

Notations:

$\{N(t)\}_{t=0}^{\infty}$: degree distribution

$N_d(t)$: Number of nodes with degree d at time t in $G(t)$

$n_d(t) = N_d(t)/n$: Proportion of nodes with degree d at time t in $G(t)$.

$n_d = \lim_{t\to\infty} \mathsf{E}[n_d(t)]$: Asymptotic expected value of $n_d(t)$

The following proposition determines the asymptotic degree distribution (i.e. t large enough) of the nodes in the independent sets for n large enough.

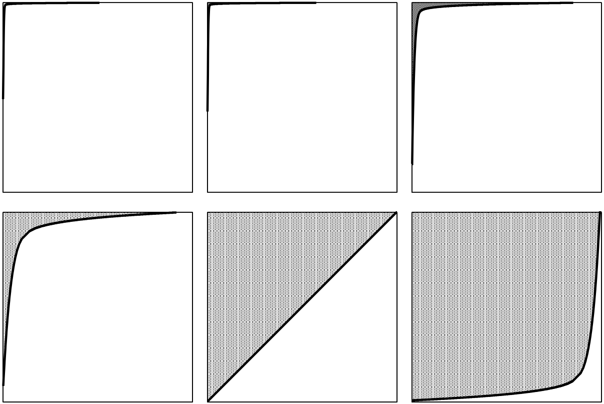
Proposition 0.4 *Let $0 \leq \alpha \leq 1/2$. Then the asymptotic expected proportion n_d of nodes in the independent sets with degrees, $d = 0, 1, ..., d^*$, for large n is given by:*

$$n_d = \frac{(1-2\alpha)}{(1-\alpha)} \left(\frac{\alpha}{1-\alpha} \right)^d$$

where

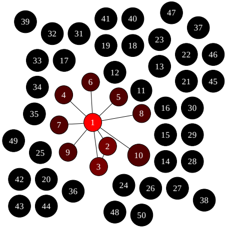
$$d^*(n, \alpha) = \frac{\ln \left(\frac{(1-2\alpha)n}{2(1-\alpha)} \right)}{\ln \left(\frac{1-\alpha}{\alpha} \right)}$$

"Expected" Adjacency Matrix

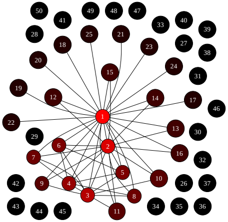


⇒ Low α, hierarchical, centralized network.

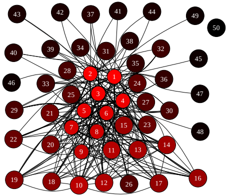
⇒ High α, highly decentralized network.



