Analisi delle Reti Sociali

http://didawiki.cli.di.unipi.it/doku.php/dm/sna.ingegneria2011

Grafi e Proprietà delle reti

Fosca Giannotti & Michele Berlingerio, KDD Lab. ISTI-CNR

kdd.isti.cnr.it/

fosca.giannotti@isti.cnr.it, michele.berlingerio@isti.cnr.it

http://didawiki.cli.di.unipi.it/doku.php/wma/start

Jure Leskovec, Stanford CS224W: Social and Information Network Analysis, http://cs224w.stanford.edu

Class Outline

- Basic network measures - recall
- Basic network measures in Real network vs Random network
  - social, technological, business, economic, content,…
- First Social science hypotheses confirmed by large scale experiments
  - Small world: by Leskovec & Watts
- Second Social science hypotheses confirmed by large scale experiments
  - Weak & strong ties
  - Clustering coefficent, triadic closure, bridges
- Centrality Measures: betweness

Biblio

1. Onnela 2007: Structure and tie strengths in mobile communication networksJ.-
2. Planetary-Scale Views on an Instant-Messaging Network∗Jure Leskovec
3. The strenght of Weak Ties, Mrk Ganovetter†
4. An Experimental Study of Search in Global Social NetworksPeter Sheridan Dodds,1 Roby Muhamad,2 Duncan J. Watts1,2*
5. An ExperimentalStudy of the Small World Problem*JEFFREY TRAVERS Harvard UniversityAND STANLEY MILGRAM

Basic measures

KEY MEASURES

Degree distribution: **P(k)**

Path length: ***l***

Clustering coefficient:

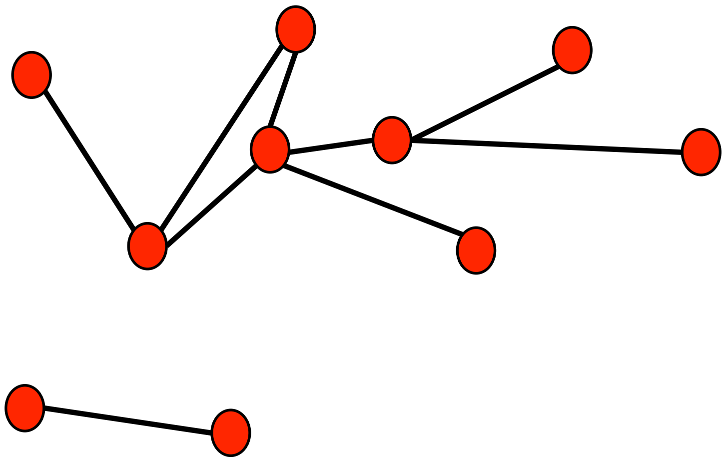
$$\boxed{C_i = \frac{2e_i}{k_i(k_i - 1)}}$$

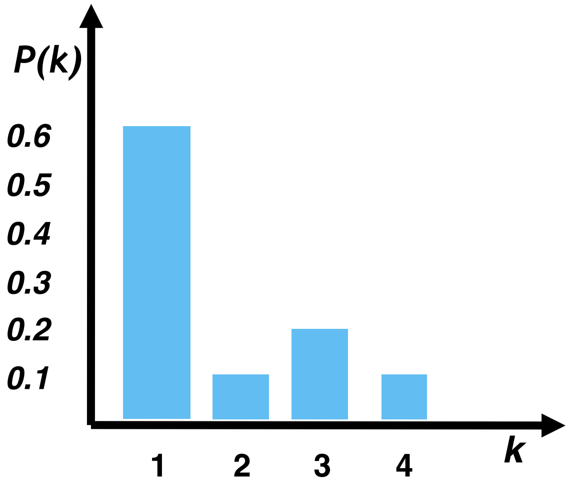
Degree distribution

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

N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot





NETWORK DIAMETER AND AVERAGE DISTANCE

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

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

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

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

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

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

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

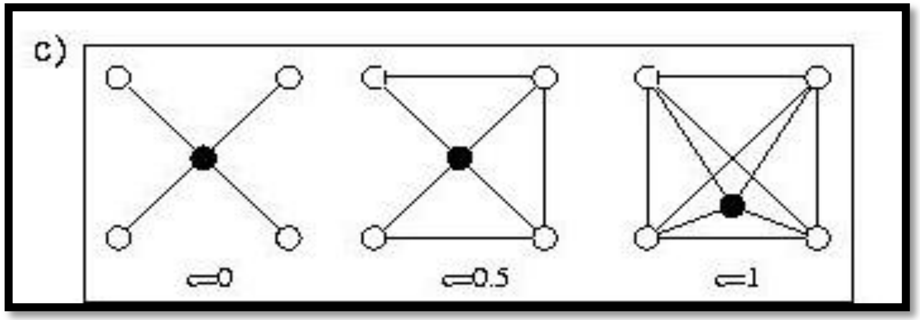
CLUSTERING COEFFICIENT

Clustering coefficient:

what portion of your neighbors are connected?

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

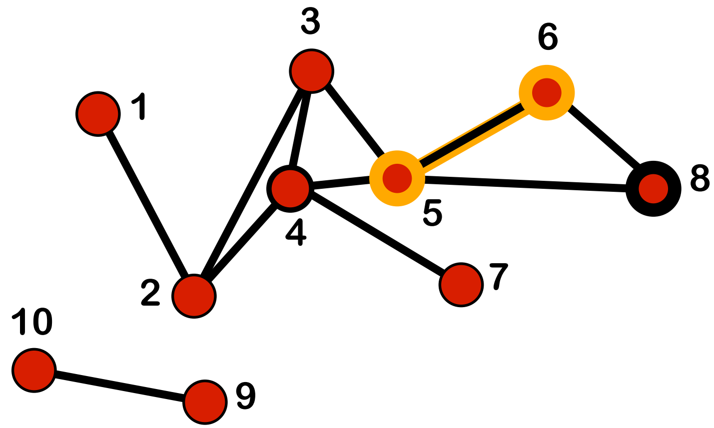
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



CLUSTERING COEFFICIENT

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

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

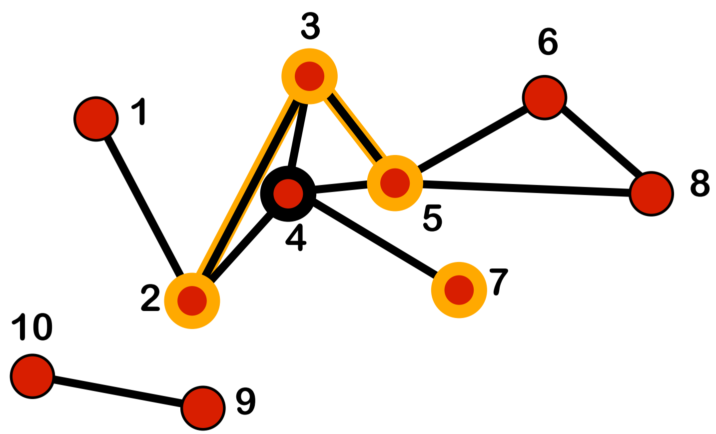


i=8: k_8=2, e_8=1, TOT=2*1/2=1 → C_8=1/1=1

CLUSTERING COEFFICIENT

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

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

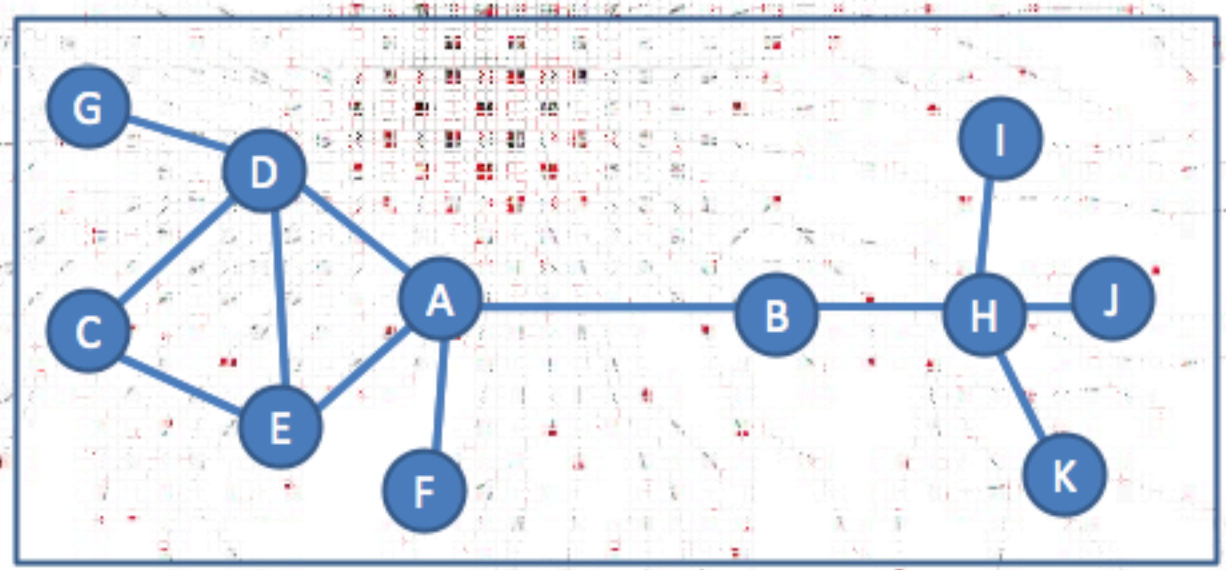


$i=4$: $k_4=4$, $e_4=2$, TOTAL=4*3/2=6 → $C_4=2/6=1/3$

Clustering Coefficient, Transitivity

$C_i = 2\Delta / k(k-1)$

$C_A = 2/12 = 1/6$ $C_C = 2/2 = 1$ $C_E = 4/6 = 2/3$



Topological Overlap
Mutual Clustering

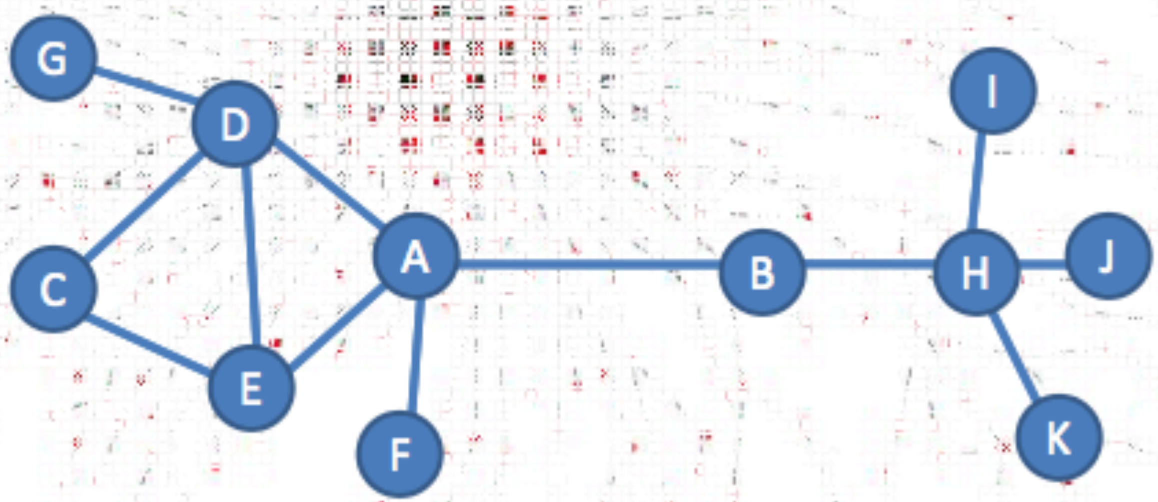
$$TO(A,B)=Overlap(A,B)/NormalizingFactor(A,B)$$

$TO(A,B)=N(A,B)/\max(k(A),k(B))$

$TO(A,B)=N(A,B)/\min(k(A),k(B))$

$TO(A,B)=N(A,B)/(k(A)\times k(B))^{1/2}$

$TO(A,B)=N(A,B)/(k(A)+k(B))$



Topological Overlap
Mutual Clustering

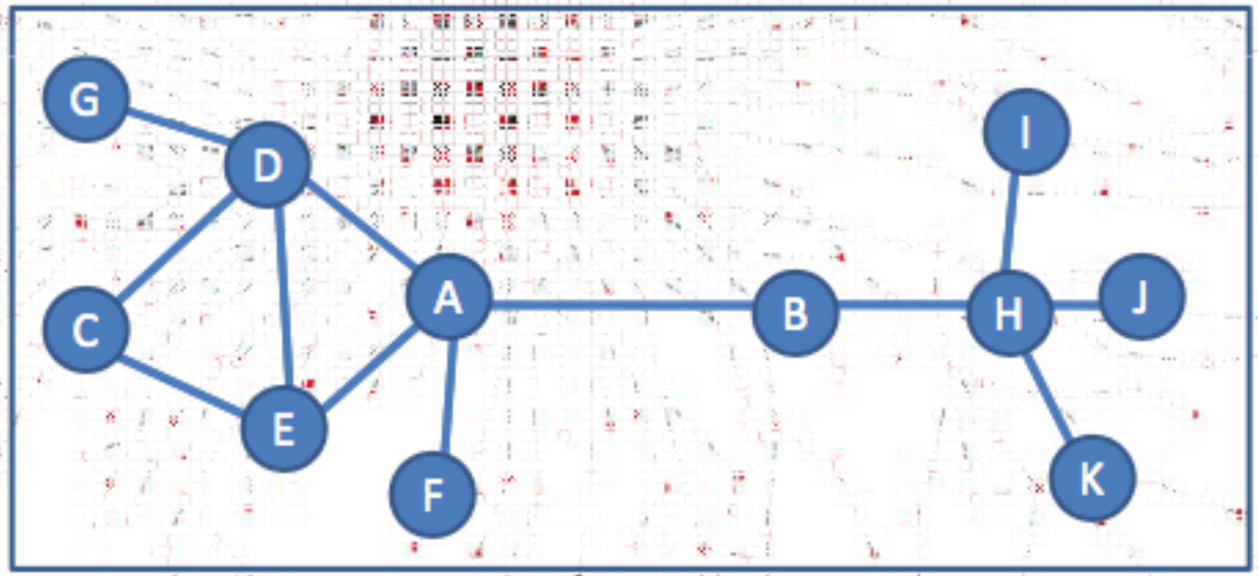
$$TO(A,B)=Overlap(A,B)/NormalizingFactor(A,B)$$

$$TO(A,B)=N(A,B)/\max(k(A),k(B))$$

$$TO(A,B)=N(A,B)/\min(k(A),k(B))$$

$$TO(A,B)=N(A,B)/\left(k(A)\times k(B)\right)^{1/2}$$

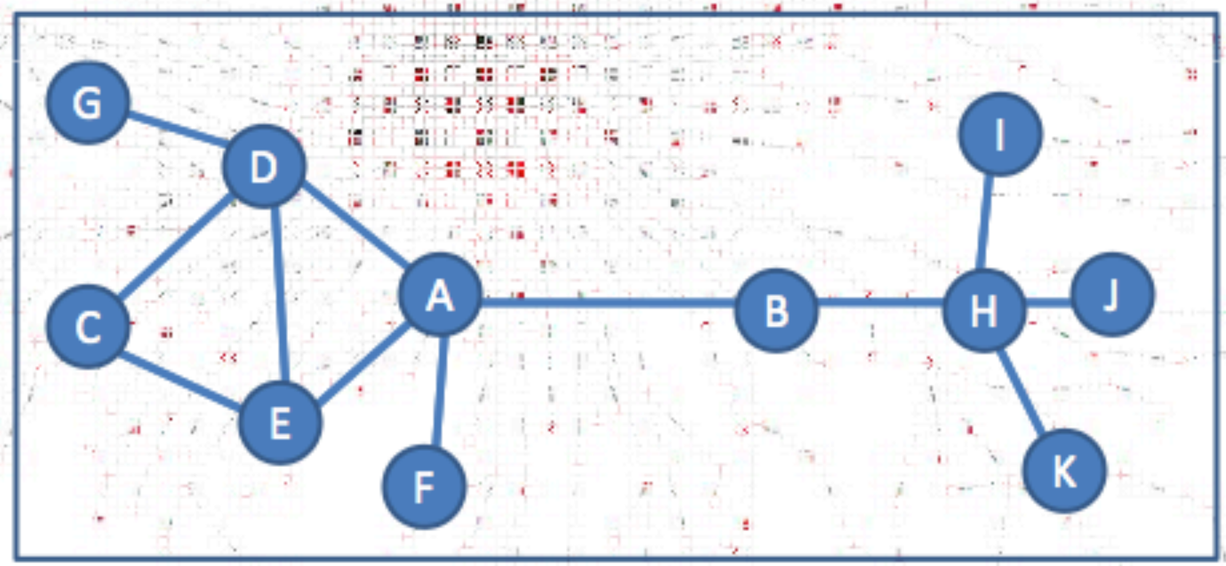
$$TO(A,B)=N(A,B)/(k(A)+k(B))$$



Topological Overlap
Mutual Clustering

TO(A,B)=N(A,B)/max(k(A),k(B))

