Complex (Social) Networks

Dino Pedreschi

KDD LAB, University of Pisa and ISTI-CNR

http://kdd.isti.cnr.it

Master MAINS

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]

–adjective

1.

composed of many interconnected parts; compound; composite: a complex highway system.

2.

characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3.

so complicated or intricate as to be hard to understand or deal with: a complex problem.

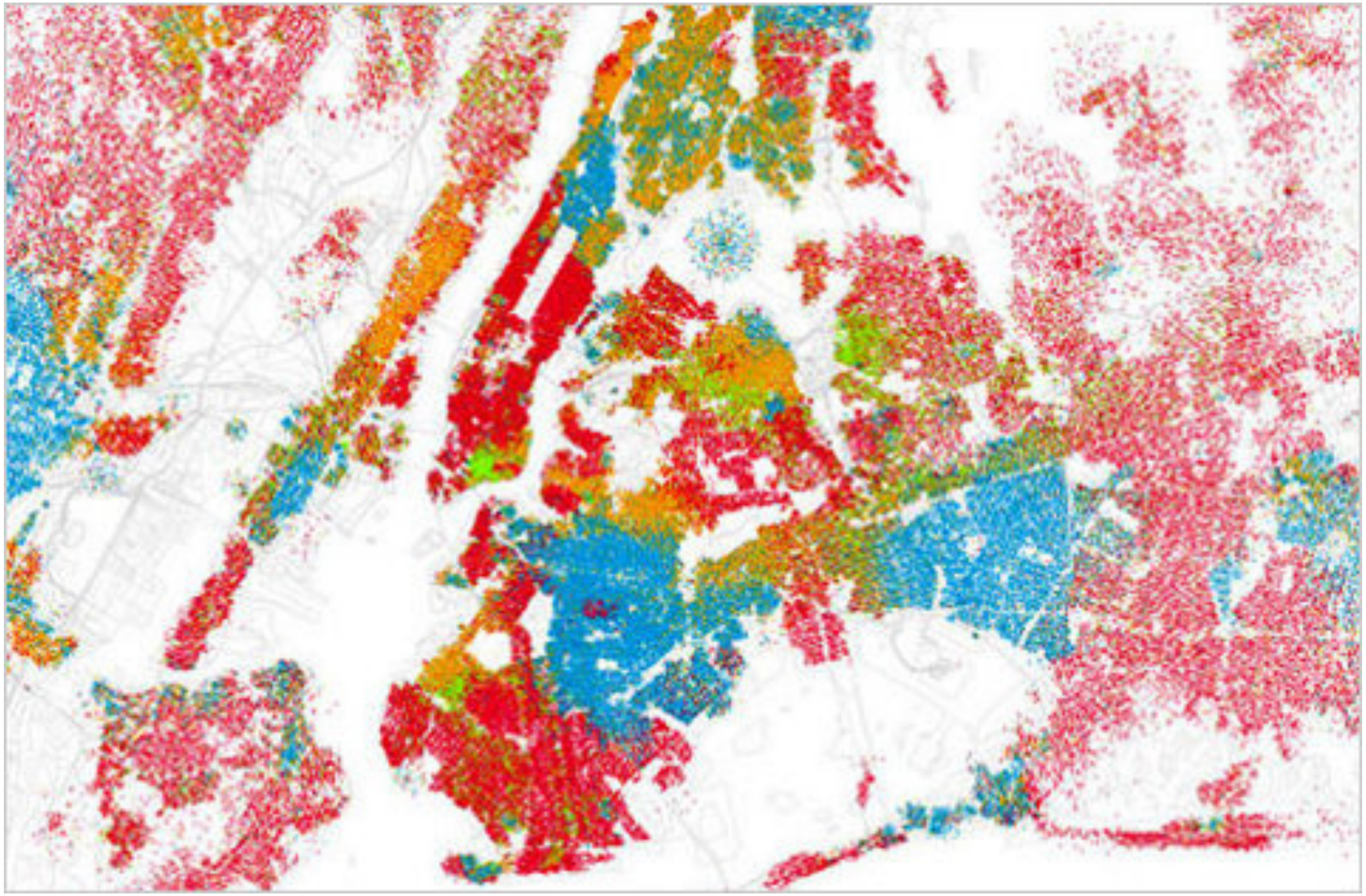
Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as **emergent behaviour**, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

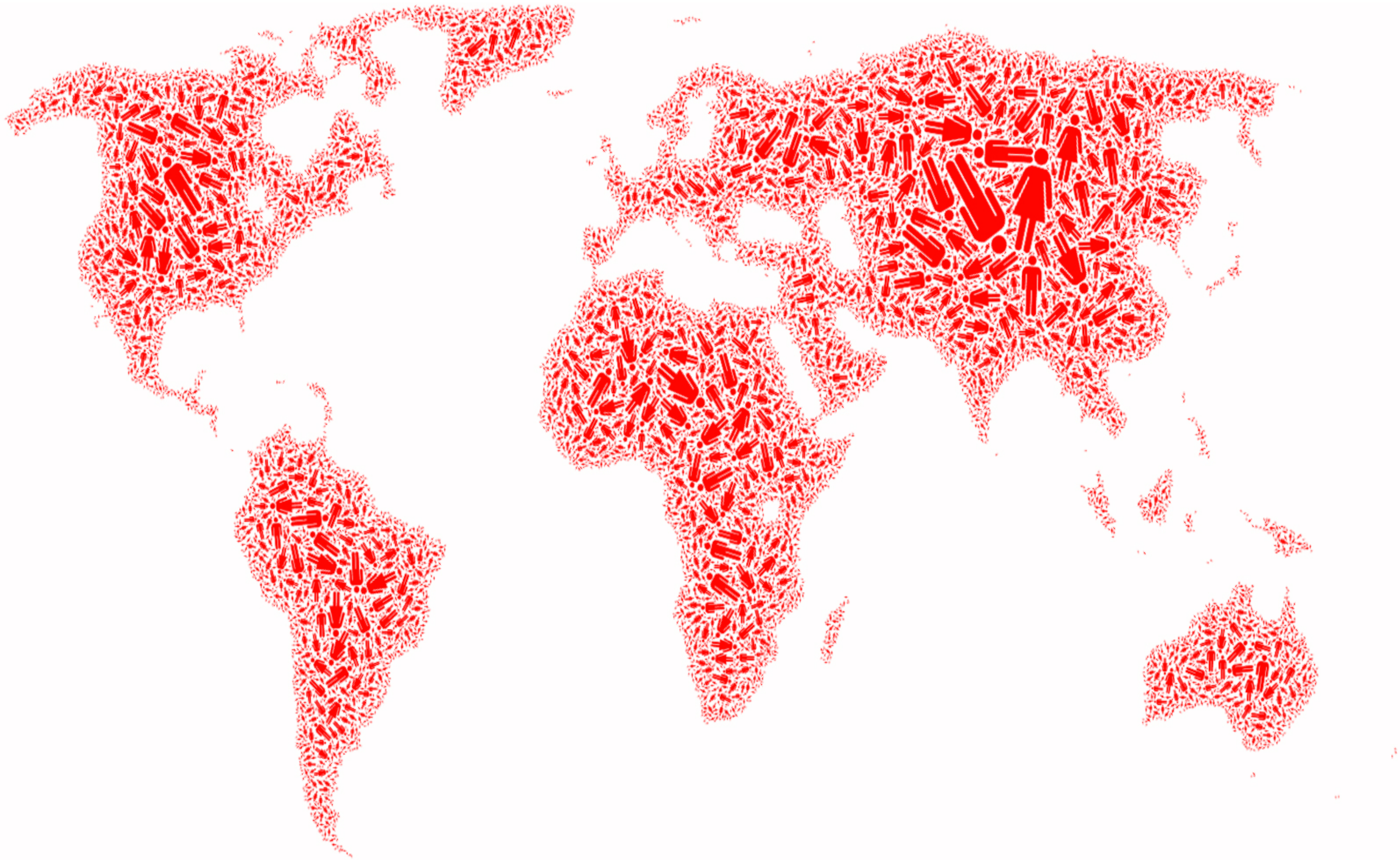
Complexity

Emergent behavior: segregation

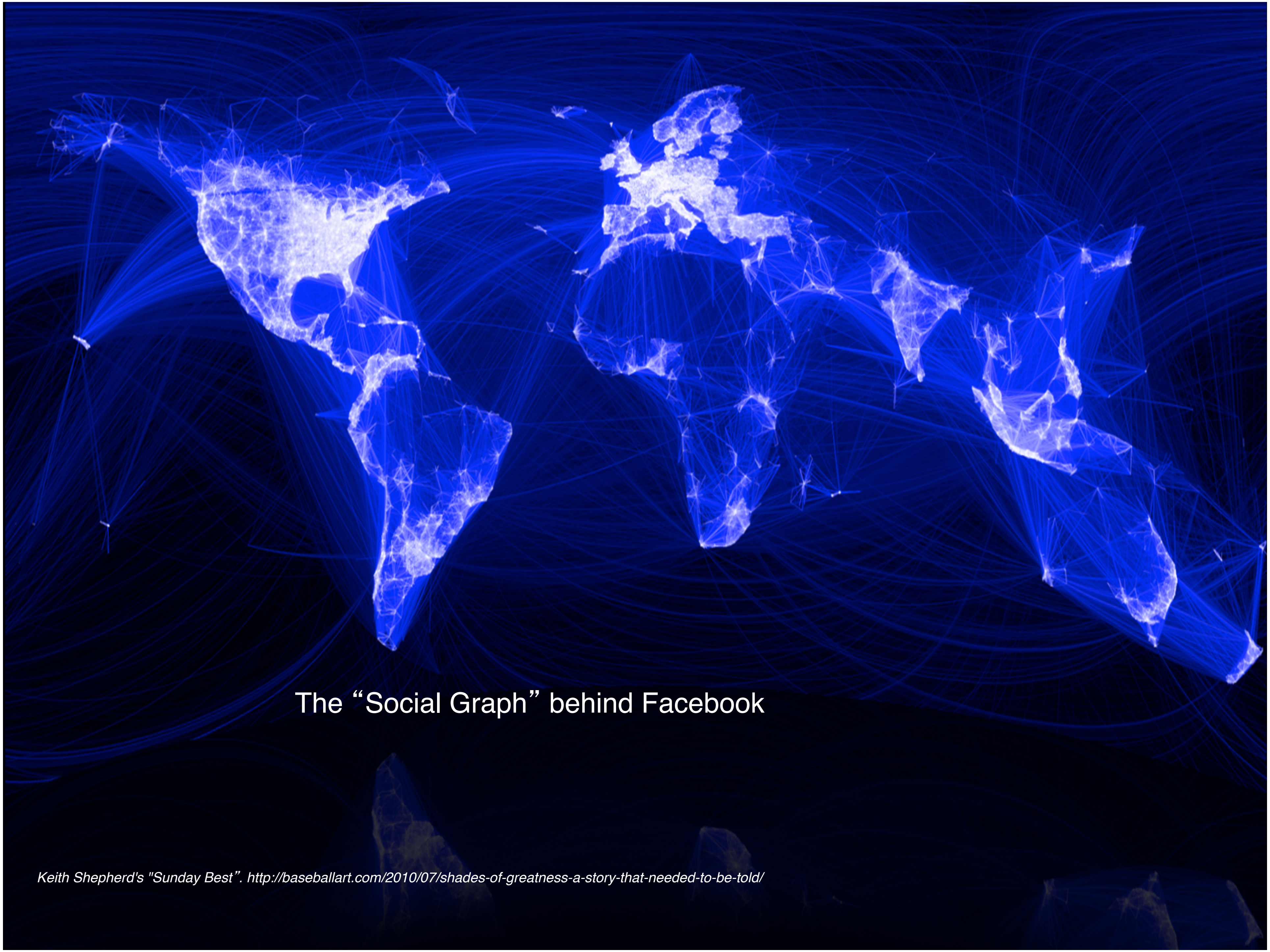


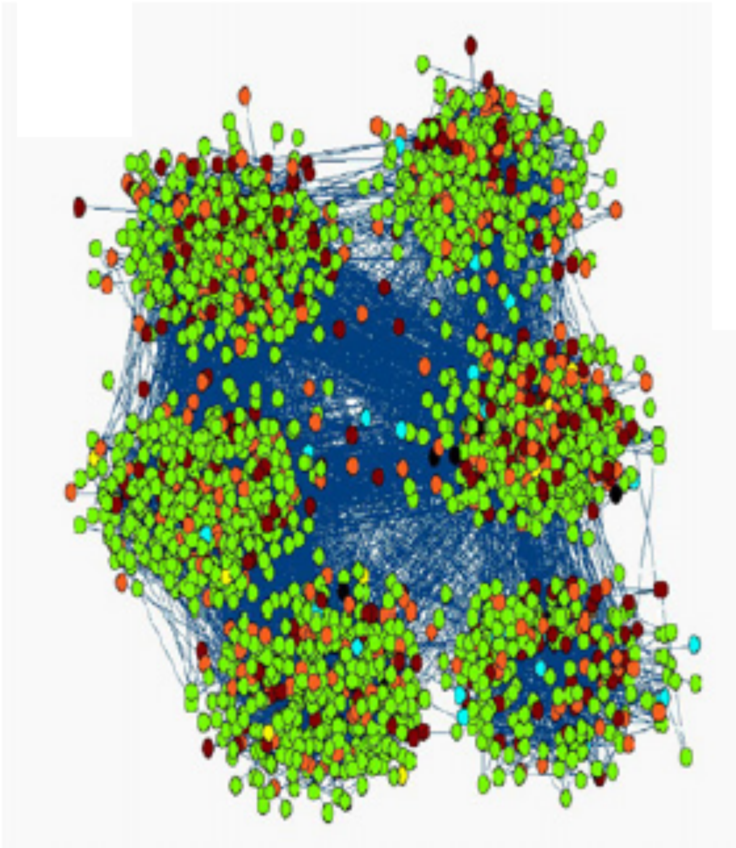
Behind each complex system there is a **network**, that defines the interactions between the components.

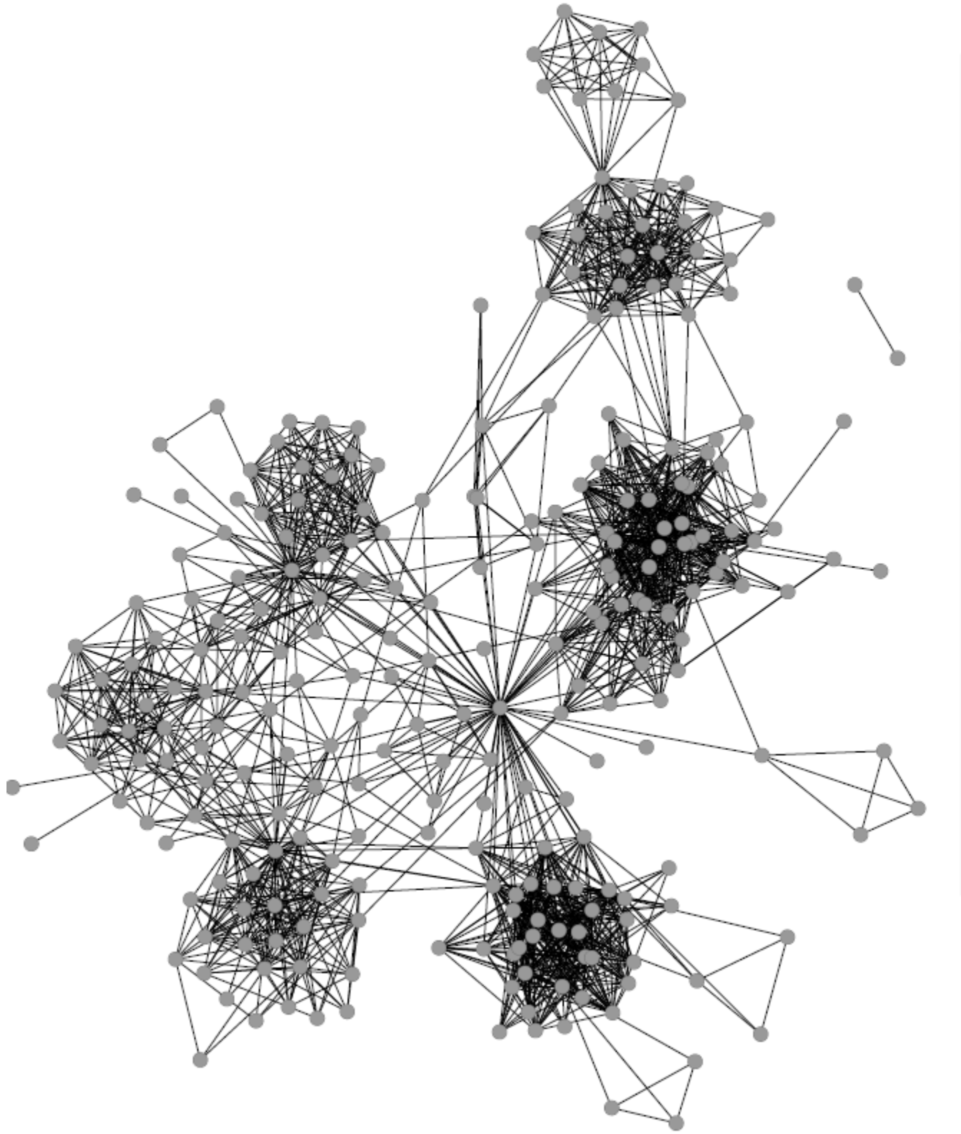
Social, informational, technological, biological networks



The "Day of 7 Billion“ has been in October 2011


The “Social Graph” behind Facebook
Keith Shepherd's "Sunday Best”. http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

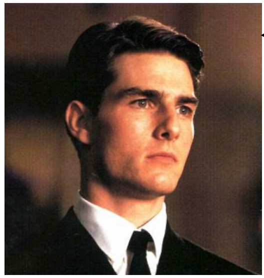




COLLABORATION NETWORKS: ACTOR NETWORK

Nodes: actors
Links: cast jointly

IMDb Internet Movie Database

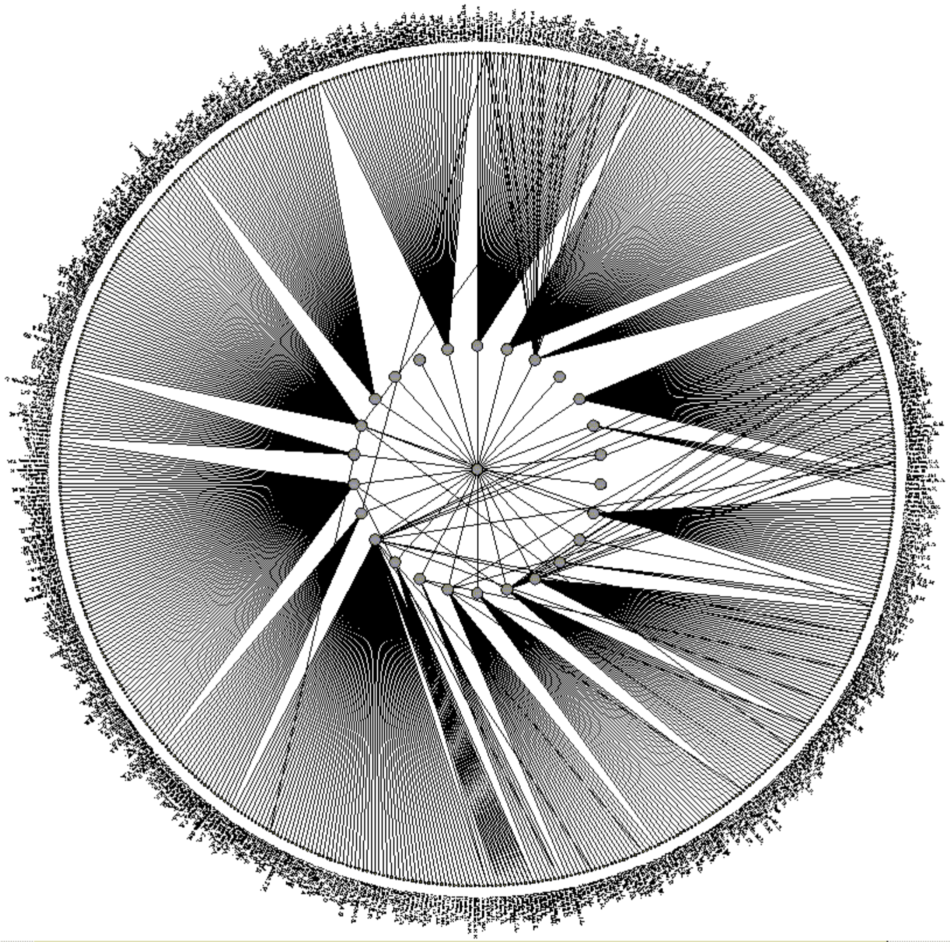


Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

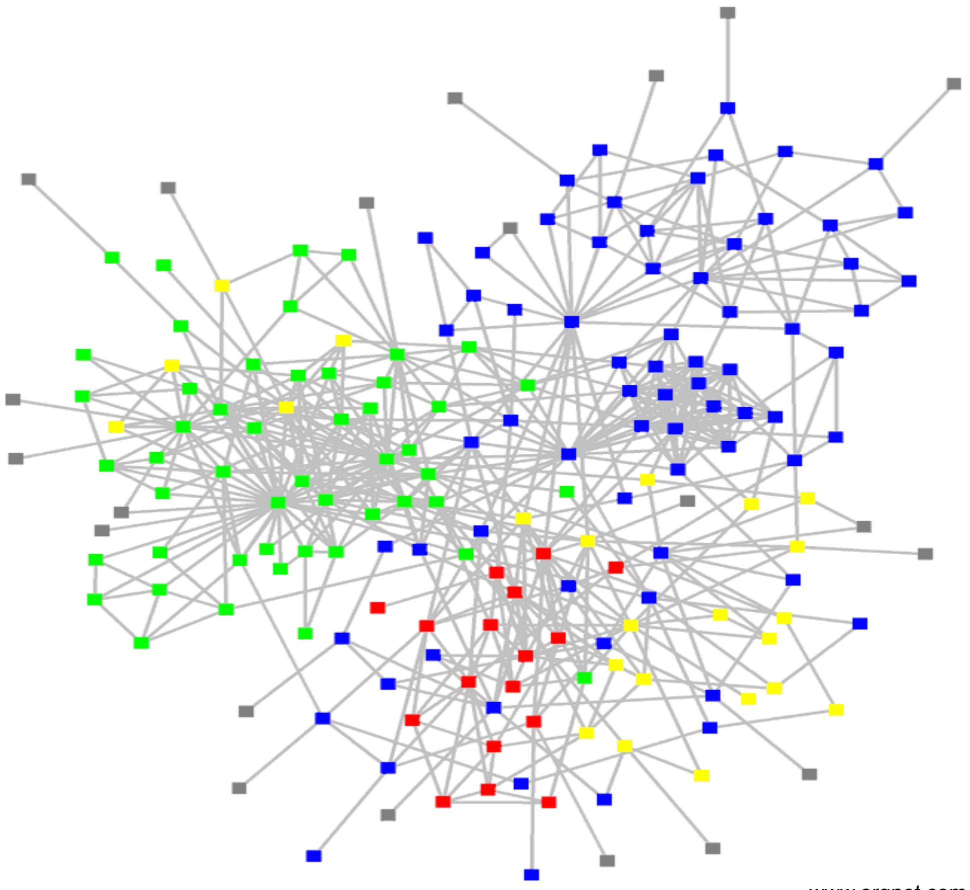
N = 212,250 actors $\langle k \rangle = 28.78$

