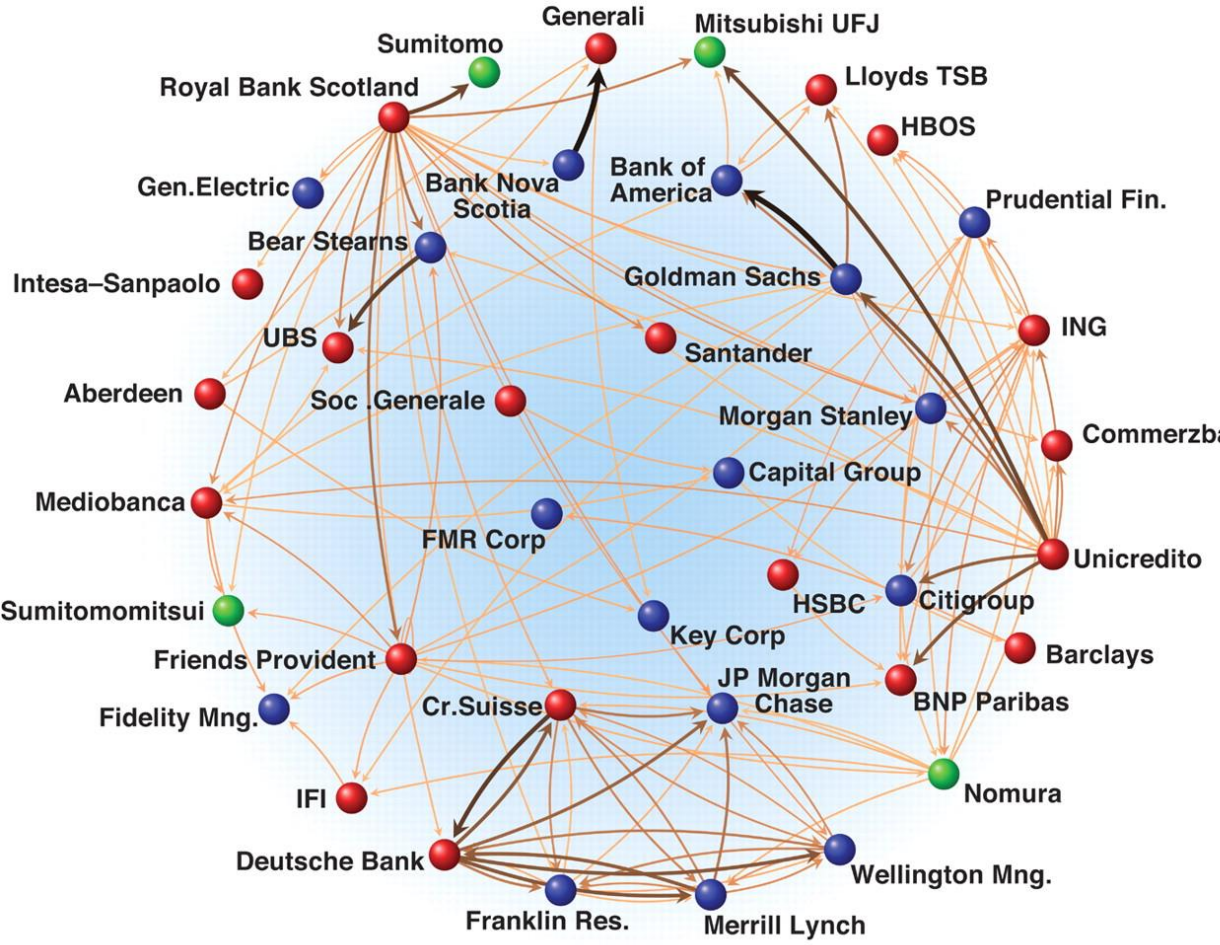
Business Network Analytics



F Schweitzer et al. Science 2009

Sep, 2017

Daning Hu
Department of Informatics
University of Zurich
Business Intelligence Research Group

What: Social Network Analysis

- Social network analysis (SNA) is a set of **metrics** and methods for systematically **describing,** modelling, and analyzing relationships among actors.

- Social network analysis (SNA)
  - is motivated by a structural intuition based on ties linking social actors
  - is grounded in systematic empirical data
  - draws heavily on graphic imagery
  - relies on the use of mathematical and/or computational models.

Social Network Analysis: Topology Analysis

- Network level topology analysis takes a macro perspective to **describe** the *physical properties of network structures*. Typical network topological measures include:
  - **Centrality Measures: Degree, Betweenness, and Closeness**
  - **Size and Density**,
  - **Average Degree**,
  - **Average Path Length**: on average, the number of steps it takes to get from one member of the network to another.
  - **Diameter**
  - **Clustering Coefficient**: a measure of an "all-my-friends-know-each-other" property; small-world feature

$$CC(i) = \frac{2E_i}{k_i(k_i - 1)}$$

$$CC = \sum_{i=1} ClusteringCoeff(i)$$

$k_i = C_d(i)$ = # of neighbors of node i

E_i = # of links actually exist between k_i nodes

Node Level Analysis: Node Centrality

- Node Centrality can be viewed as a measure of influence or importance of nodes in a network.

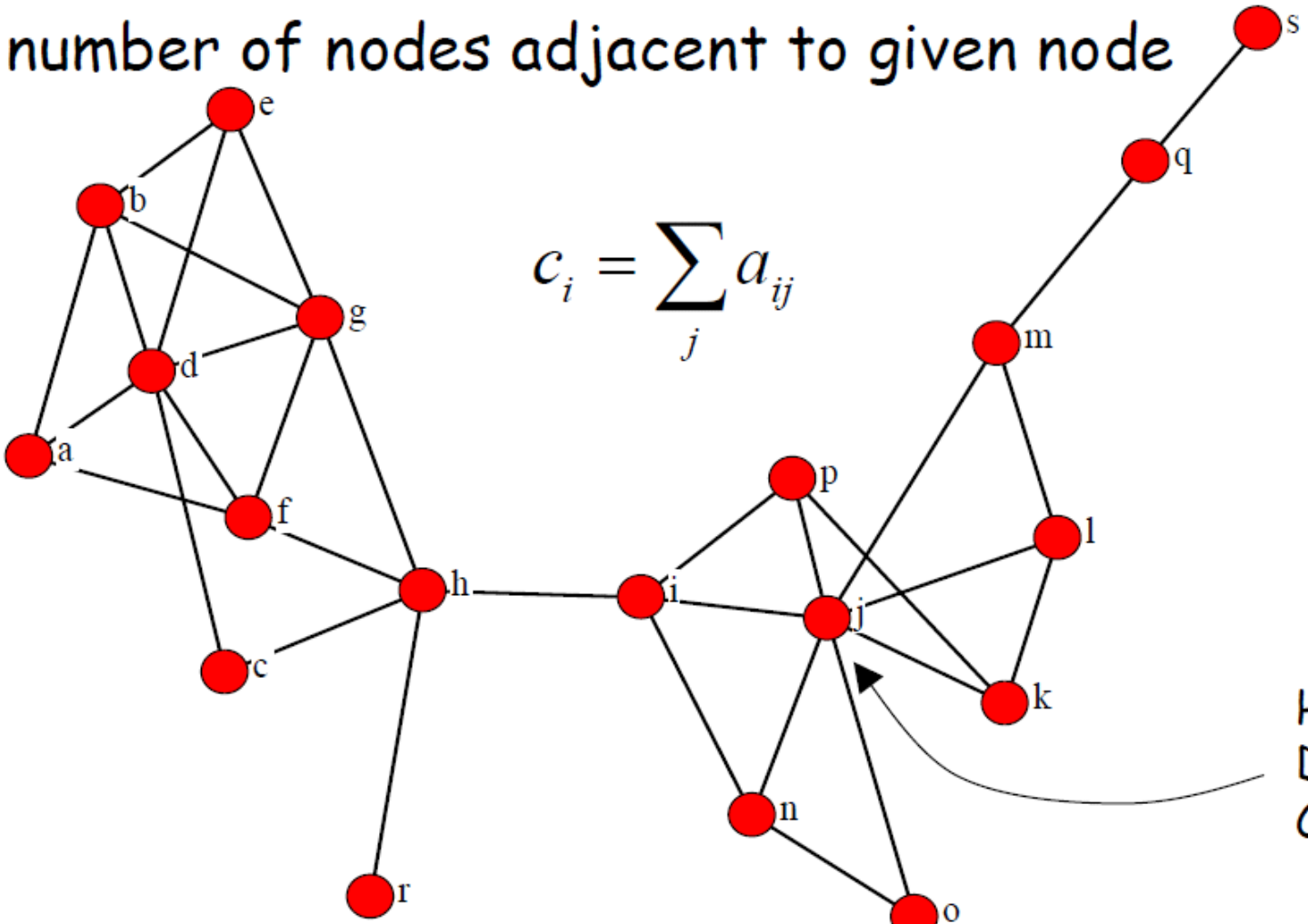
- Degree
  - the number of links that a node possesses in a network. In a **directed** network, one must differentiate between in-links and out-links by calculating in-degree and out-degree.

- Betweeness
  - the number of shortest paths in a network that traverse through that node.

- Closeness
  - the average distance that each node is from all other nodes in the network

Node Level Analysis: Degree Centrality

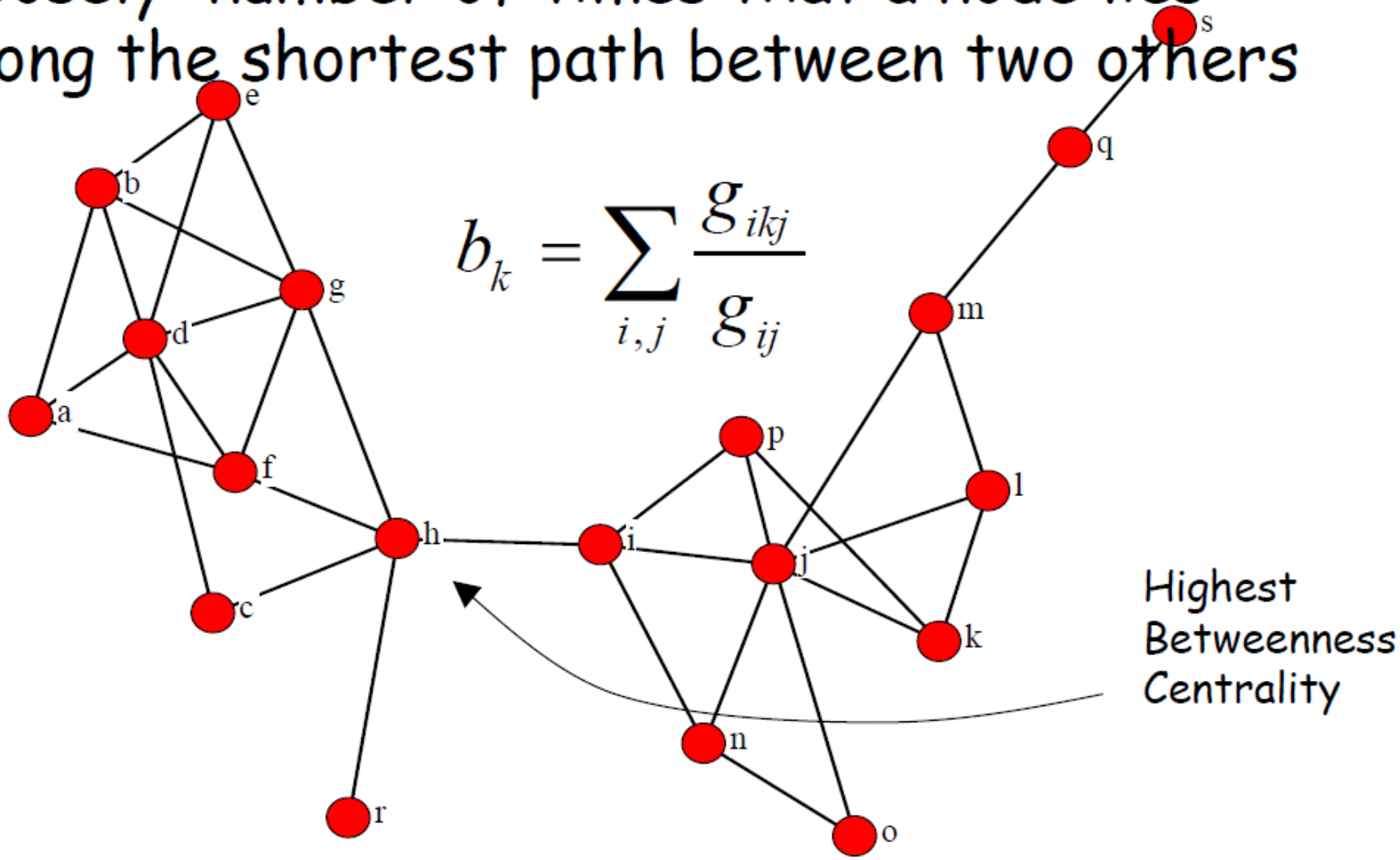
- The number of nodes adjacent to given node



From Steve Borgatti

Node Level Analysis: Betweenness Centrality

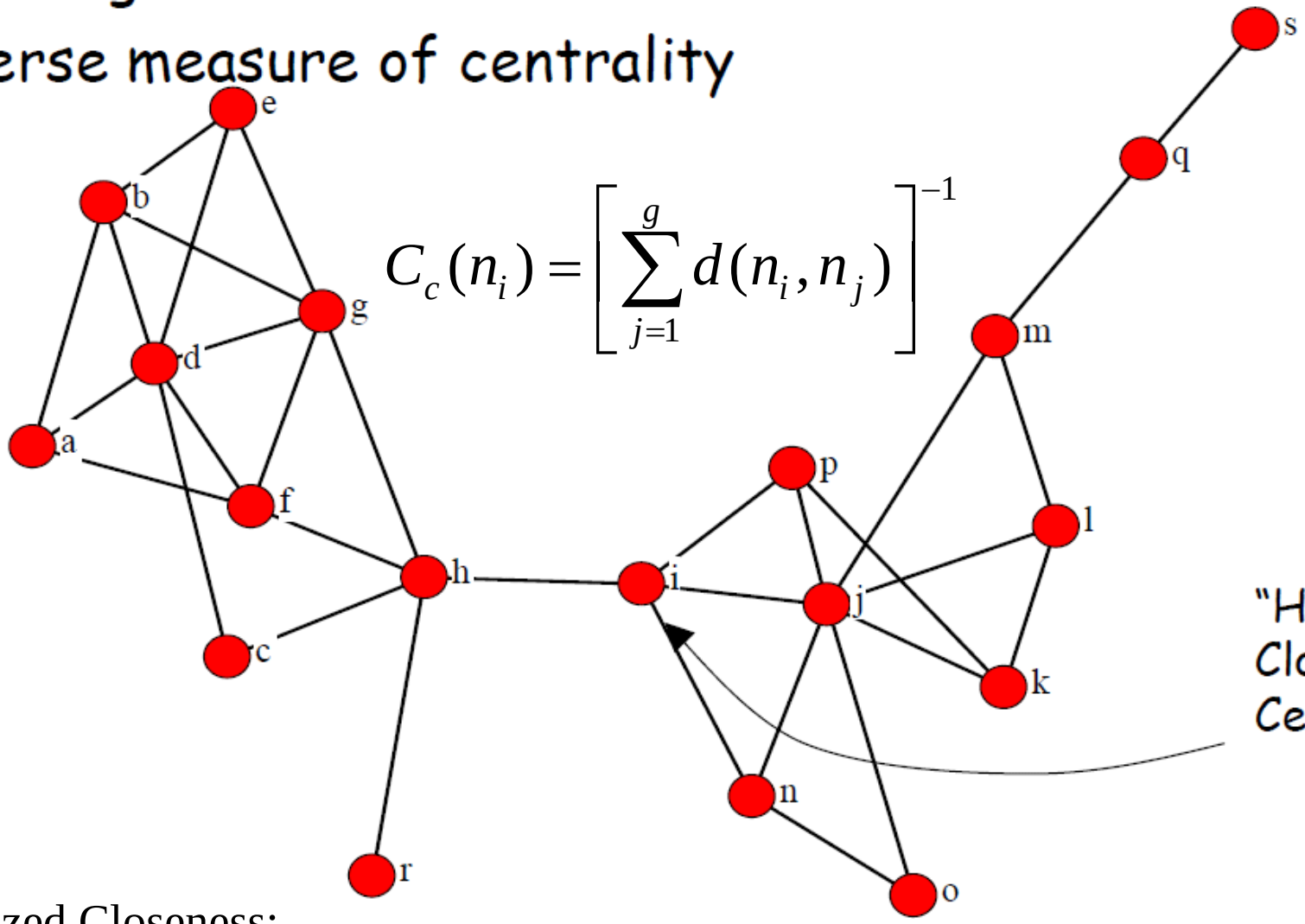
- Loosely: number of times that a node lies along the shortest path between two others