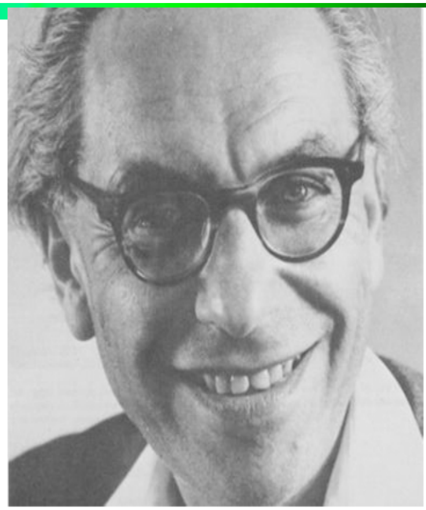
TO(A,B)=0
TO(A,D)=1/4
TO(E,D)=2/4

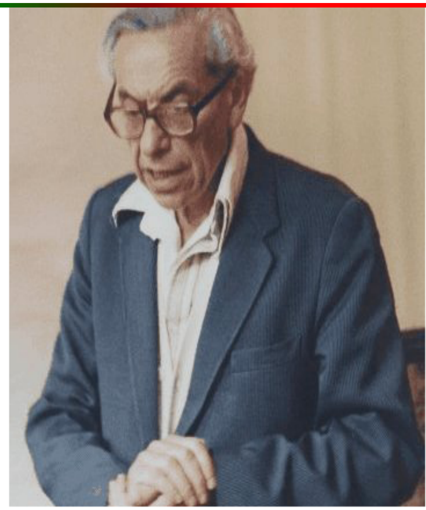


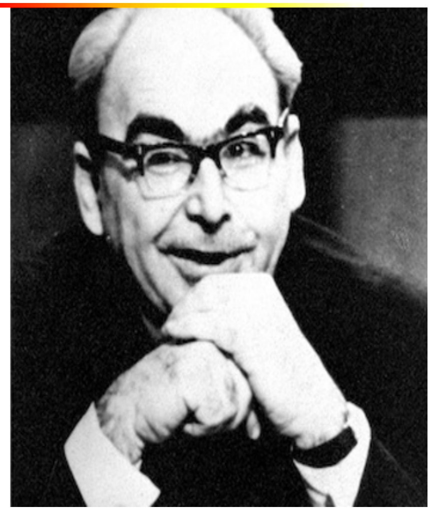
Real networks vs random networks

RANDOM NETWORK MODEL

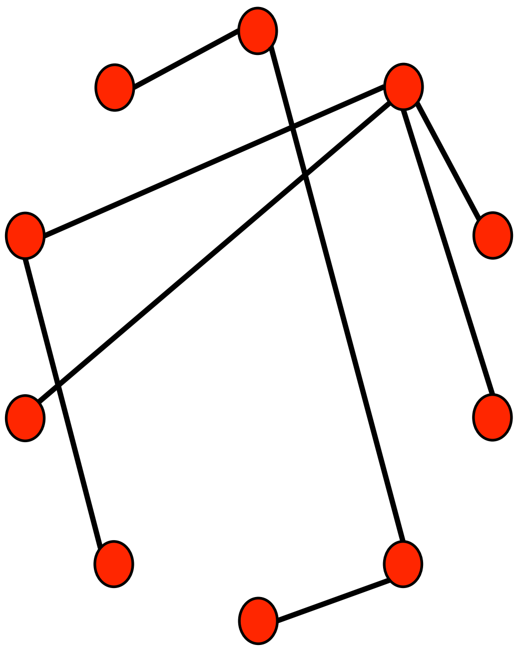
Pául Erdös
(1913-1996)







Erdös-Rényi model (1960)

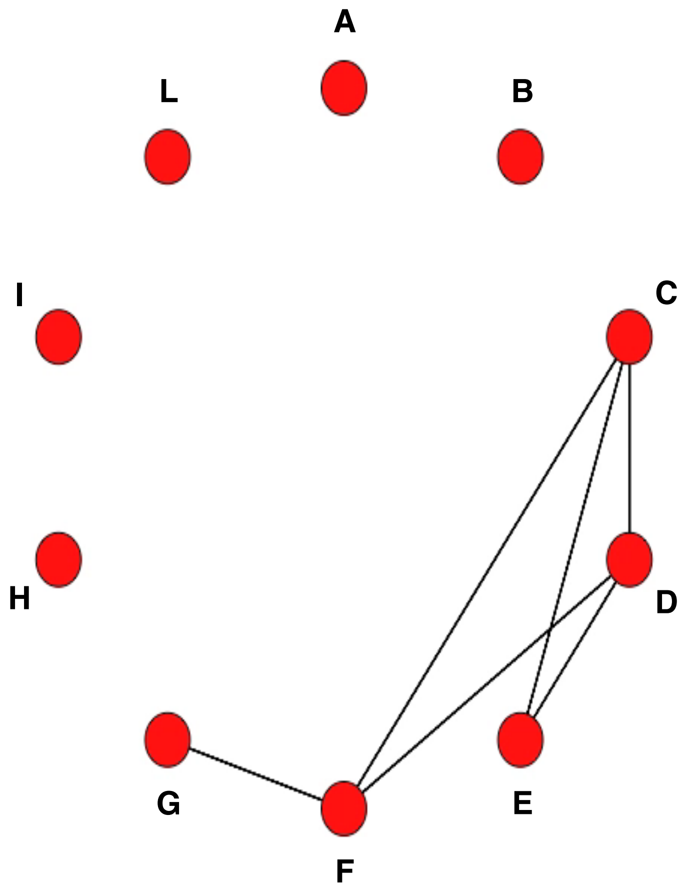


Connect with probability p

p=1/6 N=10

$\langle k \rangle \sim 1.5$

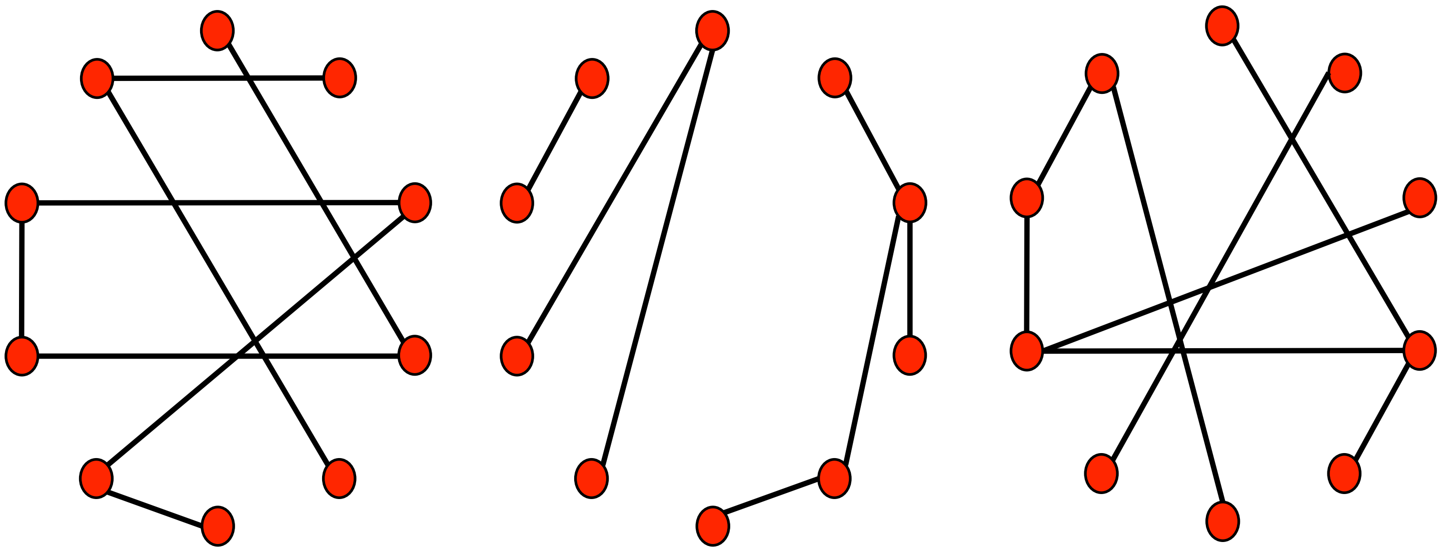
RANDOM NETWORK MODEL



Definition: A **random graph** is a graph of N labeled nodes where each pair of nodes is connected by a preset probability **p**.

RANDOM NETWORK MODEL

N and ***p*** do not uniquely define the network– we can have many different realizations of it. **How many?**



N=10
p=1/6

The probability to form a *particular* graph **G(N,L)** is

$$P(G(N,L)) = p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

That is, each graph **G(N,L)** appears with probability **P(G(N,L))**.

RANDOM NETWORK MODEL

P(L): the probability to have a network of exactly *L* links

$$P(L) = \binom{\binom{N}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

•The average number of links *<L>* in a random graph

$$<L> = \sum_{L=0}^{\frac{N(N-1)}{2}} LP(L) = p\frac{N(N-1)}{2}$$

$$<k> = 2L/N = p(N-1)$$

•The standard deviation

$$\sigma^2 = p(1-p)\frac{N(N-1)}{2}$$

RANDOM NETWORK MODEL

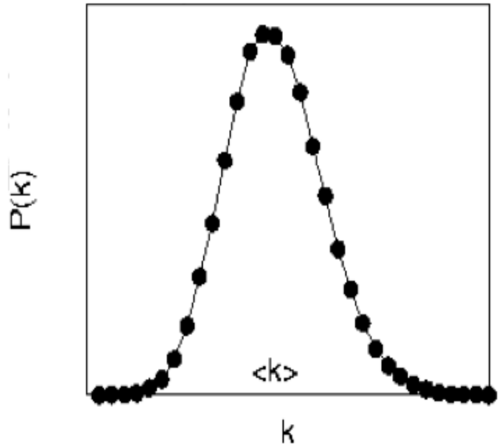
P(L): the probability to have exactly *L* links in a network of *N* nodes and probability *p*:

The maximum number of links in a network of N nodes.

$$P(L) = \underbrace{\binom{\binom{N}{2}}{L}}_{\text{Number of different ways we can choose L links among all potential links.}} p^L (1-p)^{\frac{N(N-1)}{2} - L}$$

Binomial distribution...

DEGREE DISTRIBUTION OF A RANDOM GRAPH



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

Select k
nodes from N-1

probability of
having *k* edges

probability of
missing N-1-k
edges

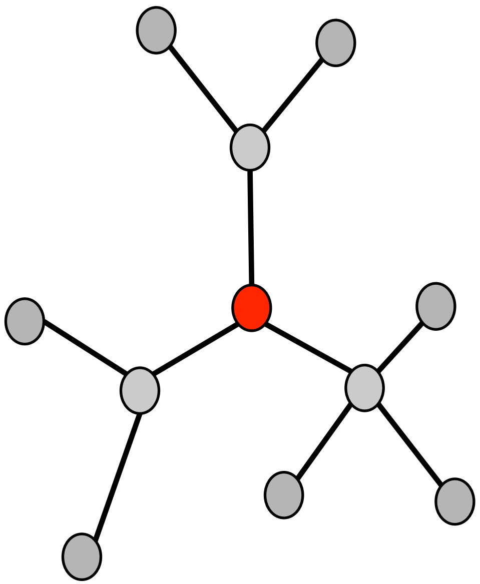
$$<k> = p(N-1) \qquad \sigma_k^2 = p(1-p)(N-1)$$

$$\frac{\sigma_k}{<k>} = \left[\frac{1-p}{p}\frac{1}{(N-1)}\right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

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>.

DISTANCES IN RANDOM GRAPHS

Random graphs tend to have a tree-like topology with almost constant node degrees.



- nr. of first neighbors: $N_1 \cong \langle k \rangle$
- nr. of second neighbors: $N_2 \cong \langle k \rangle^2$
- nr. of neighbours at distance d: $N_d \cong \langle k \rangle^d$
- estimate maximum distance:

$$1 + \sum_{l=1}^{l_{max}} \langle k \rangle^i = N \quad \Rightarrow \quad l_{max} = \frac{\log N}{\log \langle k \rangle}$$

DISTANCES IN RANDOM GRAPHS — compare with real data

$$l_{max} = \frac{\log N}{\log\langle k\rangle}$$

Network	Size	⟨k⟩	l	l_{rand}	C	C_{rand}	Reference	Nr
www, site level, undir	153127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015-6209	3.52-4.11	3.7-3.76	6.36-6.18	0.18-0.3	0.001	Yook e al., 2001a, Pastor-Satorras et al., 2001	2
Movie actors	225226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz,1998	3
LANL co-authorship	52909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE eo-authorship	1520251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabasi et al, 2001	8
Neurosci. co-authorship	209293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabasi et al, 2001	9
E. coli, sustrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Sole, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Sole, 2000	13
Words, co-occurrence	460902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Sole, 2001	14
Words, synonyms	22311	13.48	4.5	3.84	0.7	0.0006	Yook et al. 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C.Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Given the huge differences in scope, size, and average degree, the agreement is excellent.

Erdös-Rényi MODEL (1960)

•**Degree distribution**

Binomial, Poisson (exponential tails)

•**Clustering coefficient**

Vanishing for large network sizes

•**Average distance among nodes**

Logarithmically small

Are real networks like random graphs?

ARE REAL NETWORKS LIKE RANDOM GRAPHS?

As quantitative data about real networks became available, we can compare their topology with the predictions of random graph theory.

Note that once we have N and <k> for a random network, from it we can derive every measurable property. Indeed, we have:

<u>Average path length:</u>

$$< l_{rand} > \approx \frac{\log N}{\log \langle k \rangle}$$

<u>Clustering Coefficient:</u>

$$C_{rand} = p = \frac{\langle k \rangle}{N}$$

<u>Degree Distribution:</u>

$$P_{rand}(k) \cong C^{k}_{N-1} p^{k} (1-p)^{N-1-k}$$

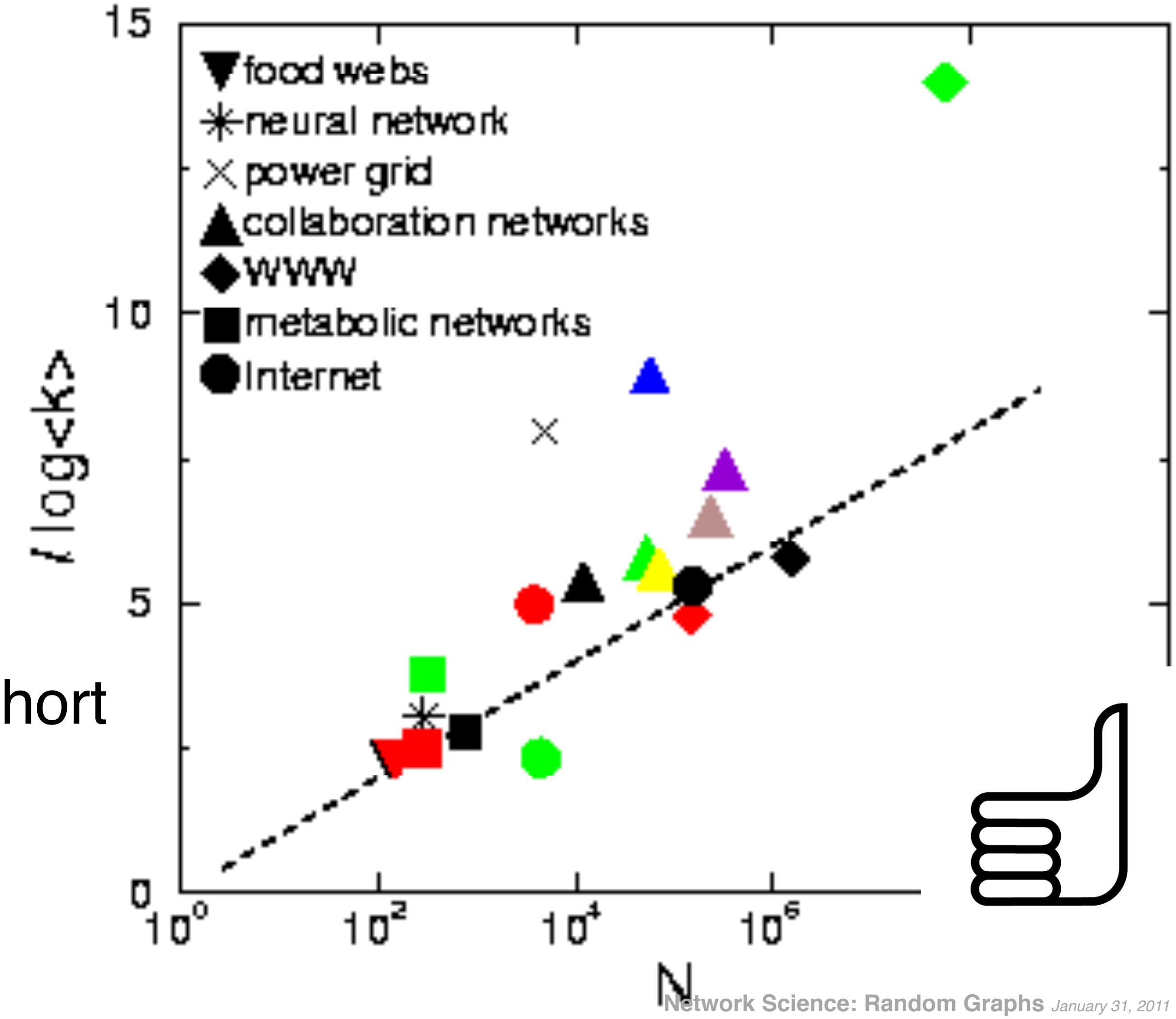
$$P(k) = e^{-<k>} \frac{<k>^{k}}{k!}$$

PATH LENGTHS IN REAL NETWORKS

Prediction:

$$l_{rand} = \frac{\log N}{\log\langle k\rangle}$$

Data:

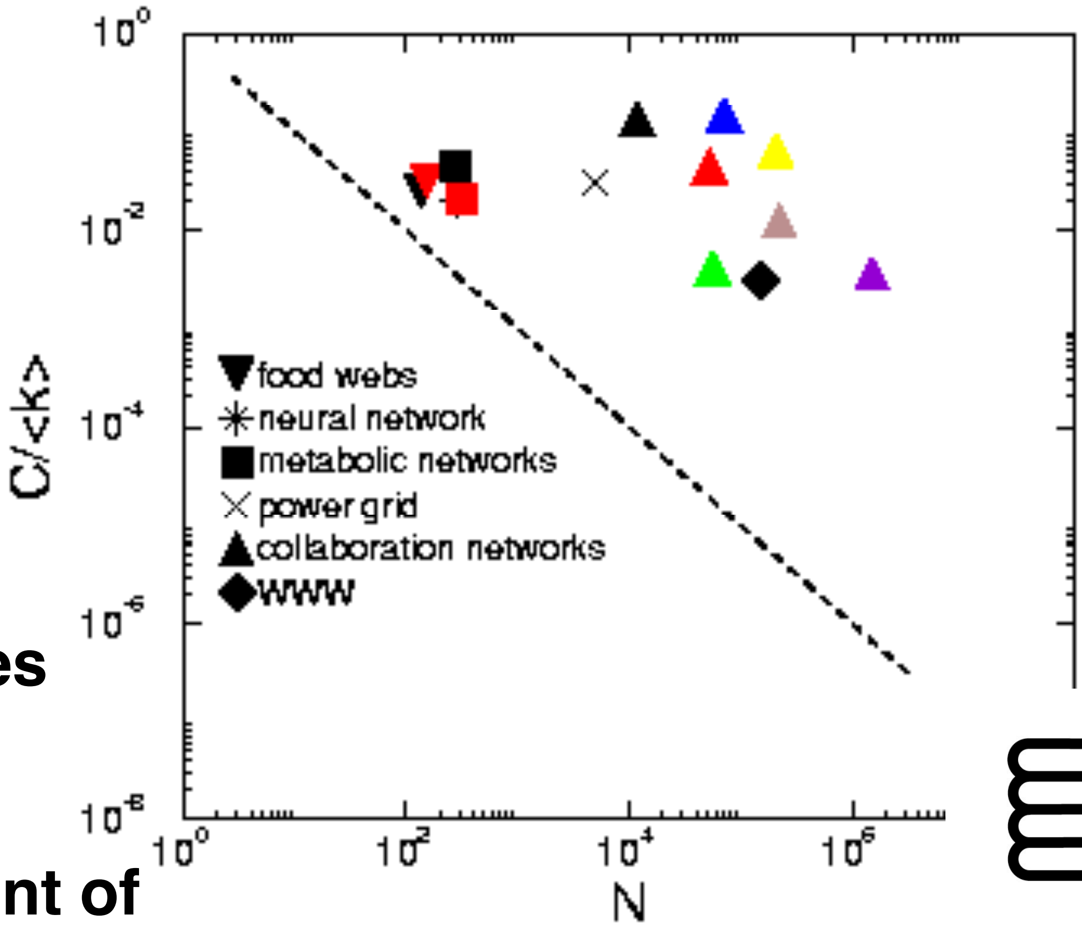


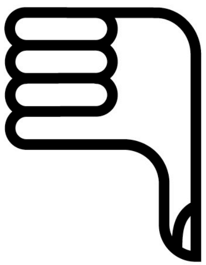
Real networks have short distances
like random graphs.

Prediction:

$$C_{rand} = \frac{\langle k \rangle}{N}$$

Data:





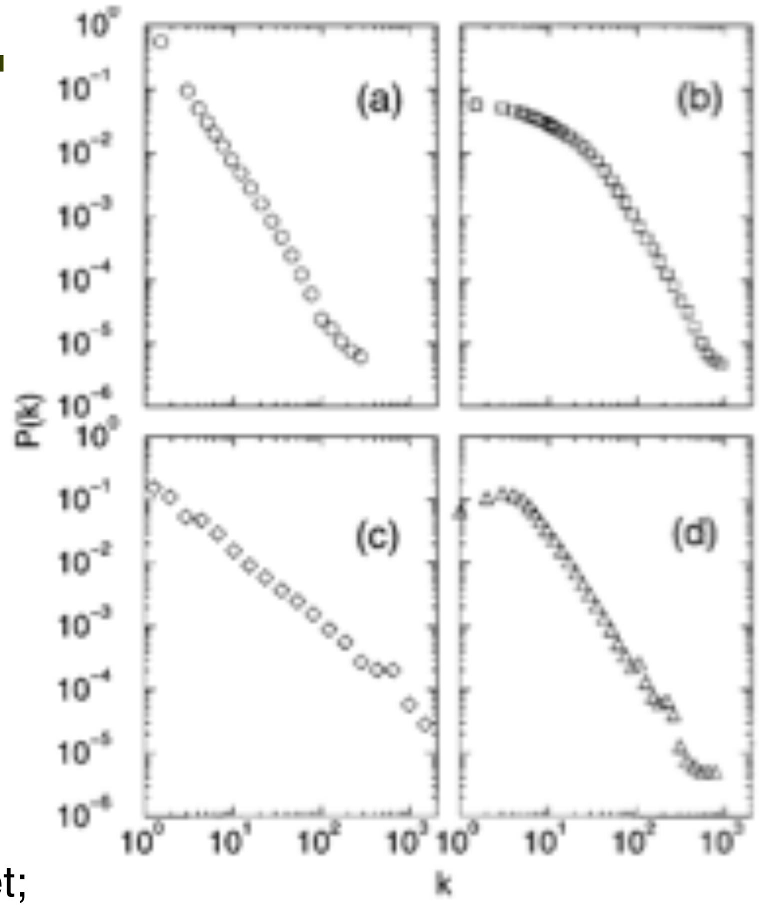
C_{rand} underestimates with orders of magnitudes the clustering coefficient of real networks.

Prediction:

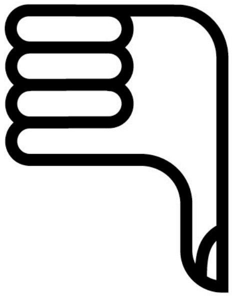
$$P_{rand}(k) \cong C^{k}_{N-1} p^{k} (1-p)^{N-1-k}$$

Data:

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



(a) Internet;
(b) Movie Actors;
(c) Coauthorship, high energy physics;
(d) Coauthorship, neuroscience



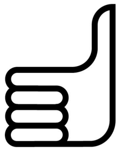
ARE REAL NETWORKS LIKE RANDOM GRAPHS?

As quantitative data about real networks became available, we can compare their topology with the predictions of random graph theory.

Note that once we have N and <k> for a random network, from it we can derive every measurable property. Indeed, we have:

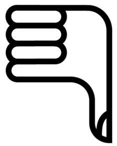
Average path length:

$$< l_{rand} > \approx \frac{\log N}{\log\langle k\rangle}$$



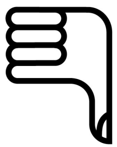
Clustering Coefficient:

$$C_{rand} = p = \frac{\langle k\rangle}{N}$$



Degree Distribution:

$$P_{rand}(k) \cong C^{k}_{N-1} p^{k} (1-p)^{N-1-k}$$

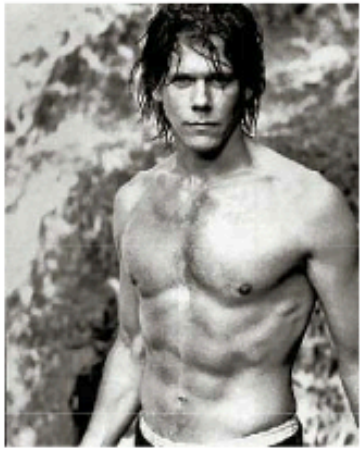


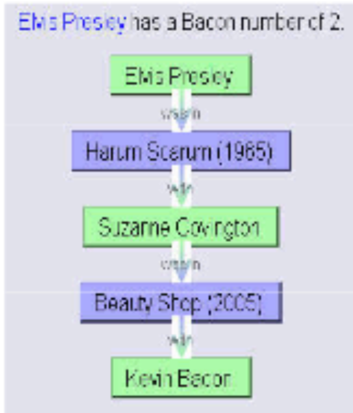
Social network as Small World

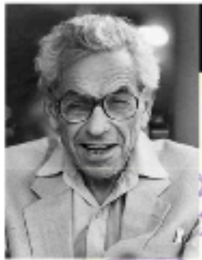
Six Degrees of Kevin Bacon

Origins of a small-world idea:

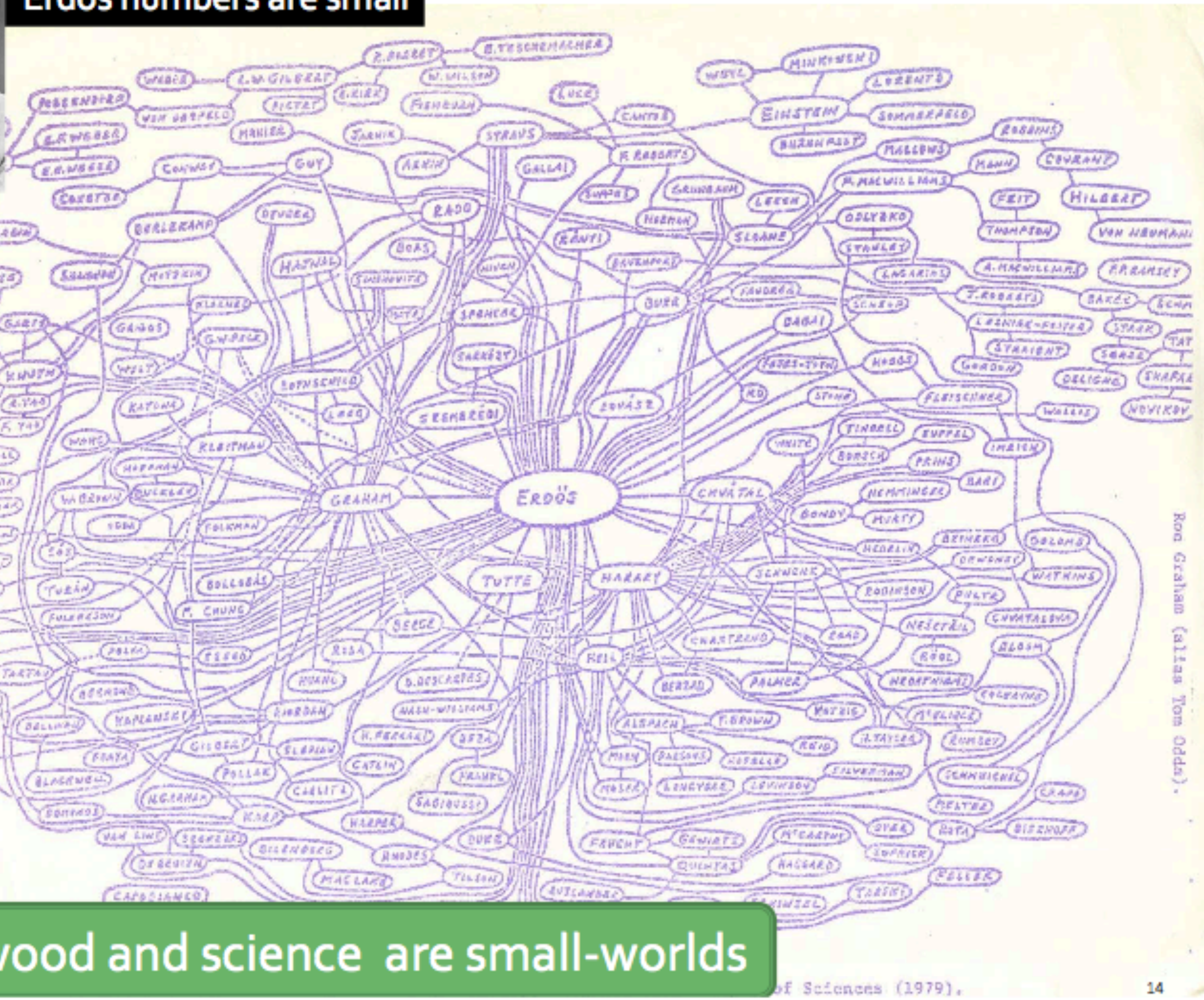
- Bacon number:
  - Create a network of Hollywood actors
  - Connect two actors if they co-appeared in the movie
  - Bacon number: number of steps to Kevin Bacon
- As of Dec 2007, the highest (finite) Bacon number reported is 8
- Only approx. 12% of all actors cannot be linked to Bacon







Erdos numbers are small



Hollywood and science are small-worlds

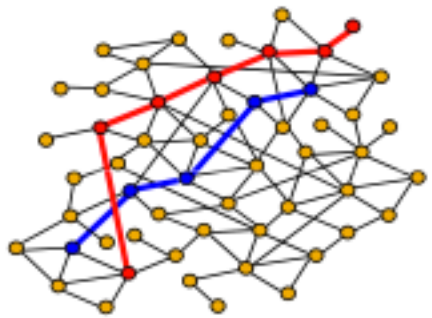
9/22/2010

14

The Small-world experiment

- What is the typical shortest path length between any two people?
  - Experiment on the global friendship network
    - Can't measure, need to probe explicitly
- The Small-world experiment [Stanley Milgram '67]
  - Picked 300 people at random
  - Ask them to get a letter to a by passing it through friends to a stockbroker in Boston
- How many steps does it take?

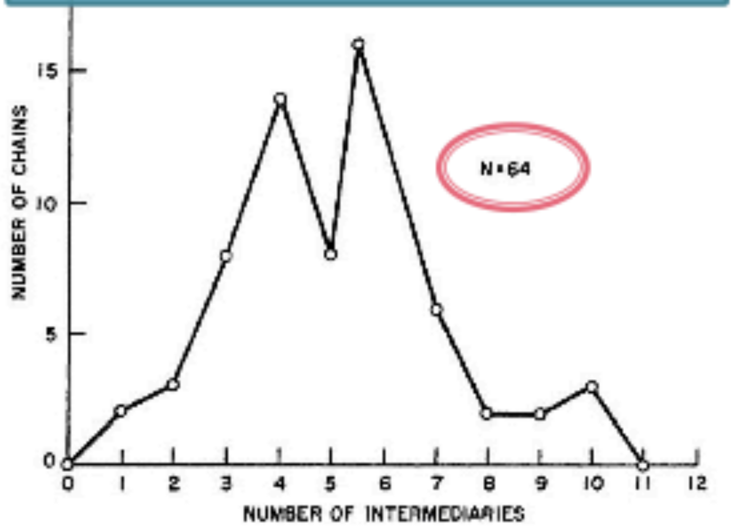
Stanley Milgram



The Small-world experiment

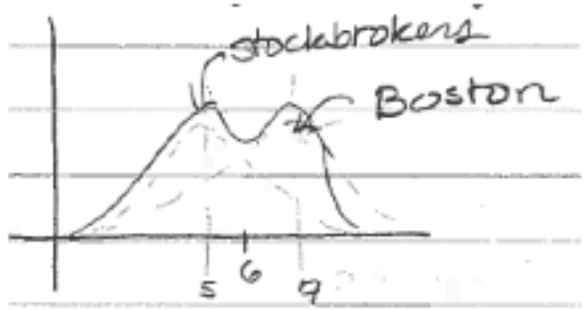
- 64 chains completed:
  - 6.2 on the average, thus "6 degrees of separation"

Milgram's small world experiment



- Further observations:
  - People what owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
  - People from the Boston area have even closer paths: 4.4

Milgram: Further observations

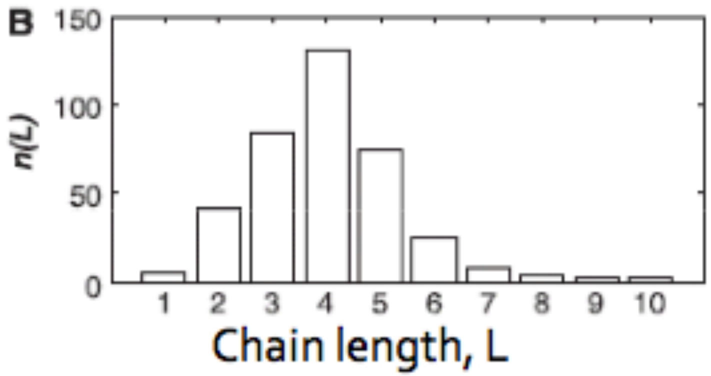


- People use different networks: Boston vs. occupation
- Criticism:
  - Funneling:
    - 31 of 64 chains passed through 1 of 3 people ass their final step → Not all links/nodes are equal
  - Choice of starting points and the target were non-random
  - People refuse to participate (25% for Milgram)
  - Some sort of social search: People in the experiment follow some strategy (e.g., geographic routing) instead of forwarding the letter to everyone. They are not finding the shortest path.
  - There are not many samples.
  - People might have used extra information resources.