COLLABORATION NETWORKS: SCIENCE CO-AUTHORSHIP

Nodes: scientist (authors)
Links: write paper together



STRUCTURE OF AN ORGANIZATION



: departments

: consultants

: external experts

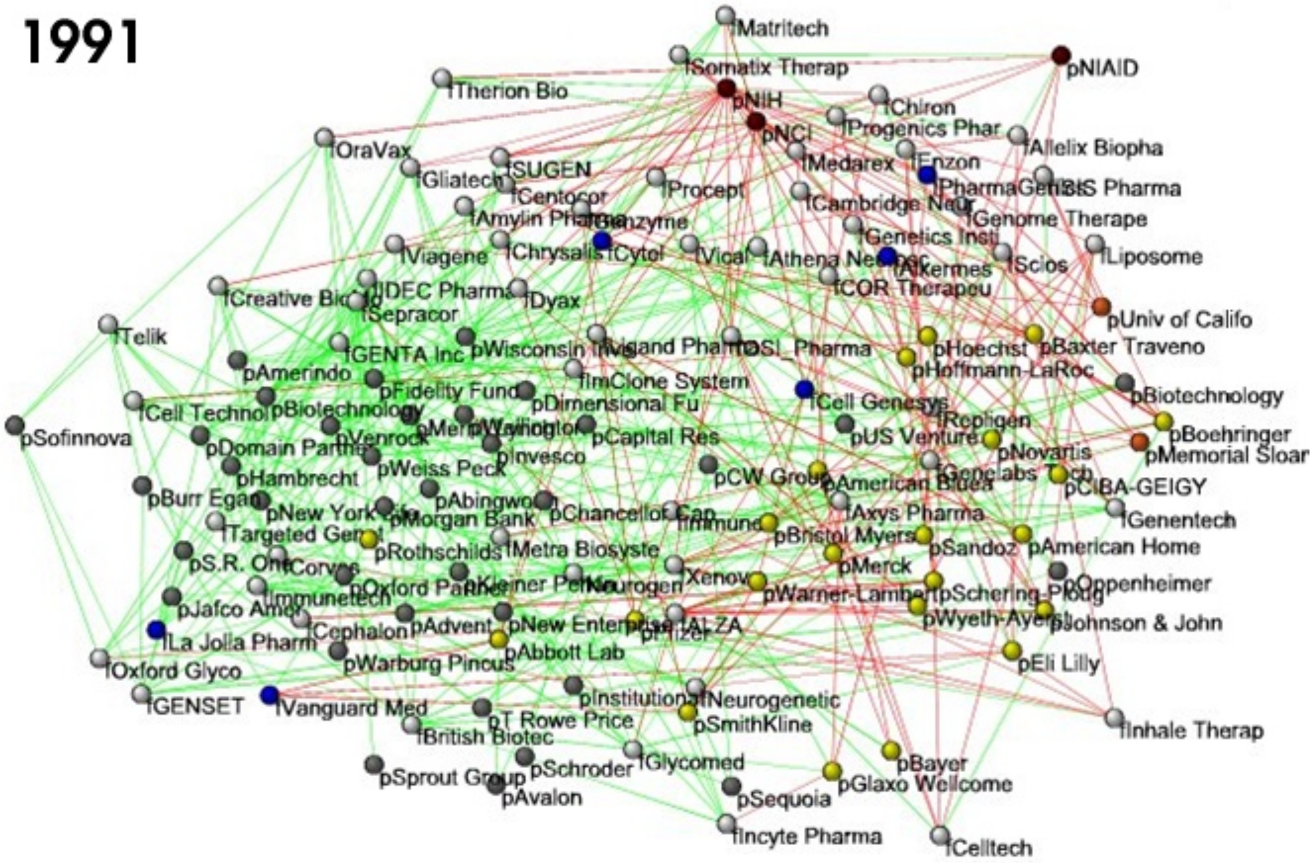
BUSINESS TIES IN US BIOTECH-INDUSTRY

Nodes:

- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

Links:

- Collaborations
- Financial
- R&D

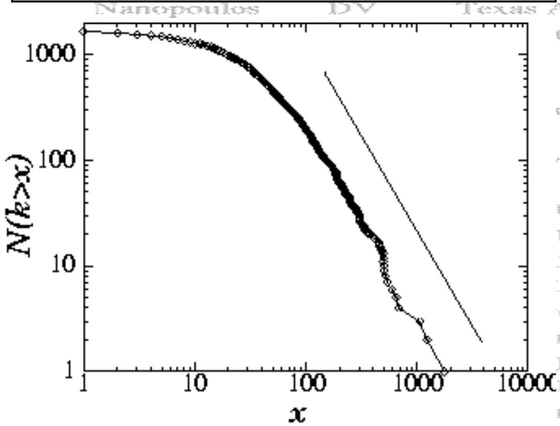


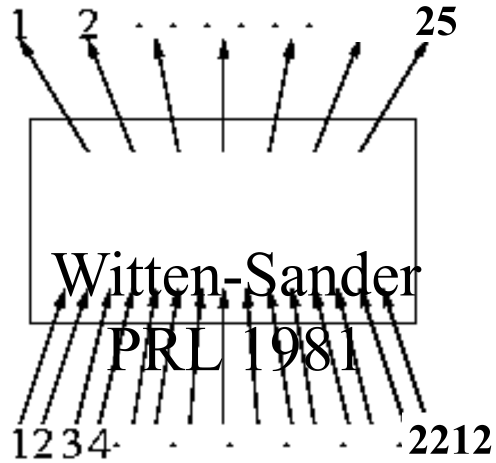
http://ecclectic.ss.uci.edu/~drwhite/Movie

Information networks: the Web and Science Citation Indexes

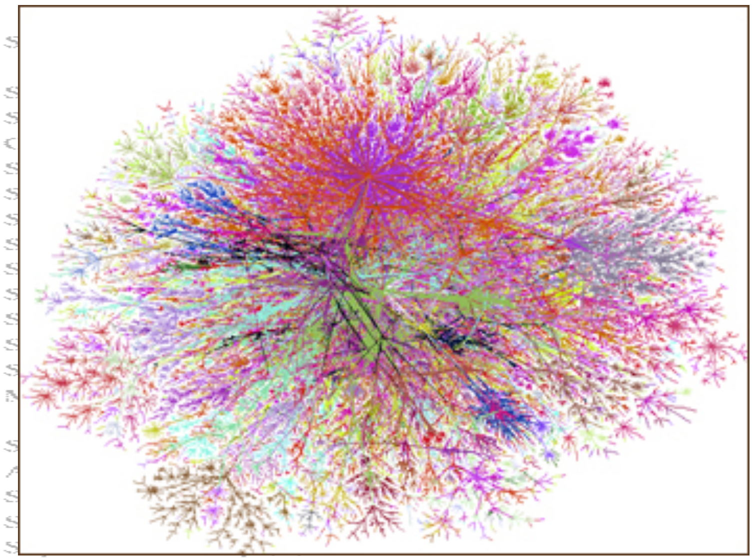
Nodes: papers
Links: citations

1736 PRL papers (1988)

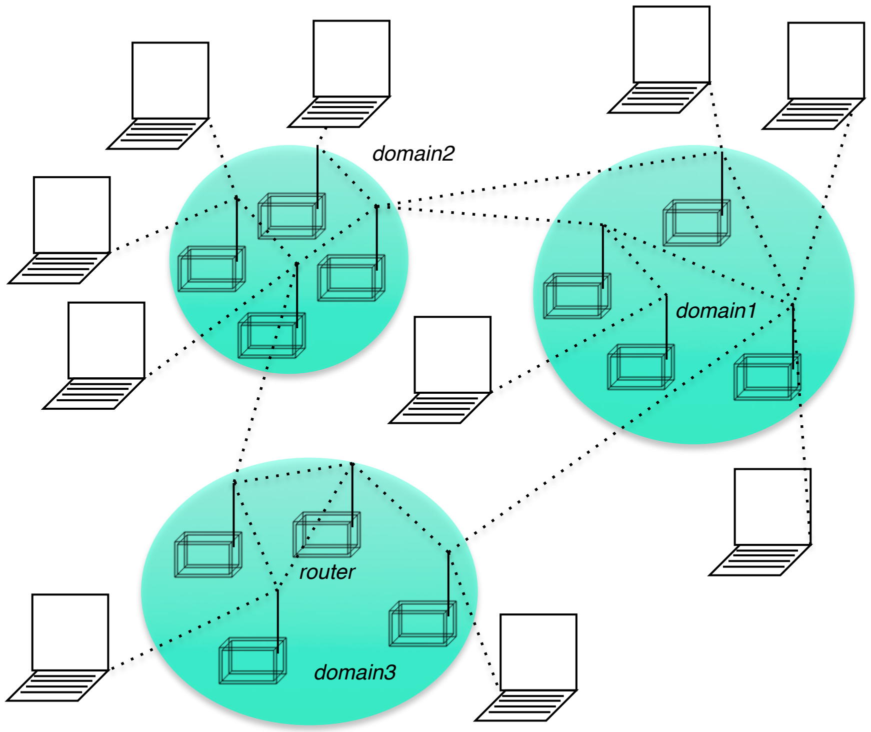


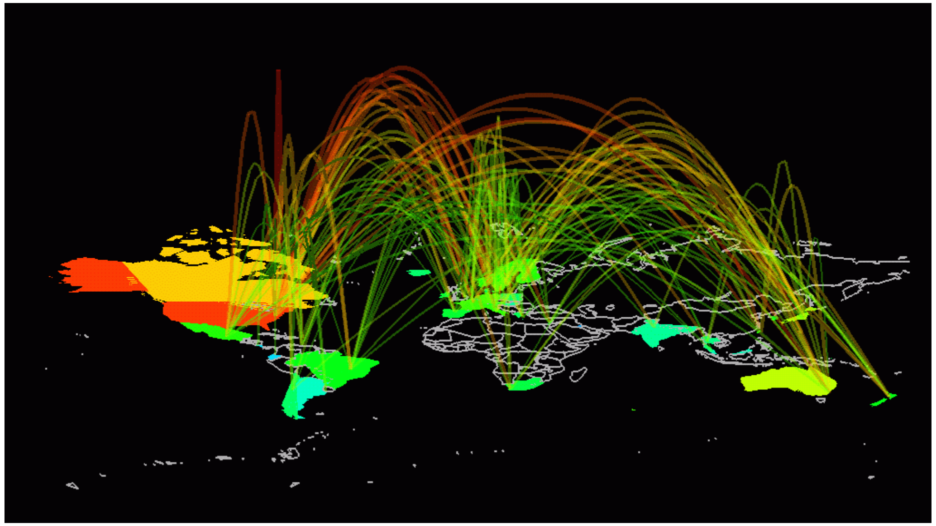


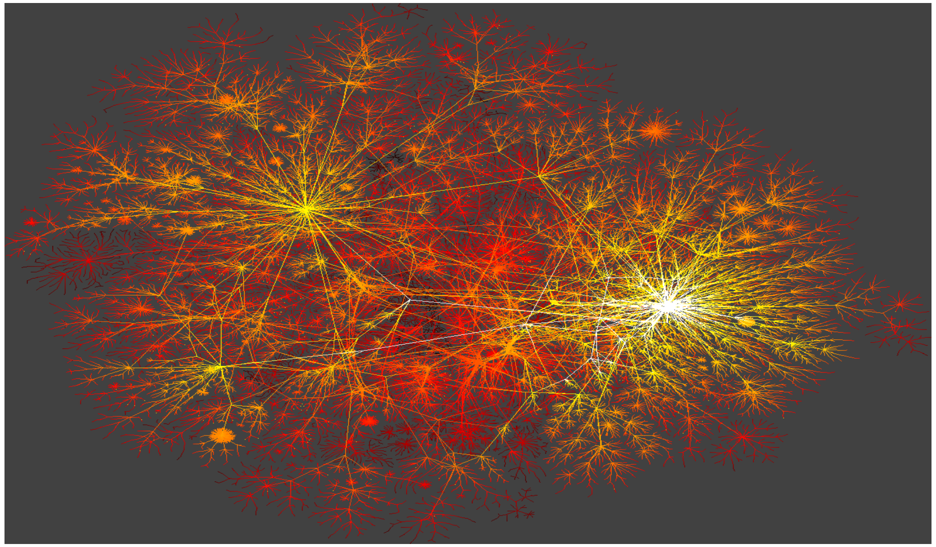
Nodes: web pages
Links: ditto ;-)

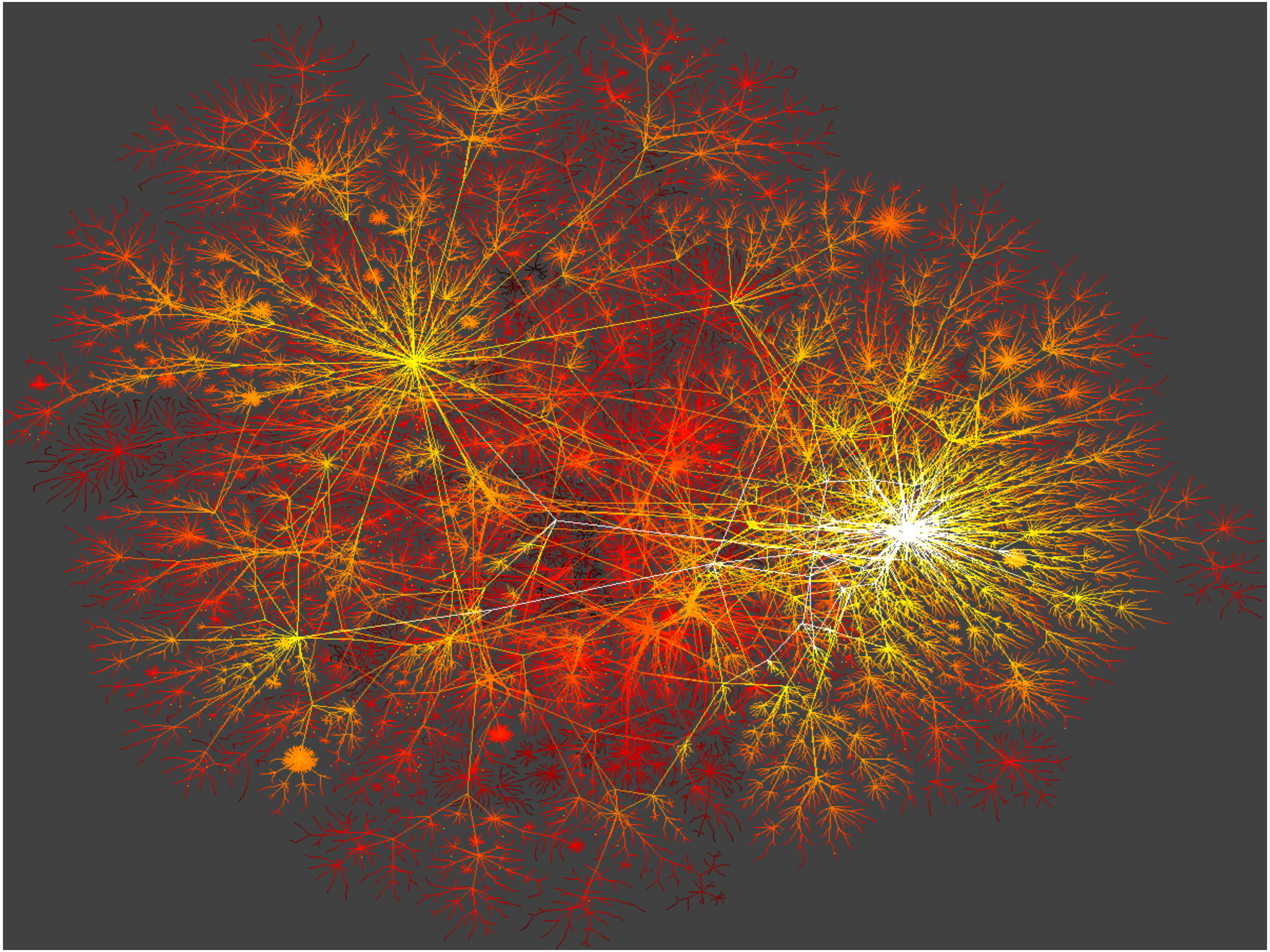


INTERNET

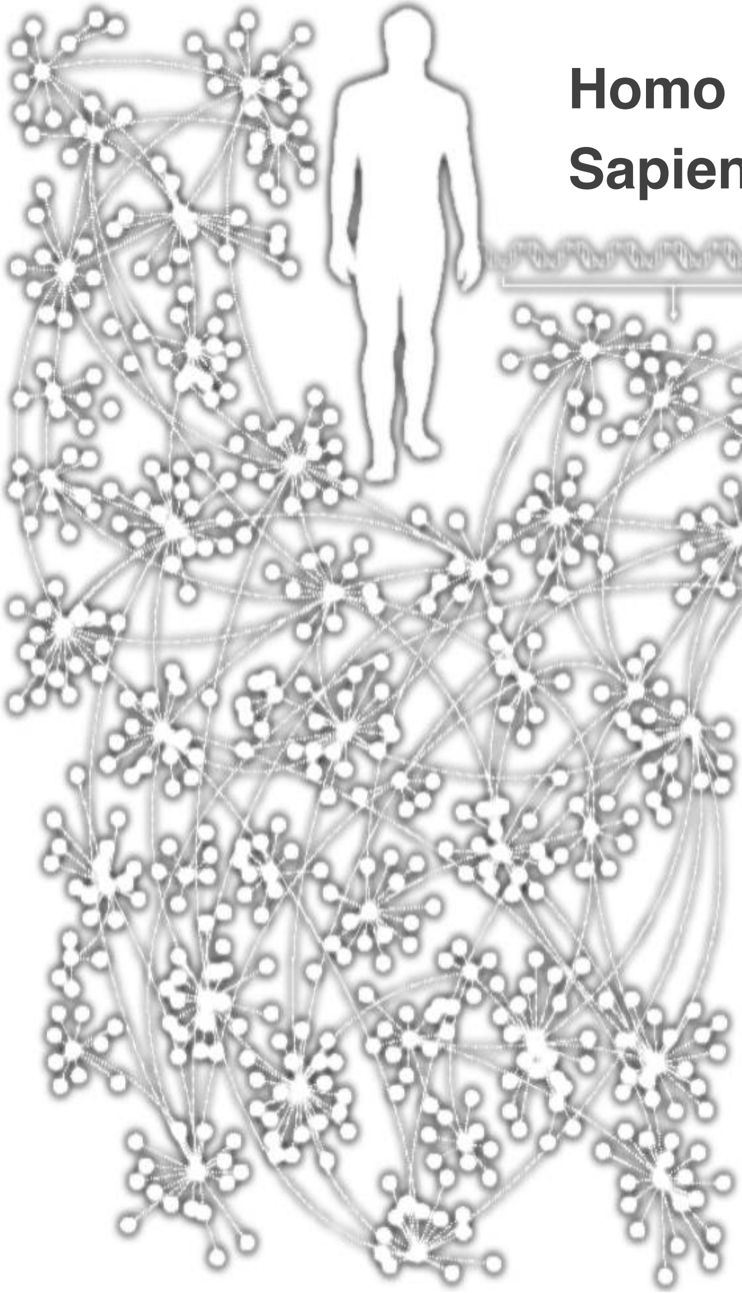
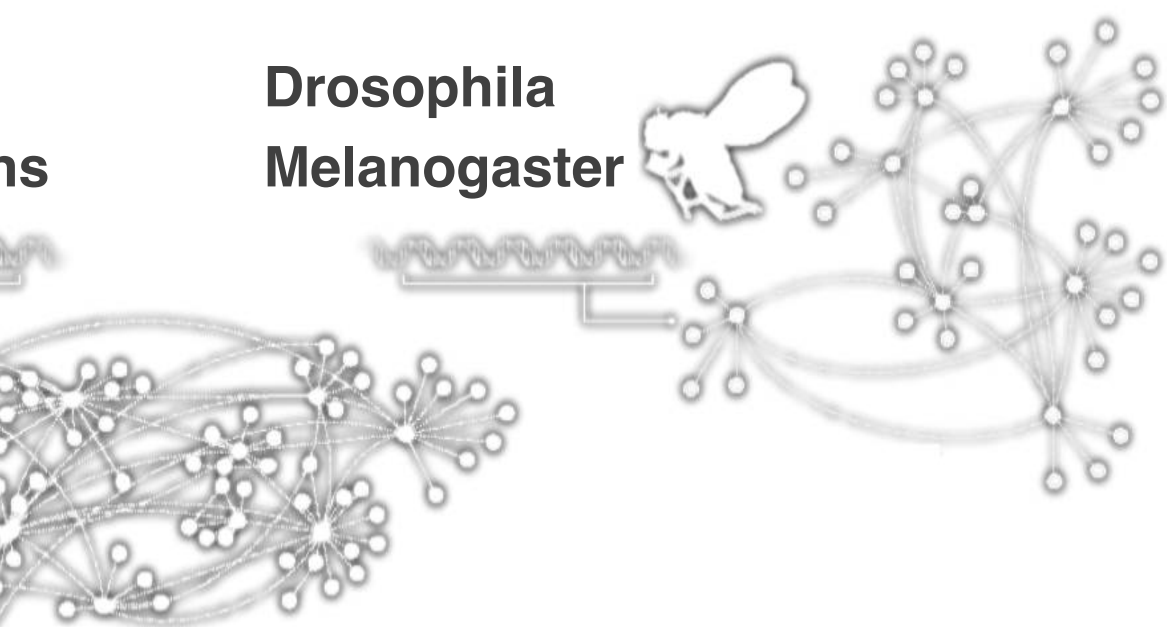