From Steve Borgatti

Node Level Analysis: Closeness Centrality

- Sum of geodesic distances to all other nodes
- Inverse measure of centrality



$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j) \right]^{-1}$$

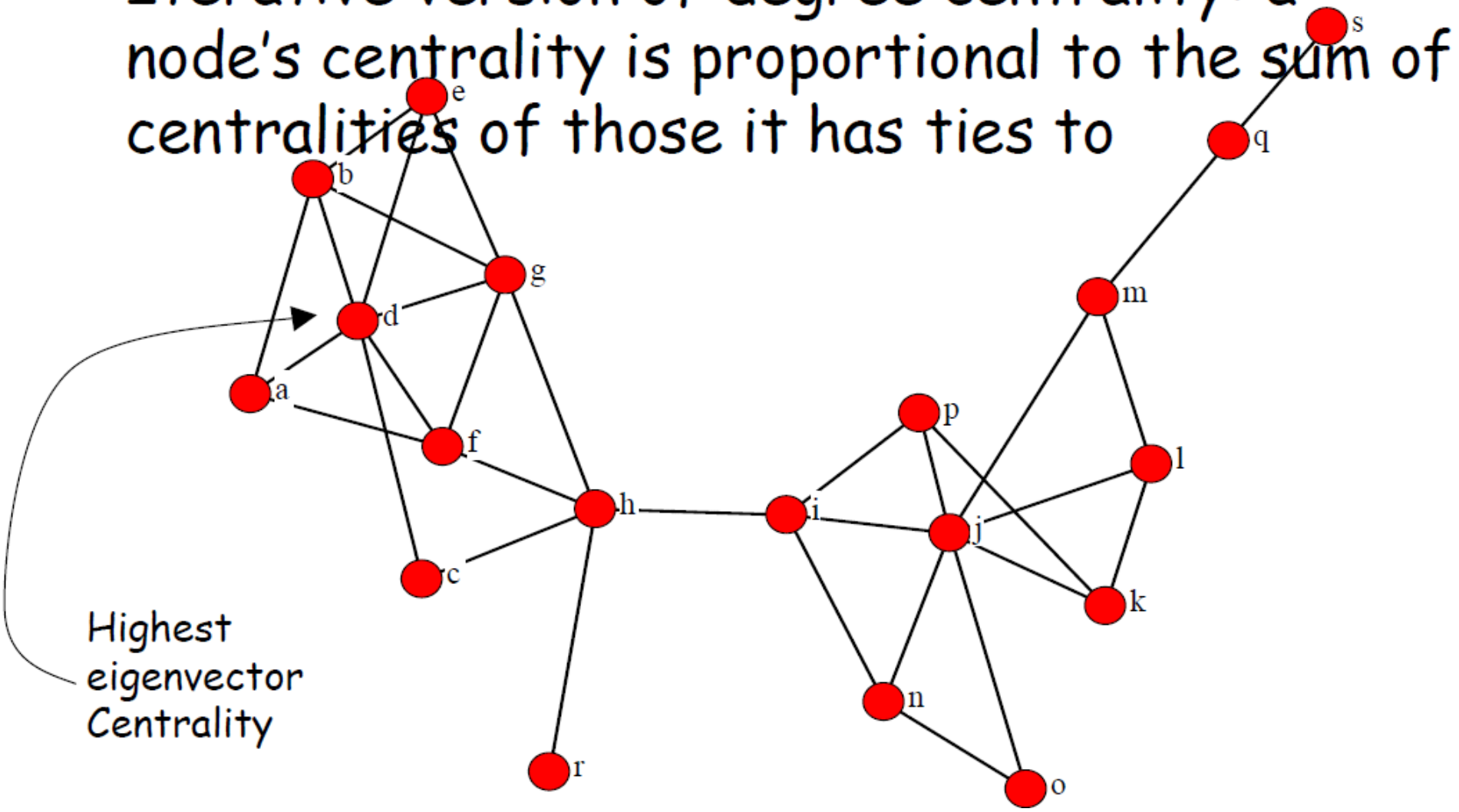
Normalized Closeness:

$$C_C'(n_i) = (C_C(n_i))(g-1)$$

From Steve Borgatti

Node Level Analysis: Eigenvector Centrality

- Iterative version of degree centrality: a node's centrality is proportional to the sum of centralities of those it has ties to

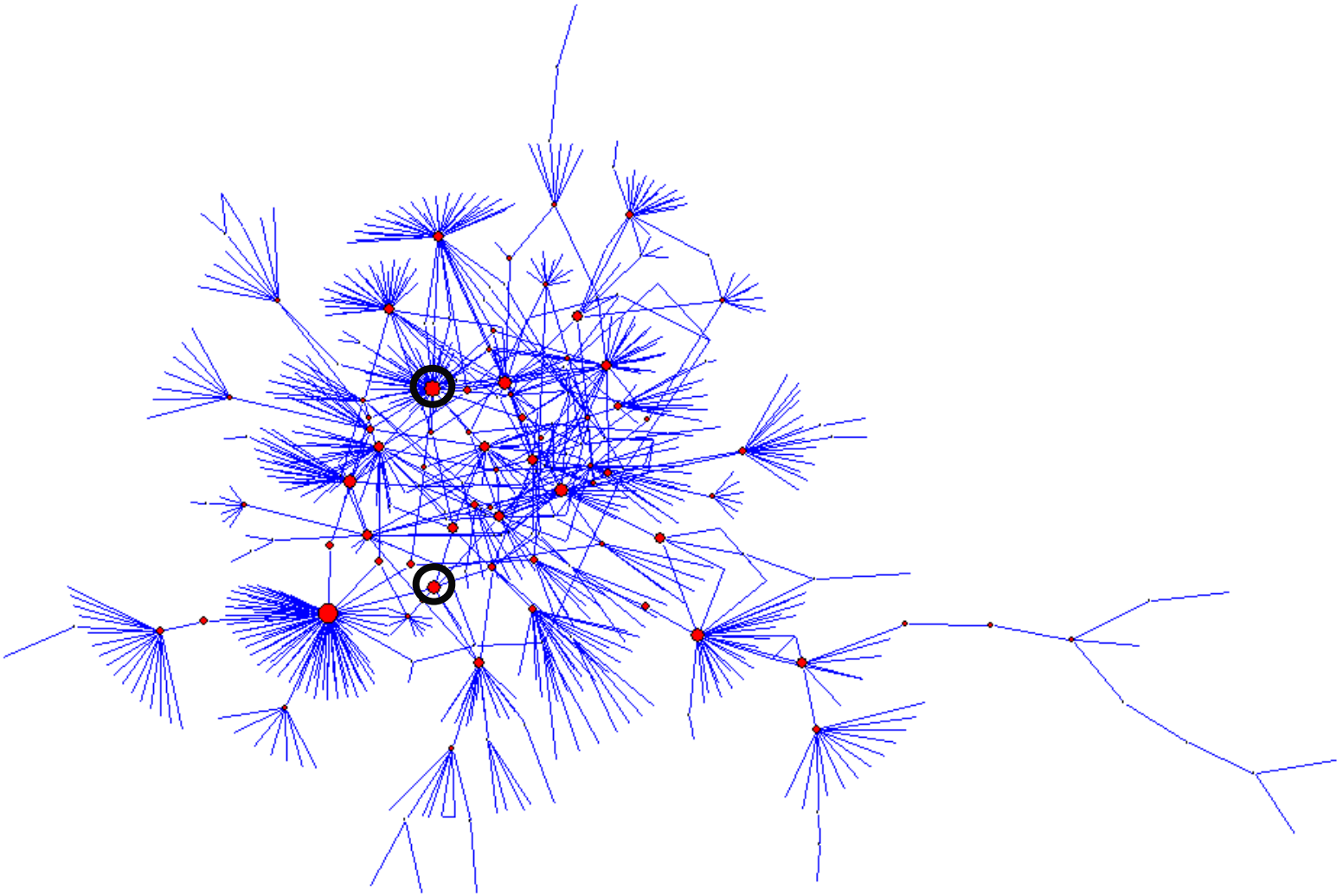


From Steve Borgatti

Comparison of Centrality Measures

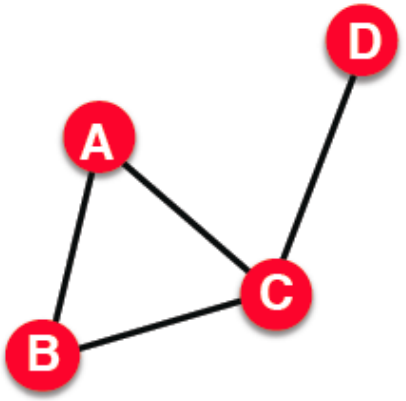
	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant - communication bypasses him/her
High Closeness	Key player tied to important important/active alters		Ego is near many people, but so are many others
High Betweenness	**Ego's few ties are crucial for network flow**	**Very rare cell. Would mean that ego monopolizes the ties from a small number of people to many others.**	

Identifying Key Players



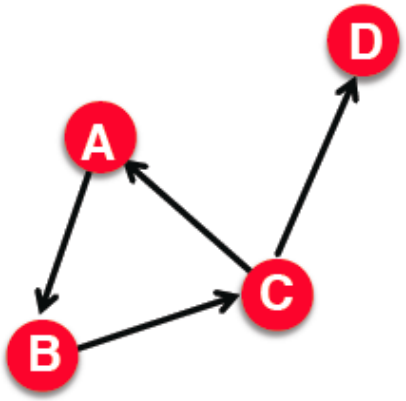
(Node size proportional to betweenness centrality)

Link Level Analysis: Length and Distance



$h_{B,D} = 2$

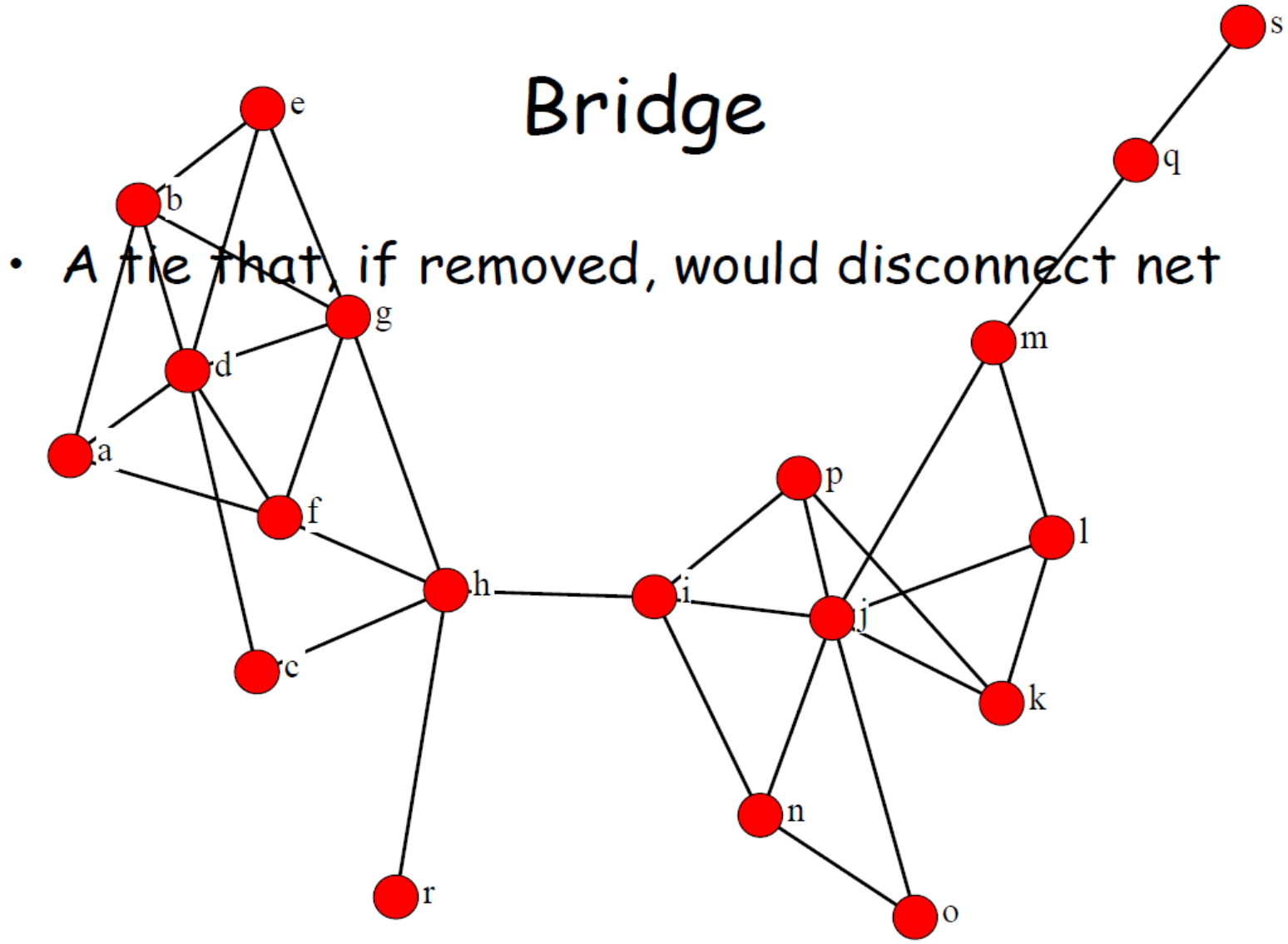
- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - *If the two nodes are disconnected, the distance is usually defined as infinite



$h_{B,C} = 1$, $h_{C,B} = 2$

- In **directed graphs** paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{A,C} \neq h_{C,A}$