[Dodds-Muhamad-Watts, '03]

Columbia small-world study

- In 2003 Dodds, Muhamad and Watts performed the experiment using email:
  - 18 targets of various backgrounds
  - 24,000 first steps (~1,500 per target)
  - 65% dropout per step
  - 384 chains completed (1.5%)



Avg. chain length = 4.01

PROBLEM: Huge drop-out rate, i.e., longer chains are less likely to complete

Correcting for the drop-out rate

- Huge drop-out rate:
  - Longer chains don't complete

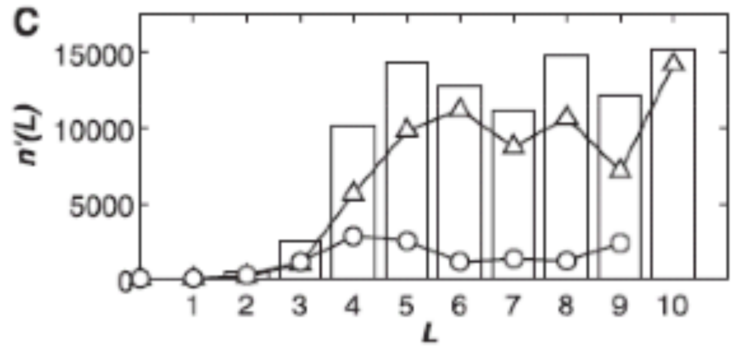
Correction proposed by Harrison-White. Let:

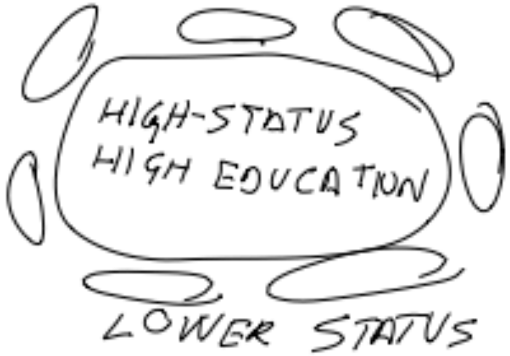
  - f_j = true (unobserved) fraction of chains that would have length j
  - N = total # of starters
  - N_j = # starters who reached target in j steps
  - Then: $f_j^* := N_j/N$
  - Assume drop-out rate 1-α in each step, so $f_j^* := f_j\,\alpha^j$
  - $\sum_j f_j=1 \rightarrow \sum_j f_j^*\,\alpha^j=1$
  - Observe f_j^*, calculate the average dropout rate 1-α and

then $f_j = f_j^* \cdot \alpha^{-j}$

Small-world in soc. networks

- After the correction:
  - Typical path length L=7
    (MEDIAN)
- Some not well understood phenomena in social networks:
  - Funneling effect: some target's friends are more likely to be the final step.
    - Conjecture: High reputation/authority
  - Effects of target's characteristics: structurally why are high-status target easier to find
    - Conjecture: Core-periphery net structure





18 target persons: Status/Authority

Target	*City*	*Country*	*Occupation*	*Gender*	*N*	N_c (%)	r (n_r)	$\langle L \rangle$
1	Novosibirsk	Russia	PhD student	F	8234	20(0.24)	64 (76)	4.05
2	New York	USA	Writer	F	6044	31 (0.51)	65 (73)	3.61
3	Bandung	Indonesia	Unemployed	M	8151	0	66 (76)	n/a
4	New York	USA	Journalist	F	5690	44 (0.77)	60 (72)	3.9
5	Ithaca	USA	Professor	M	5855	168 (2.87)	54 (71)	3.84
6	Melbourne	Australia	Travel Consultant	F	5597	20 (0.36)	60 (71)	5.2
7	Bardufoss	Norway	Army veterinarian	M	4343	16 (0.37)	63 (76)	4.25
8	Perth	Australia	Police Officer	M	4485	4 (0.09)	64 (75)	4.5
9	Omaha	USA	Life Insurance Agent	F	4562	2 (0.04)	66 (79)	4.5
10	Welwyn Garden City	UK	Retired	M	6593	1 (0.02)	68 (74)	4
11	Paris	France	Librarian	F	4198	3 (0.07)	65 (75)	5
12	Tallinn	Estonia	Archival Inspector	M	4530	8 (0.18)	63(79)	4
13	Munich	Germany	Journalist	M	4350	32 (0.74)	62 (74)	4.66
14	Split	Croatia	Student	M	6629	0	63 (77)	n/a
15	Gurgaon	India	Technology Consultant	M	4510	12 (0.27)	67 (78)	3.67
16	Managua	Nicaragua	Computer analyst	M	6547	2 (0.03)	68 (78)	5.5
17	Katikati	New Zealand	Potter	M	4091	12 (0.3)	62 (74)	4.33
18	Elderton	USA	Lutheran Pastor	M	4438	9 (0.21)	68 (76)	4.33
Totals					98,847	384 (0.4)	63 (75)	4.05

HIGH STATUS

- N… # people assigned to correspond to target
- N_c…# completed chains
- r… frac. of people who did not forward
- L… mean path length

18 target persons: Status/Authority

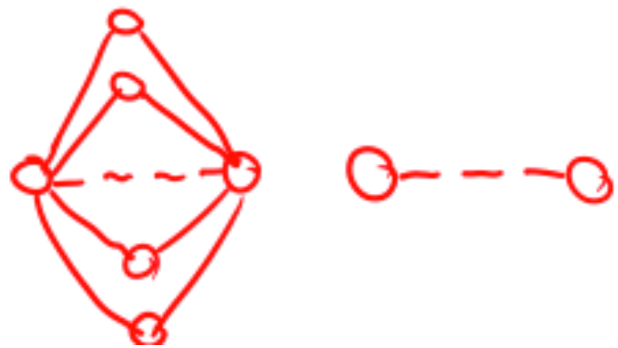
Target	City	Country	Occupation	Gender	N	N_c (%)	r (n_r)	$\langle L \rangle$
1	Novosibirsk	Russia	PhD student	F	8234	20(0.24)	64 (76)	4.05
2	New York	USA	Writer	F	6044	31 (0.51)	65 (73)	3.61
3	Bandung	Indonesia	Unemployed	M	8151	0	66 (76)	n/a
4	New York	USA	Journalist	F	5690	44 (0.77)	60 (72)	3.9
5	Ithaca	USA	Professor	M	5855	168 (2.87)	54 (71)	3.84
6	Melbourne	Australia	Travel Consultant	F	5597	20 (0.36)	60 (71)	5.2
7	Bardufoss	Norway	Army veterinarian	M	4343	16 (0.37)	63 (76)	4.25
8	Perth	Australia	Police Officer	M	4485	4 (0.09)	64 (75)	4.5
9	Omaha	USA	Life Insurance Agent	F	4562	2 (0.04)	66 (79)	4.5
10	Welwyn Garden City	UK	Retired	M	6593	1 (0.02)	68 (74)	4
11	Paris	France	Librarian	F	4198	3 (0.07)	65 (75)	5
12	Tallinn	Estonia	Archival Inspector	M	4530	8 (0.18)	63(79)	4
13	Munich	Germany	Journalist	M	4350	32 (0.74)	62 (74)	4.66
14	Split	Croatia	Student	M	6629	0	63 (77)	n/a
15	Gurgaon	India	Technology Consultant	M	4510	12 (0.27)	67 (78)	3.67
16	Managua	Nicaragua	Computer analyst	M	6547	2 (0.03)	68 (78)	5.5
17	Katikati	New Zealand	Potter	M	4091	12 (0.3)	62 (74)	4.33
18	Elderton	USA	Lutheran Pastor	M	4438	9 (0.21)	68 (76)	4.33
Totals					98,847	384 (0.4)	63 (75)	4.05

HIGH STATUS

- N… # people assigned to correspond to target
- N_c…# completed chains
- r… frac. of people who did not forward
- L… mean path length

6-degrees: Should we be surprised?

- Assume each human is connected to 100 other people:
- So:
  - In step 1 she can reach 100 people
  - In step 2 she can reach 100*100 = 10,000 people
  - In step 3 she can reach 100*100*100 = 100,000 people
  - In 5 steps she can reach 10 billion people
- What's wrong here?
  - Many edges are local ("short"): friend of a friend

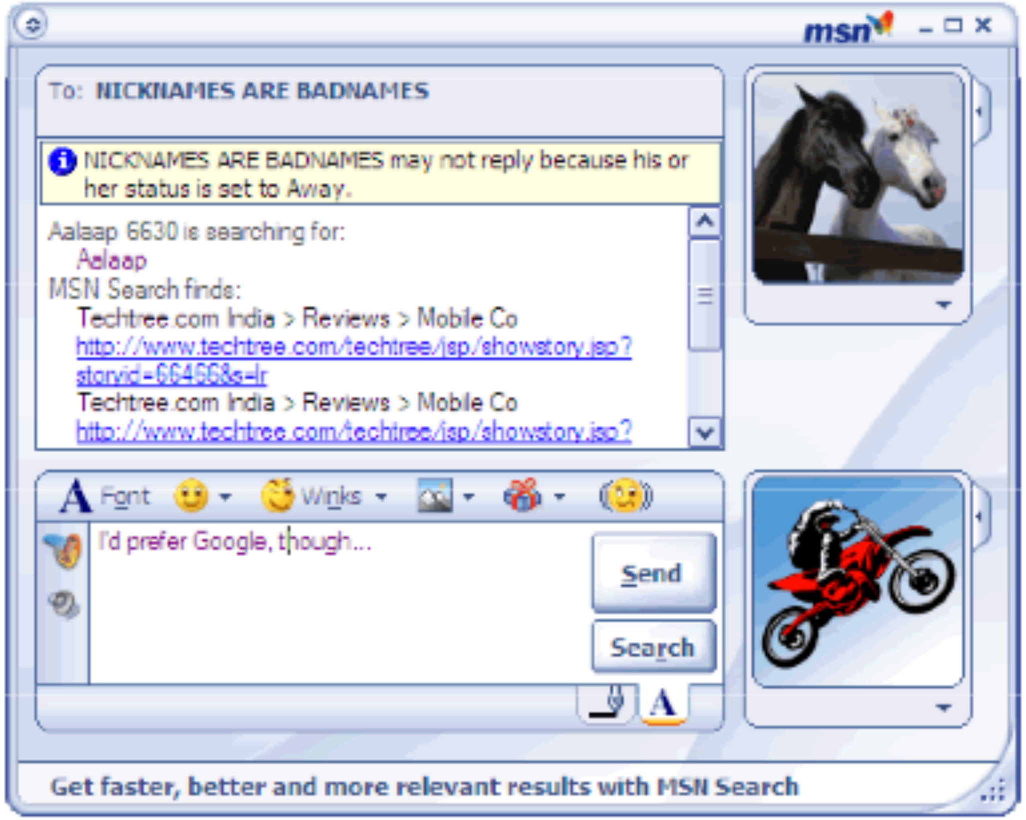


Planetary-Scale Views on an Instant-Messaging Network

∗Jure Leskovec†

Machine Learning DepartmentCarnegie Mellon University Pittsburgh, PA, USAEric HorvitzMicrosoft Research Redmond, WA, USAMicrosoft Research Technical Report MSR-TR-2006-186June 2007

IM experiment



- Contact (buddy) list
- Messaging window

Data statistics

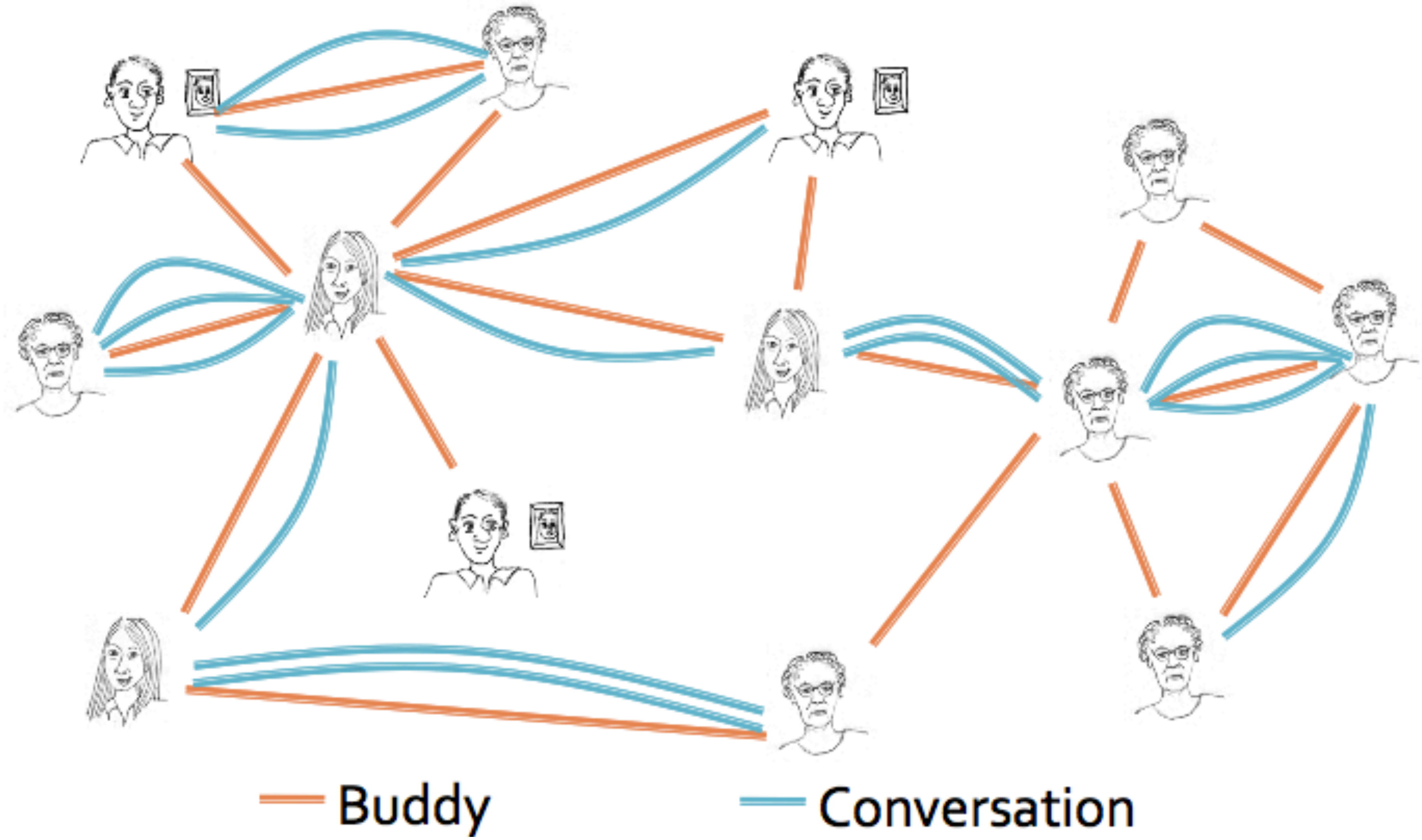
- Data for June 2006
- Log size:
  150Gb/day (compressed)
- Total: 1 month of communication data:
  4.5Tb of compressed data
- Activity over June 2006 (30 days)
  - 245 million users logged in
  - 180 million users engaged in conversations
  - 17,5 million new accounts activated
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

Data statistics: typical day

Activity on a typical day (June 1 2006):

- 1 billion conversations
- 93 million users login
- 65 million different users talk (exchange messages)
- 1.5 million invitations for new accounts sent