HUMANS GENES

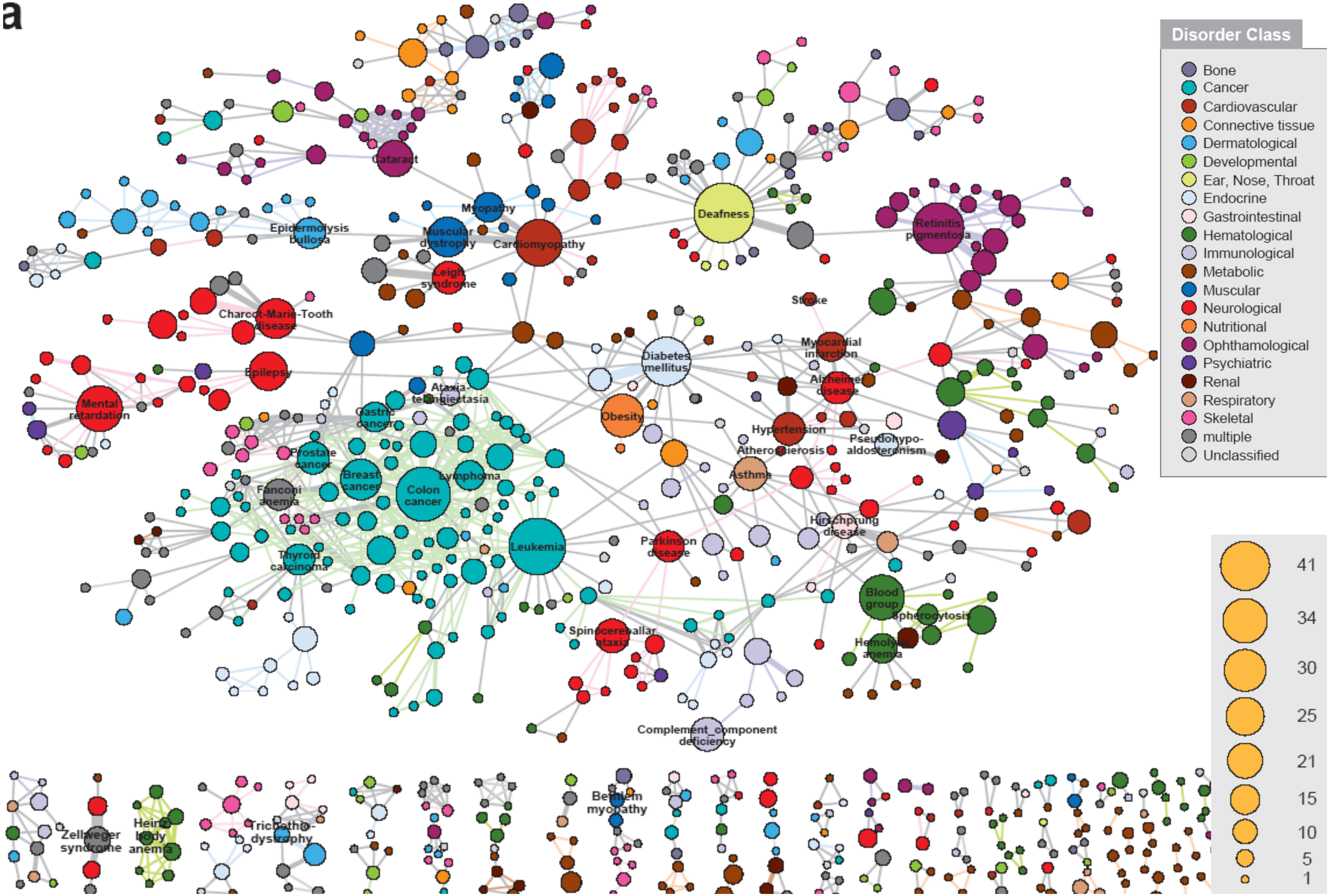



Complex systems

Made of many non-identical **elements** connected by diverse **interactions**.

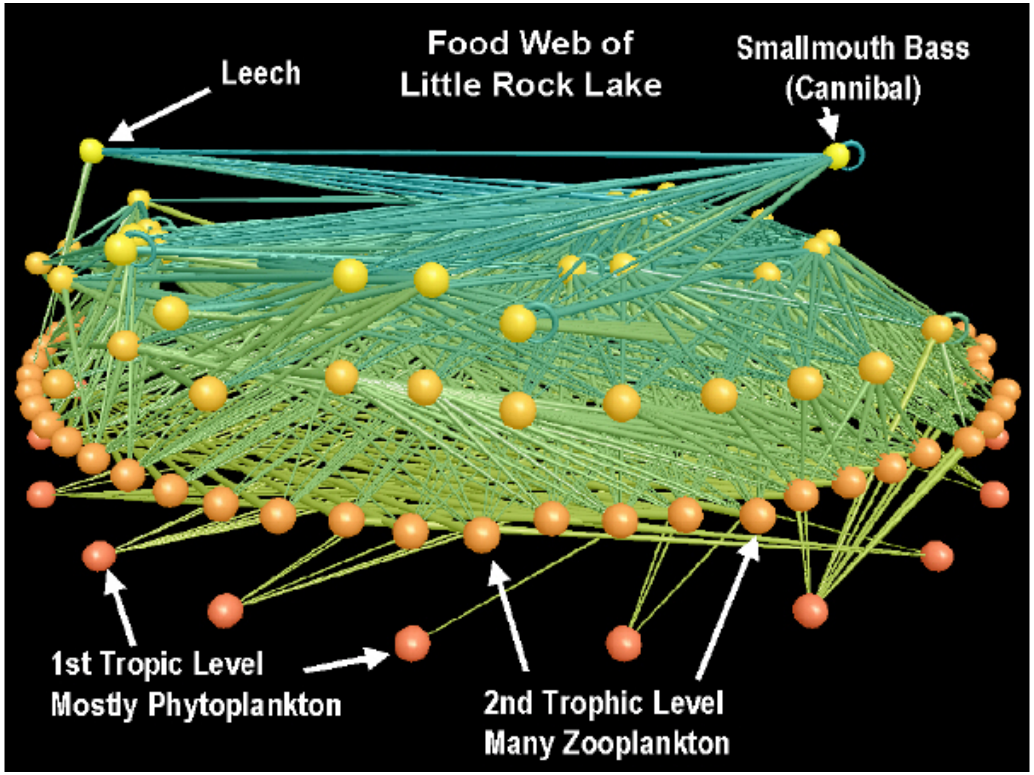
NETWORK

HUMAN DISEASE NETWORK



Biological networks: Food Web

Nodes: species
Links: trophic interactions



R. Sole (cond-mat/0011195) R.J. Williams, N.D. Martinez *Nature* (2000)

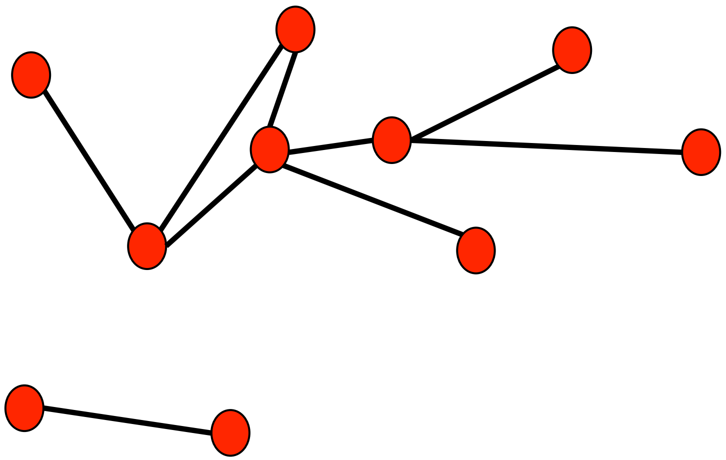
Basic network measures

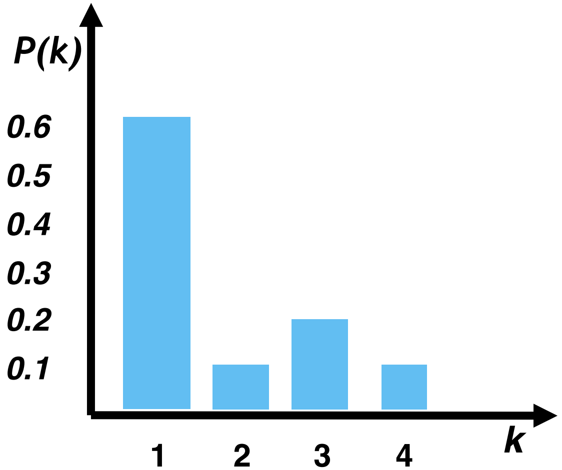
Degree of a node
Distance between two nodes
Clustering among three nodes

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

N_k = # nodes with degree k

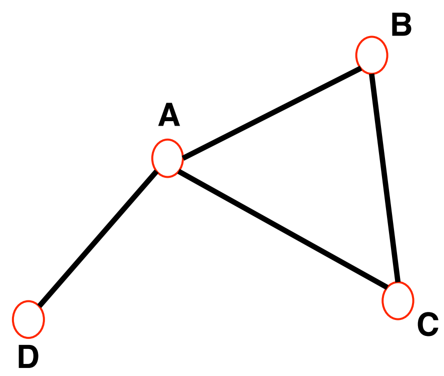
$P(k) = N_k / N$ ➔ plot





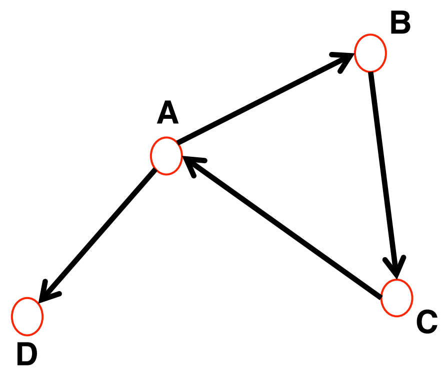
DISTANCE IN A GRAPH

Shortest Path, Geodesic Path



The *distance (shortest path, geodesic path)* between two nodes is defined as the **number of edges along the shortest path connecting them**.

*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

NETWORK DIAMETER AND AVERAGE DISTANCE

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

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

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

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

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

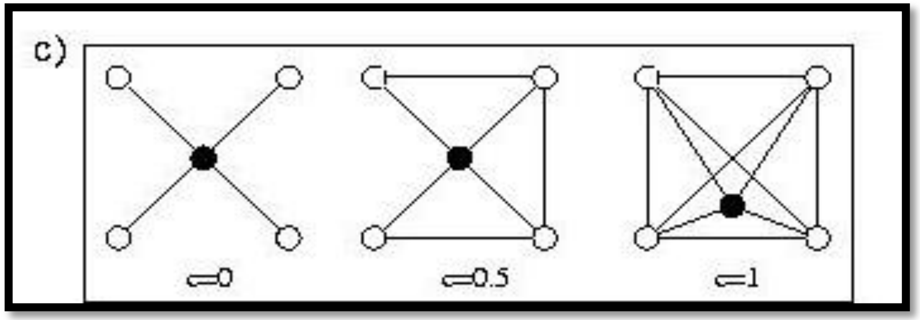
$$< l > \equiv \frac{1}{L_{\max}} \sum_{i,j > i} l_{ij}$$

Clustering coefficient:

what portion of your neighbors are connected?

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

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



KEY MEASURES

Degree distribution: **P(k)**

Path length: ***l***

Clustering coefficient:

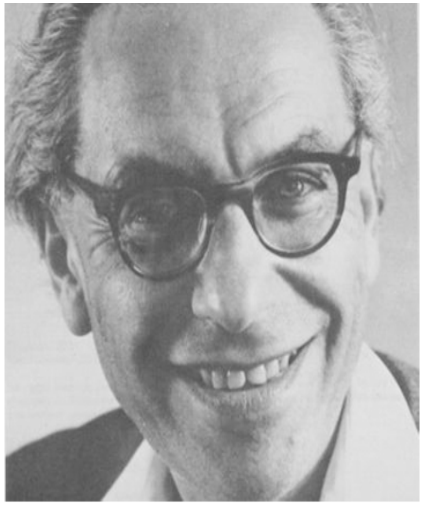
$$\boxed{C_i = \frac{2e_i}{k_i(k_i - 1)}}$$

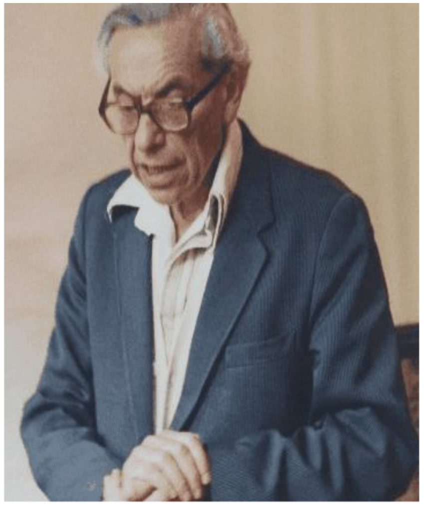
Random graphs

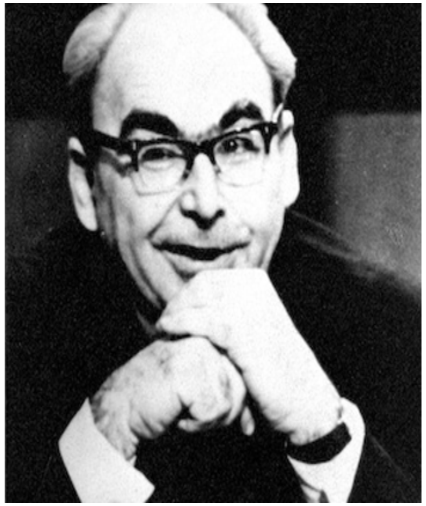
What are the expected basic measures emerging from random?

RANDOM NETWORK MODEL

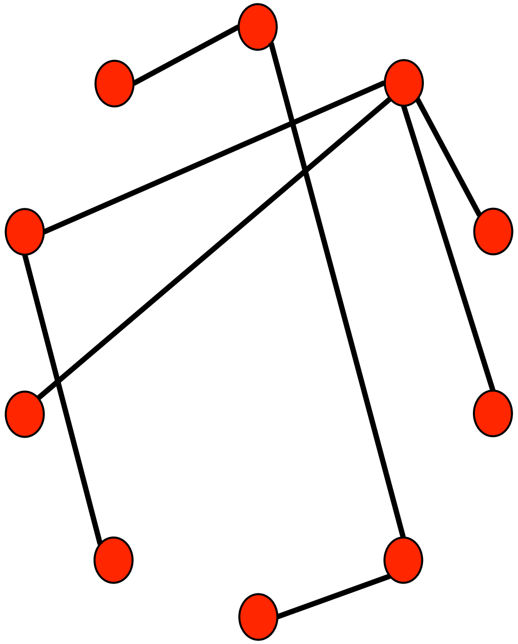
Pául Erdös
(1913-1996)







Erdös-Rényi model (1960)



Connect with probability p

p=1/6 N=10

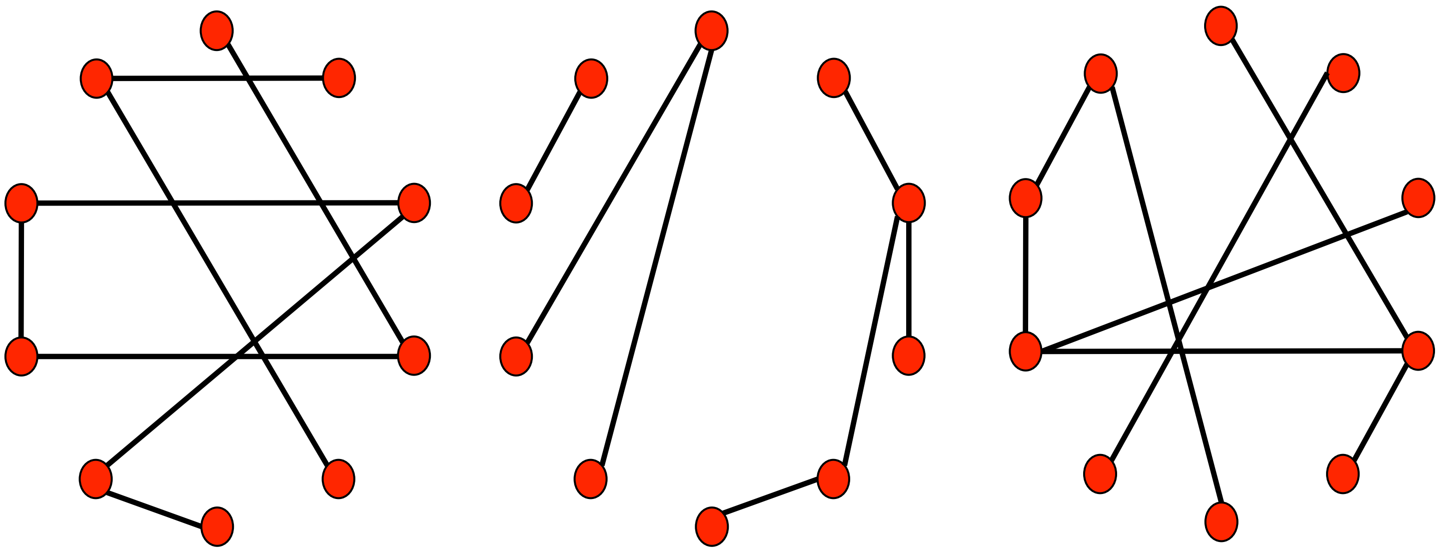
$\langle k \rangle \sim 1.5$

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?**



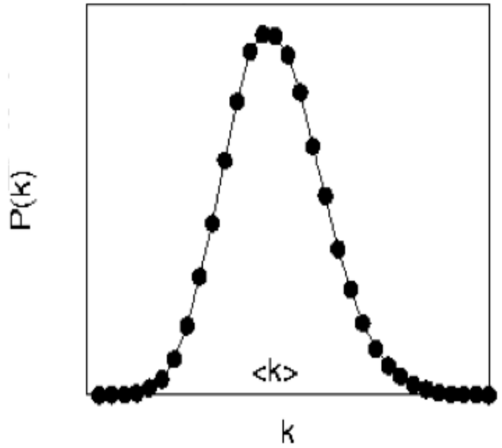
N=10
p=1/6

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))**.

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

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

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

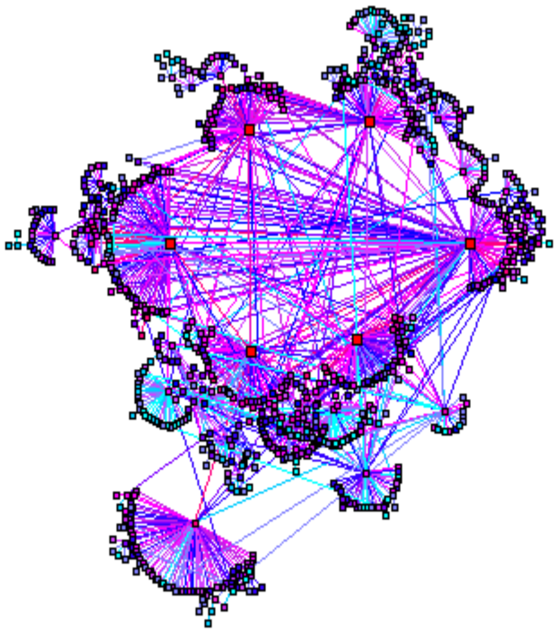
As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of <k>.