Link Level Analysis: Cut Points and Bridge



From Steve Borgatti

The Strength of Weak Tie (Granovetter 1973)

- Strong ties create transitivity
  - Two nodes linked by a strong tie will have mutual acquaitances
- Ties that are part of transitive triples cannot be bridges
- Therefore, only weak ties can be bridges
  - the value of weak ties!!
- Strong ties embeded in tight homophilous clusters, while weak ties connect to diversity
  - Weak ties is a major source of novel information

Network Level Analysis

- Network Topology Analysis takes a macro perspective to study the physical properties of network structures. Network topological measures include:
  - **Size**, i.e., number of nodes and links, **Average Degree**,
  - **Degree Distribution**
  - **Average Path Length**, **Diameter**
  - **Clustering Coefficient**: a measure of an "all-my-friends-know-each-other" property; small-world feature
  - **Centralization and Density**

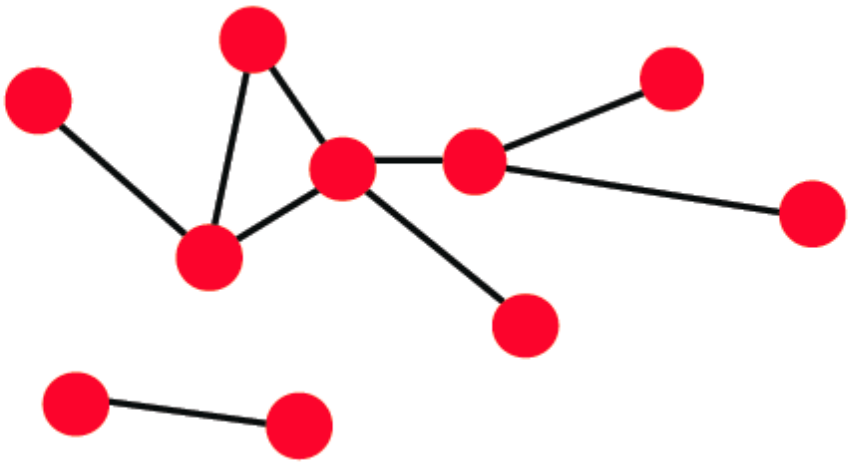
Network Level Analysis: Degree Distribution

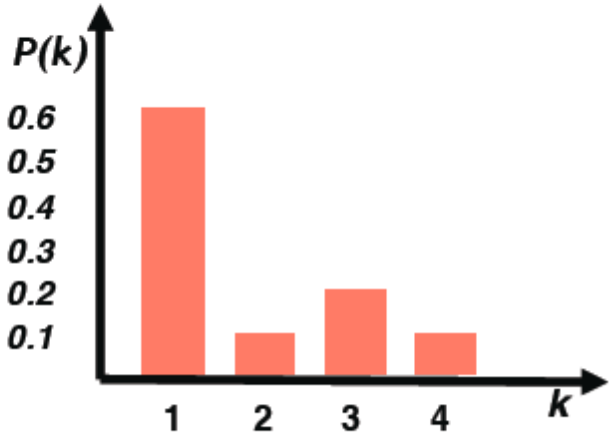
- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

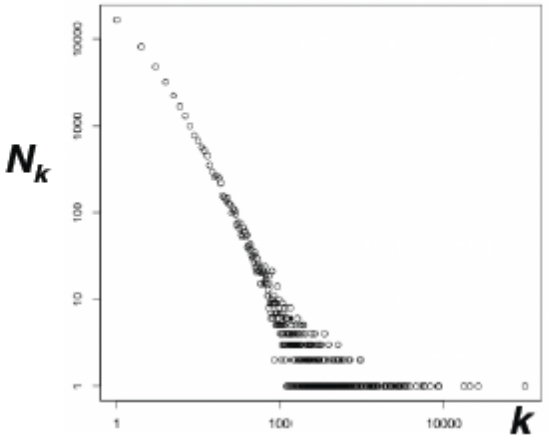
  N_k = # nodes with degree k

- Normalized histogram:

  $P(k) = N_k / N$ → plot







Network Level Analysis: Diameter and Average Path Length

- **Diameter:** the maximum (shortest path) distance between any pair of nodes in a graph

- **Average path length** for a connected graph (component) or a strongly connected (component of a) directed graph

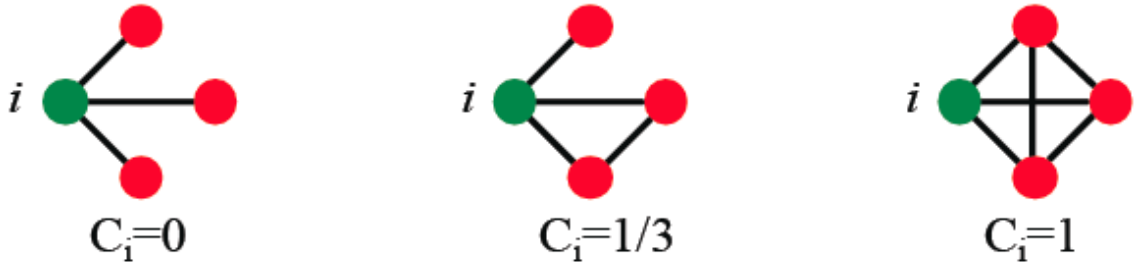
$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{ij}$$

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

  - Many times we compute the average only over the connected pairs of nodes (that is, we ignore "infinite" length paths)

Network Level Analysis: Clustering Coefficient

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



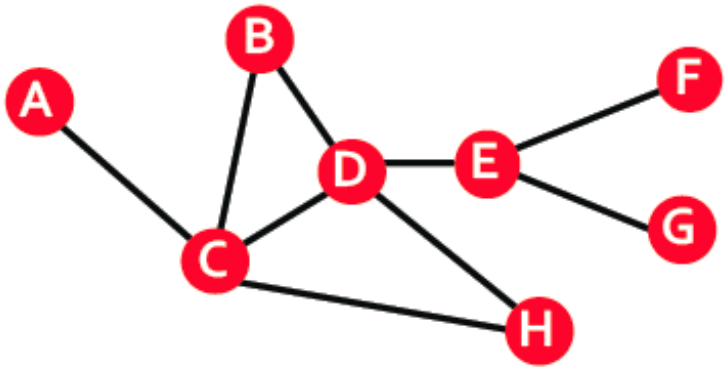
C_i=0 C_i=1/3 C_i=1

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

Network Level Analysis: Clustering Coefficient

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

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



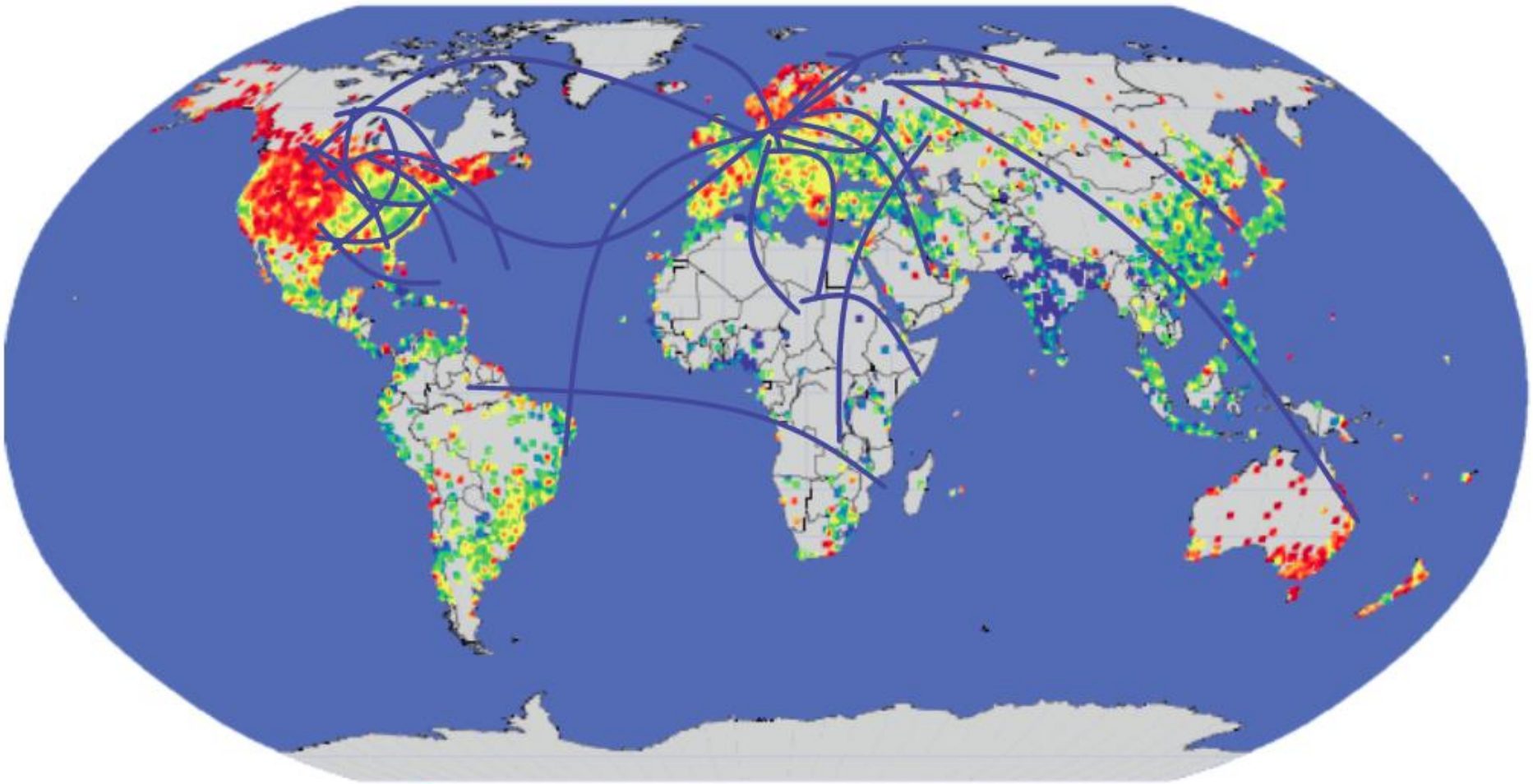
$k_B=2,\ e_B=1,\ C_B=2/2 = 1$

$k_D=4,\ e_D=2,\ C_D=4/12 = 1/3$

A Real Network: The MSN Messenger

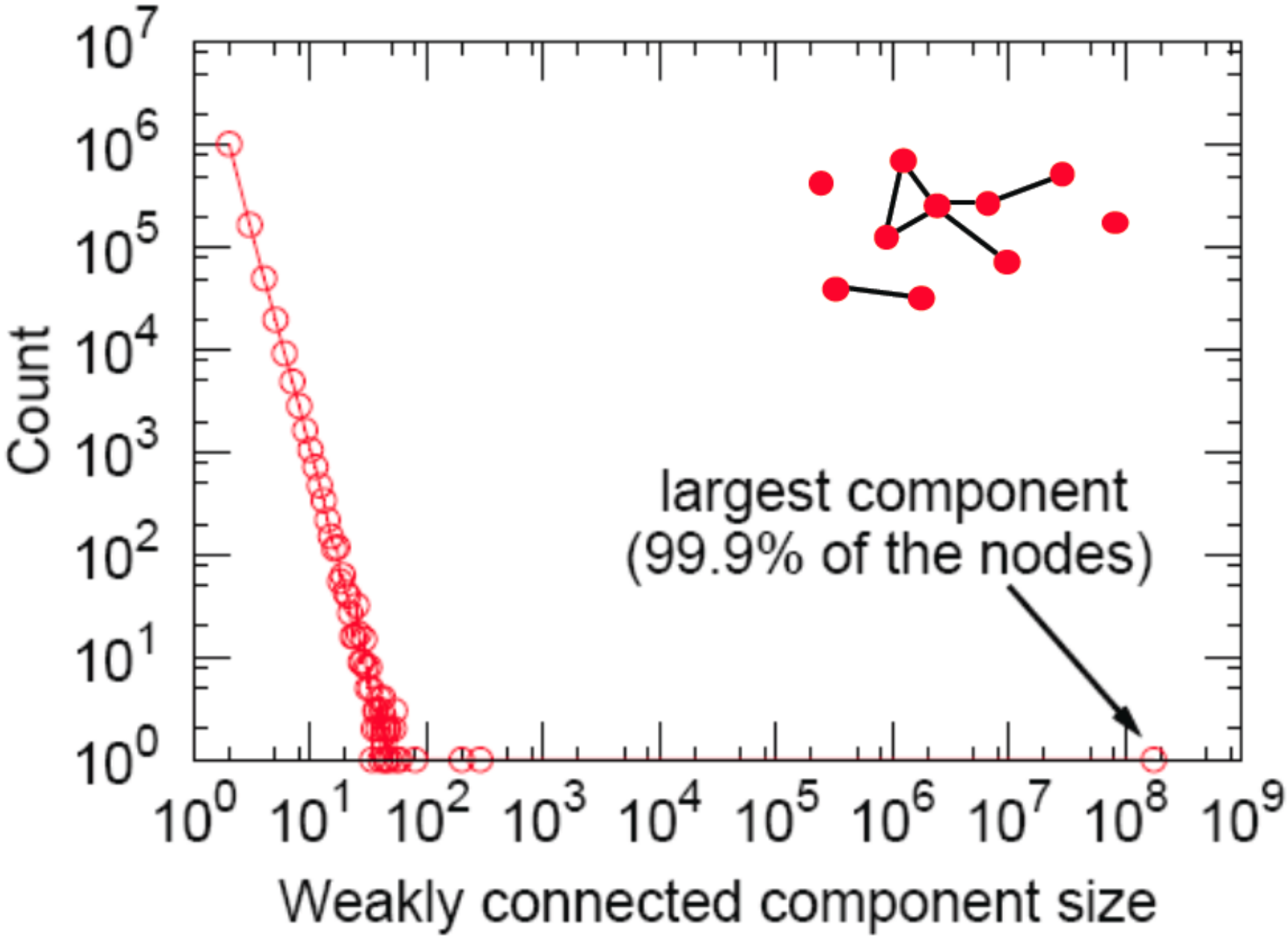
- **MSN Messenger activity in June 2006:**
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