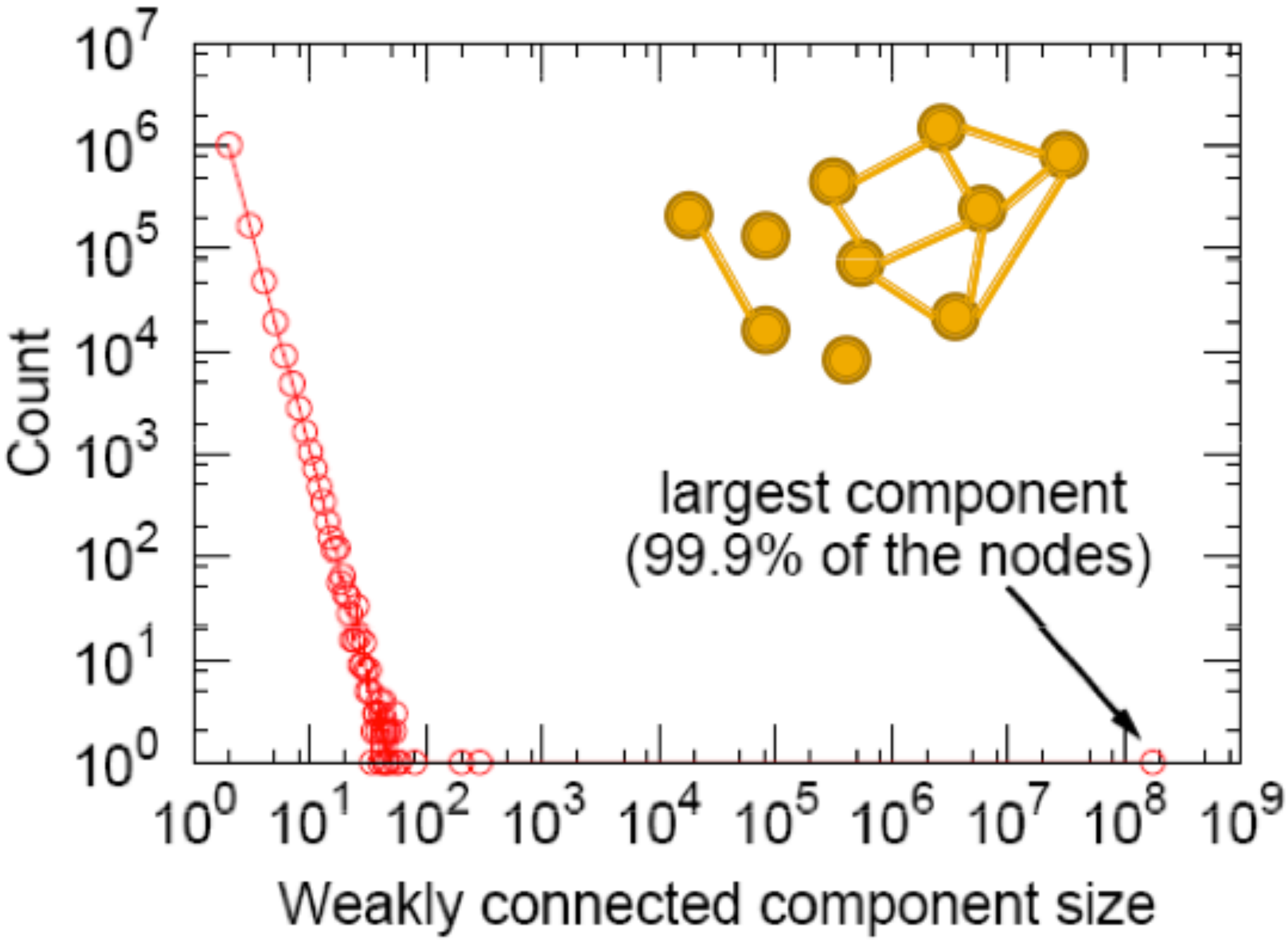
Messaging as a network



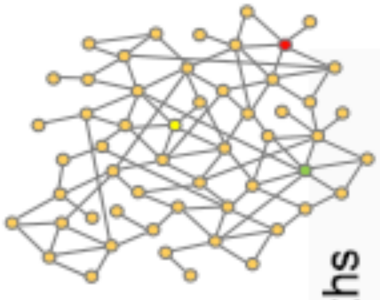
IM communication network

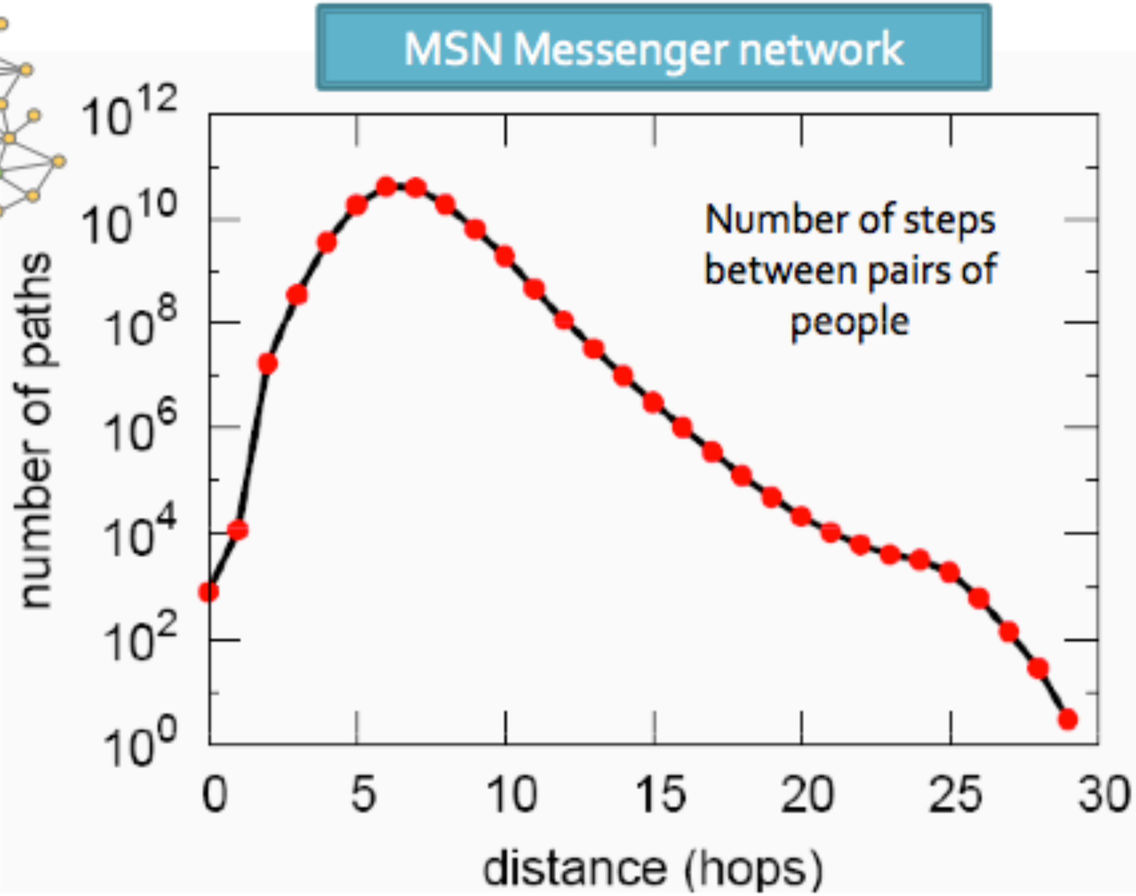
- Buddy graph
  - 240 million people (people that login in June '06)
  - 9.1 billion buddy edges (friendship links)
- Communication graph (take only 2-user conversations)
  - Edge if the users exchanged at least 1 message
  - 180 million people
  - 1.3 billion edges
  - 30 billion conversations

Network connectivity



MSN Network: Small world





Avg. path length **6.6**

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

Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Strenght of weak ties in Social Networks

Networks: Flow of information

- How information flows through the network?
- How different nodes can play structurally distinct roles in this process?
- How different links (short range vs. long range) play different roles in diffusion?

[Granovetter '73]

Strength of weak ties

- How people find out about new jobs?
  - Mark Granovetter, part of his PhD in 1960s
  - People find the information through personal contacts
- But: Contacts were often acquaintances rather than close friends
  - This is surprising:
    - One would expect your friends to help you out more than casual acquaintances when you are between the jobs
- Why is it that distance acquaintances are most helpful?

[Granovetter '73]

Granovetter's answer

- Two perspectives on friendships:
  - Structural:
    - Friendships span different portions of the network
  - Interpersonal:
    - Friendship between two people is either strong or weak

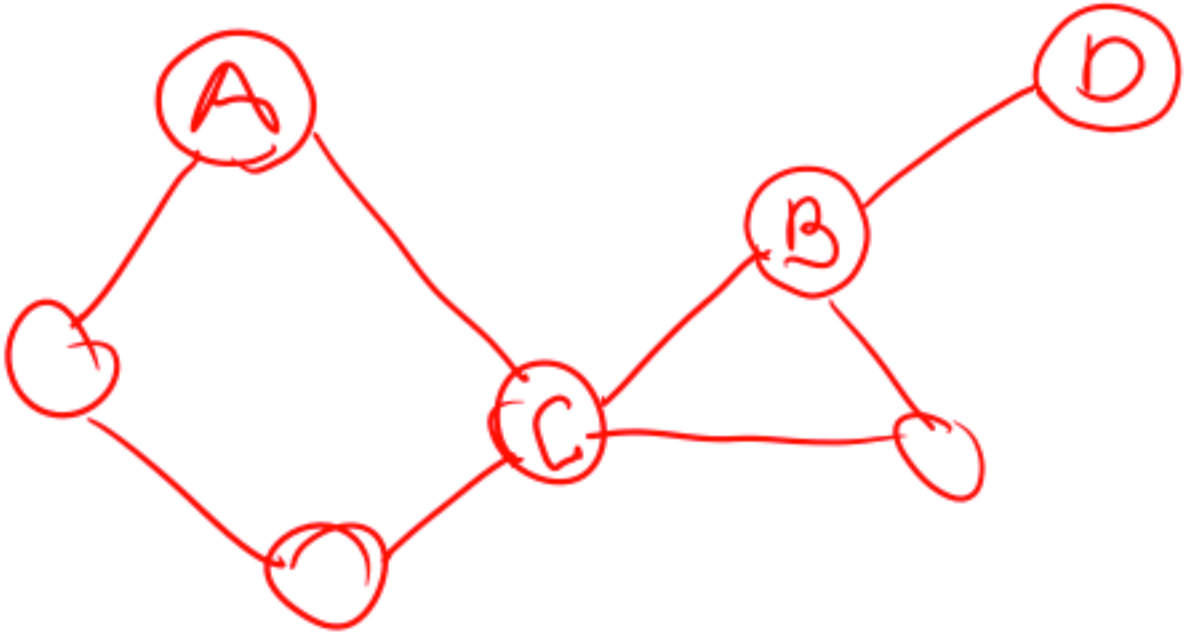
[Granovetter '73]

Granovetter's answer

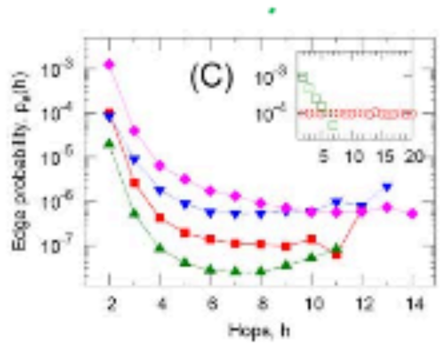
- Two perspectives on friendships:
  - Structural:
    - Friendships span different portions of the network
  - Interpersonal:
    - Friendship between two people is either strong or weak

Triadic closure

- Which edge is more likely A-B or A-D?



- Triadic closure: If two people in a network have a friend in common there is an increased likelihood they will become friends themselves



Triadic closure

- Triadic closure == High clustering coefficient

Reasons for triadic closure:

- If B and C have a friend A in common, then:
  - B is more likely to meet C
    - (since they both spend time with A)
  - B and C trust each other
    - (since they have a friend in common)
  - A has incentive to bring B and C together
    - (as it is hard for A to maintain two disjoint relationships)

- Empirical study by Bearman and Moody:
  - Teenage girls with low clustering coefficient are more likely to contemplate suicide

Triadic closure

- Triadic closure == High clustering coefficient

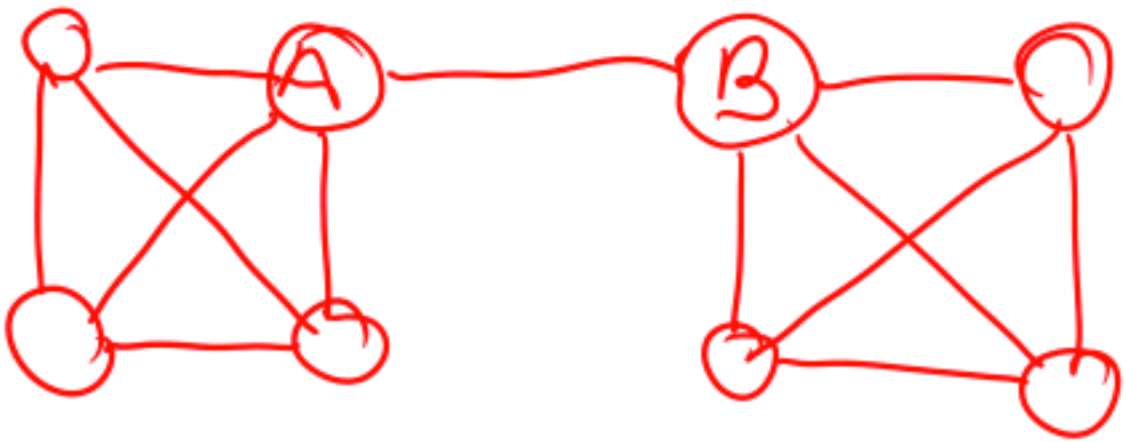
Reasons for triadic closure:

- If B and C have a friend A in common, then:
  - B is more likely to meet C
    - (since they both spend time with A)
  - B and C trust each other
    - (since they have a friend in common)
  - A has incentive to bring B and C together
    - (as it is hard for A to maintain two disjoint relationships)

- Empirical study by Bearman and Moody:
  - Teenage girls with low clustering coefficient are more likely to contemplate suicide

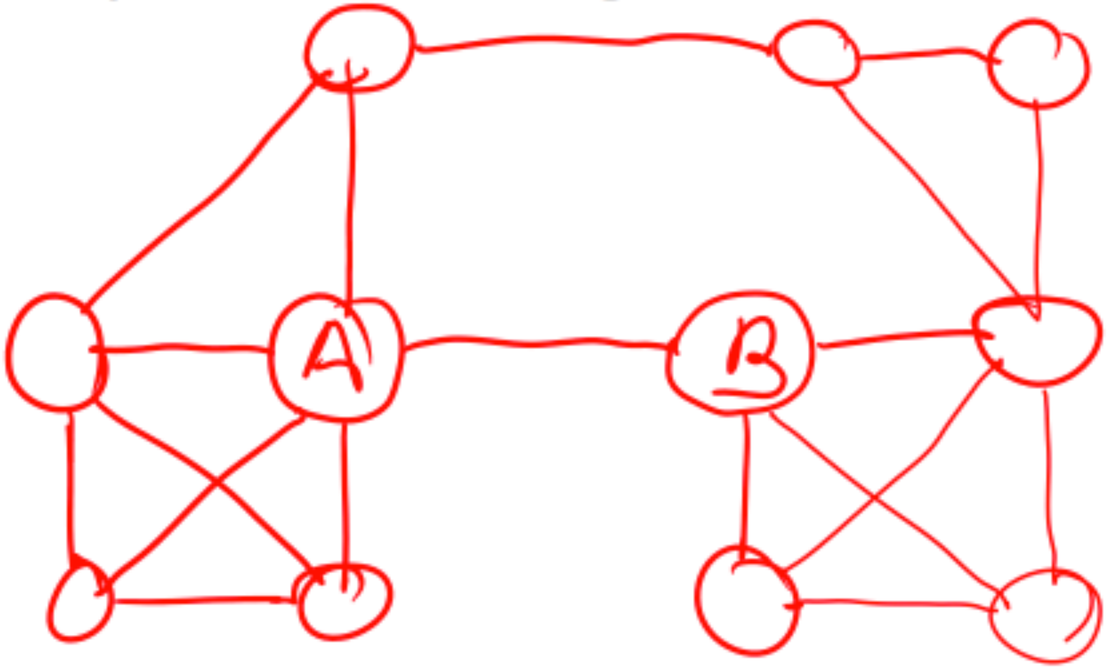
Bridges and Local Bridges

- Edge (A,B) is a bridge if deleting it would make A and B in be in two separate connected components.



Bridges and Local Bridges

- Edge (A,B) is a local bridge A and B have no friends in common
- Span of a local bridge is the distance of the edge endpoints if the edge is deleted



(local bridges with long span are like real bridges)

Strong Triadic Closure

- Links in networks have strength:
  - Friendship
  - Communication
- We characterize links as either Strong (friends) or Weak (acquaintances)
- Def: **Strong Triadic Closure Property:**
  If A has strong links to B and C, then there must be a link (B,C) (that can be strong or weak)

Local Bridges and Weak ties

- Claim: If node A satisfies Strong Triadic Closure and is involved in at least two strong ties, then any local bridge adjacent to A must be a weak tie.
- Proof by contradiction:
  - A satisfies Strong Triadic Closure
  - Let A-B be local bridge and a strong tie
  - Then B-C must exist because of Strong Triadic Closure
  - But then (A,B) is not a bridge

Summary of what we just did

- Defined Local Bridges:
  - Edges not in triangles
- Set two types of edges:
  - Strong and Weak Ties
- Defined Strong Triadic Closure:
  - Two strong ties imply a third edge
- → Local bridges are weak ties

Tie strength in real data

- For many years the Granovetter's theory was not tested
- But, today we have large who-talks-to-whom graphs:
  - Email, Messenger, Cell phones, Facebook
- Onnela et al. 2007:
  - Cell-phone network of 20% of country's population