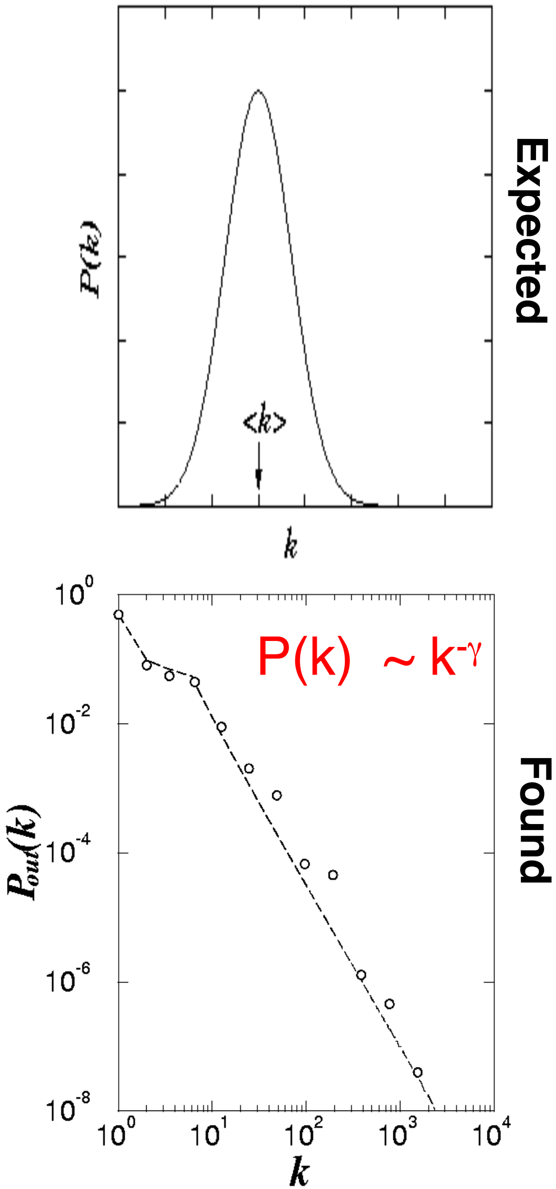
WORLD WIDE WEB

Nodes: **WWW documents**
Links: **URL links**

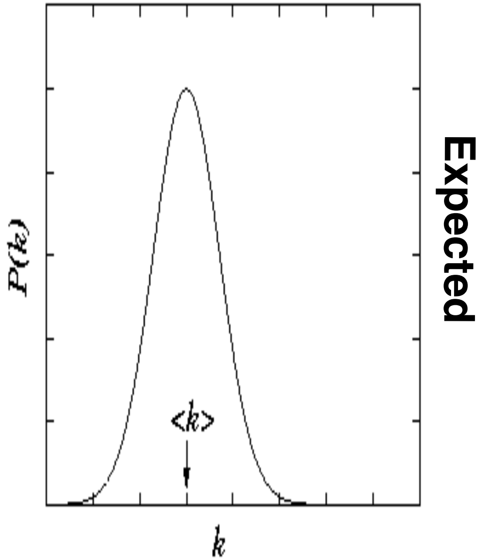
Over 3 billion documents

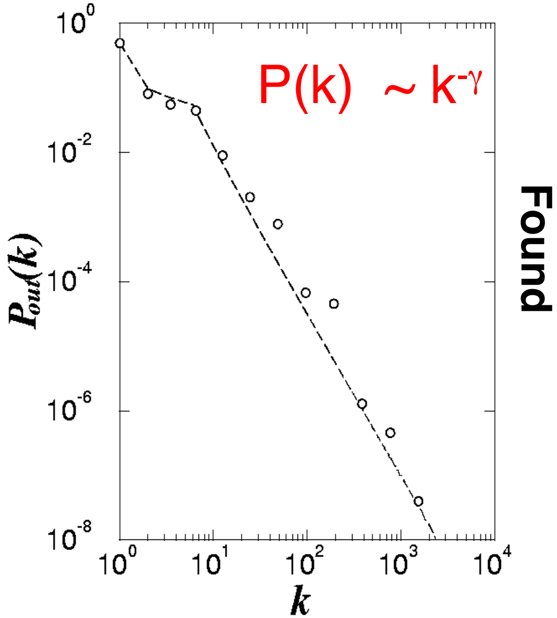
ROBOT: collects all URL's found in a document and follows them recursively

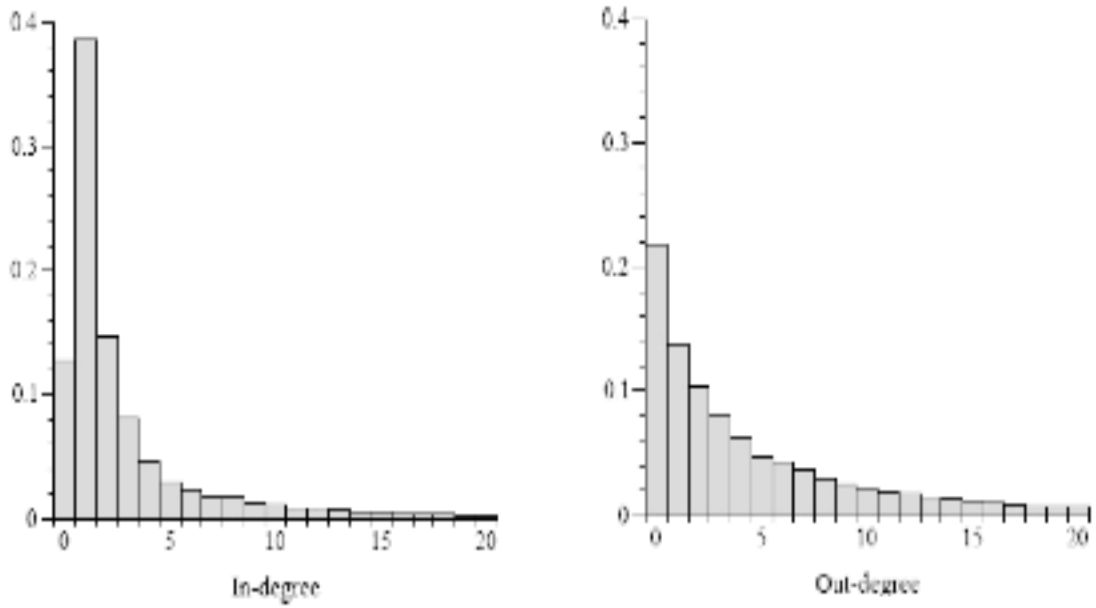




Degree distribution of the WWW

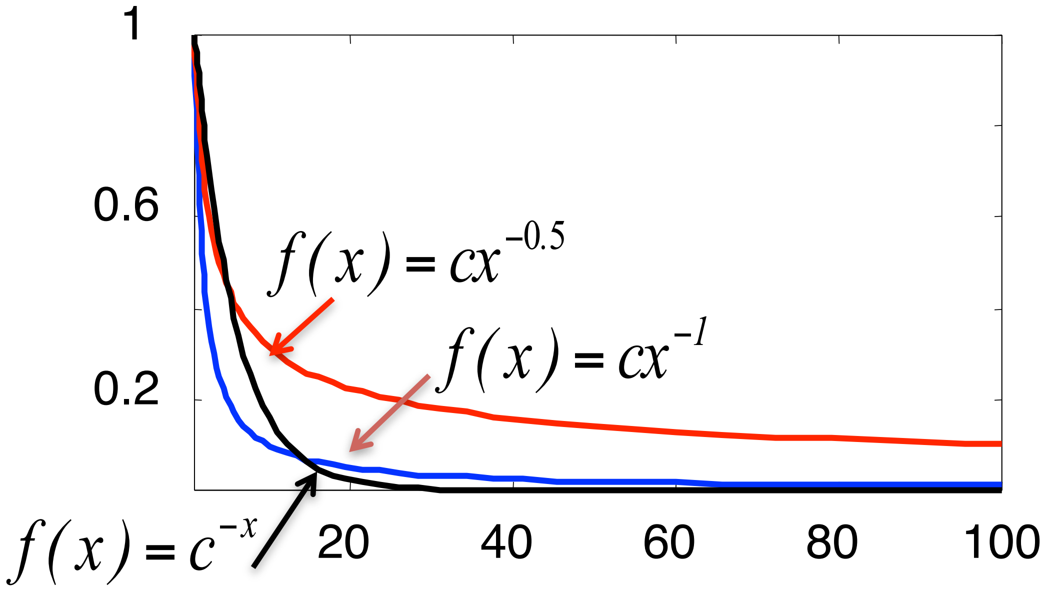






R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

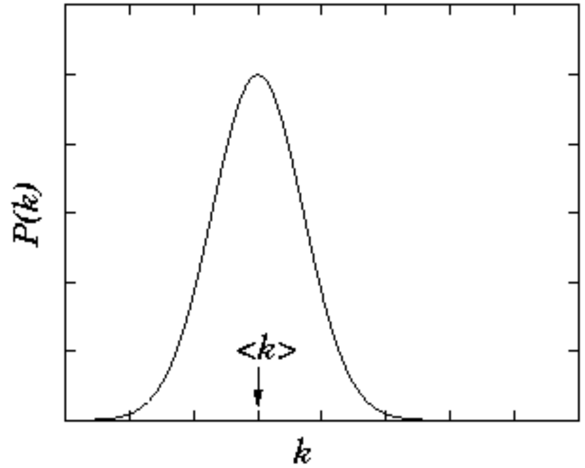
The difference between a power law and an exponential distribution



Above a certain x value, the power law is always higher than the exponential.

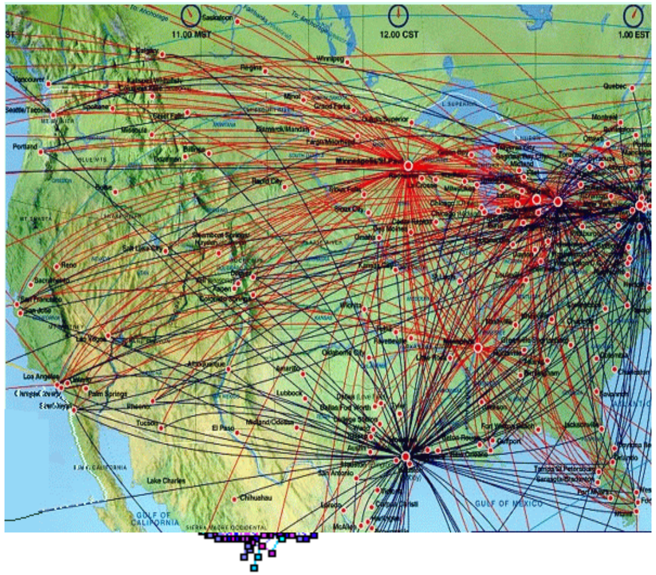
What does the difference mean? Visual representation.

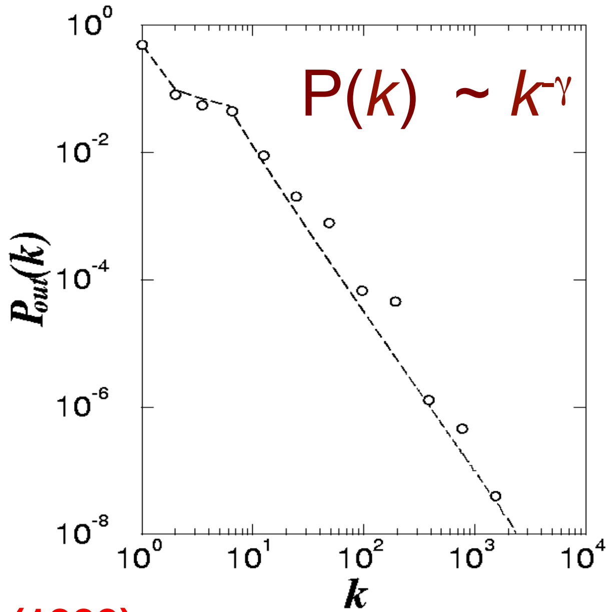
Exponential Network



Expected

Scale-free Network

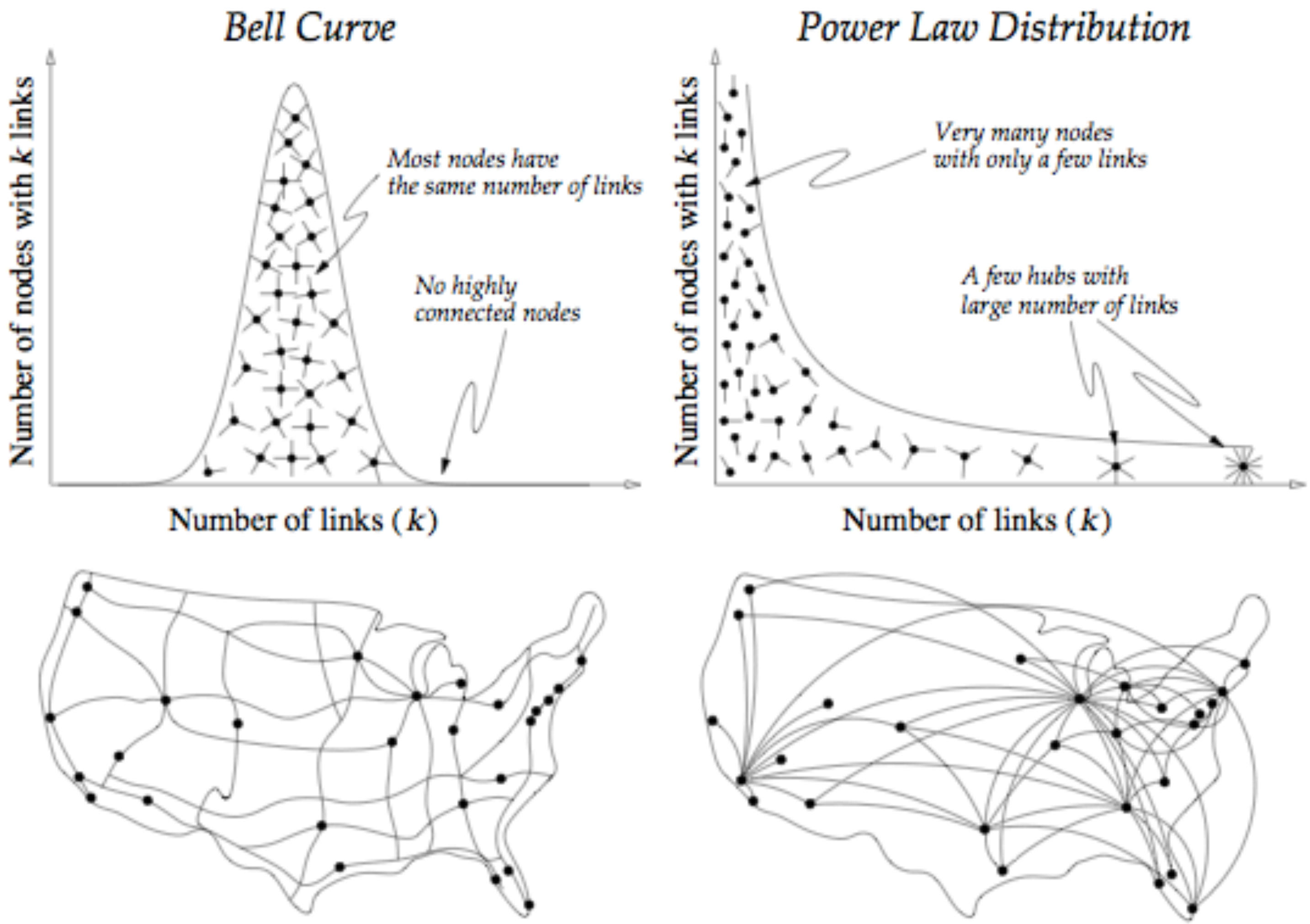




Found

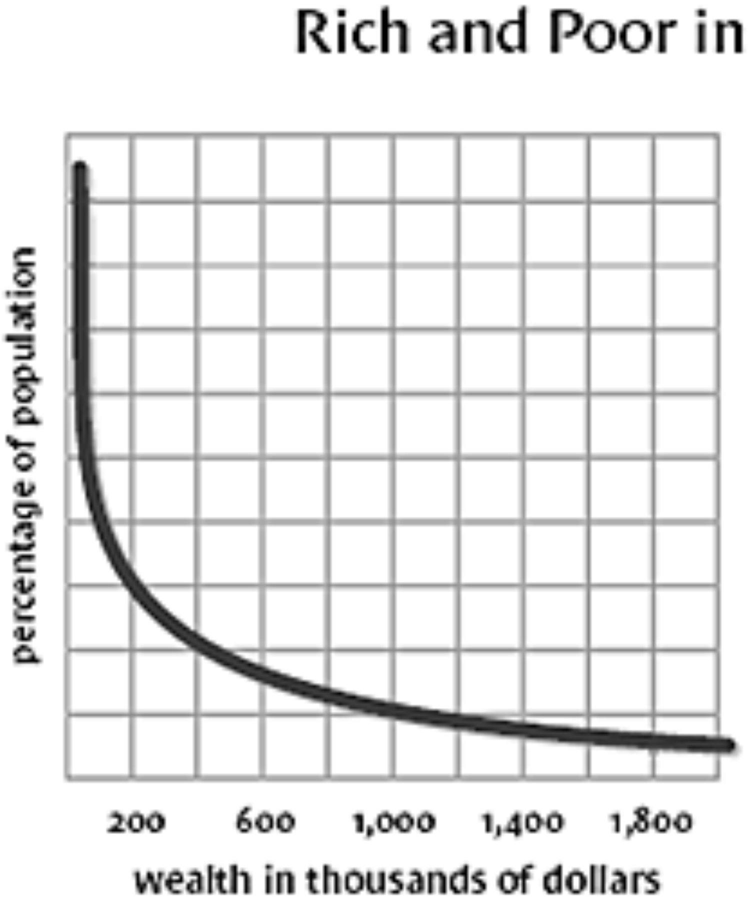
R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

WORLD WIDE WEB

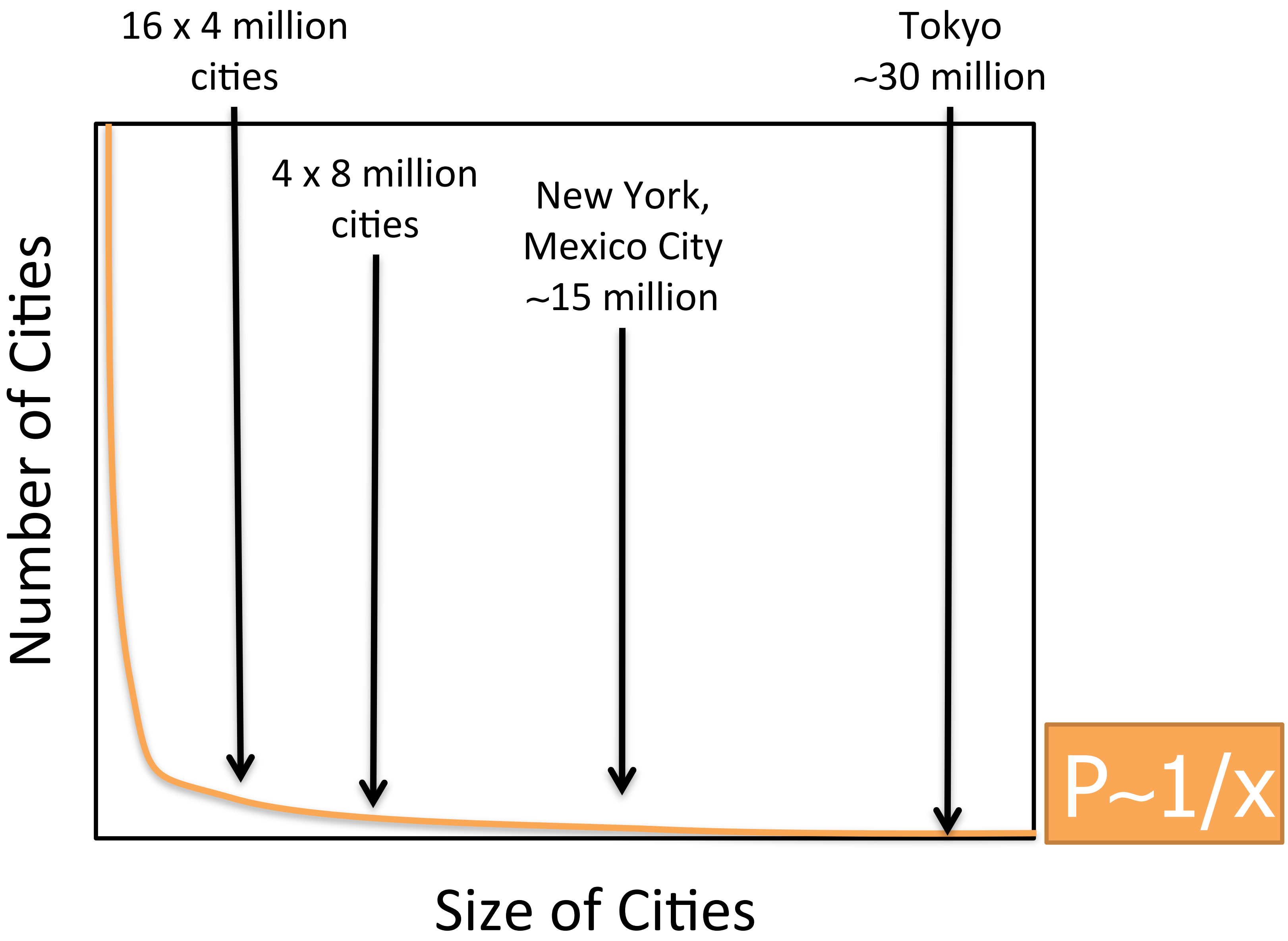


PARETO DISTRIBUTION OF WEALTH

Vilfredo Pareto (1848-1923)



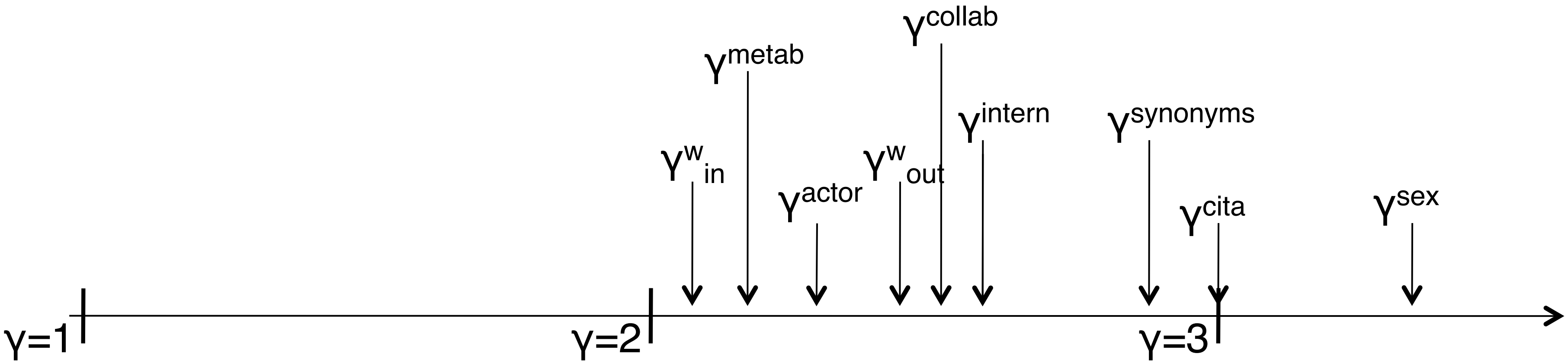
This plot of household wealth in the United States, taken from 1998 census figures, clearly shows a distribution of rich and poor forming a Pareto curve. The highest percentage of households fall at the lower levels of wealth, but at the higher end, the curve drops off relatively slowly, displaying Pareto's "fat-tailed" pattern.


16 x 4 million cities
4 x 8 million cities
New York, Mexico City ~15 million
Tokyo ~30 million
Number of Cities
Size of Cities
P~1/x

After Bill enters the arena the average income of the public ~ USD $1,000,000

~ $50 billion

VARIANCE DIVERGES!



$<k^2>$ diverges		$<k^2>$ finite
Regime full of anomalies…	The scale-free behavior is relevant	Behaves like a random network

Why are most exponents in this regime?

UNIVERSALITY

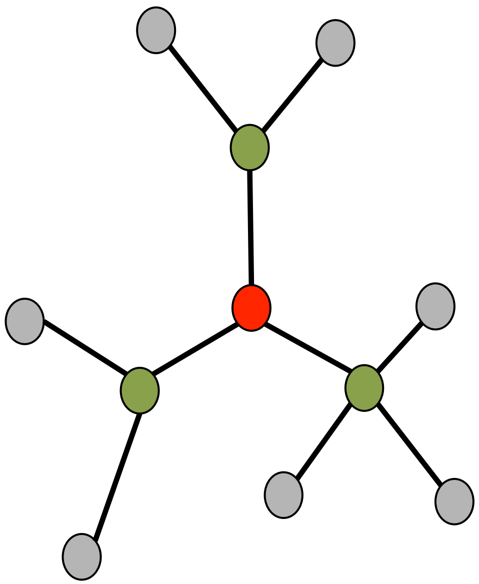
Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}
WWW	325 729	4.51	900	2.45	2.1
WWW	4×10^7	7		2.38	2.1
WWW	2×10^8	7.5	4000	2.72	2.1
WWW, site	260 000				1.94
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2
Internet, router*	3888	2.57	30	2.48	2.48
Internet, router*	150 000	2.66	60	2.4	2.4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

Networks:
The exponents vary from system to system.
Most are between 2 and 3

Universality:
the emergence of common features across different networks. Like the scale-free property.