The MSN Global Communication Network

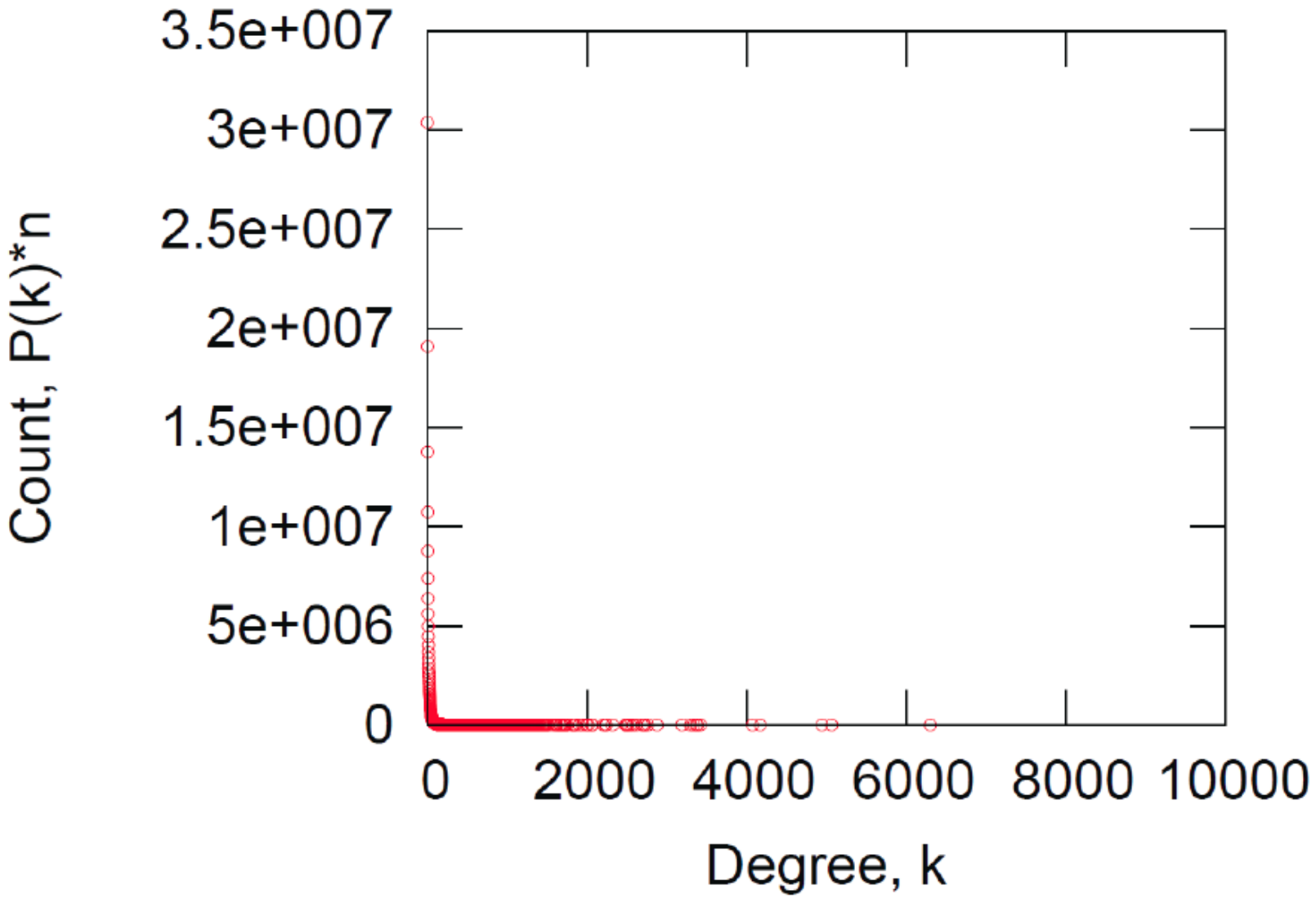


Network: 180M people, 1.3B edges

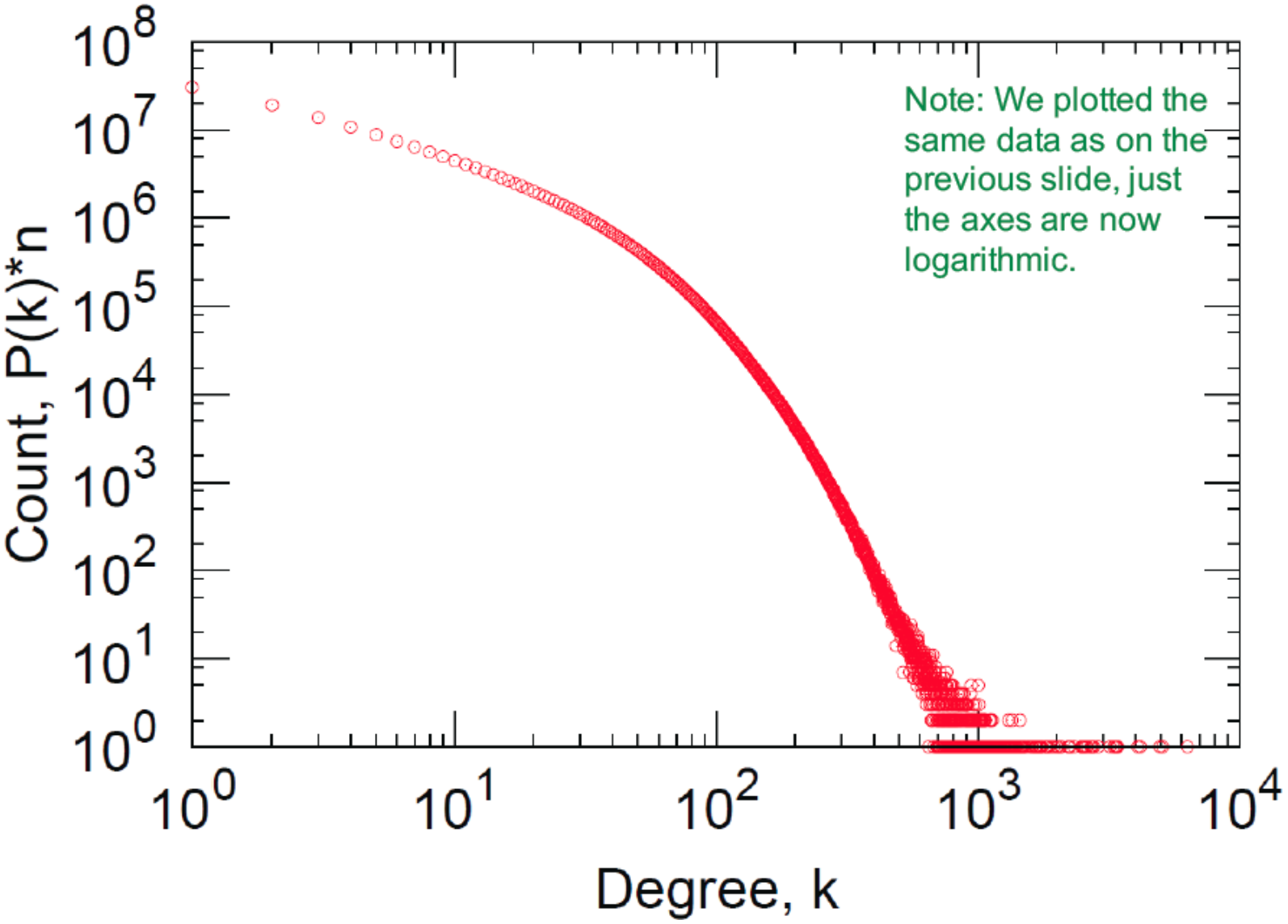
The MSN Global Communication Network



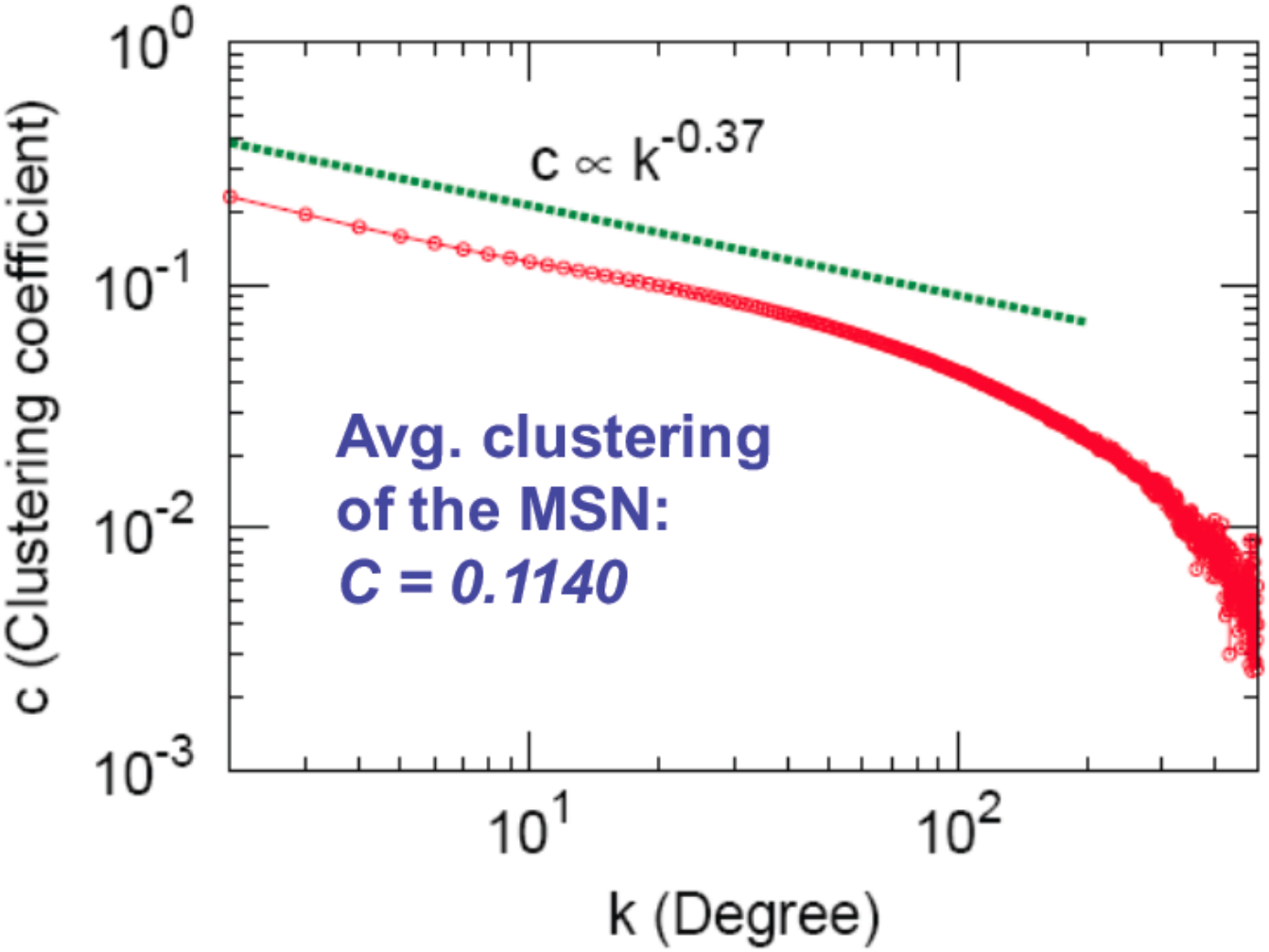
Degree Distribution



Log-Log Degree Distribution

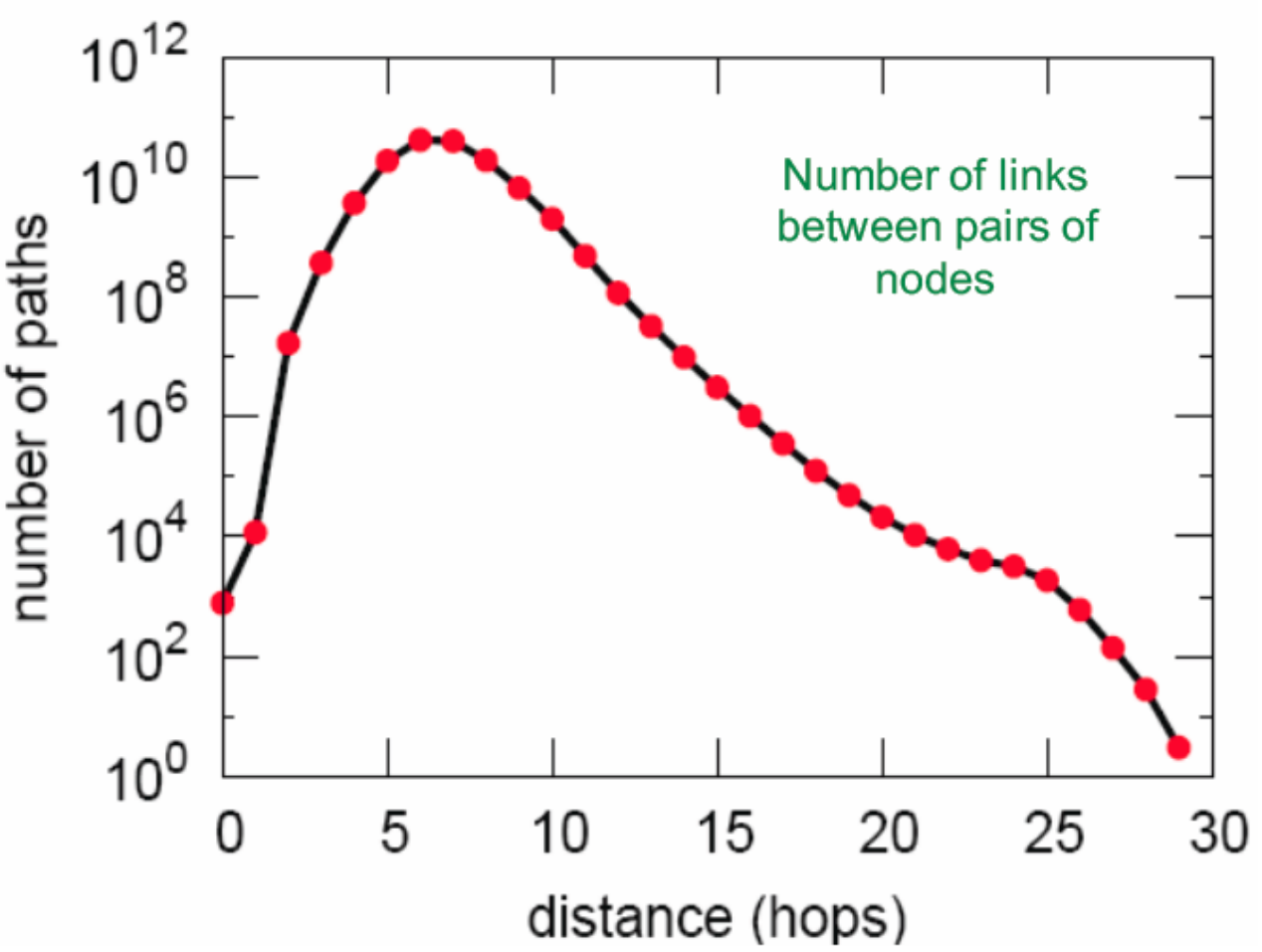


Clustering



C_k: average C_i of nodes i of degree k: $C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i$

Diameter



Avg. path length **6.6**

90% of the nodes can be reached in < 8 hops

To Sum Up

Degree distribution: *Heavily skewed*
avg. degree= 14.4

Path length: *6.6*

Clustering coefficient: *0.11*

Are these values "expected"?
Are they "surprising"?

To answer this we need a null-model!