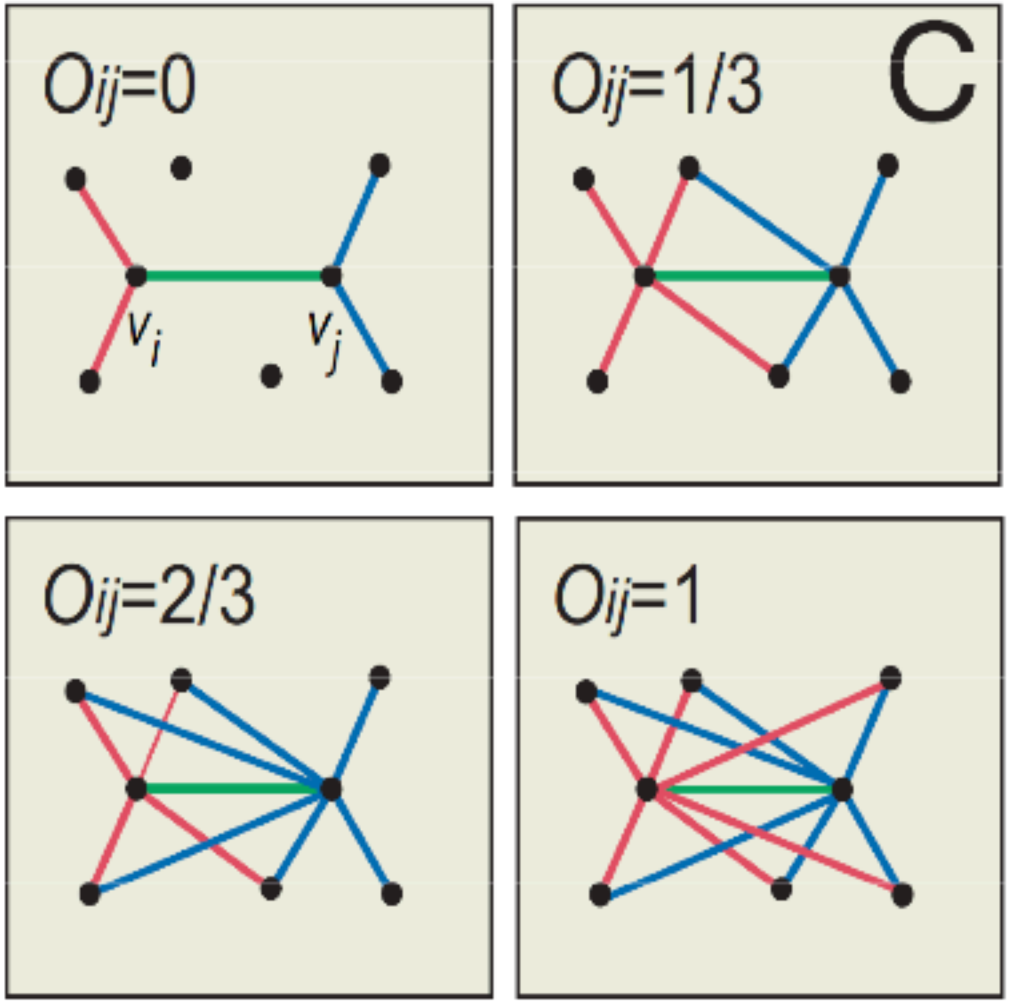
Neighborhood Overlap

- Overlap:

  $$O_{ij} = \frac{n(i) \cap n(j)}{n(i) \cup n(j)}$$

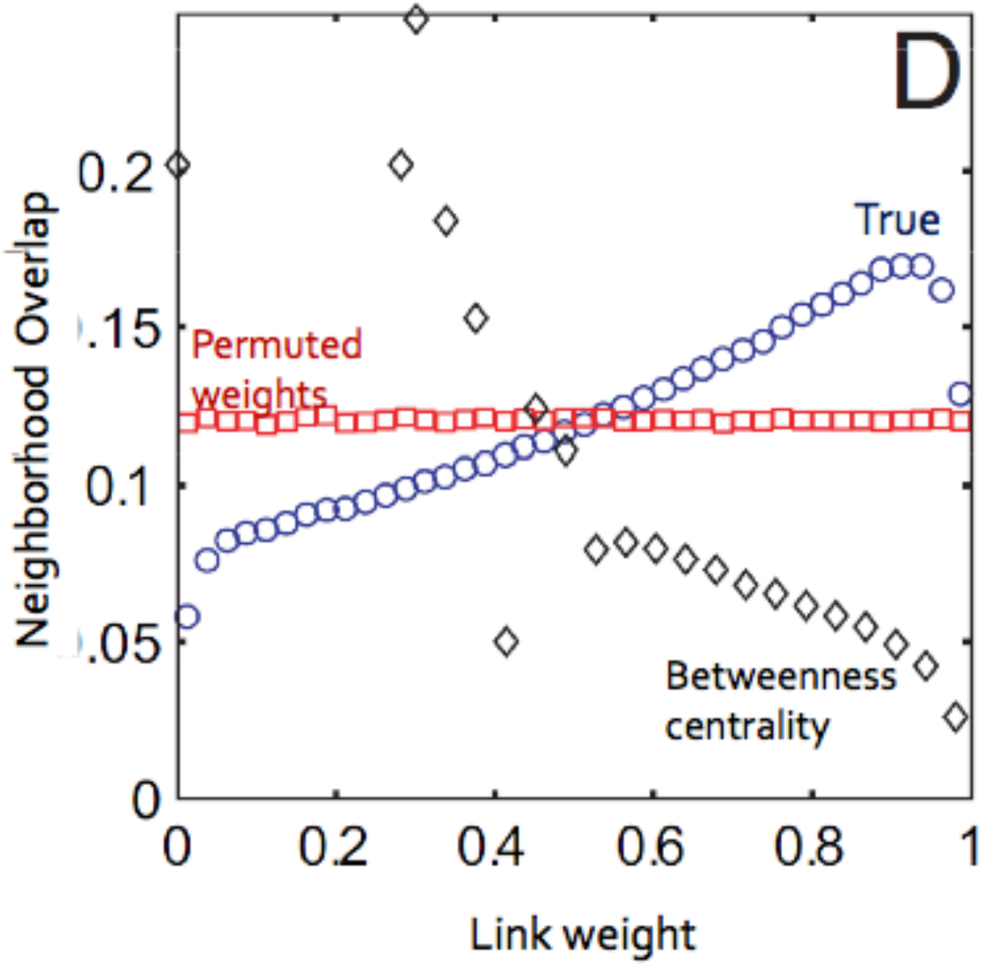
  - n(i) … set of neighbors of A

- Overlap $= 0$ when an edge is a local bridge



[Onnela et al. '07]

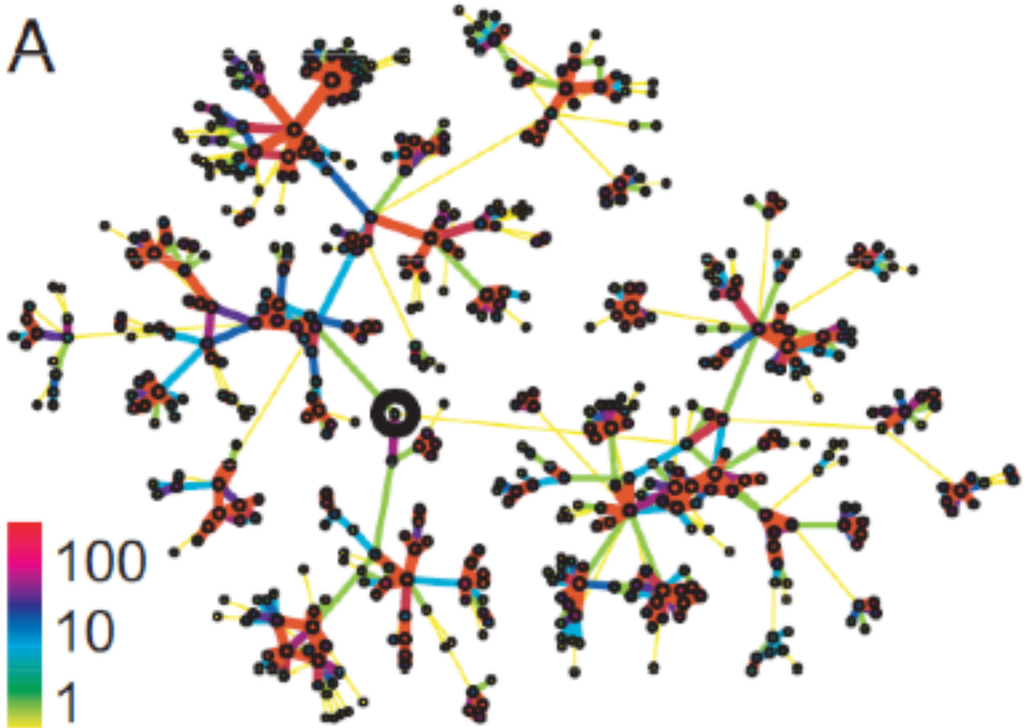
Mobile phones: Overlap vs. Weight



- Permuted weights: Keep the structure but randomly reassign edge weights
- Betweenness centrality: Number of shortest paths going through an edge

Real network tie strengths

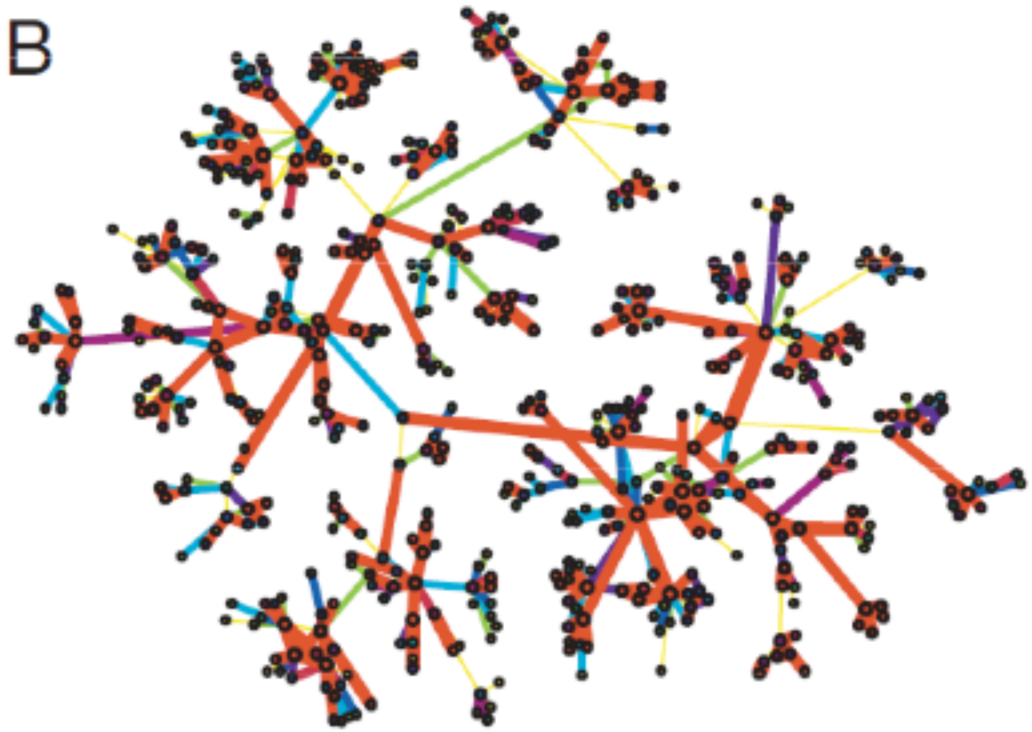
[Onnela et al. '07]



- Real edge strengths in mobile call graph

Permuted tie strenghts

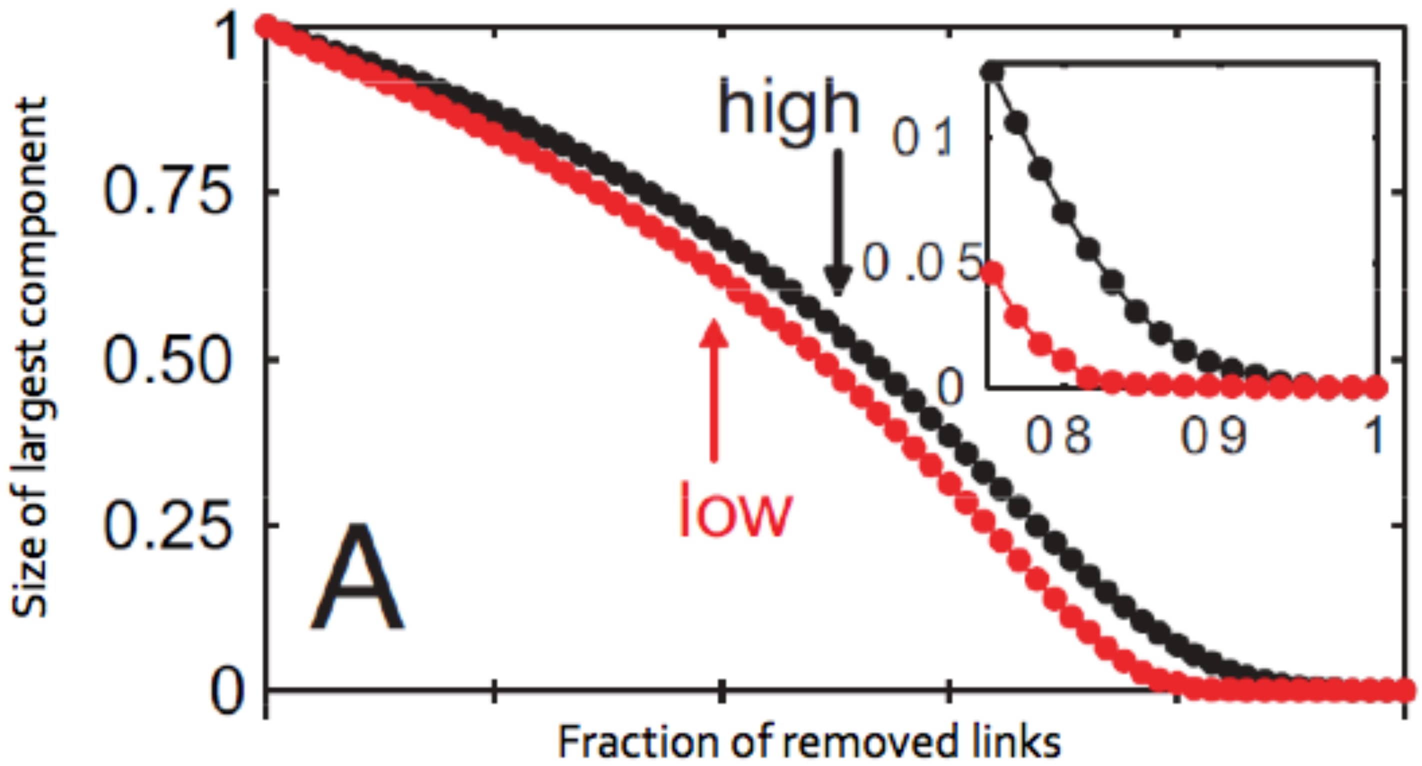
[Onnela et al. '07]



- Same network, same set of edge strengths
- But now strengths are randomly shuffled over the edges

[Onnela et al. '07]

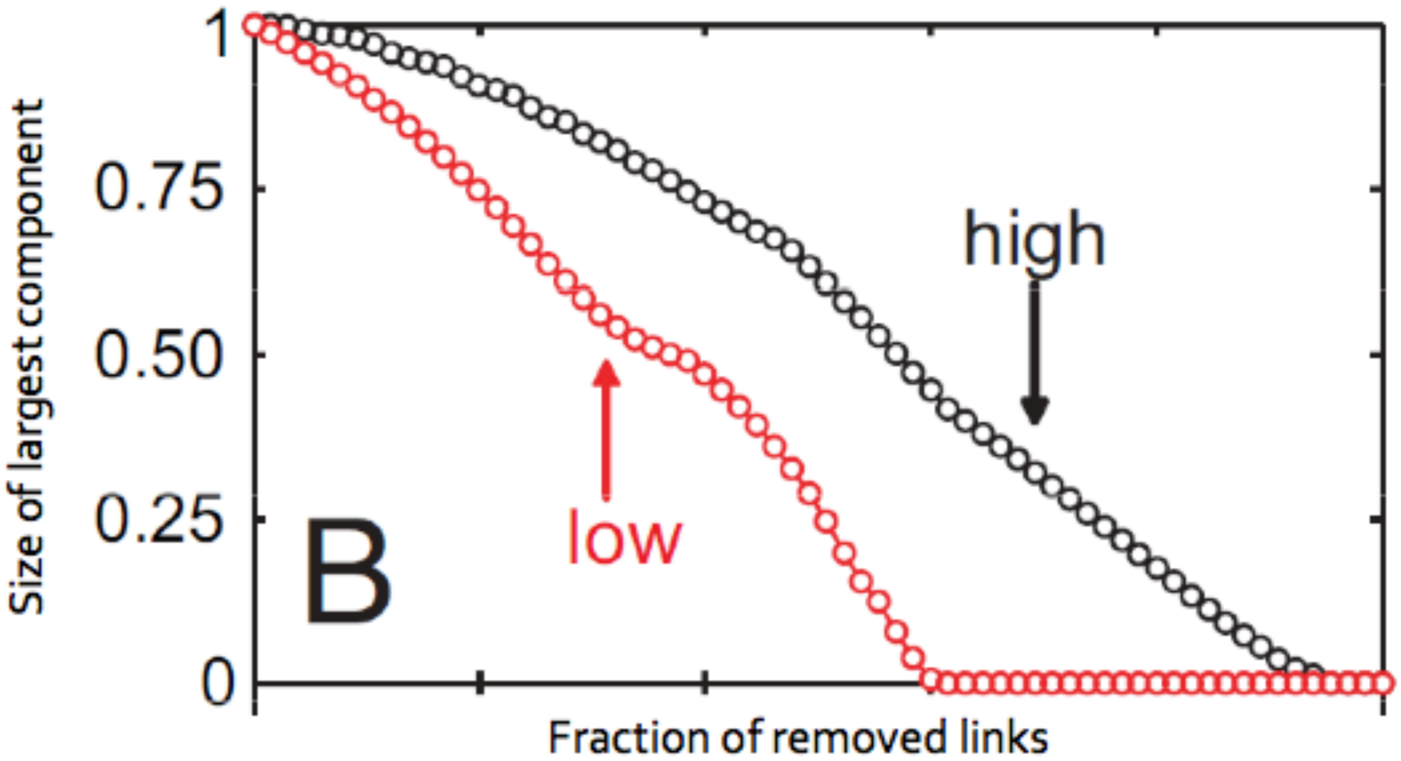
Link removal: Weight



- Removing links based on **strength** (# conversations)
  - Low to high
  - High to low

[Onnela et al. '07]