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	$\langle k \rangle$	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.

CLUSTERING COEFFICIENT

$$C_i \equiv \frac{2n_i}{k_i(k_i - 1)}$$

Since edges are independent and have the same probability *p*,

$$n_i \cong p\frac{k_i(k_i - 1)}{2} \quad \Rightarrow \quad C \cong p = \frac{<k>}{N}$$

The clustering coefficient of random graphs is small.

For fixed degree C decreases with the system size N.

13.47 from Newman 2010

CLUSTERING IN RANDOM GRAPHS compare with real data

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

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?

NO!

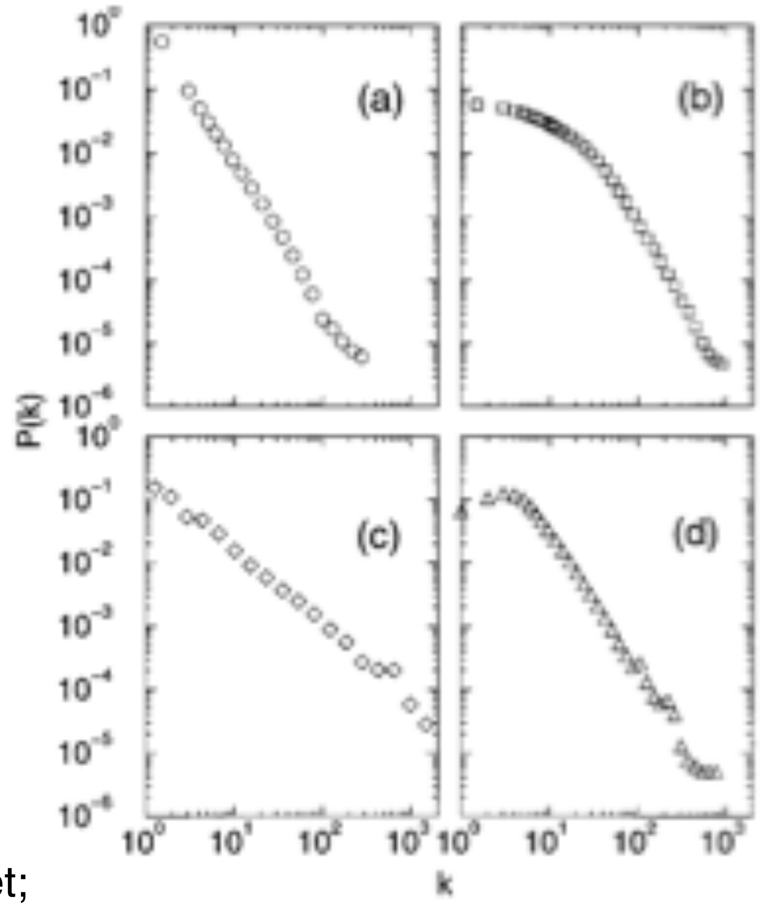
THE DEGREE DISTRIBUTION

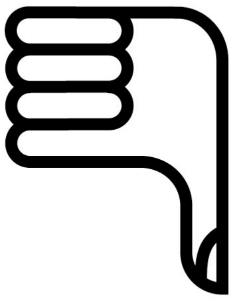
Prediction:

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

Data:

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





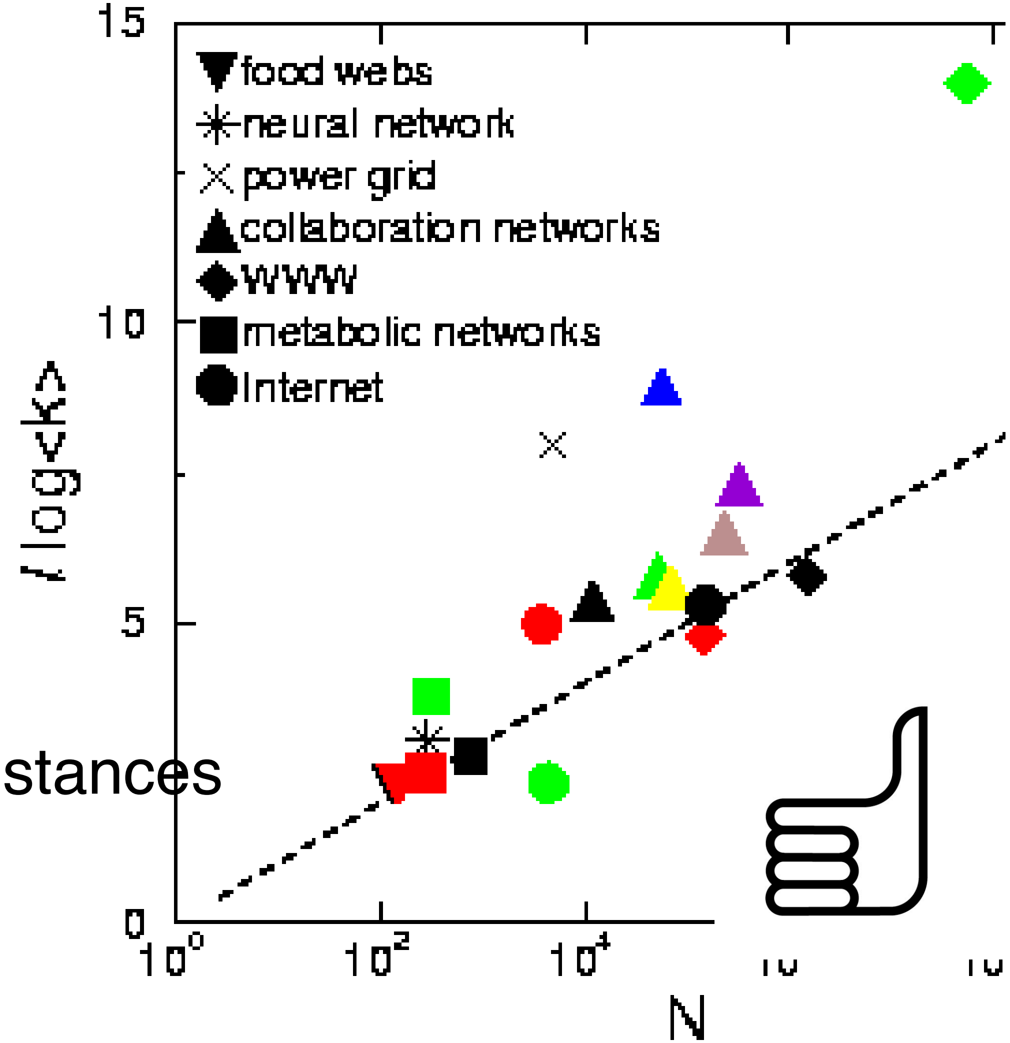
(a) Internet;
(b) Movie Actors;
(c) Coauthorship, high energy physics;
(d) Coauthorship, neuroscience

PATH LENGTHS IN REAL NETWORKS

Prediction:

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

Data:



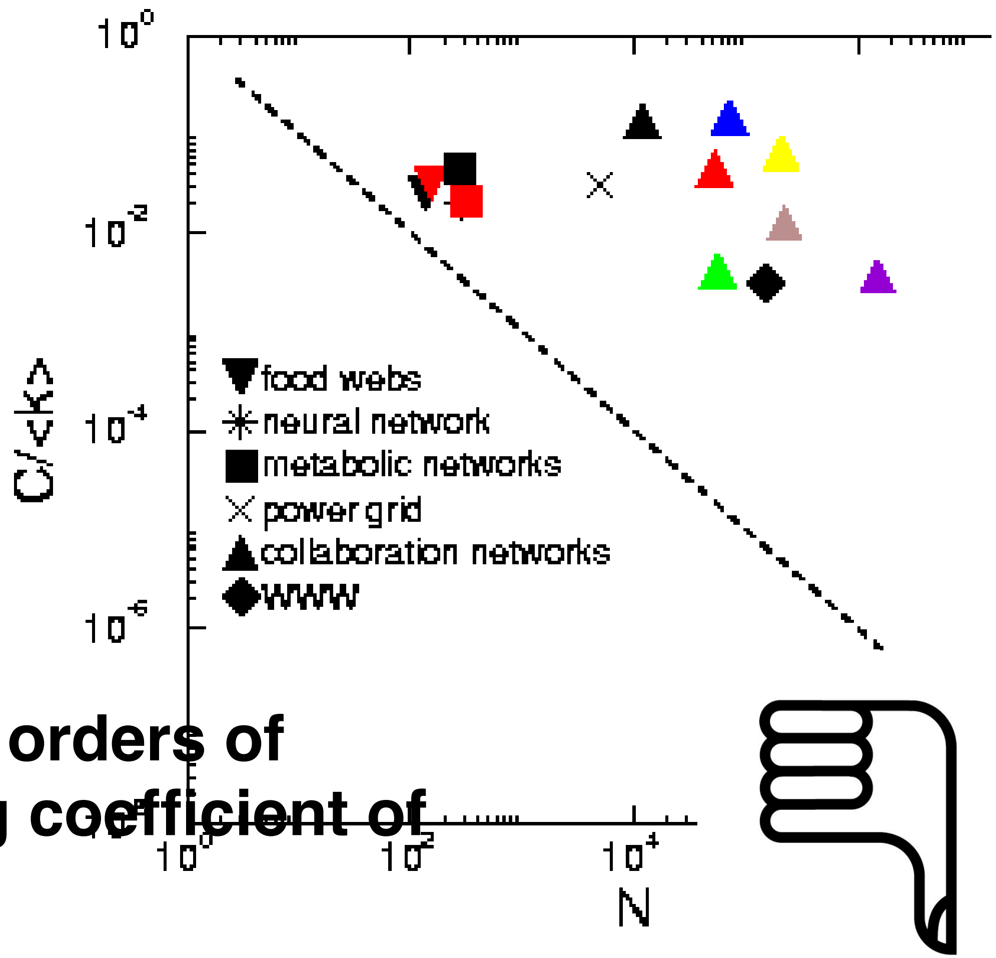
Real networks have short distances like random graphs.

CLUSTERING COEFFICIENT

Prediction:

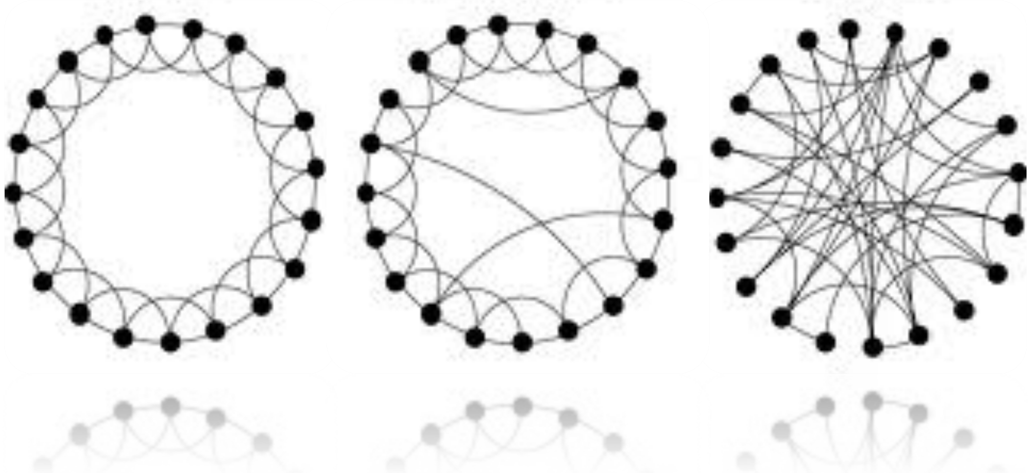
$$C_{rand} = \frac{\langle k \rangle}{N}$$

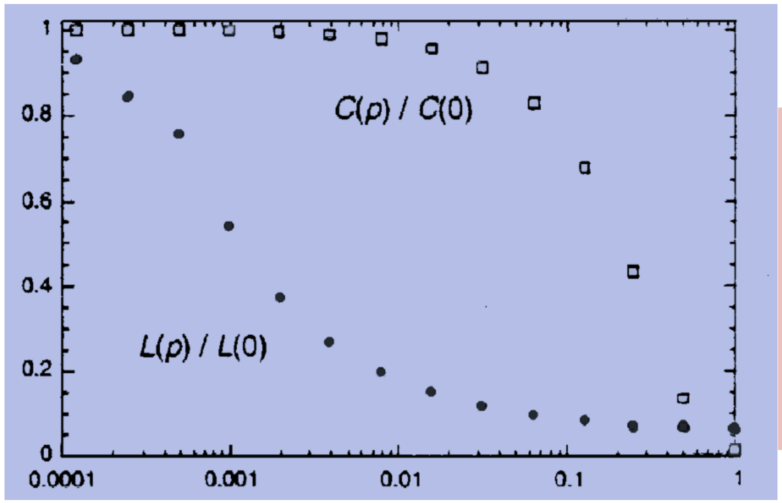
Data:



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

Models for «real» networks: small world





The Watts Strogatz Model:
It takes a lot of randomness to ruin the clustering, but a very small amount to overcome locality

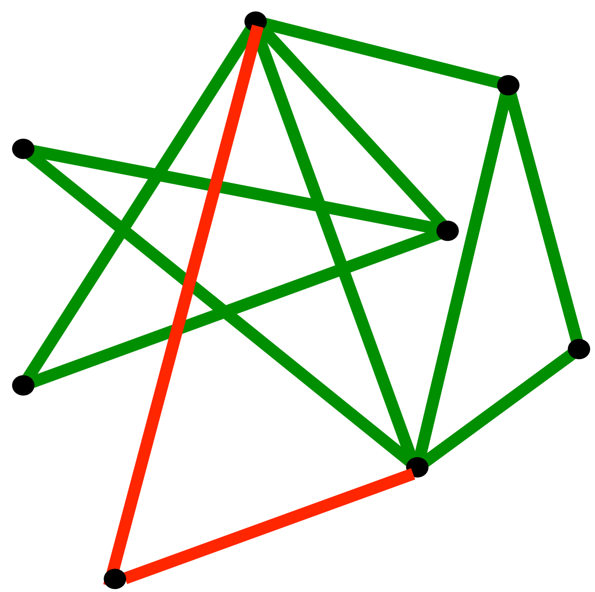
Models for real networks: Preferential Attachment

Where will the new node link to?
ER, WS models: choose randomly.

New nodes prefer to link to highly connected nodes (www, citations, IMDB).

PREFERENTIAL ATTACHMENT:

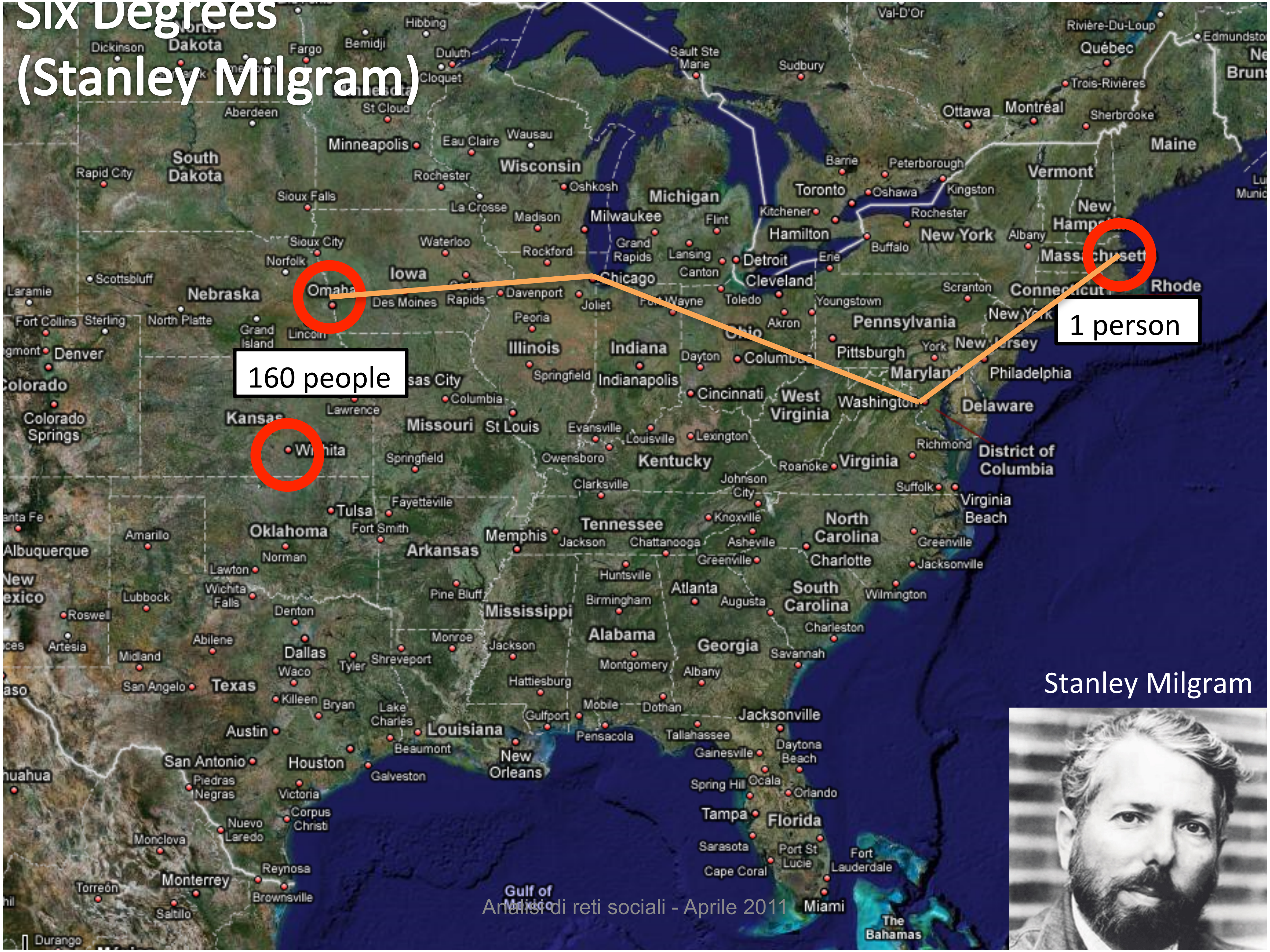
the probability that a node connects to a node with *k* links is proportional to *k*.



$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$

Barabási & Albert, *Science* **286,** 509 (1999)

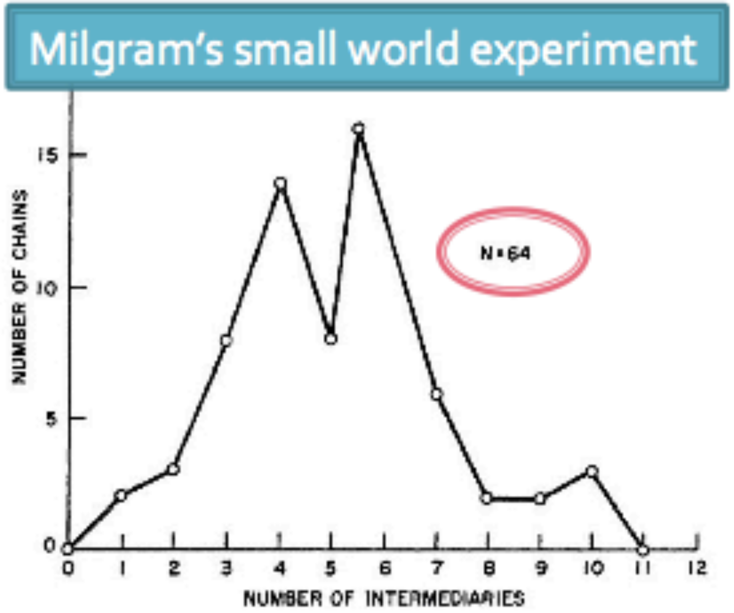
Six Degrees (Stanley Milgram)



Stanley Milgram

The Small-world experiment

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



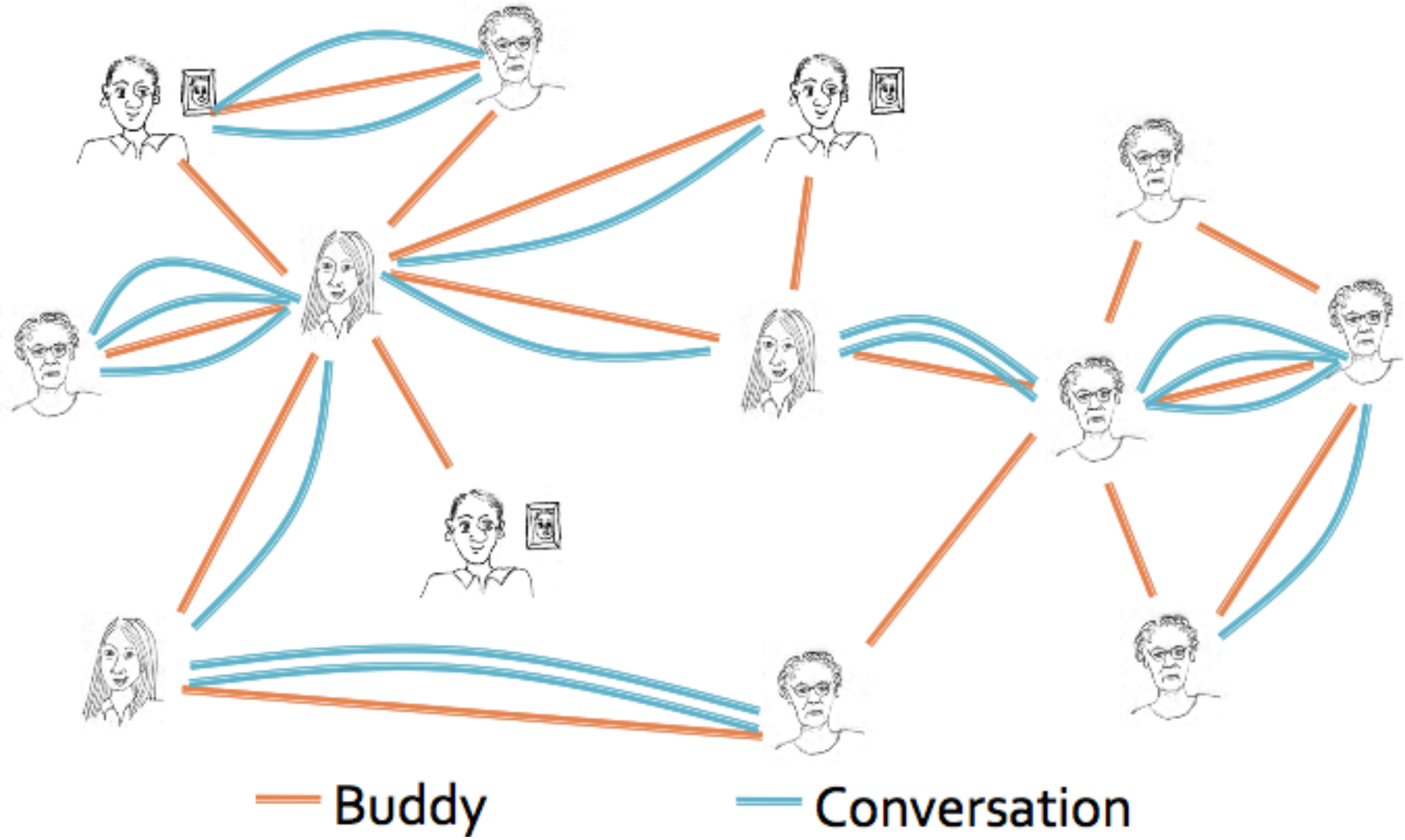
- Further observations:
  - People what owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
  - People from the Boston area have even closer paths: 4.4