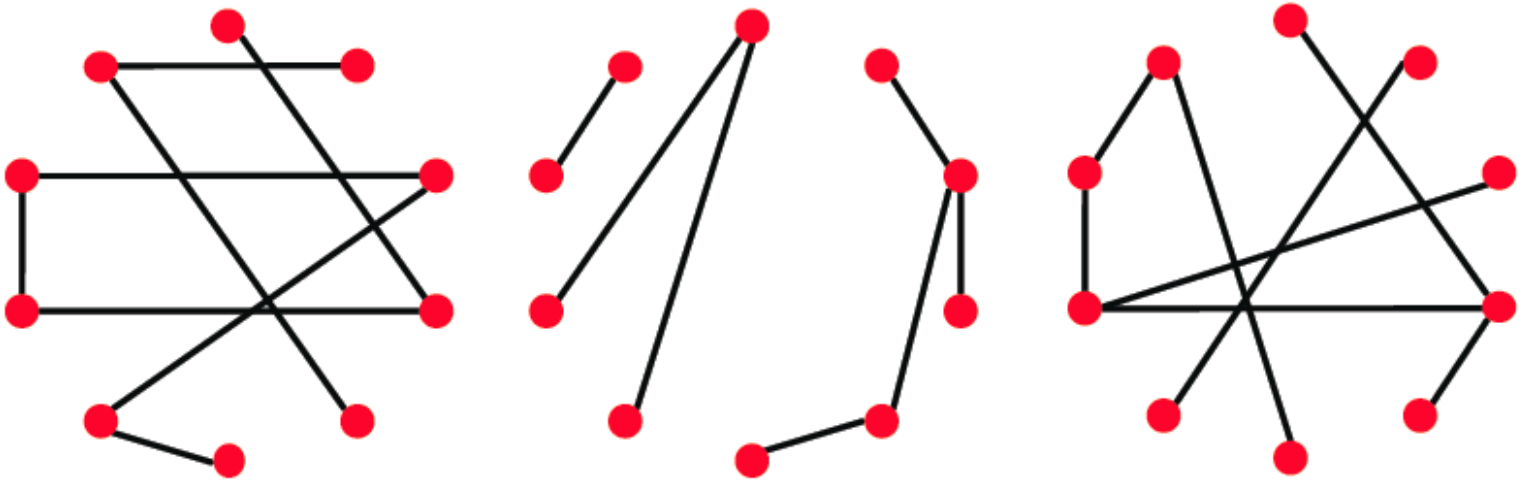
The Null Model: Random Network (Graph)

- **Erdös-Renyi Random Graphs** [Erdös-Renyi, '60]
- **Two variants:**
  - $G_{n,p}$: undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p
  - $G_{n,m}$: undirected graph with n nodes, and m uniformly at random picked edges

What kinds of networks does such model produce?

The Null Model: Random Network (Graph)

- ***n* and *p* do not uniquely determine the graph!**
  - The graph is a result of a random process
- We can have many different realizations given the same ***n*** and ***p***



n = 10
p= 1/6

Degree Distribution of Gnp

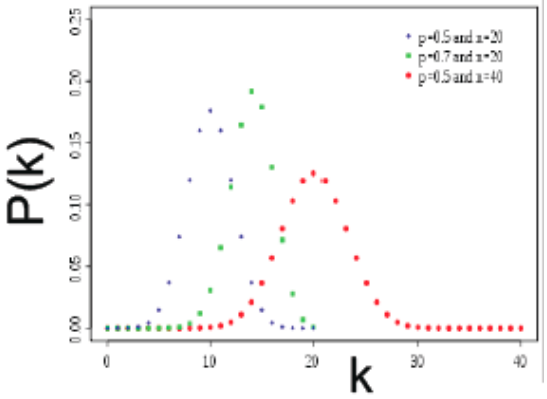
- **Fact: Degree distribution of G_{np} is Binomial.**
- Let ***P(k)*** denote a fraction of nodes with degree ***k***:

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

Select *k* nodes out of *n-1*

Probability of having *k* edges

Probability of missing the rest of the *n-1-k* edges



Mean, variance of a binomial distribution

$$\bar{k} = p(n-1)$$

$$\sigma^2 = p(1-p)(n-1)$$

$$\frac{\sigma}{\bar{k}} = \left[\frac{1-p}{p}\frac{1}{(n-1)}\right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$

By the law of large numbers, as the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of *k*.

Clustering Coefficient of Gnp

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

  Where e_i is the number of edges between i's neighbors

- Edges in G_{np} appear i.i.d. with prob. p

- **So:** $e_i = p\frac{k_i(k_i - 1)}{2}$

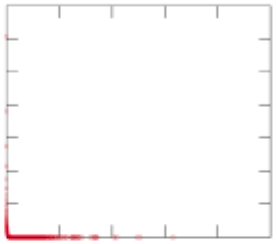
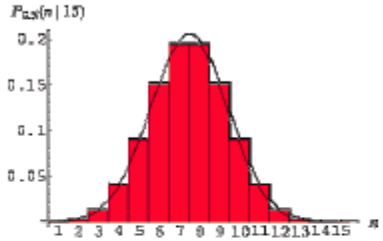
  Each pair is connected with prob. p

  Number of distinct pairs of neighbors of node i of degree k_i

- **Then:** $C = \frac{p \cdot k_i(k_i - 1)}{k_i(k_i - 1)} = p = \frac{\bar{k}}{n-1} \approx \frac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
For a fixed avg. degree (that is p=1/n), *C* decreases with the graph size *n*.

The MSN Network vs. Gnp

	MSN	G_{np}
Degree distribution:		
Path length:	**6.6**	$O(\log n)$ ≈ 8.2
Clustering coefficient:	0.11	$\overline{k} / n$ ≈ $8 \cdot 10^{-8}$

Real (MSN) Networks vs. Gnp

- **Are real networks like random graphs?**
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random networks model:**
  - Degreed distribution differs from that of real networks
  - Giant component in most real network does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

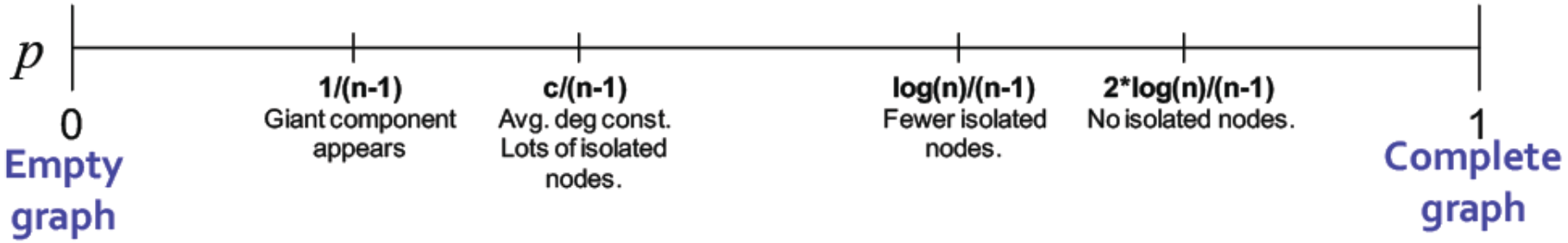
Real (MSN) Networks vs. Gnp

- **If G_{np} is wrong, why did we spend time on it?**
  - It is the reference model for the rest of the class.
  - It will help us calculate many quantities, that can then be compared to the real data
  - It will help us understand to what degree is a particular property the result of some random process

So, while G_{np} is WRONG, it will turn out to be extremly USEFUL!

Evolution of a Random Network (Graph)

- **Graph structure of G_{np} as p changes:**

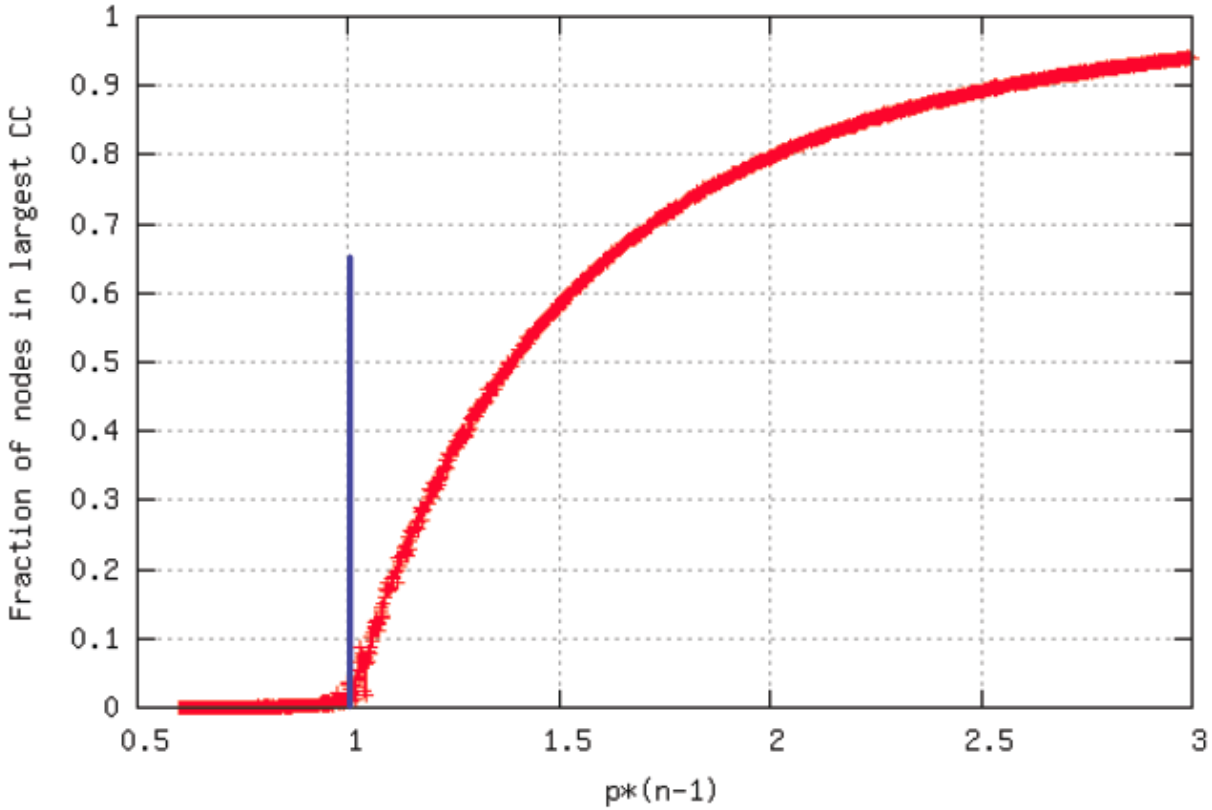


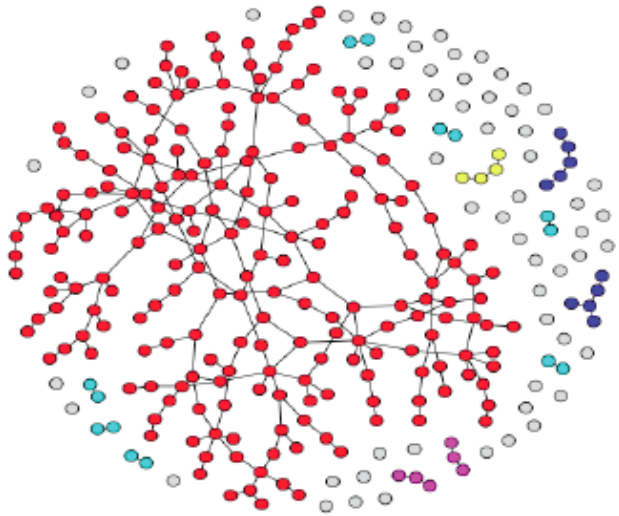
- **Emergence of a Giant Component:**

  avg. degree $k=2E/n$ or $p=k/(n-1)$

  - $k=1-\varepsilon$: all components are of size $\Omega(\log n)$
  - $k=1+\varepsilon$: 1 component of size $\Omega(n)$, others have size $\Omega(\log n)$

Gnp Simulation Experiment





Fraction of nodes in the largest component

- G_{np}, n=100k, $p(n-1)$ = 0.5 … 3

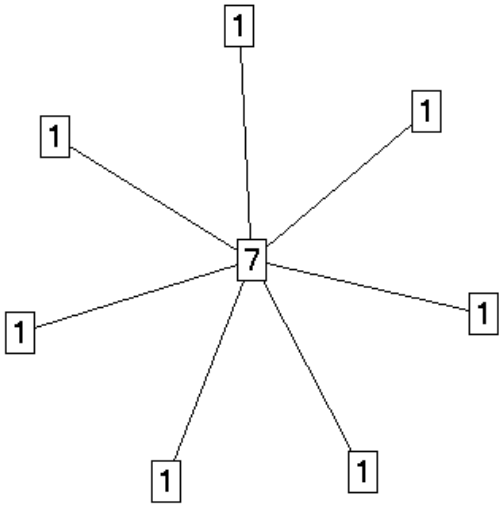
Network Level Analysis: Centralization

- Variance of the individual centrality scores.

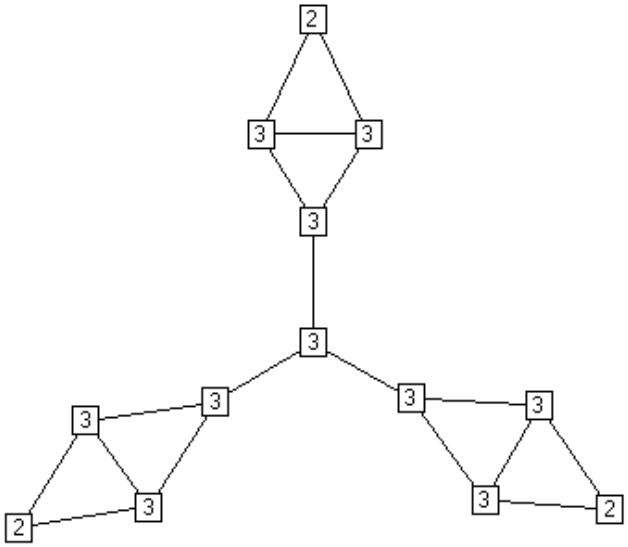
$$S_D^2 = \left[\sum_{i=1}^{g} (C_D(n_i) - \overline{C}_d)^2 \right] / g$$

- Freeman's general formula for centralization (which ranges from 0 to 1):

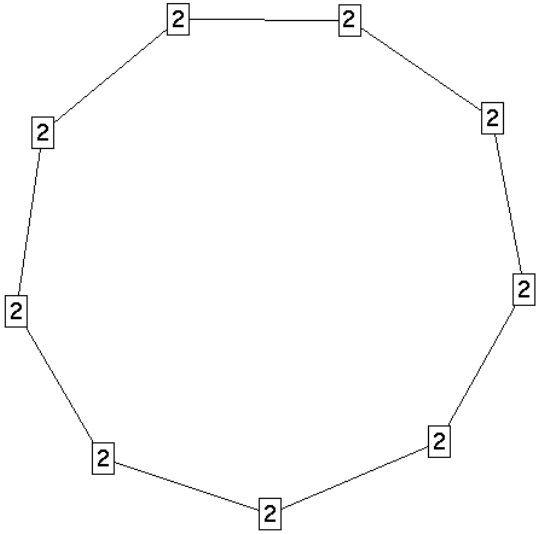
$$C_D = \frac{\sum_{i=1}^{g} \left[C_D(n^*) - C_D(n_i) \right]}{[(g-1)(g-2)]}$$