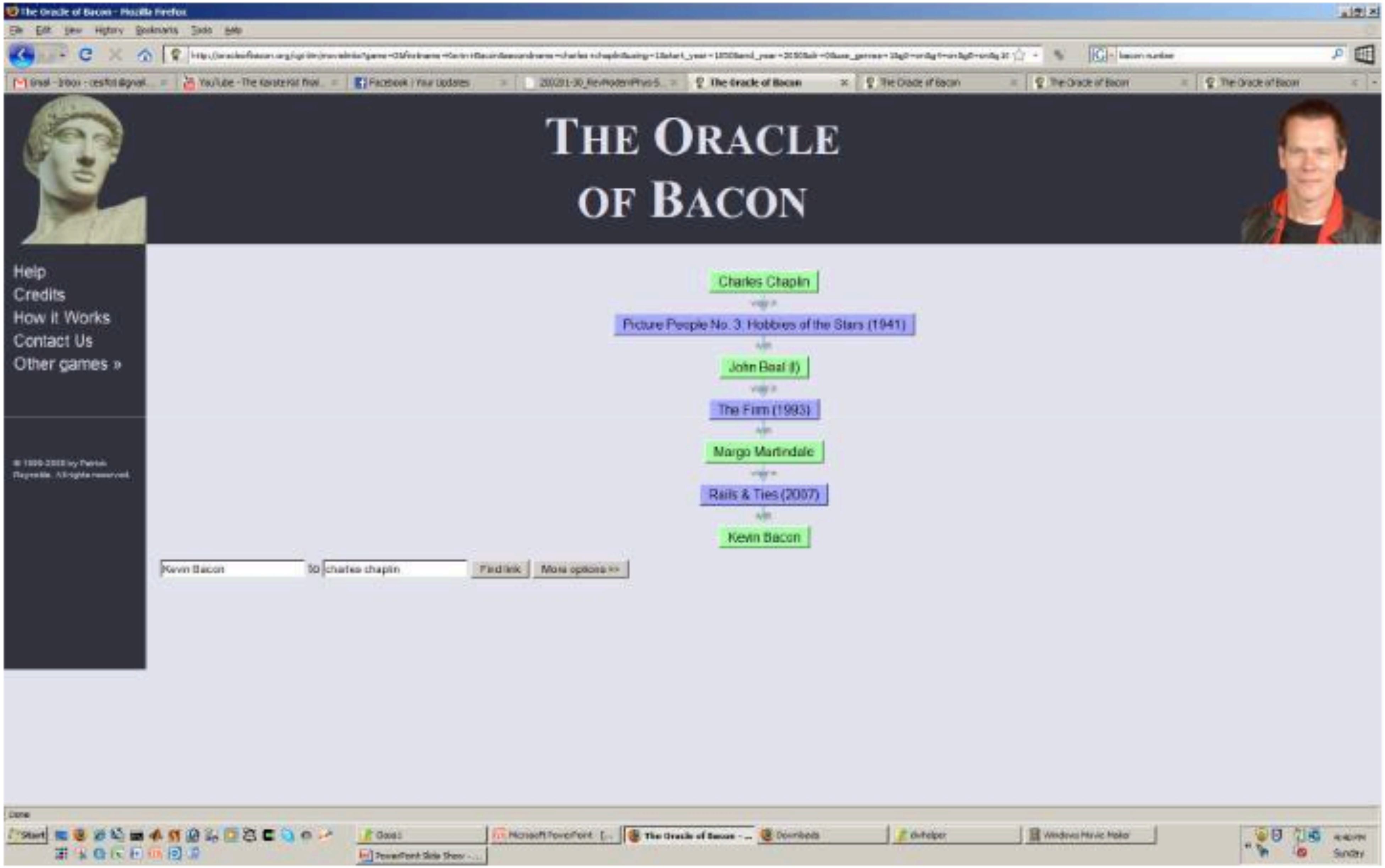
Link removal: Overlap



- Removing links based on **overlap**
  - Low to high
  - High to low

Centrality Measures

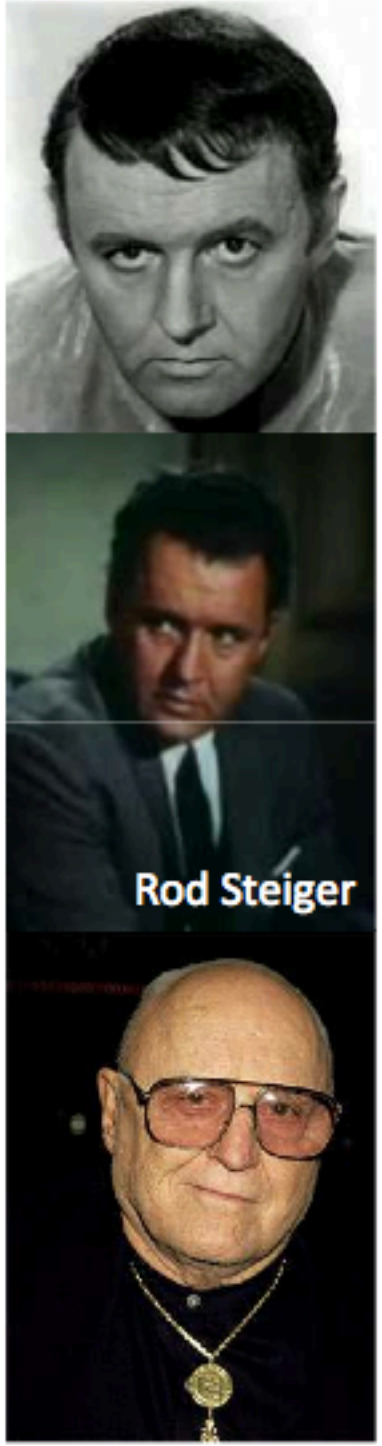
Measures of the "importance" of a node in a network


The Oracle of Bacon
Help
Credits
How it Works
Contact Us
Other games »
Charles Chaplin
Picture People No. 3: Hobbies of the Stars (1941)
John Beal (I)
The Firm (1993)
Margo Martindale
Rails & Ties (2007)
Kevin Bacon
Kevin Bacon
to
charles chaplin
Find link
More options >>

Hollywood Revolves Around

Click on a name to see that person's table.

Steiger, Rod (2.678695)
Lee, Christopher (I) (2.684104)
Hopper, Dennis (2.698471)
Sutherland, Donald (I) (2.701850)
Keitel, Harvey (2.705573)
Pleasence, Donald (2.707490)
von Sydow, Max (2.708420)
Caine, Michael (I) (2.720621)
Sheen, Martin (2.721361)
Quinn, Anthony (2.722720)
Heston, Charlton (2.722904)
Hackman, Gene (2.725215)
Connery, Sean (2.730801)
Stanton, Harry Dean (2.737575)
Welles, Orson (2.744593)
Mitchum, Robert (2.745206)
Gould, Elliott (2.746082)
Plummer, Christopher (I) (2.746427)
Coburn, James (2.746822)
Borgnine, Ernest (2.747229)



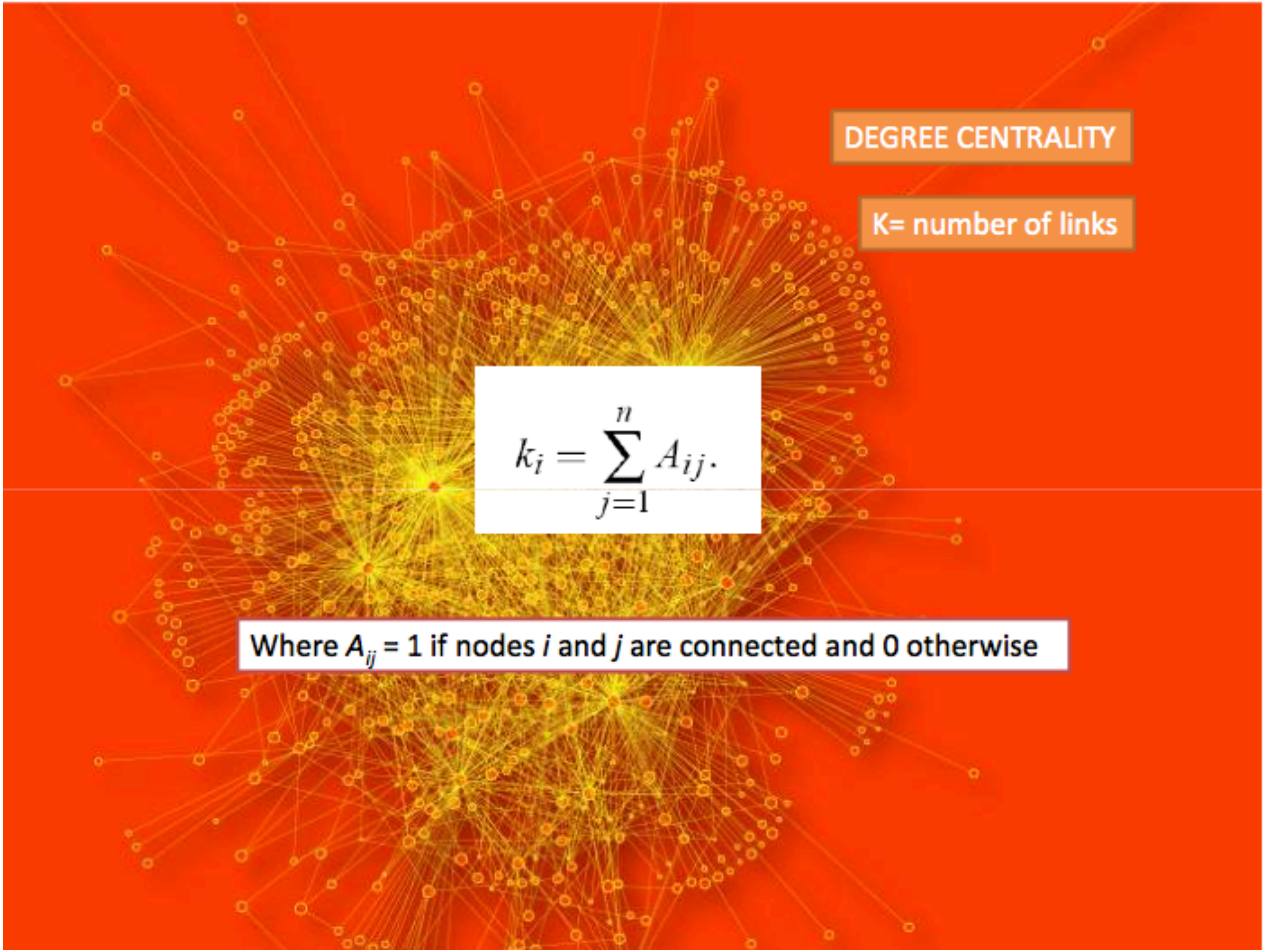
Most Connected Actors in Hollywood

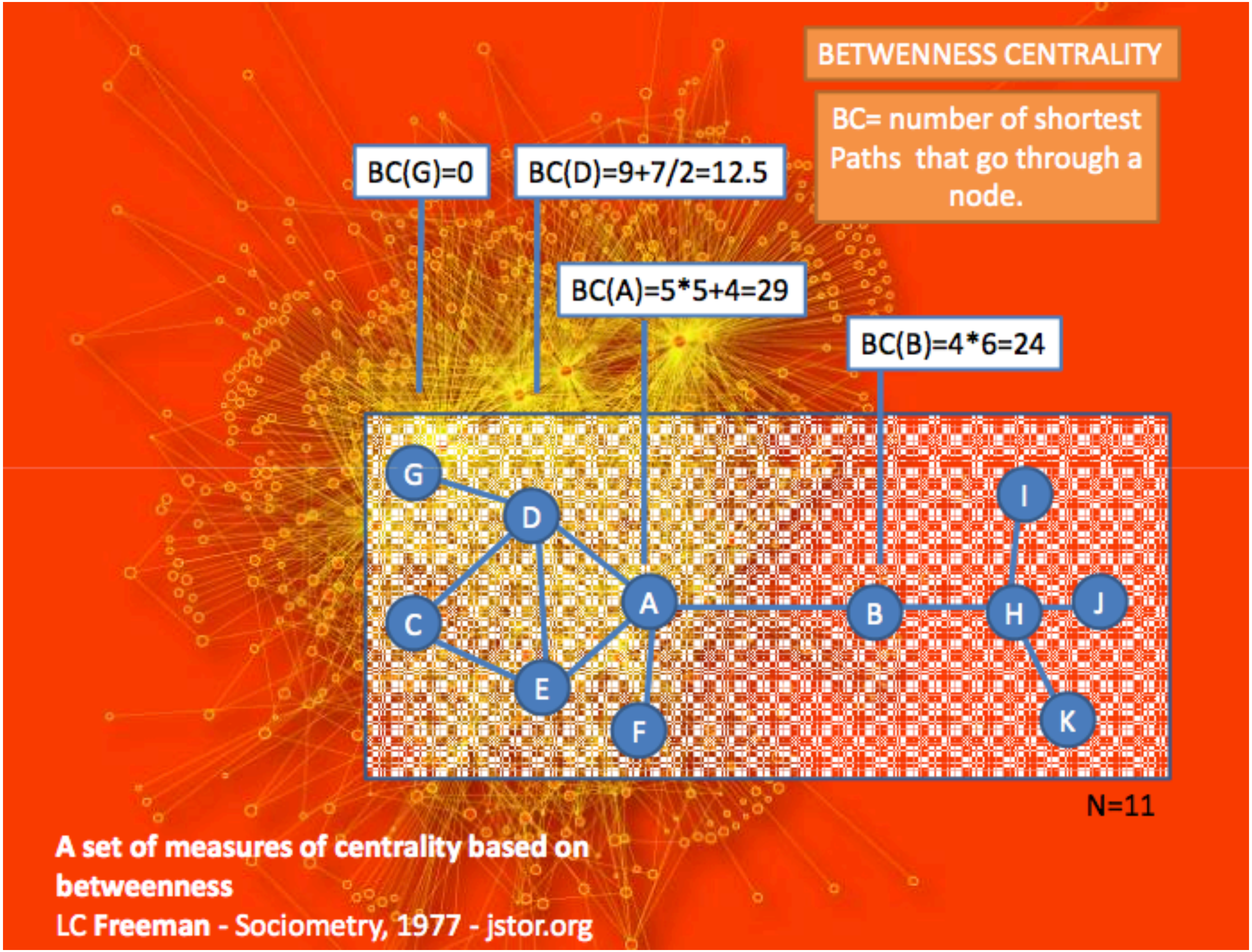
(measured in the late 90's)

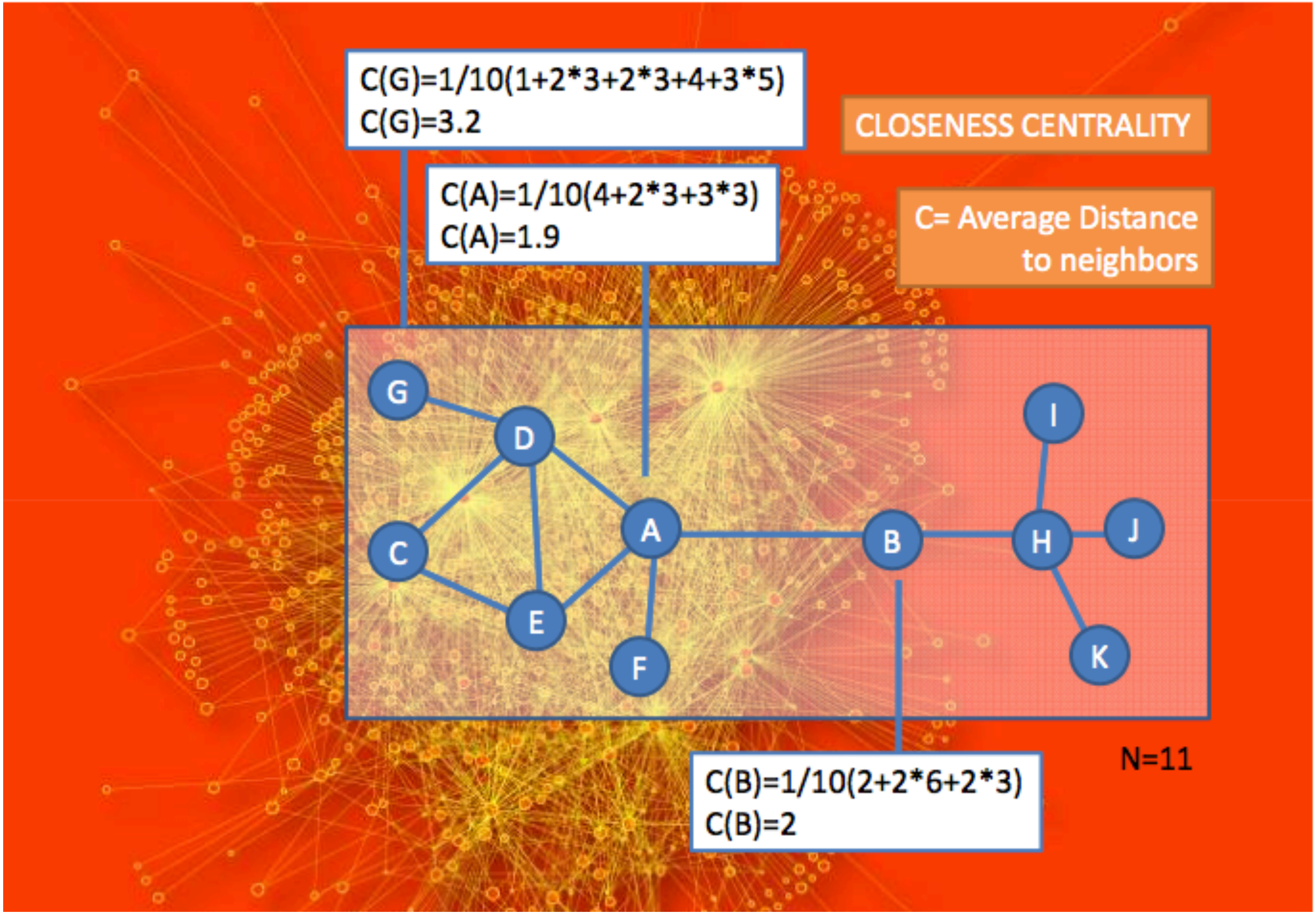
Mel Blanc 759
Tom Byron 679
Marc Wallice 535
Ron Jeremy 500
Peter North 491
TT Boy 449
Tom London 436
Randy West 425
Mike Horner 418
Joey Silvera 410