Planetary-Scale Views on an Instant-Messaging Network

Jure Leskovec & Eric Hirvitz

Microsoft Research Technical Report MSR-TR-2006-186 June 2007

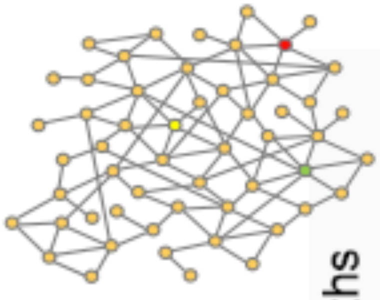
Messaging as a network



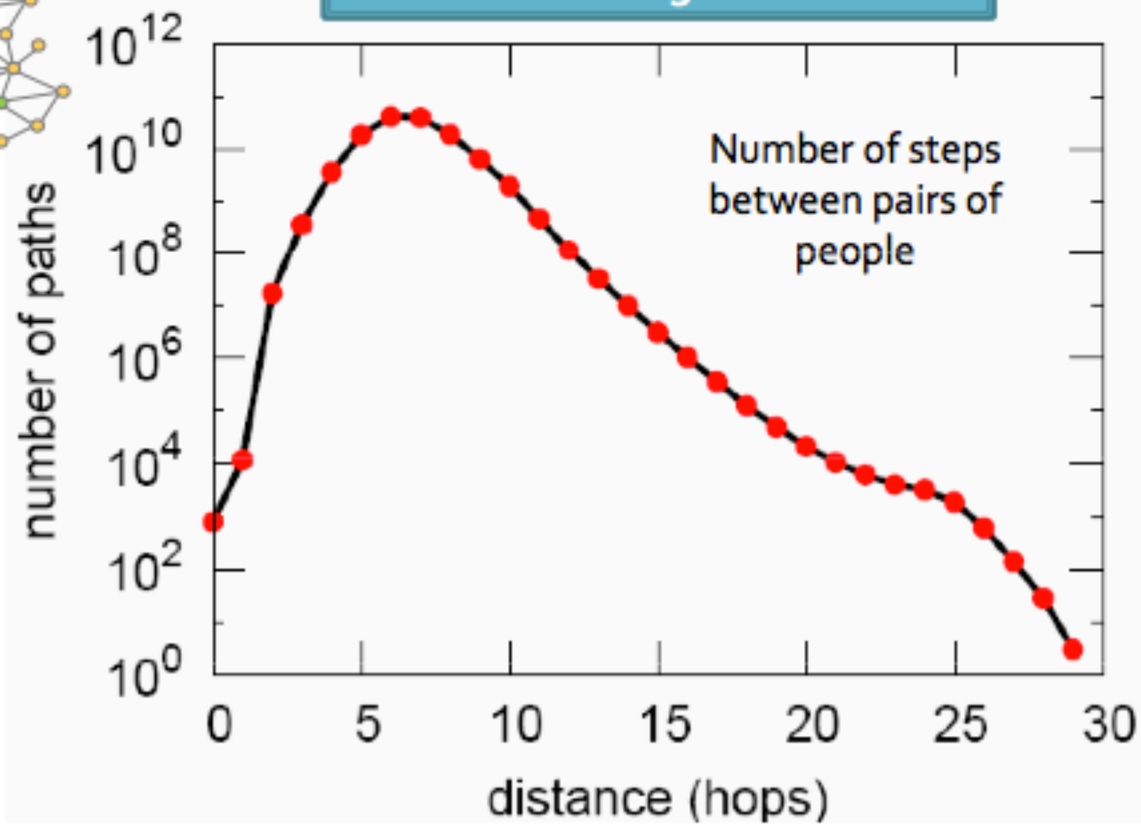
IM communication network

- **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

MSN Network: Small world



MSN Messenger network

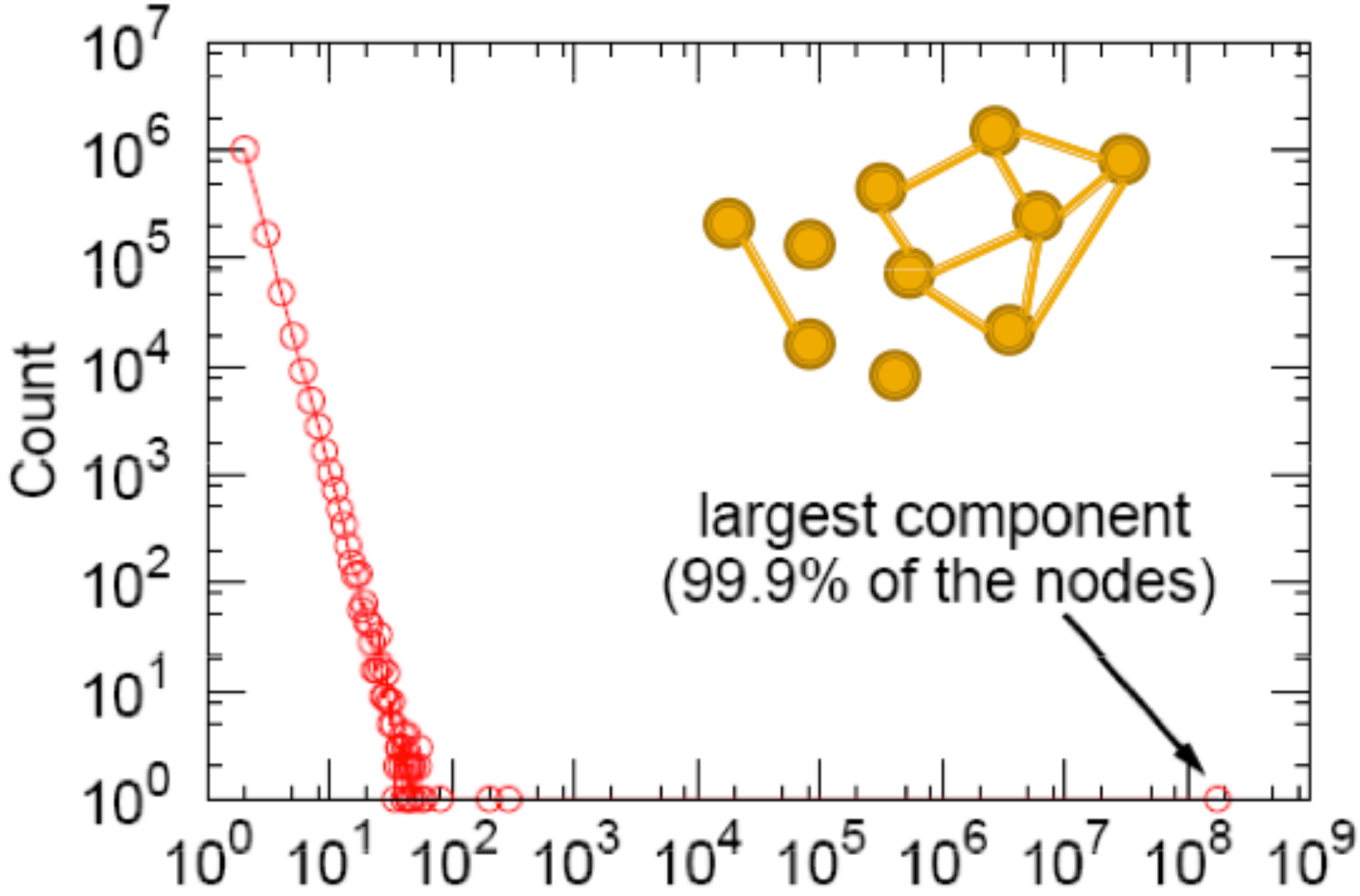


Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Avg. path length **6.6**

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

The giant connected component



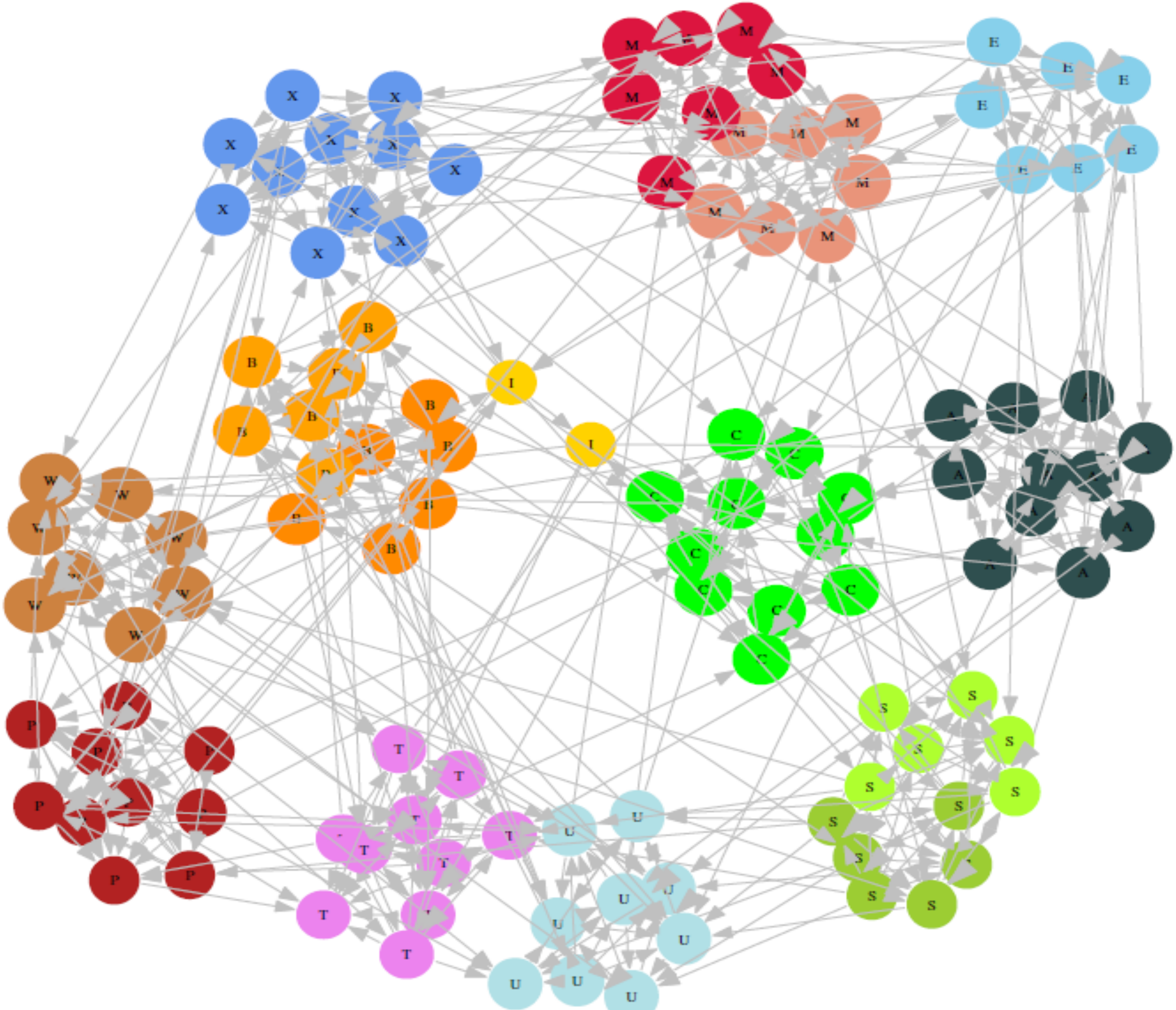
The strength of weak ties

- Mark S. **Granovetter**, 1973
- His PhD thesis: how people get to know about new jobs?
- Through personal contacts
- Surprise: often acquaintances, **not** close friends

The Strength of Weak Ties

Mark S. Granovetter

American Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380.


M
E
X
B
I
W
C
A
P
T
U
S

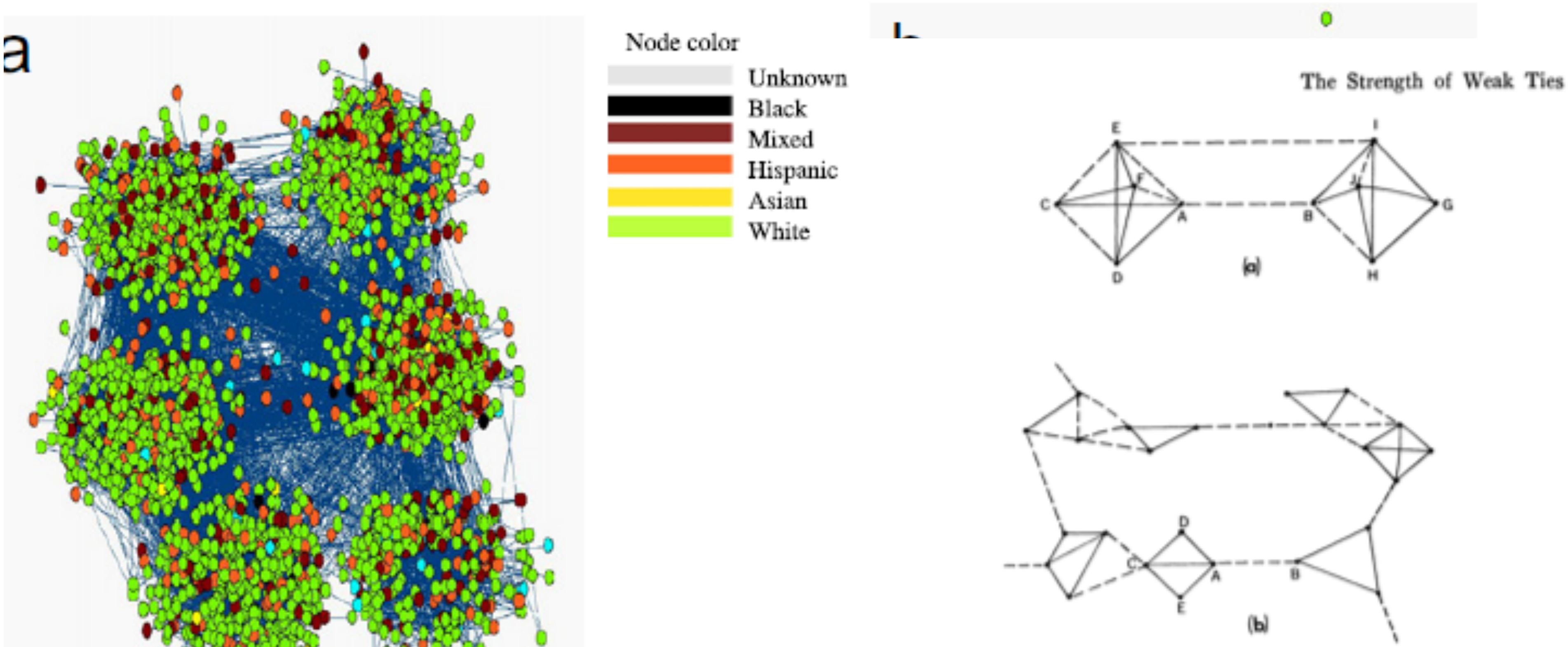
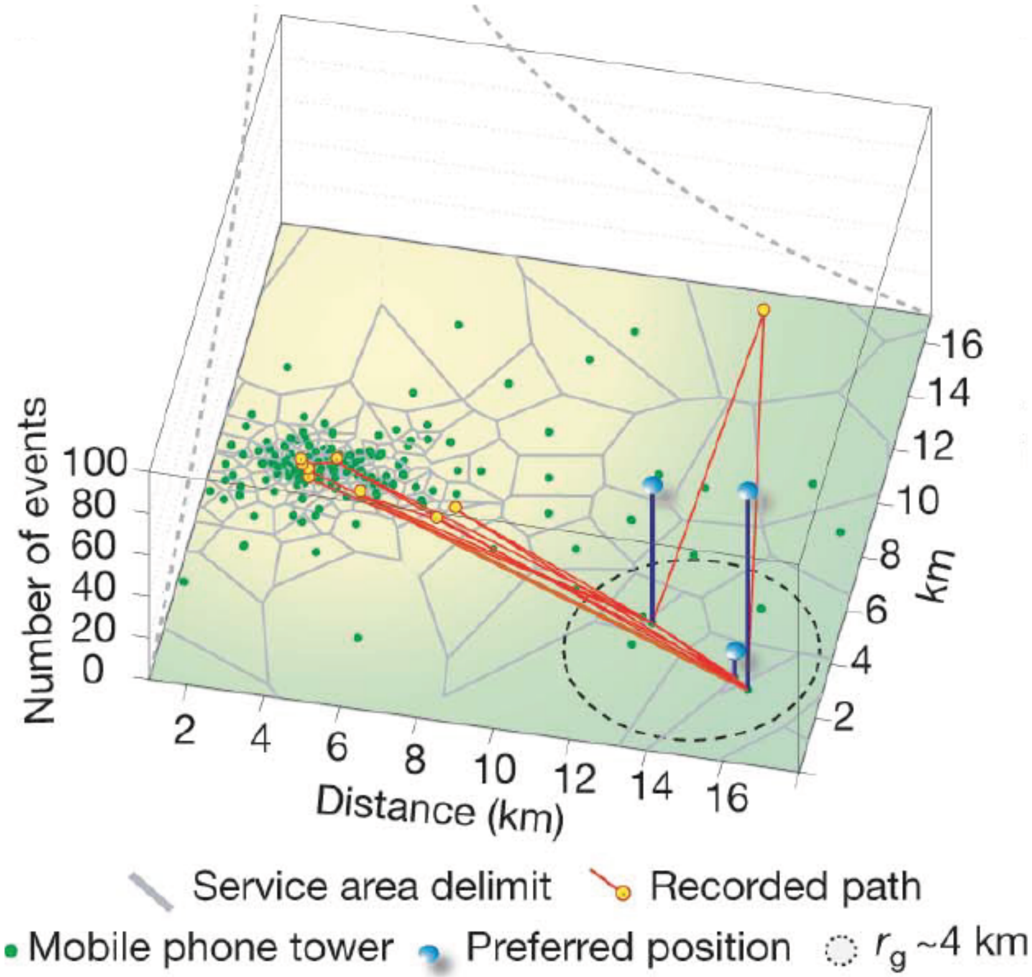


FIG. 2.—Local bridges. *a*, Degree 3; *b*, Degree 13. ——— = strong tie; – – – = weak tie.

Country-wide mobile phone data



when
you
call

where
you
call

who
you
call

Social proximity and tie strength

- How connected are u and v in the social network.
  - Various well-established **measures of network proximity**, based on the common neighbors (Jaccard, Adamic-Adar) or the structure of the paths (Katz) connecting u and v in the who-calls-whom network.
- How intense is the interaction between u and v.
  - Number of calls as **strength of tie**

Strength of weak ties

- Large scale empirical validation of Granovetter's theory
  - Social proximity increases with tie strength
  - Weak ties span across different communities

- J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. **Structure and tie strengths in mobile communication networks**. PNAS 104 (18), 7332-7336 (2007).