Freeman: 1.0
Variance: 3.9



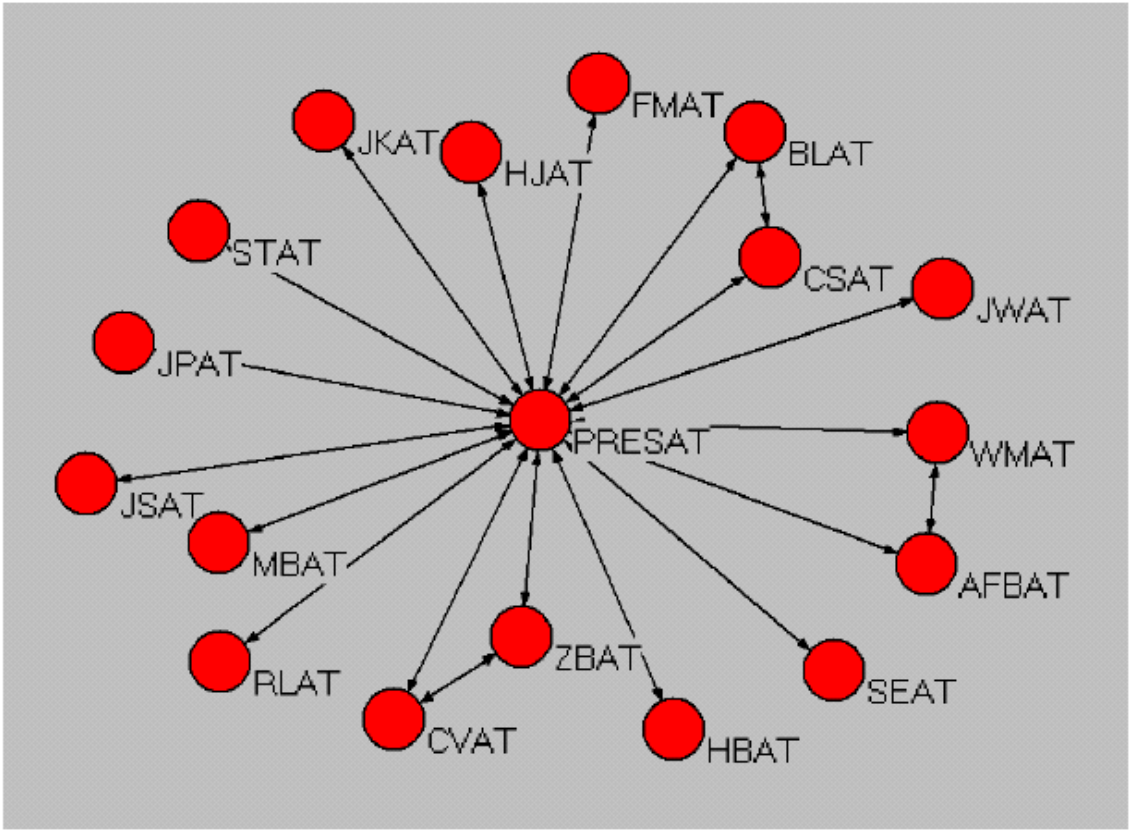
Freeman: .02
Variance: .17



Freeman: 0.0
Variance: 0.0

Network Level Analysis: Centralization

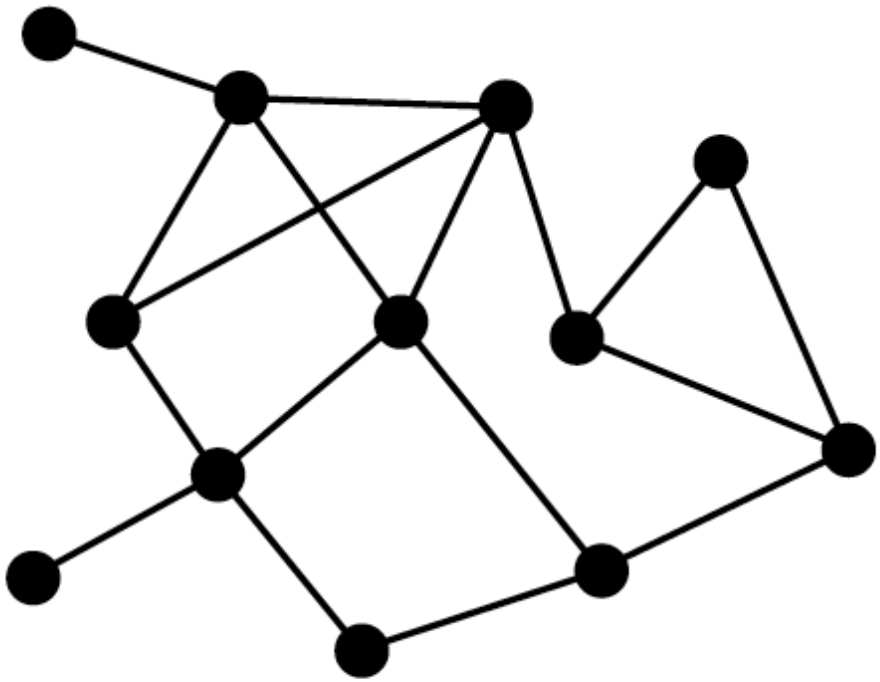
- Centralization: Degree to which network revolves around a single node.



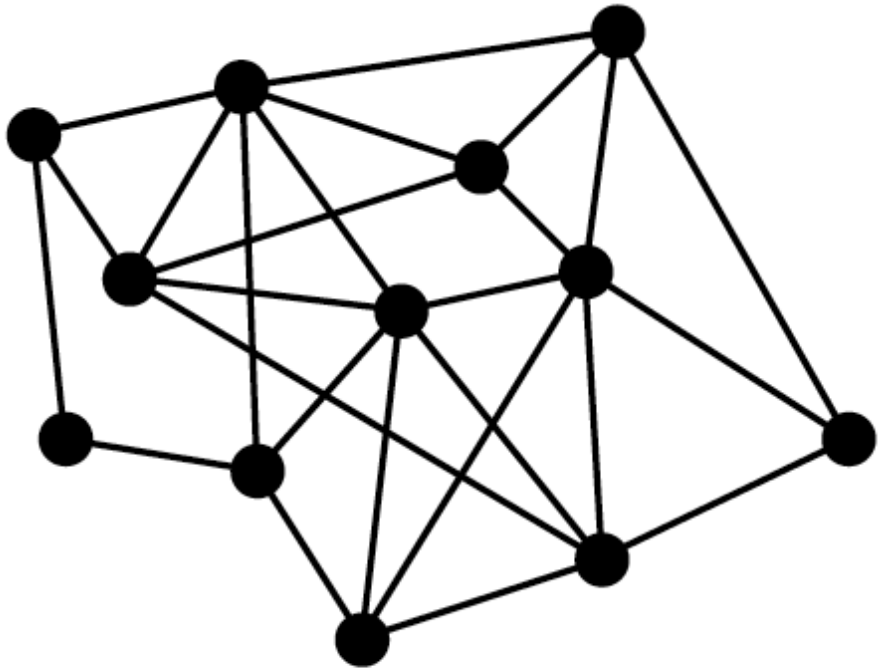
From Steve Borgatti

Network Level Analysis: Network Density

- The density D of a network is the percentage of the links L divided by the number of all possible links N(N-1)/2



Low Density (25%)
Avg. Dist. = 2.27



High Density (39%)
Avg. Dist. = 1.76

Our Approach

What

- Social network analysis (Metrics)
- **Describe** the changes in network evolution
  - Temporal changes in network topological measures
- Dynamic network recovery
- (Relational) data mining

Why

- Econometric **identification** of casual Social and Economic influence
  - Distinguishing homophily
  - Confounding factors
  - PSM, DID, RD, etc.
  - Explanations

How

- **Combine** social science methods, data mining, machine learning with econometric analysis
- **Predict** link formation
- **Simulate** the evolution of networks

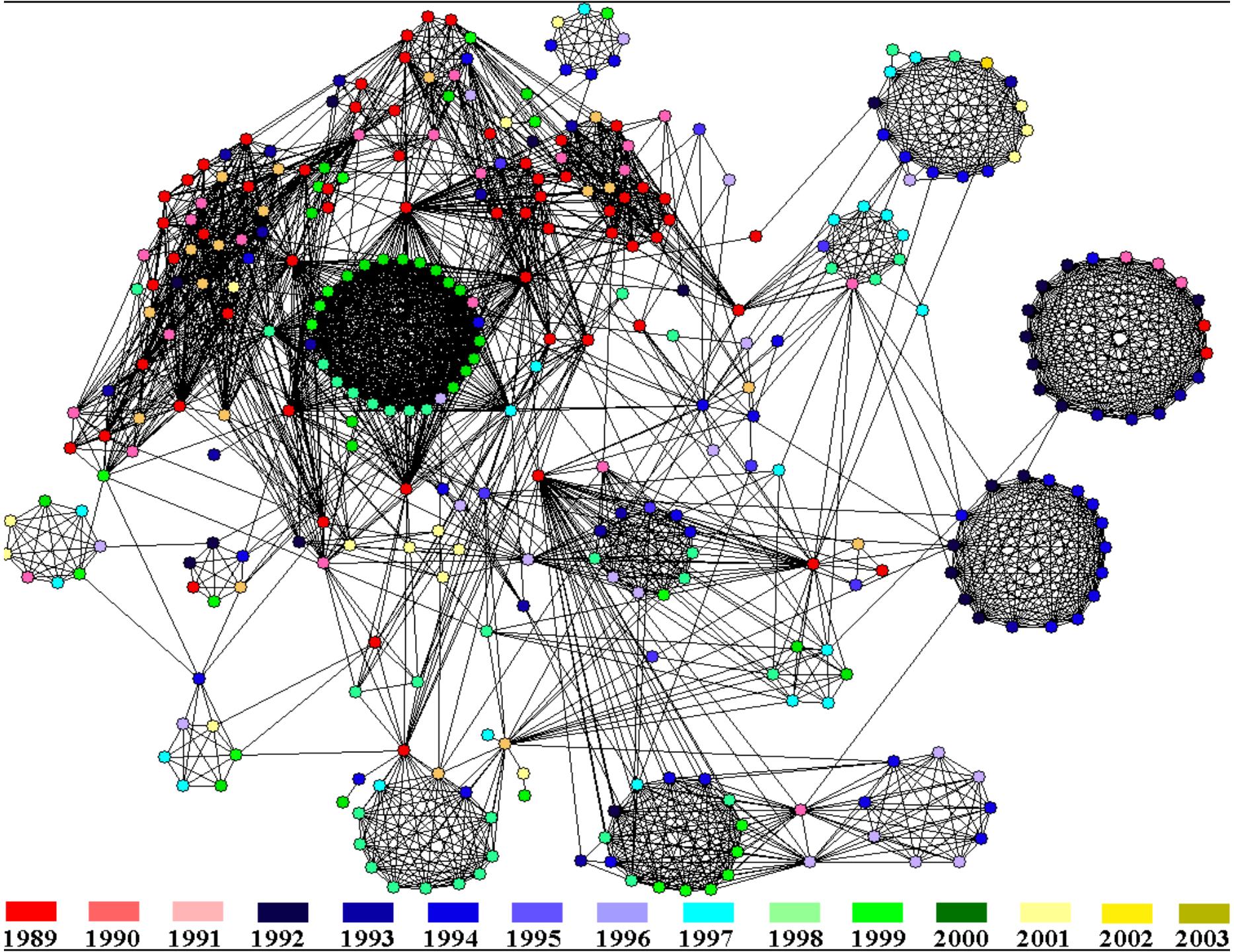
Describe Network Evolution: A Global Terrorist Network

- The Global Salafi Jihad (GSJ) network data is compiled by a former CIA operation officer Dr. *Marc Sageman* - **366 terrorists**
  - *friendship*, *kinship*, same *religious* leader, *operational* interactions, etc.
  - geographical origins, socio-economic status, education, etc.
  - **when** they join and leave GSJ

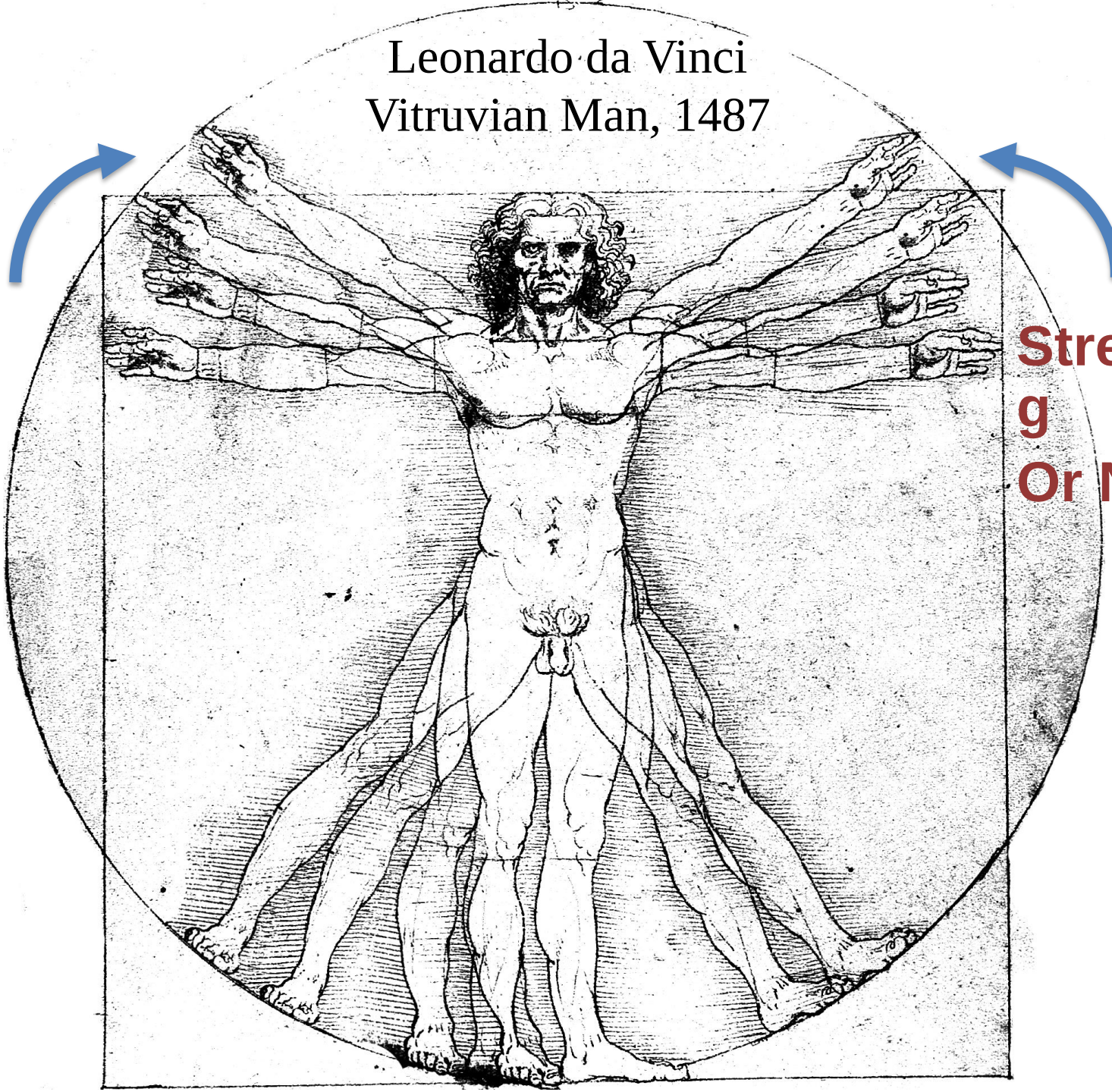
- The goal of dynamic analysis
  - gain insights about the evolution of GSJ network
  - develop effective attack strategies to break down GSJ network

name	YrLeft	DOB	Cluster	RelBack	School	Edu	EduType	Occ	Married	Kids	CrimBack	YrJoin	AgeJoin	Acq	Friends	Fam
Banshiri	1996		1	2	2	4	4	2	1	1	3	1989		8,34	1,2,4	
M Atef	2001	1957	1	2	2	4		2	1	1	3	1989	32	6,357,	1,2,3	
Khadr	2003	1948	1	2	2	4	4	2	1	1	0	1989	41	2		1
mithim	1996	1972	2		2	1	1	1			0	1993	21		46,47,48	
Hajri	1996	1972	2		2	1	1	1			0	1992	20		45,47,48	