XXX

A-L Barabasi, "Linked", 2002

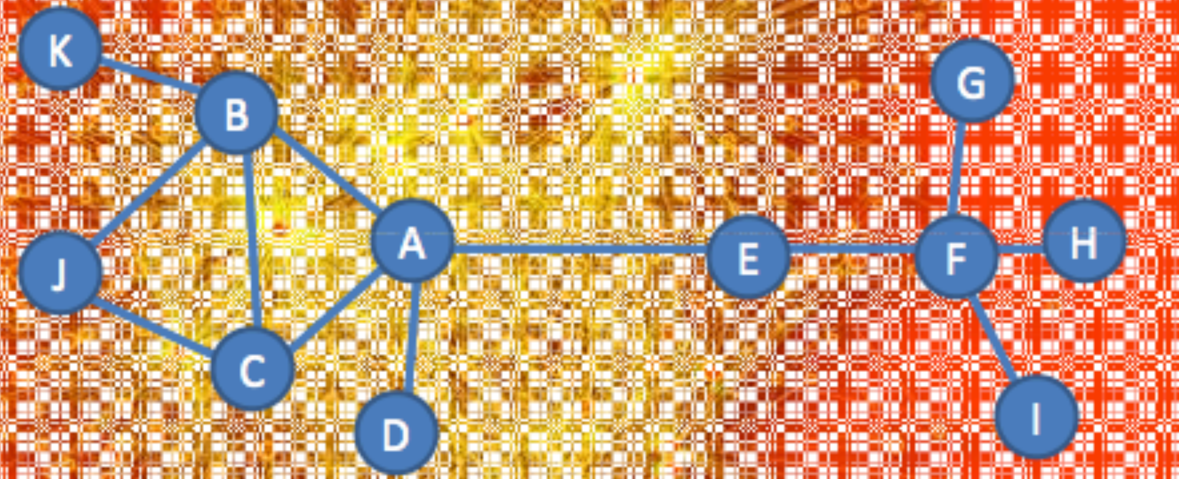

DEGREE CENTRALITY
K= number of links
$$k_i = \sum_{j=1}^{n} A_{ij}.$$
Where A_{ij} = 1 if nodes *i* and *j* are connected and 0 otherwise


BETWENNESS CENTRALITY
BC= number of shortest Paths that go through a node.
BC(G)=0
BC(D)=9+7/2=12.5
BC(A)=5*5+4=29
BC(B)=4*6=24
G
D
C
E
A
F
B
I
H
J
K
N=11
A set of measures of centrality based on betweenness
LC Freeman - Sociometry, 1977 - jstor.org


C(G)=1/10(1+2*3+2*3+4+3*5)
C(G)=3.2
CLOSENESS CENTRALITY
C(A)=1/10(4+2*3+3*3)
C(A)=1.9
C= Average Distance to neighbors
G
D
A
C
E
F
B
H
I
J
K
N=11
C(B)=1/10(2+2*6+2*3)
C(B)=2

PAGE RANK

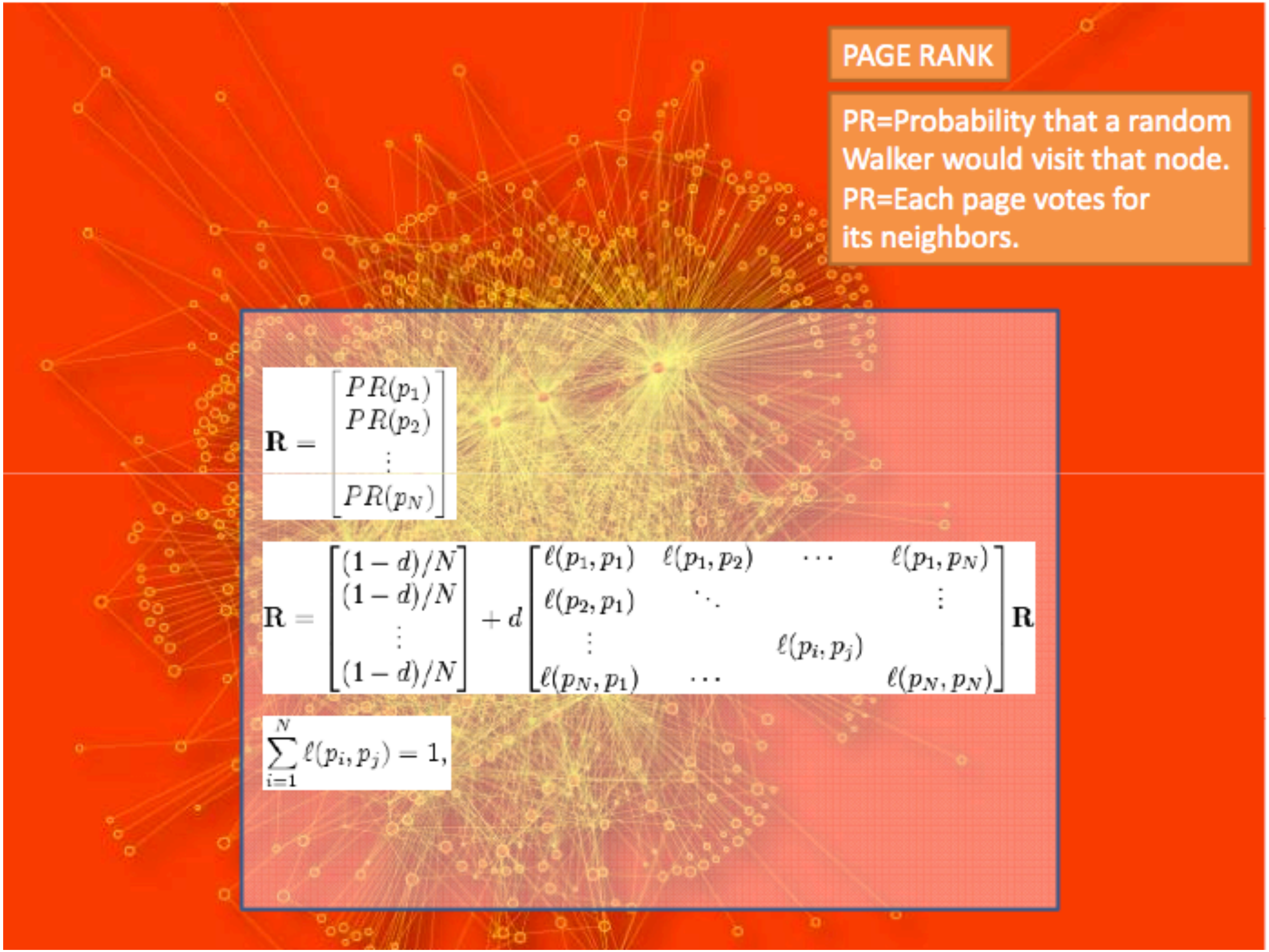
PR=Probability that a random walker with interspersed Jumps would visit that node.
PR=Each page votes for its neighbors.



$PR(A)=PR(B)/4 + PR(C)/3 + PR(D)+PR(E)/2$

A random surfer eventually stops clicking

$PR(X)=(1-d)/N + d(\sum PR(y)/k(y))$



PAGE RANK

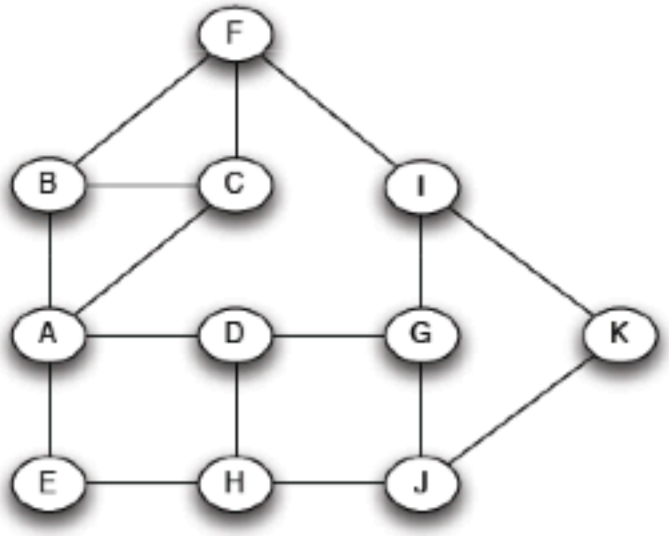
PR=Probability that a random Walker would visit that node.
PR=Each page votes for its neighbors.

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

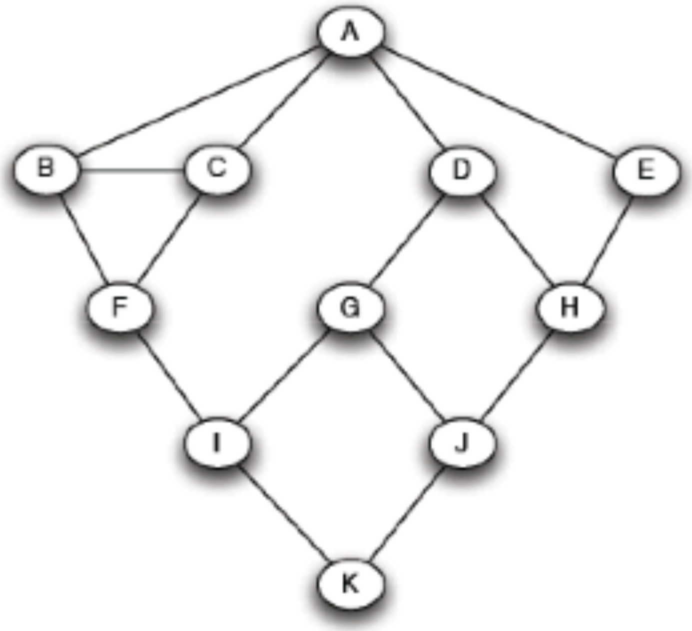
$$\sum_{i=1}^{N} \ell(p_i,p_j) = 1,$$

How to compute betweenness?



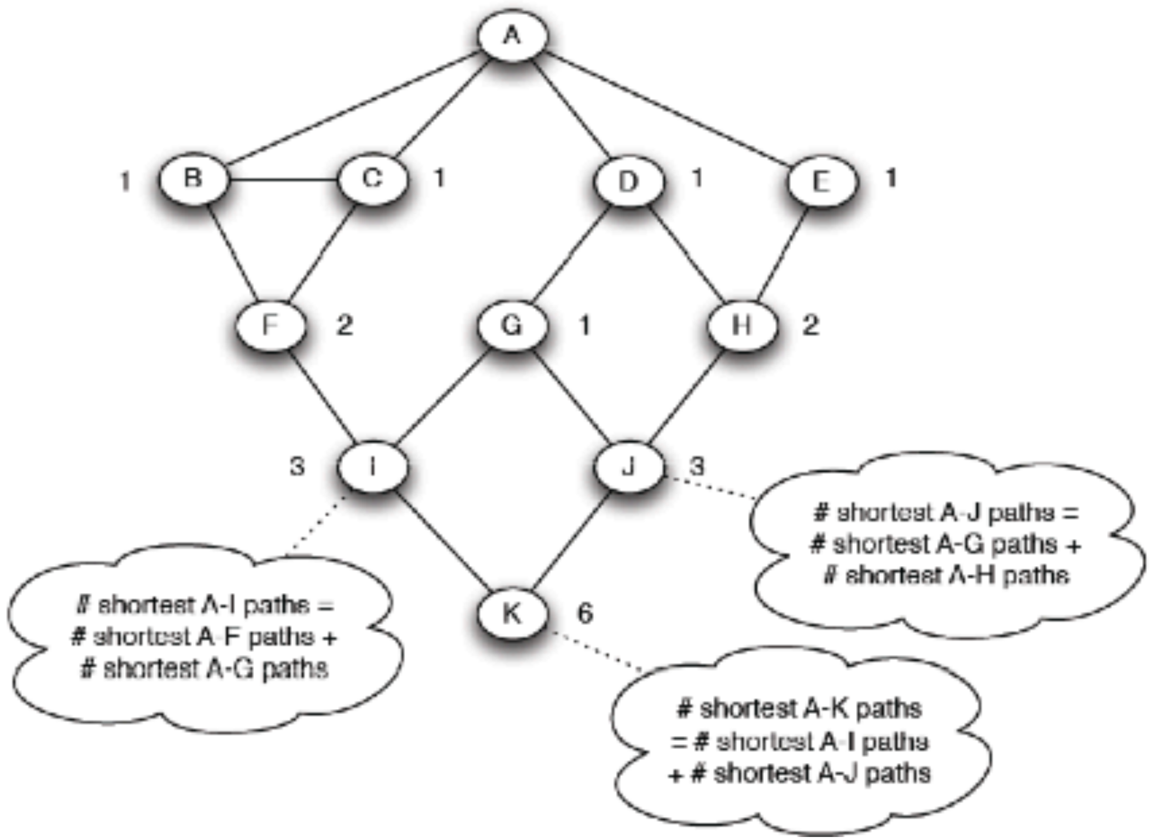
- Want to compute betweenness of paths starting at node A

- Breath first search starting from A:



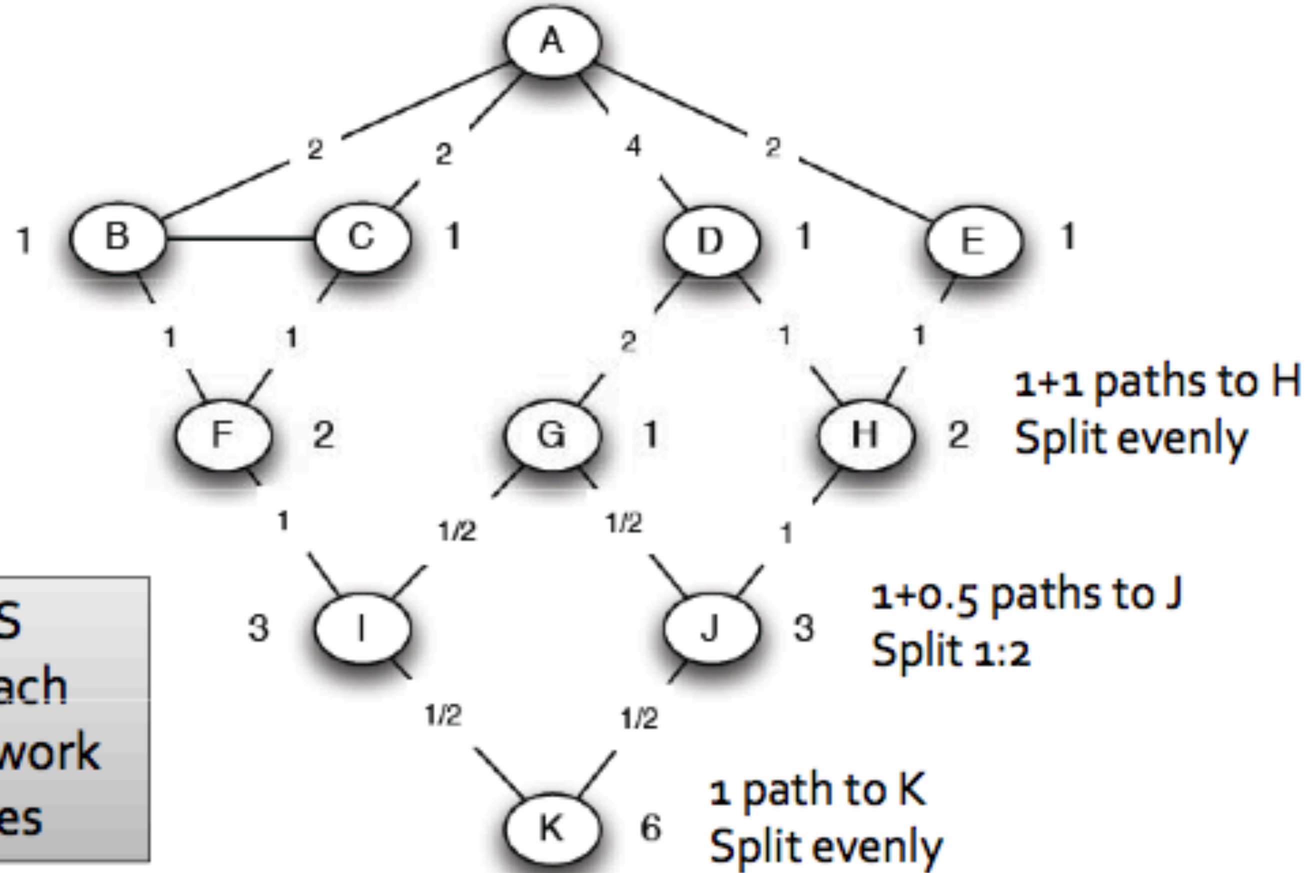
How to compute betweenness (2)

- Count the number of shortest paths from A to all other nodes of the network:



How to compute betweenness (3)

- Compute betweenness by working up the tree: If there are multiple paths count them fractionally



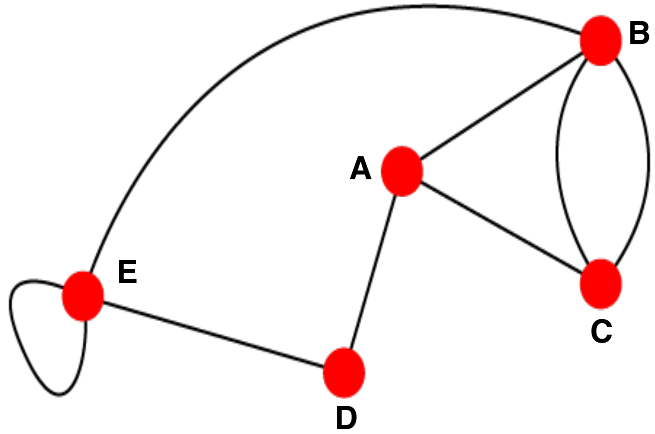
- Repeat the BFS procedure for each node of the network
- Add edge scores

PATHS

A *path is* a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

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



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

•A legitimate path on the graph on the right:
ABCBCADEEBA

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