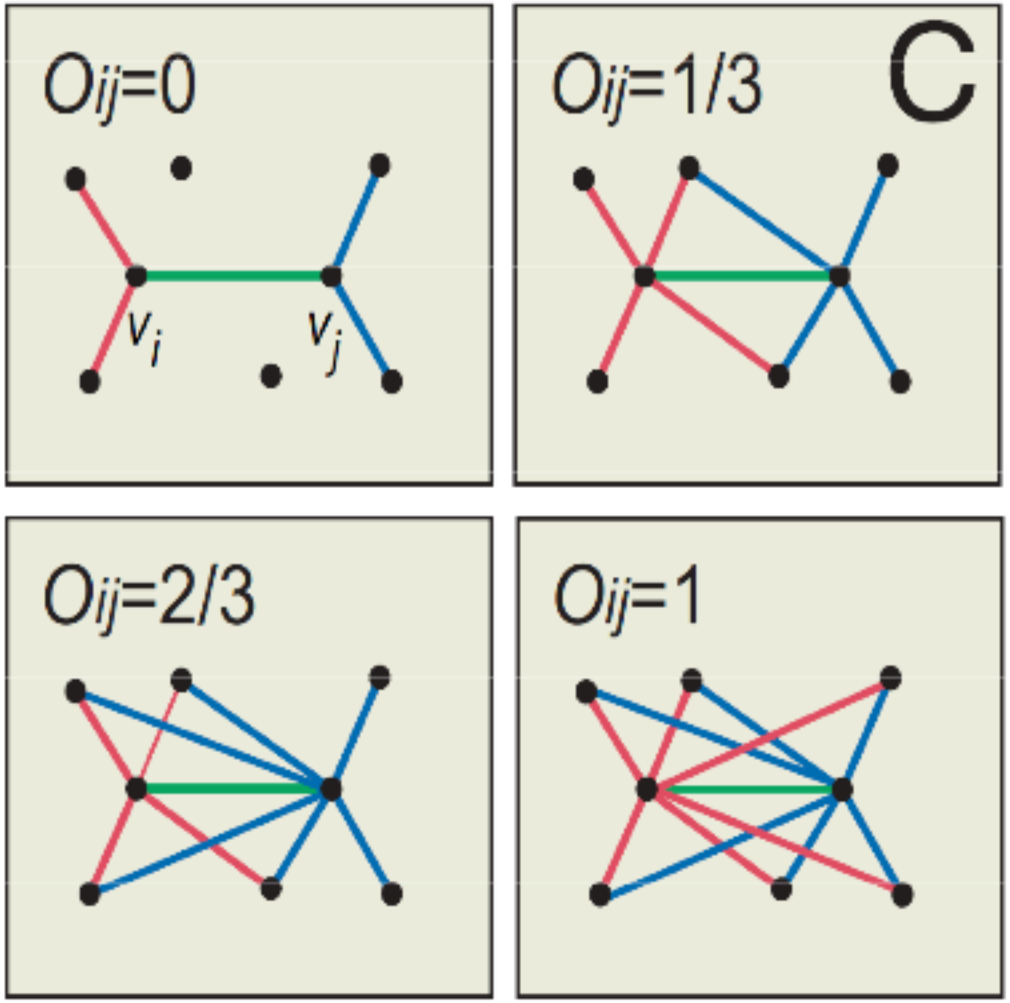
Neighborhood Overlap

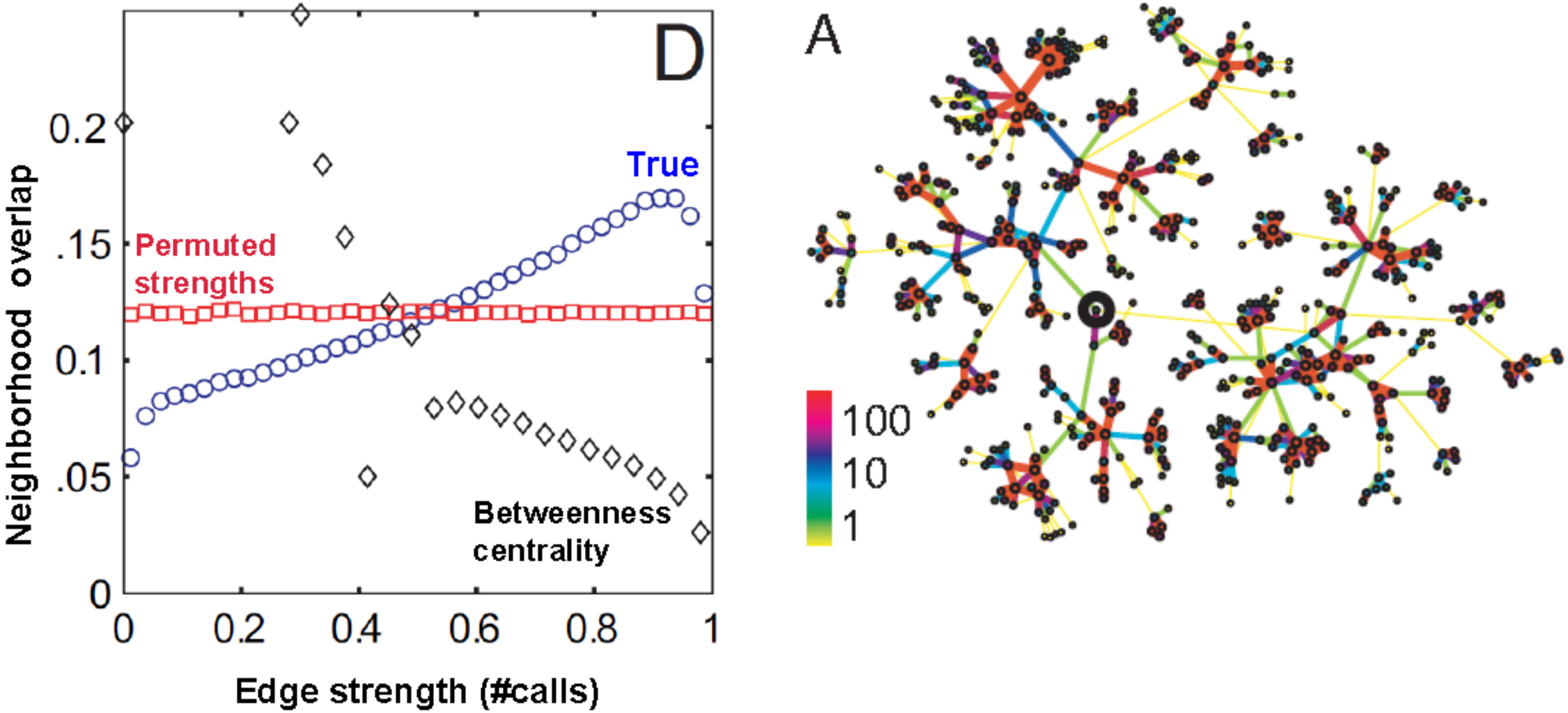
- Overlap:

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

  - n(i) … set of neighbors of A

- Overlap = 0 when an edge is a local bridge

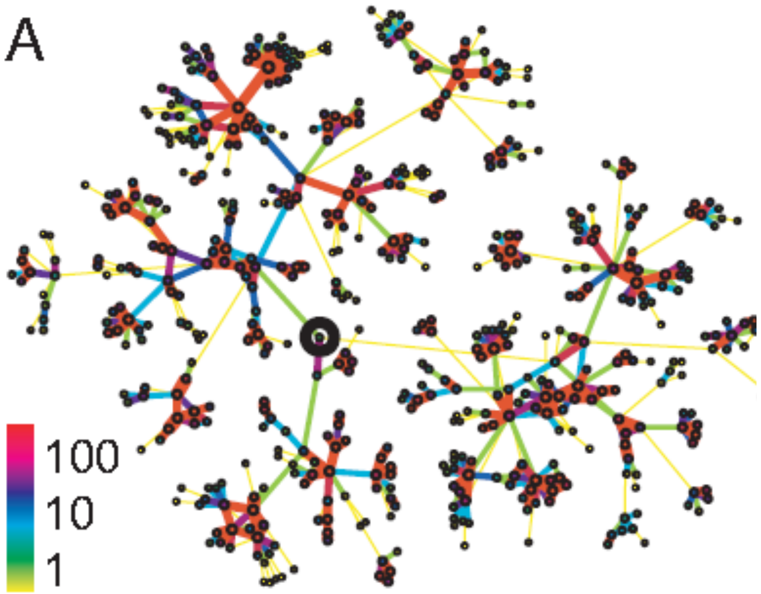


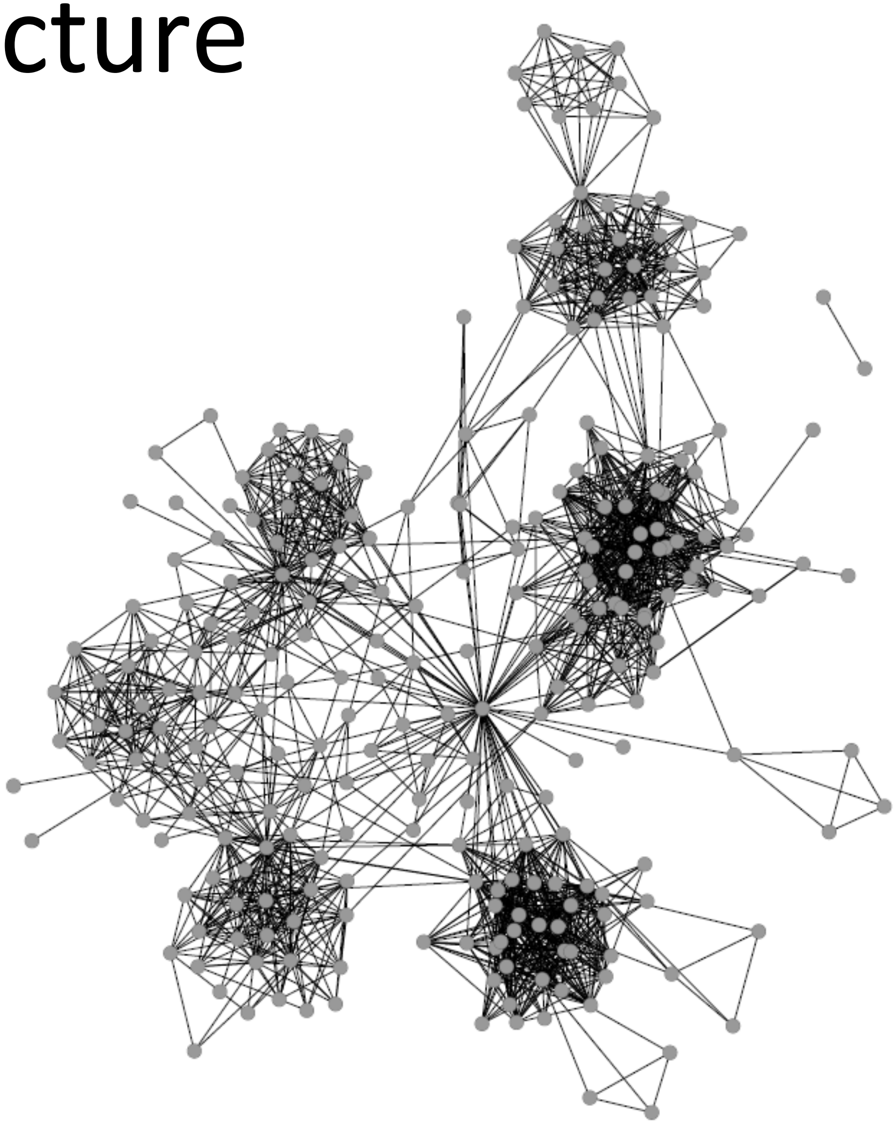

D
A
Neighborhood overlap
Edge strength (#calls)
True
Permuted
strengths
Betweenness
centrality
0.2
.15
0.1
.05
0
0
0.2
0.4
0.6
0.8
1
100
10
1

Social network mining: community discovery

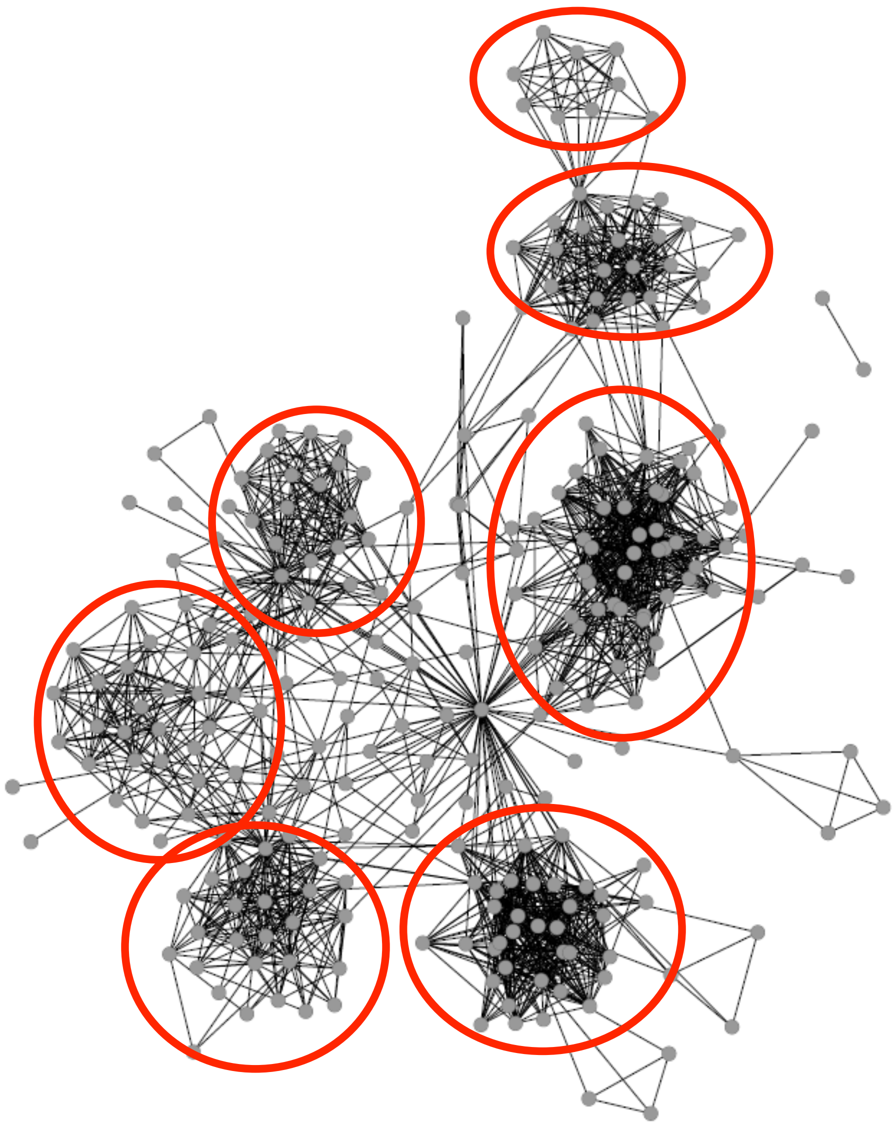
How to highlight the modular structure of a network?