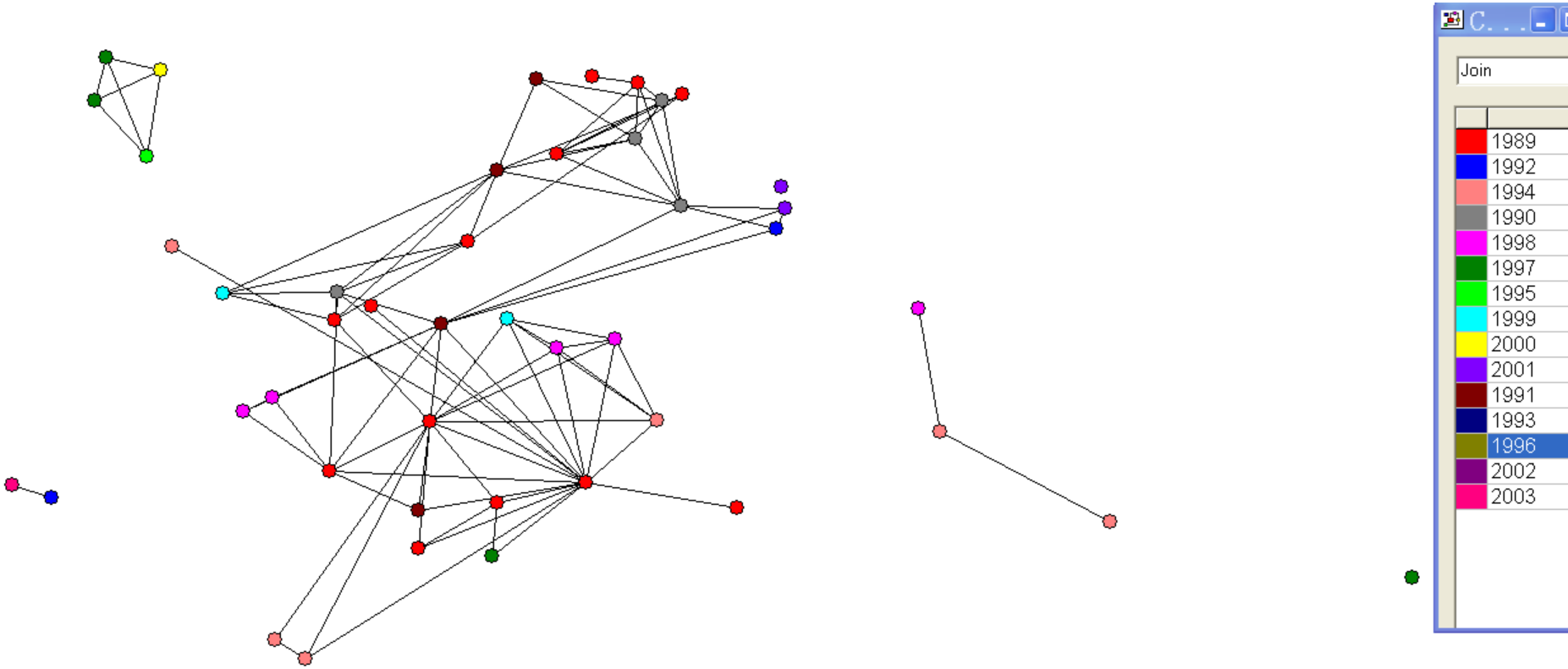
Sample data of GSJ terrorists


1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003

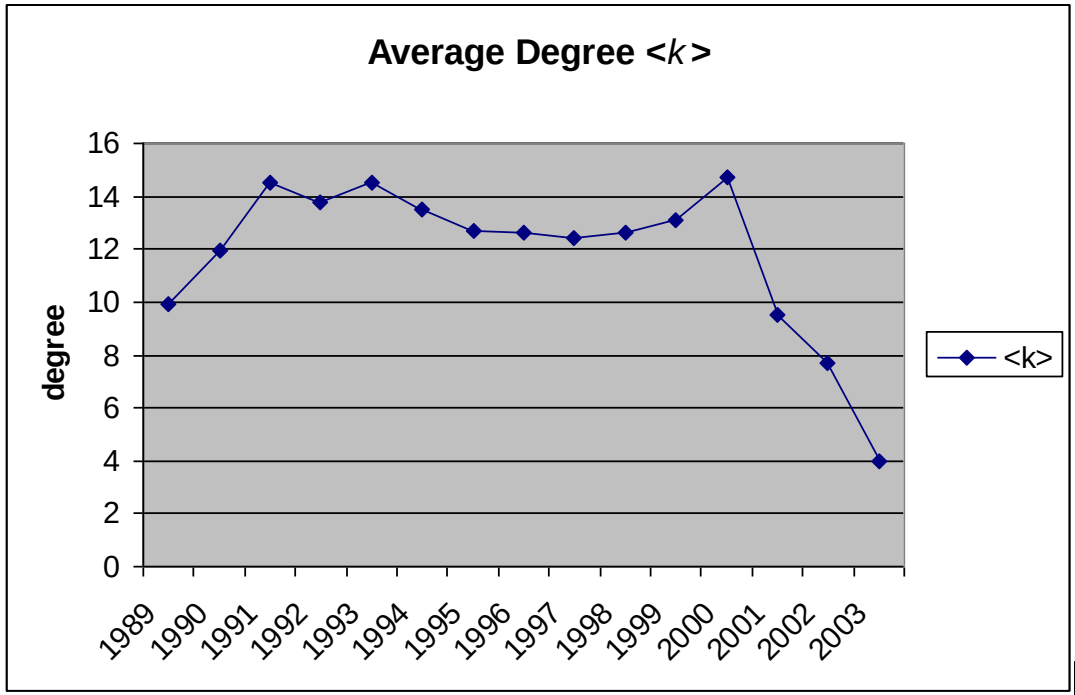
Leonardo da Vinci
Vitruvian Man, 1487



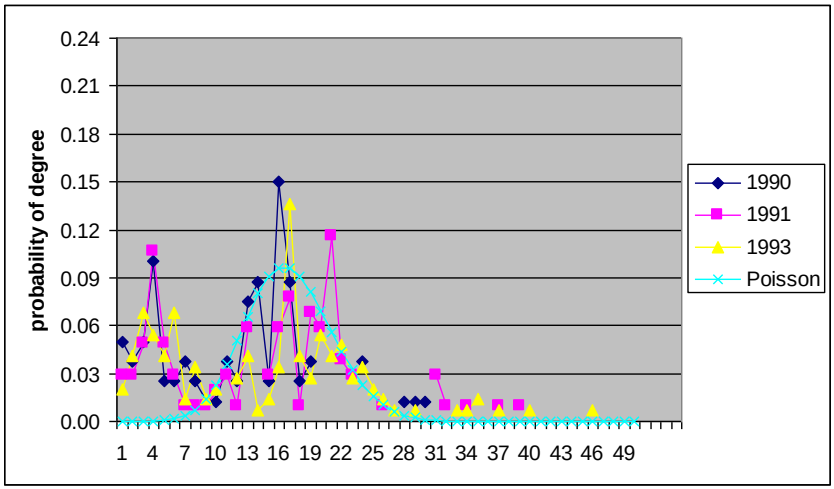
Stretching Or Not?


C. . .
Join
1989
1992
1994
1990
1998
1997
1995
1999
2000
2001
1991
1993
1996
2002
2003

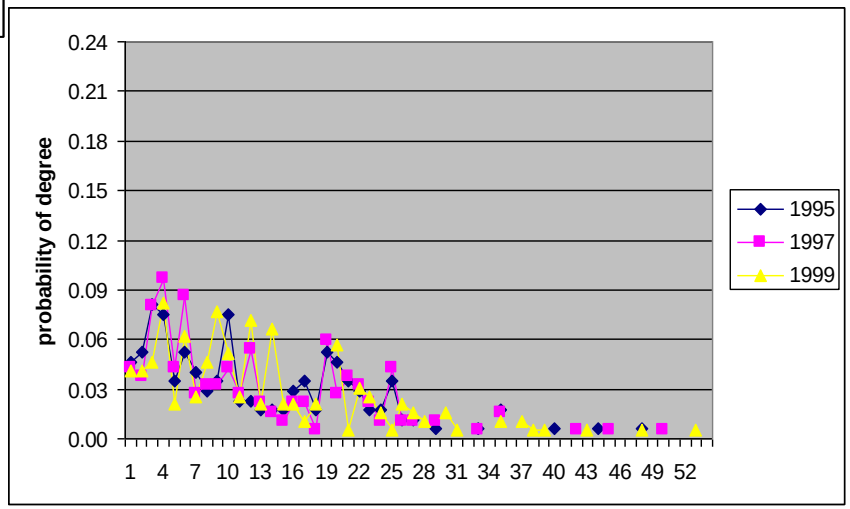
Temporal Changes in Network-level Measures



a



b



c

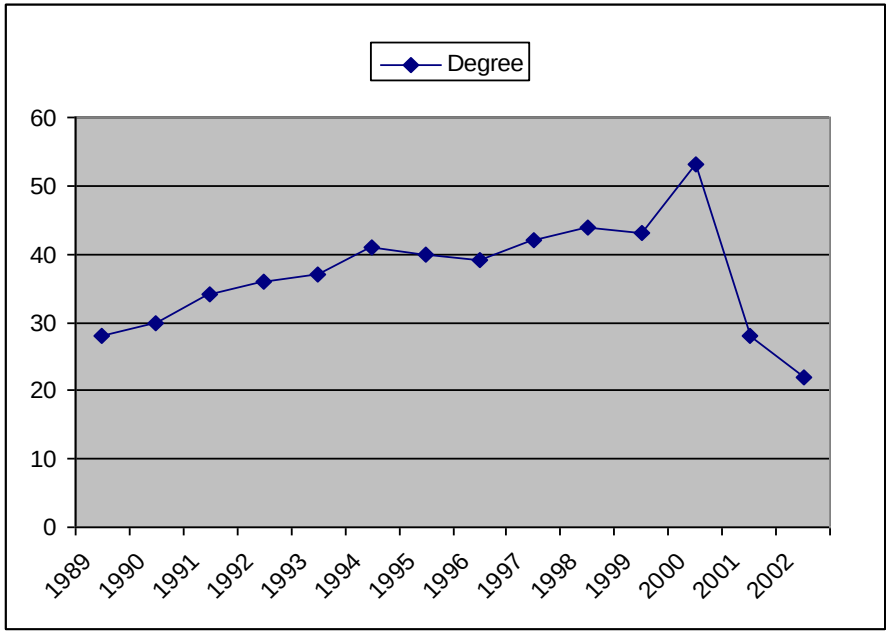
- Fig.1. The temporal changes in the (a) average degree, (b) and (c) degree distribution

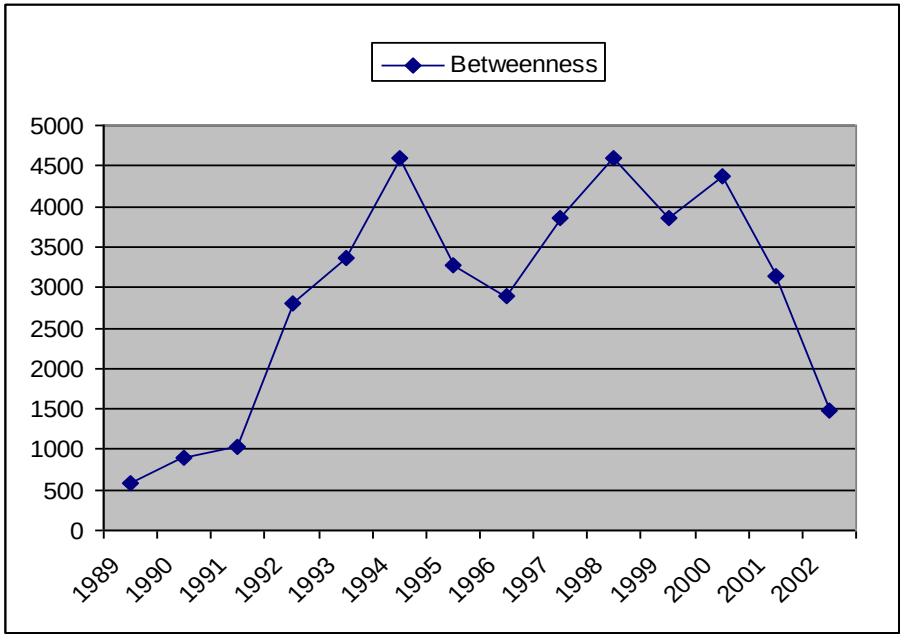
- Degree = number of links a node has

Findings

- There are three stages for the evolution of the GSJ network:
  - 1989 - 1993 The **emerging** stage:
    - The network grows in size
    - **Accelerated Growth** - No. of edges increases faster than nodes
    - **Random** network topology (**Poisson** degree distribution)
  - 1994 - 2000 The **mature** stage:
    - The size of the network reached its peak in 2000
    - **Scale-free** topology (**Power-law** degree distribution)
  - 2001 - 2003 The **disintegration** stage:
    - Falling into small disconnected components after 9/11

Temporal Changes in Node Centrality Measures





- Figure.2. Temporal changes in Degree and Betweenness centrality of ***Osama Bin Laden***

- Degree: No. of links a node has
- Betweenness of a node *i*
  - No. of shortest paths from all nodes to all others that pass through node *i*
  - Measure *i's* influence on the traffic (information, resource) flowing through it