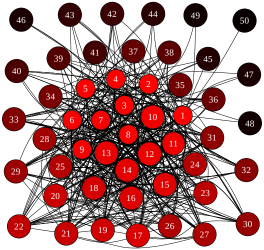
$\alpha = 0.2$



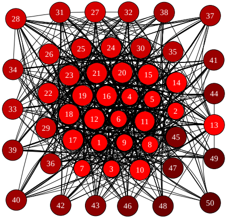
$\alpha = 0.4$



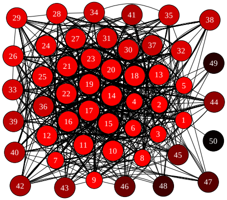
$\alpha = 0.48$



$\alpha = 0.495$



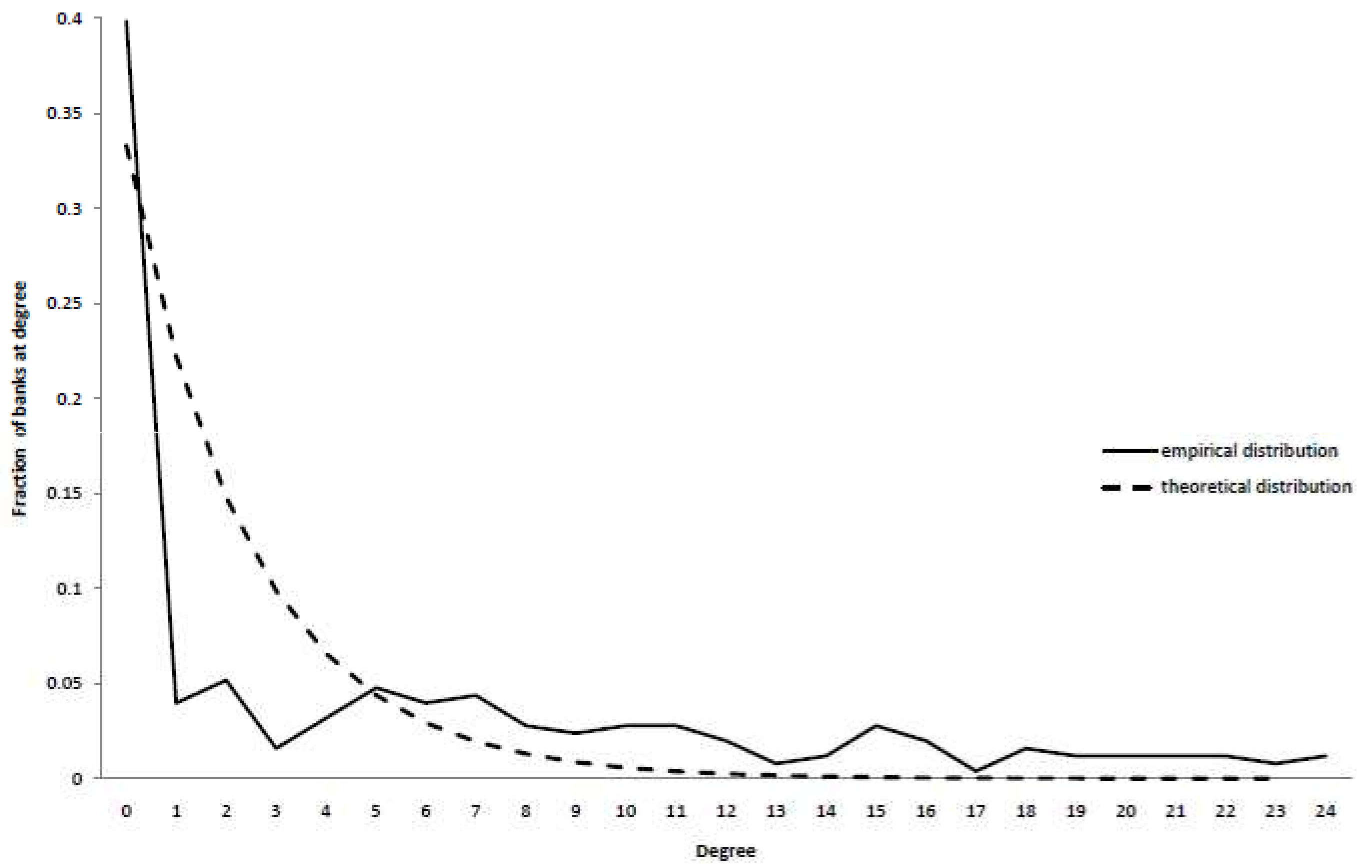
$\alpha = 0.5$



$\alpha = 0.52$

Example of Empirical v Theoretical Distribution

Figure 5a: Empirical vs Theoretical Degree Distribution – January 2002



Example of Empirical v Theoretical Distribution

Figure 5b: Empirical vs Theoretical Degree Distribution – December 2009

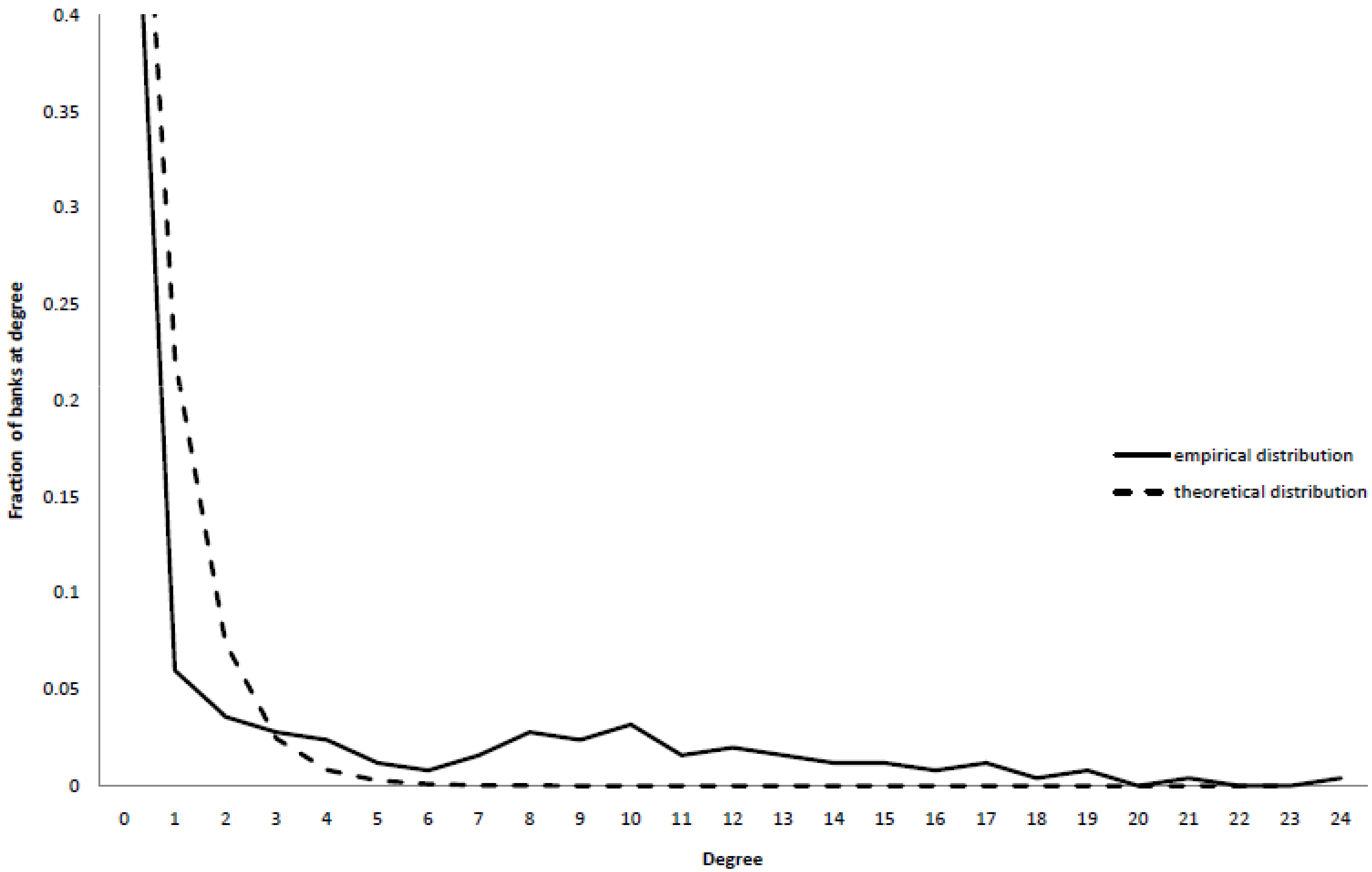


Figure 5b shows the empirical degree distribution of the network that existed on December 31, 2009. On the same figure, we plot the theoretical distribution generated by the empirical link formation probability on that day.

Whither now?

- Bridging random/mechanical – economic/strategic
- Networks in Applications
  - Diffusion of information, technology– relate to network structure
  - Labor, mobility, voting, trade, collaboration, crime, www, ...
- Empirical/Experimental
  - case studies lack economic variables, tie networks to outcomes,
  - enrich modeling of social interactions from a structural perspective
- Furthering game theoretic modeling, and random modeling
- Foundations and Tools– centrality, power, allocation rules, community structures, ...

Jackson, M.O. (2008), *Social and Economic Networks, PUP.*

Easley, D. and J. Kleinberg (2010), *Networks, Crowds, and Markets: Reasoning About a Highly Connected World,* CUP.

Goyal, S. (2007), *Connections: An Introduction to the Economics of Networks, PUP.*

Vega-Redondo, F. (2007), *Complex Social Networks,* CUP.