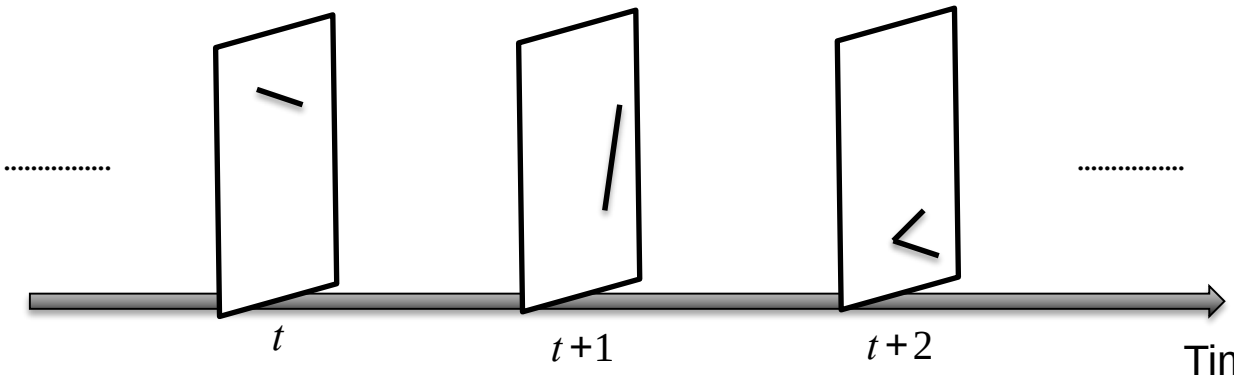
Findings and Possible Explanations

- 1994 – 1996: A sharp decrease in Bin Laden's *Betweenness*
  - 1994: Saudi revoked his citizenship and expelled him
  - 1995: Went to Sudan and was expelled again under U.S. pressure
  - 1996: Went to Afghanistan and established camps there

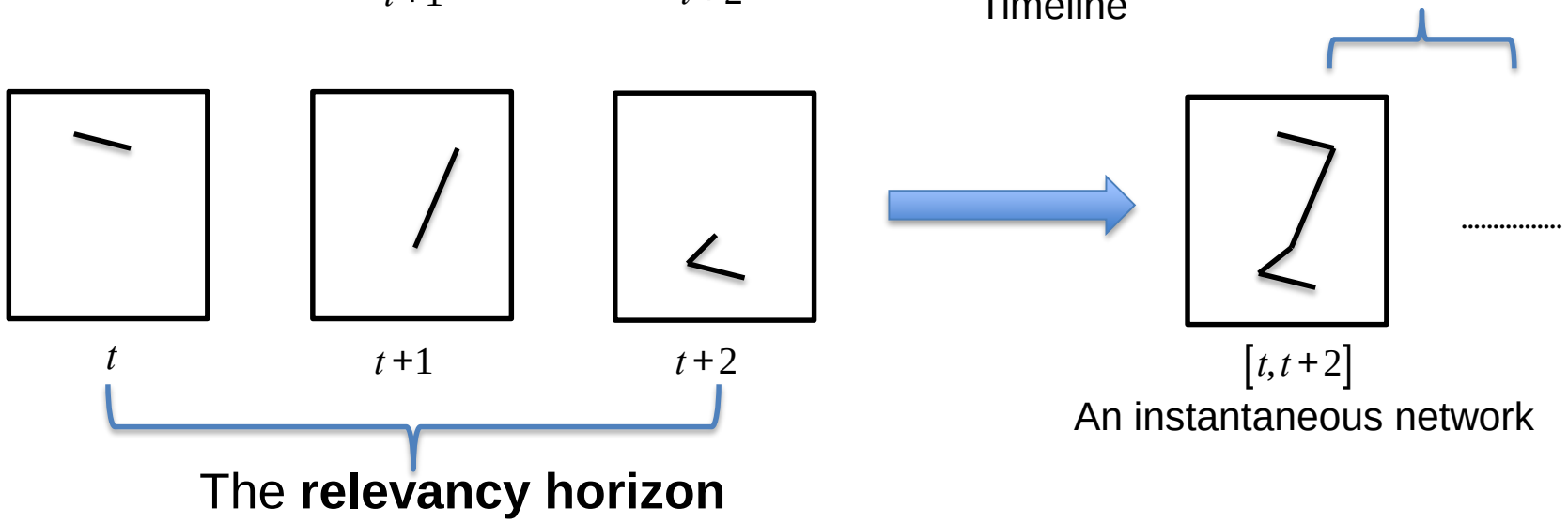
- 1998 –1999: Another sharp decrease in his *Betweenness*
  - After 1998 bombings of U.S. embassies, Bill Clinton ordered a freeze on assets linked to bin Laden (top 10 most wanted)
  - August 1998: A failed assassination on him from U.S.
  - 1999: UN imposed sanctions against Afghanistan to force the Taliban to extradite him

Dynamic Network Recovery

- The GSJ Network: Small, Low Frequency (Yearly), **AdHoc**!
- How about *large*, *high-frequency* longitudinal network data?
  - Email communication network



- How frequent to sample?



- Which past links are relevant to the current state of the network?

Our Approach

What

- Social network analysis (Metrics)
- **Describe** the changes in network evolution
  - Temporal changes in network topological measures
- Dynamic network recovery
- (Relational) data mining

Why

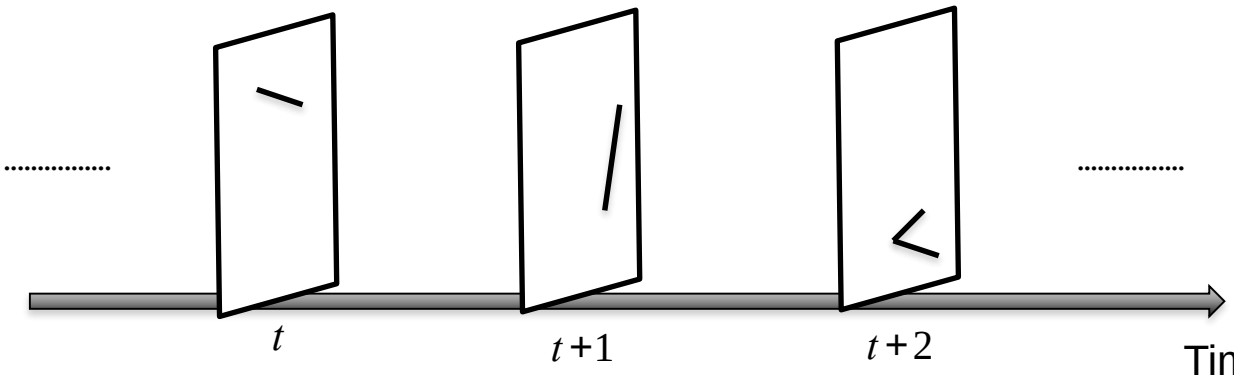
- Econometric **identification** of casual Social and Economic influence
  - Distinguishing homophily
  - Confounding factors
  - PSM, DID, RD, etc.
  - Explanations

How

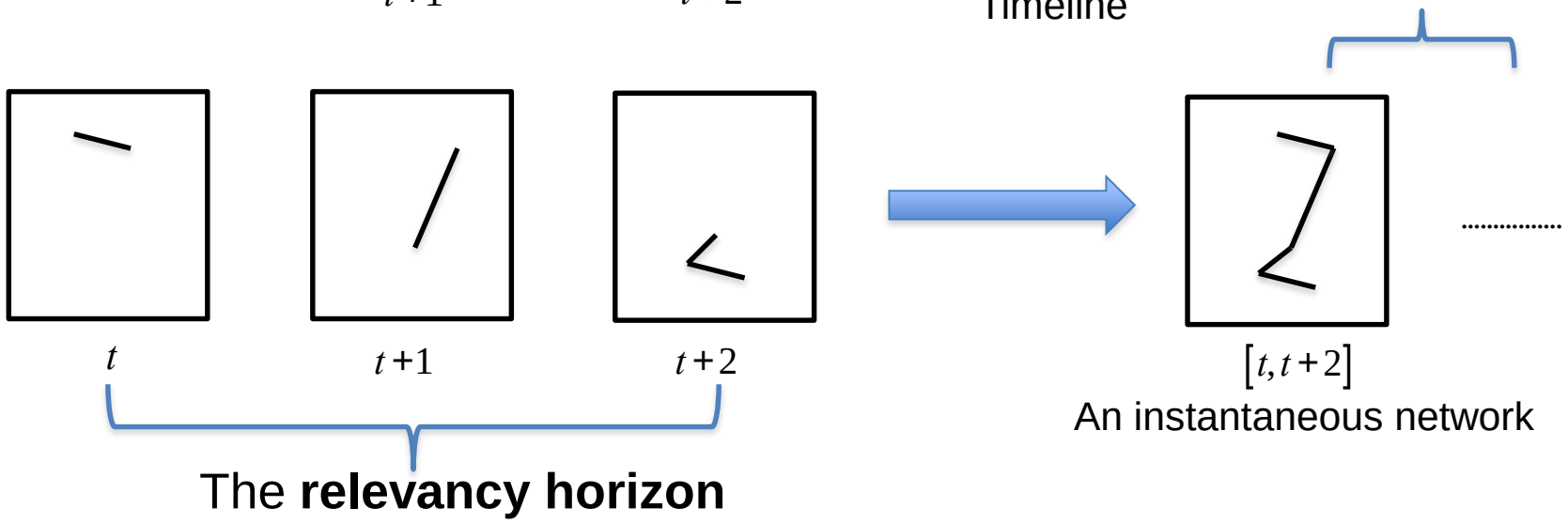
- **Combine** social science methods, data mining, machine learning with econometric analysis
- **Predict** link formation
- **Simulate** the evolution of networks

Dynamic Network Recovery

- The GSJ Network: Small, Low Frequency (Yearly), **AdHoc**!
- How about *large*, *high-frequency* longitudinal network data?
  - Email communication network



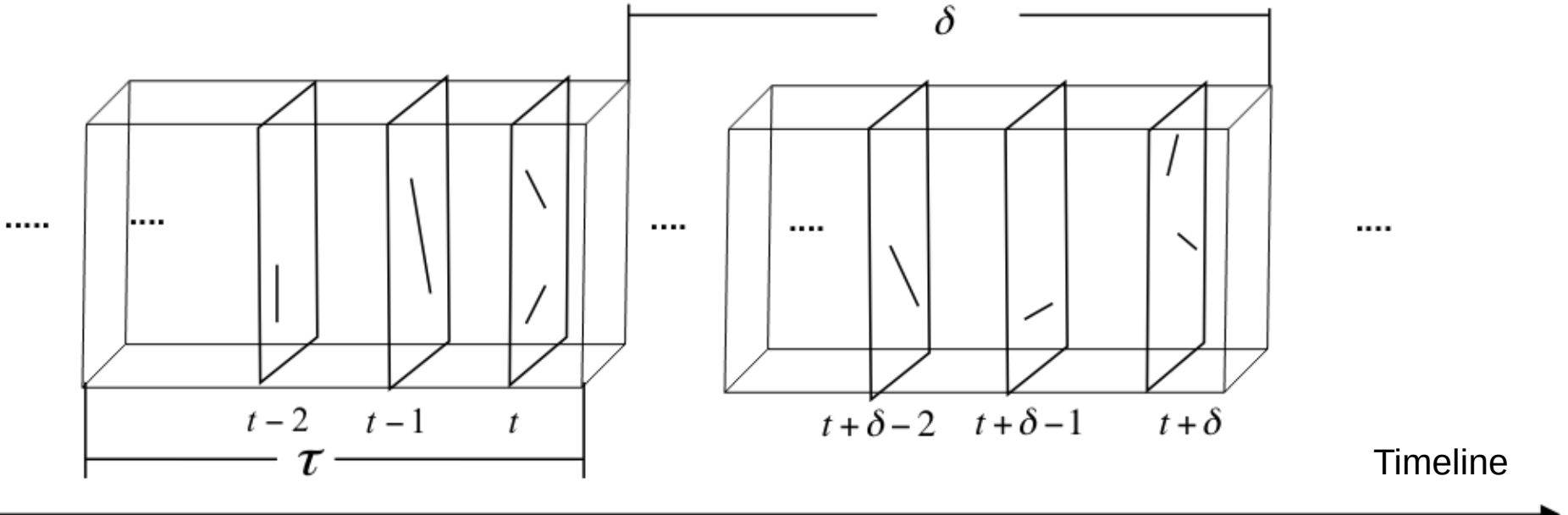
- How frequent to sample?



- Which past links are relevant to the current state of the network?

Relevancy Horizon and Sampling Period

- Recovering a set of instantaneous social networks from longitudinal network data by setting a ***sliding window*** **filter**
  - The **relevancy horizon** τ: the maximum time length that a past event (link) has impact on current network.
  - The **sampling period** δ : determines which events were considered to be simultaneous and independent of each other



Research Testbed: A Narcotic Criminal Network

- The COPLINK dataset contains 3 million police incident reports from the Tucson Police Department (1990 to 2006).
  - 3 million incident reports and 1.44 million individuals
  - Their personal and sociological information (age, ethnicity, etc.)
  - Time information: when two individuals co-offend
  - AZ Inmate affiliation data: when and where an inmate was housed
- A Narcotic Criminal Network
  - 19,608 individuals involved in **organized narcotic crimes**
  - 29,704 co-offending pairs (links)

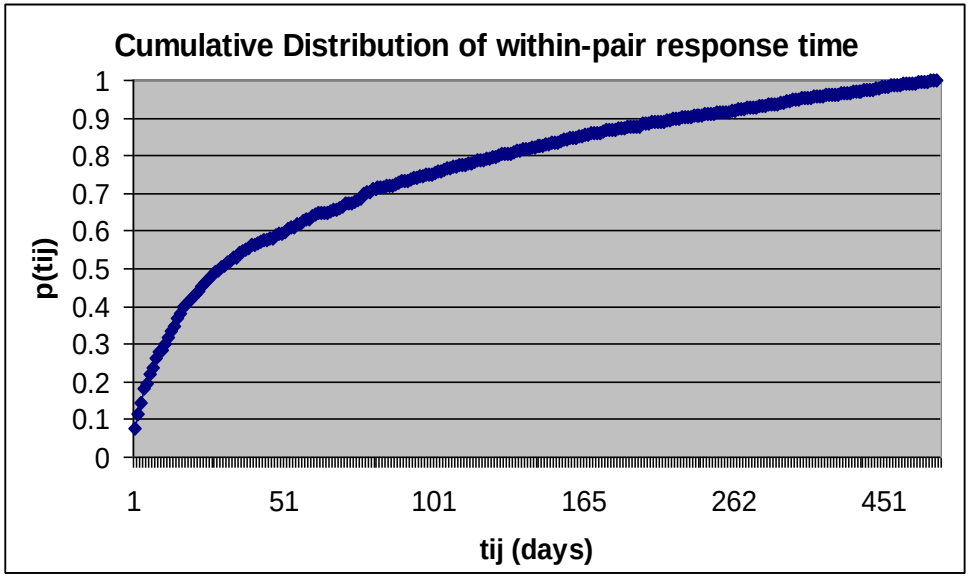
Table 1. Summary of the COPLINK dataset and the Arizona inmate dataset

	COPLINK Narcotic Data	**Arizona Inmate Data**	**Overlapped (identified by first name, last name and DOB)**
Number of People	36,548	165,540	19,608
Time Span	1990 - 2006	1985 - 2006	17 years

Determine Sampling Period and Relevancy Horizon

- The sampling period d can be calculated based on
  - **Nyquist–Shannon sampling theorem**: $\delta = 1/2 f_{\max}$, where $f_{\max}$ is the maximum frequency of link formation (i.e., co-offending a crime).

- The **relevancy horizon** is determined by
  - **Within-pair Response Time** t_{ij}: the time gap between two subsequent co-offendings by i and j.



- 90% of the response time gaps are within $t_{0.90}$= 210 days.

- Set 210 days as the **relevancy horizon** for our empirical analysis.