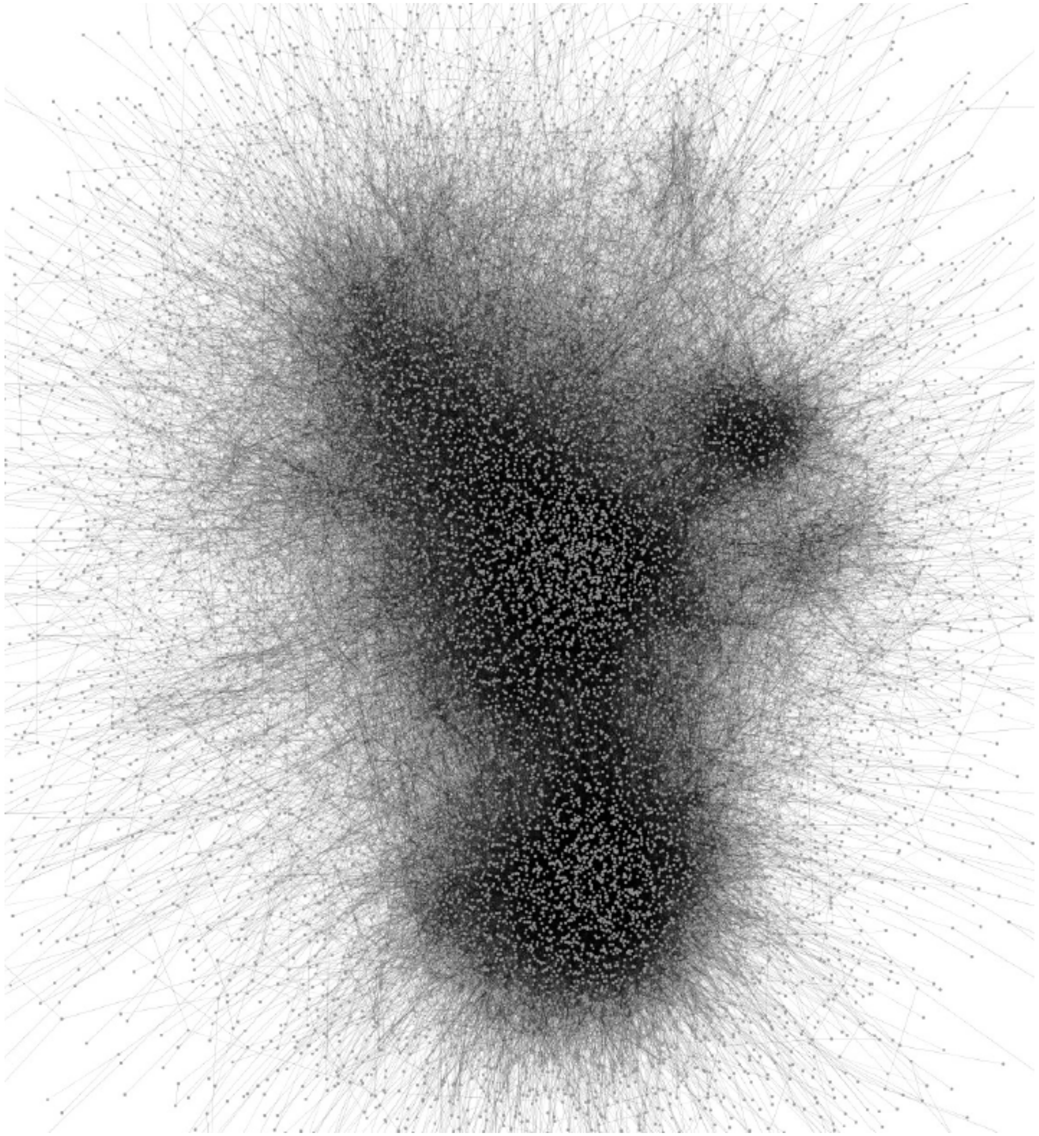
Community structure

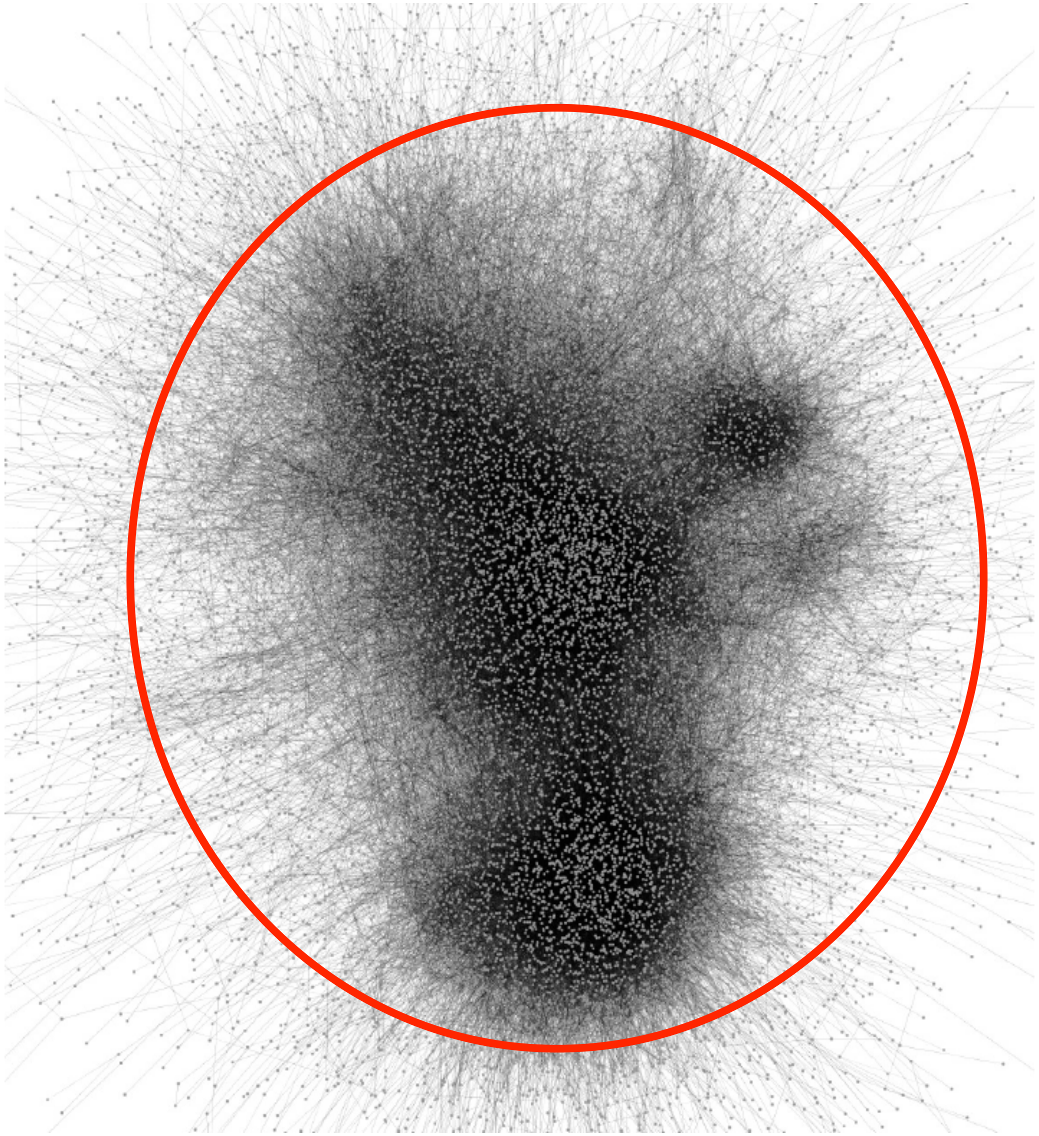




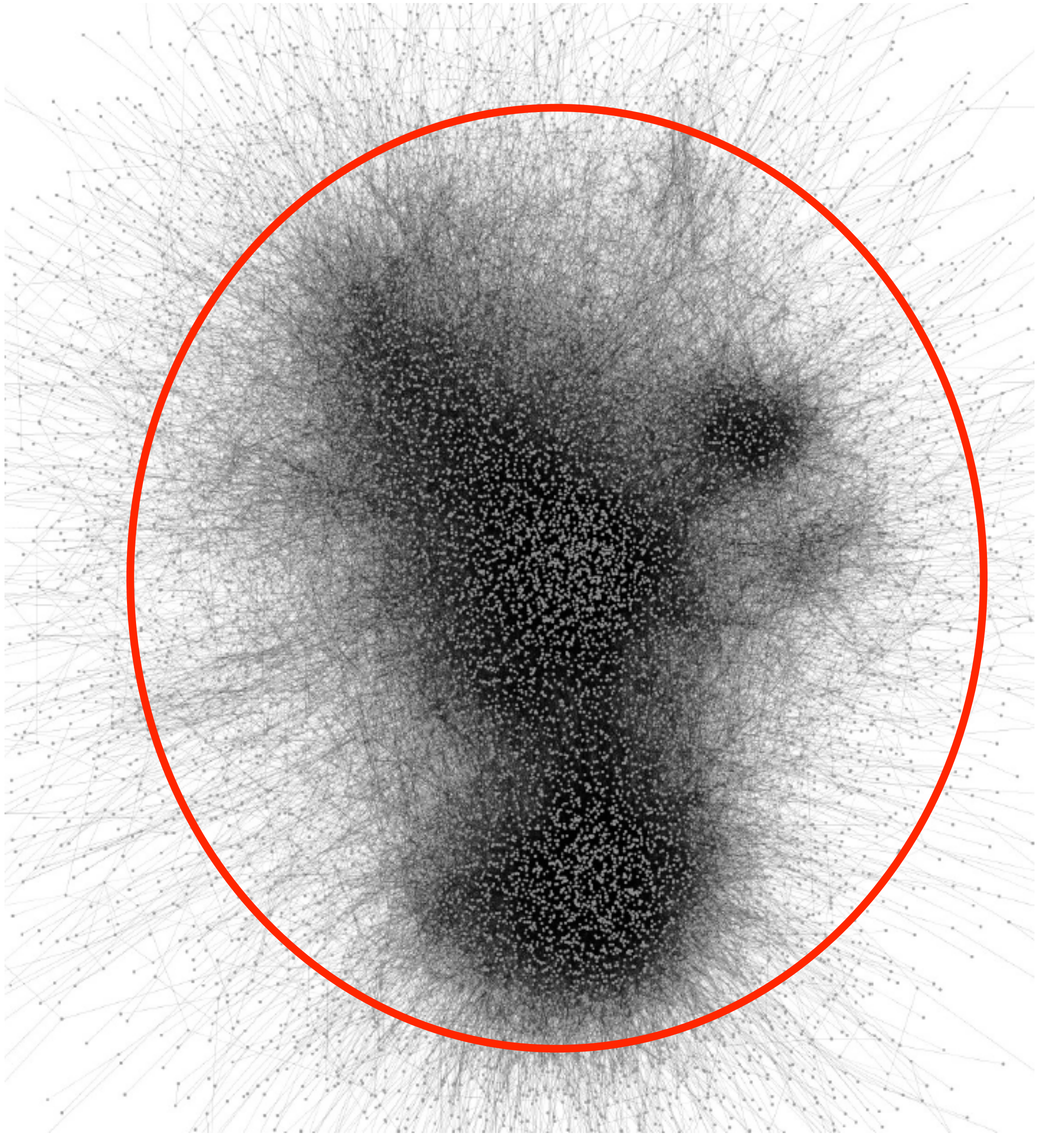
Communities



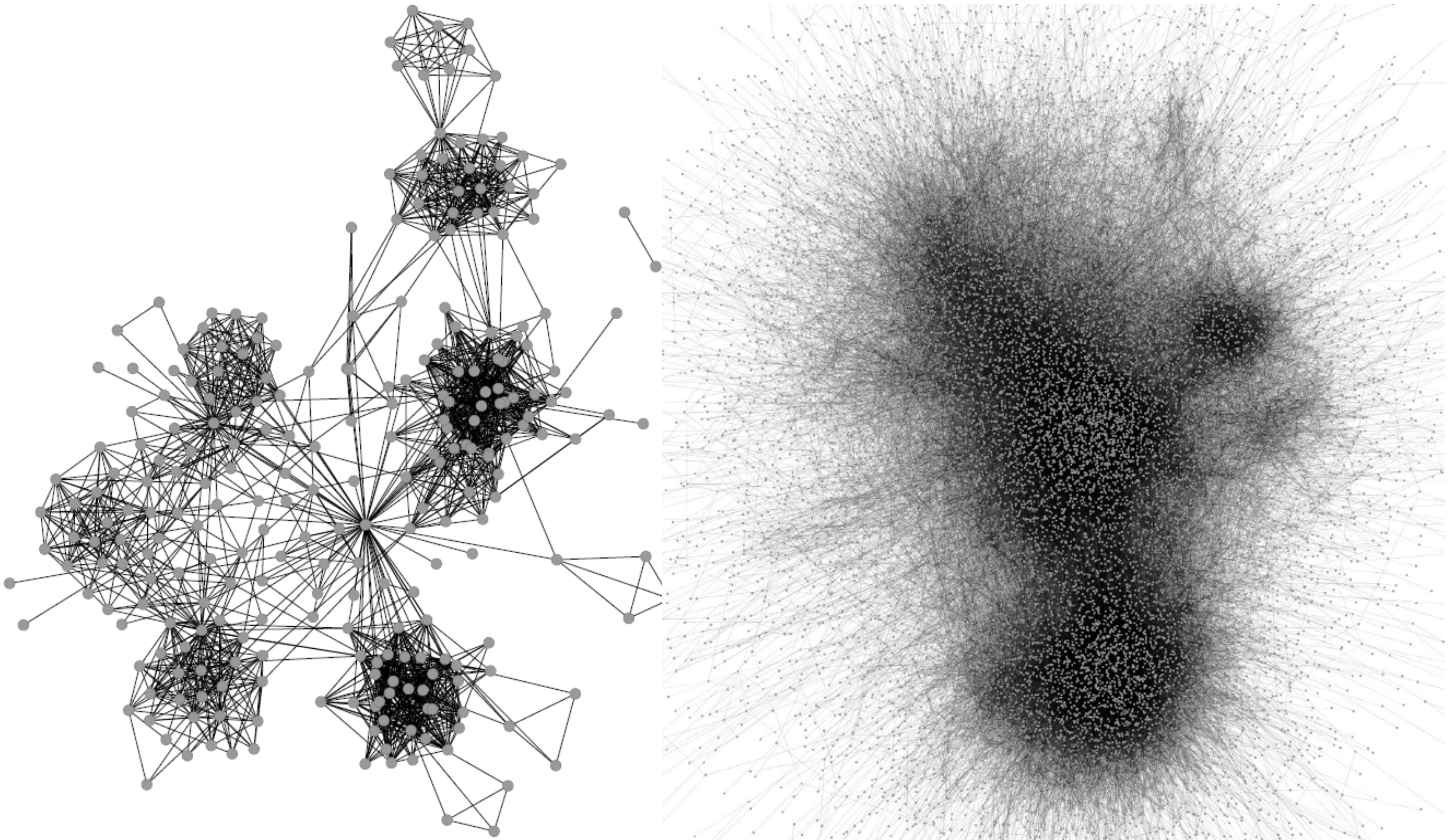




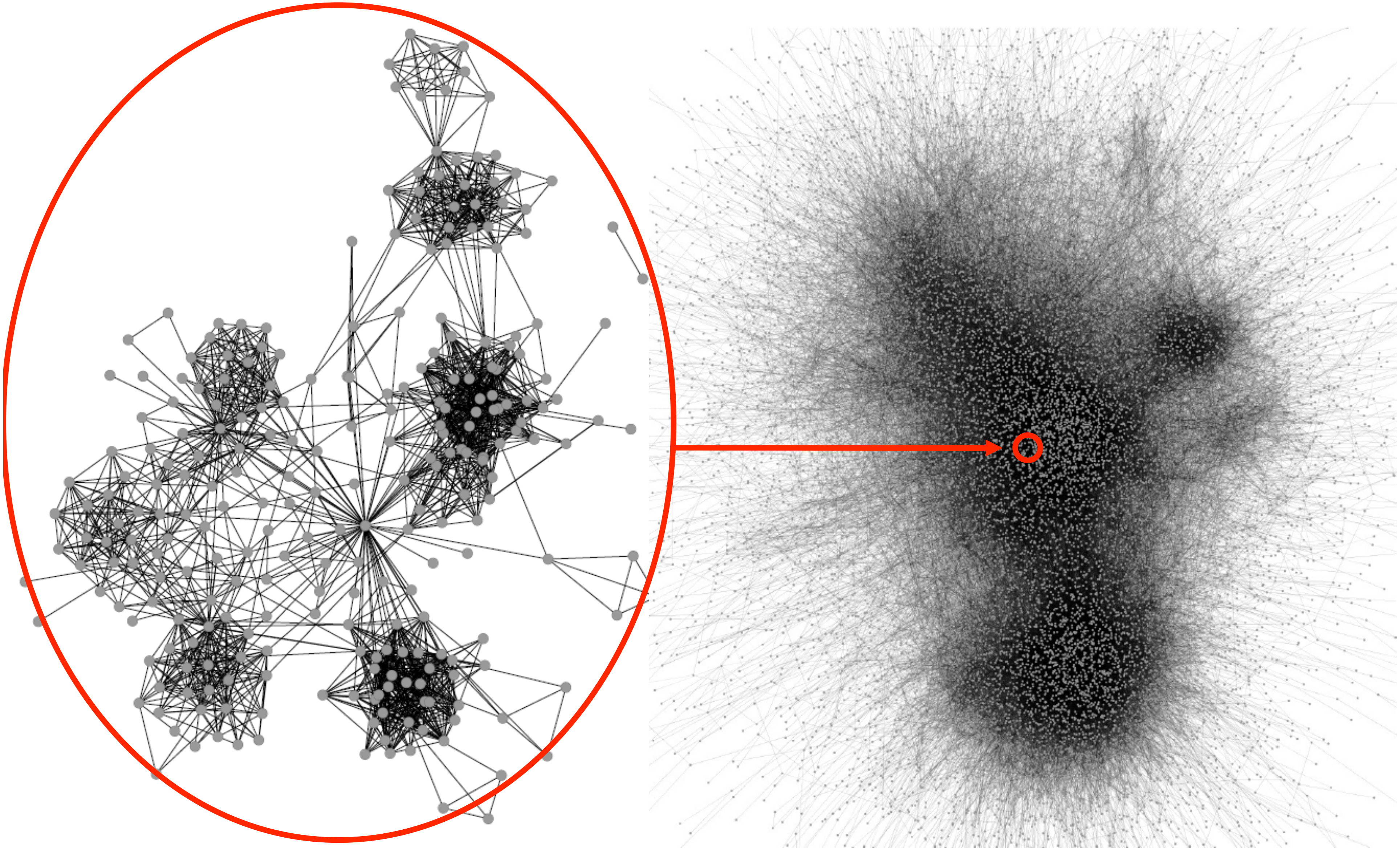
?



Are these two different networks?



No!



DEMON
A Local-first Discovery Method For Overlapping Communities

Giulio Rossetti[1,2] ,Michele Coscia[3], Fosca Giannotti[2], Dino Pedreschi[1,2]

[1] Computer Science Dep., University of Pisa, Italy

[2] ISTI - CNR KDDLab, Pisa, Italy

[3] Harvard Kennedy School, Cambridge, MA, US

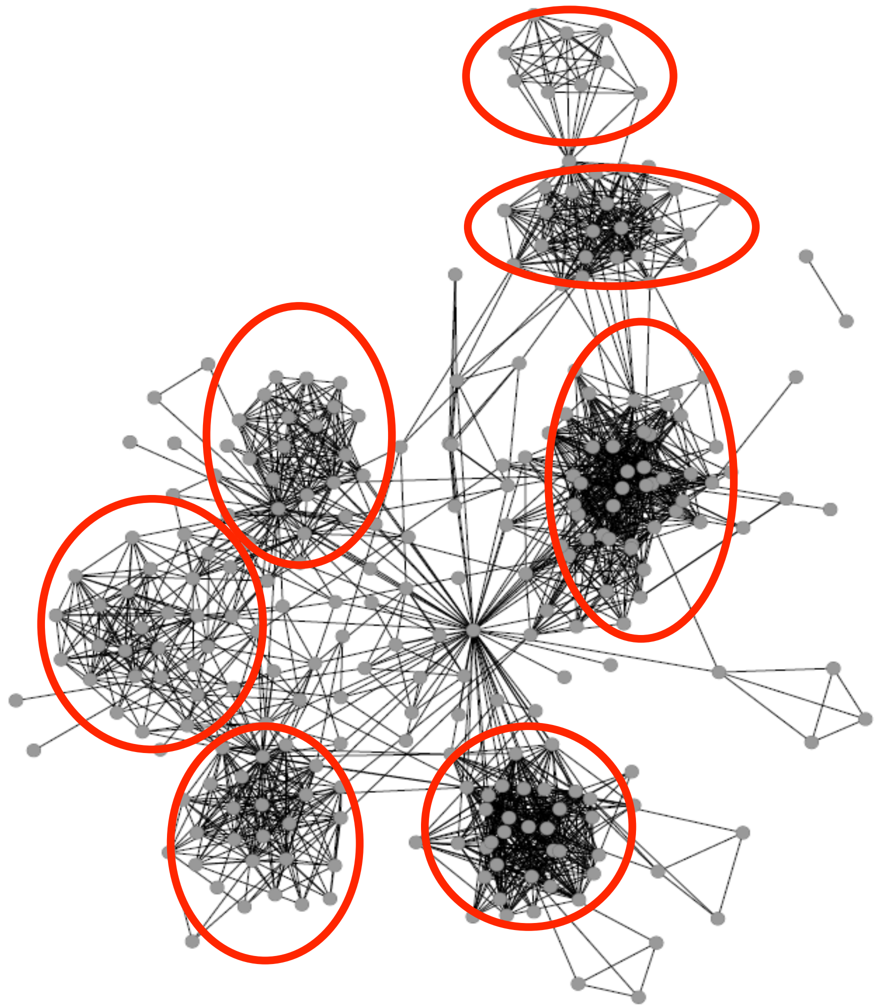
DEMON

Democratic Estimate of the Modular Organization of a Network

Communities in (Social) Networks

- Communities can be seen as the basic bricks of a (social) network

- In simple, small, networks it is easy identify them by looking at the structure..



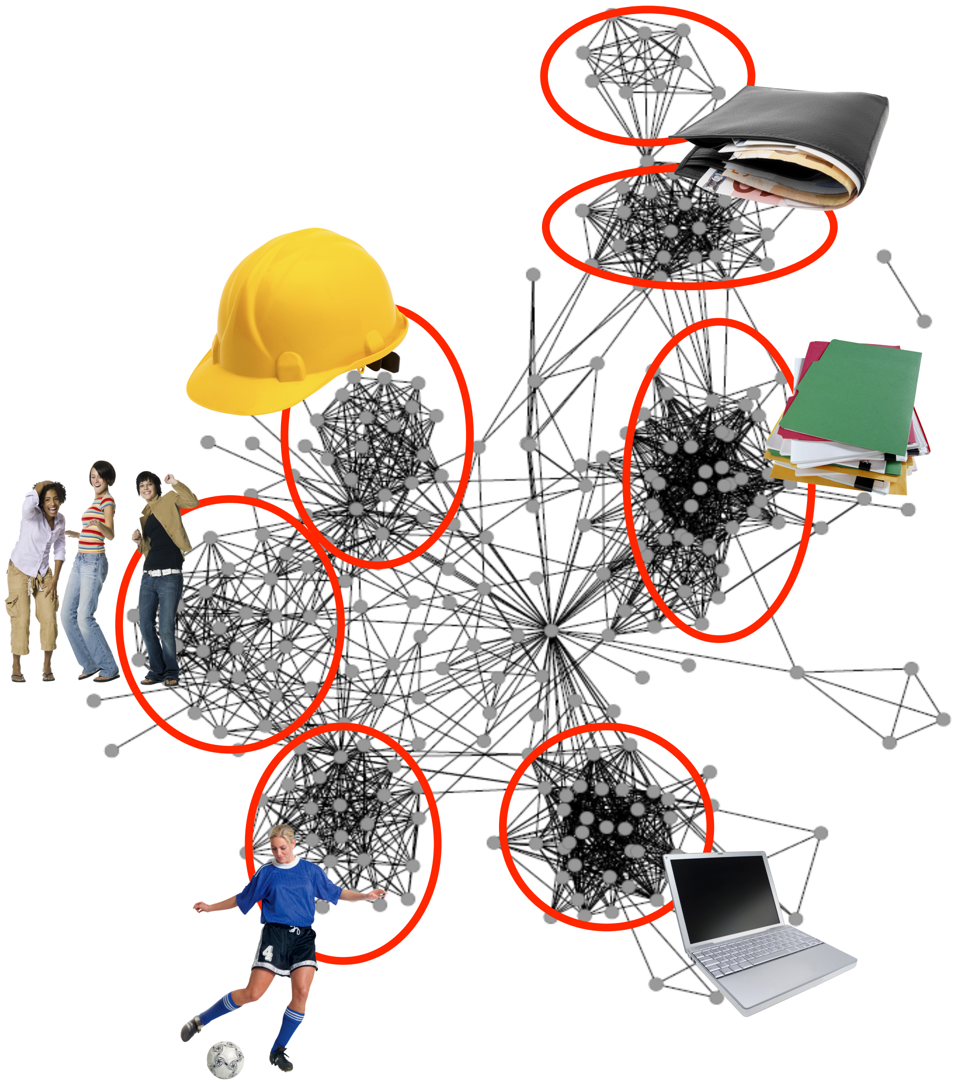
Reducing the complexity

Real Networks are Complex Objects

Can we make them “simpler”?

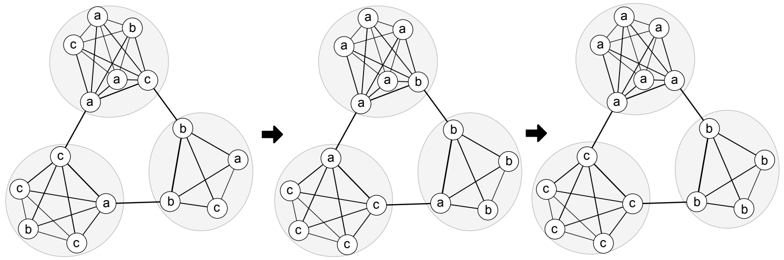
Ego-Networks

*(*networks builded upon a focal node , the "*ego*”, and the nodes to whom *ego* is directly connected to plus the ties, if any, among the alters)



DEMON Algorithm

- For each node n:
  1. Extract the Ego Network of n
  2. Remove n from the Ego Network
  3. Perform a Label Propagation[1]
  4. Insert n in each community found
  5. Update the raw community set C

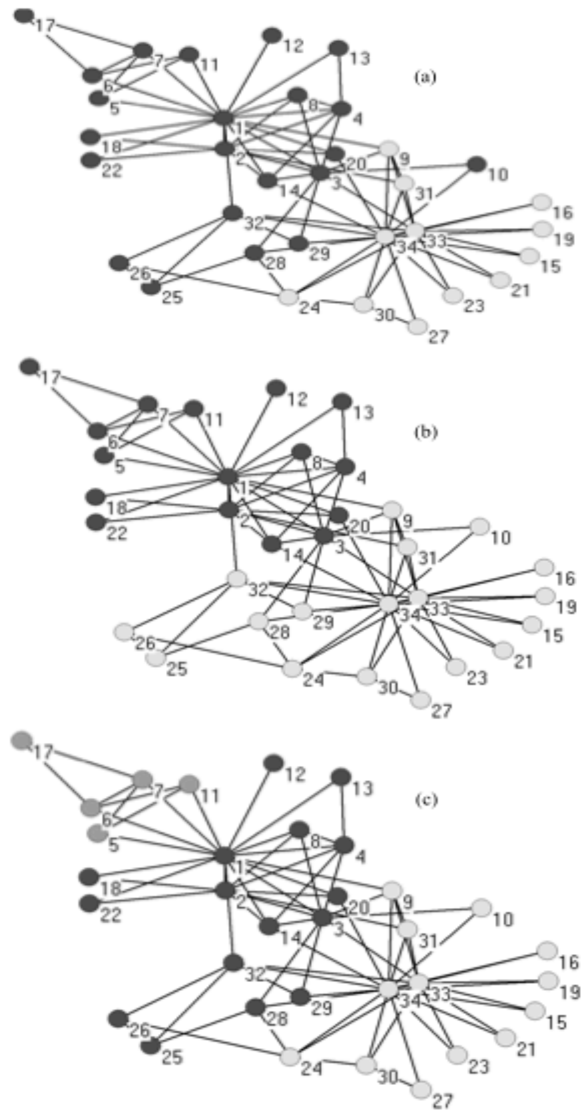


- For each raw community c in C
  1. Merge with "similar" ones in the set (given a threshold)
     (i.e. merge iff at most the ε% of the smaller one is not included in the bigger one)

[1] Usha N. Raghavan, R´eka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E

Label Propagation – The idea

- Each node has an unique label (i.e. its id)
- In the first (setup) iteration each node, with probability α, change its label to one of the labels of its neighbors;
- At each subsequent iteration each node adopt as label the one shared *(at the end of the previous iteration)* by the majority of its neighbors;
- We iterate untill consensus is reached.



DEMON - Two nice properties

- **Incrementality**:

  Given a graph G, an initial set of communities C and an incremental update ΔG consisting of new nodes and new edges added to G, where ΔG contains the entire ego networks of all new nodes and of all the preexisting nodes reached by new links, then

$$DEMON(\Delta G \cup G, C) = DEMON(\Delta G, DEMON(G, C))$$

- **Compositionality**:

  Consider any partition of a graph G into two subgraphs G1, G2 such that, for any node v of G, the entire ego network of v in G is fully contained either in G1 or G2. Then, given an initial set of communities C:

$$DEMON(G_1 \cup G_2, C) = Max(DEMON(G_1, C), DEMON(G_2, C))$$

Those property makes the algorithm highly parallelizable: it can run independently on different fragments of the overall network with a relatively small combination work

DEMON @ Work

DEMON was successfully applied to different networks and its communities were validated against their semantics

Social Networks

- Skype, Facebook, Twitter, Last.fm, 20lines

Colocation Networks

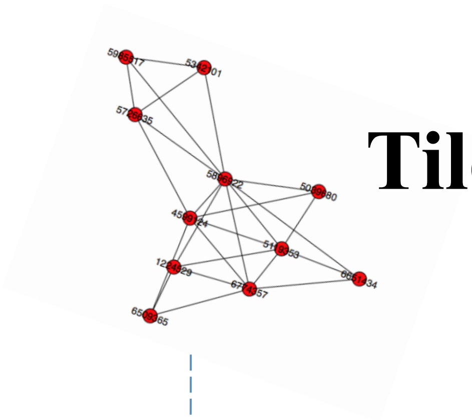
- Foursquare

Collaboration Networks

- DBLP, IMDb, US Congress

Product Networks

- Amazon

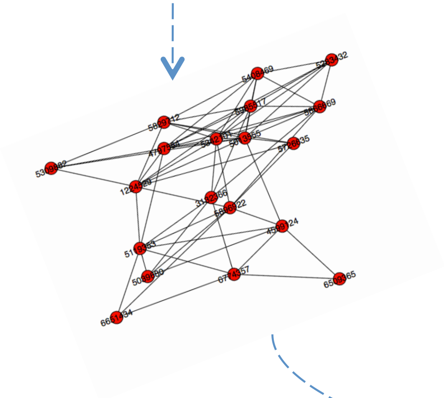


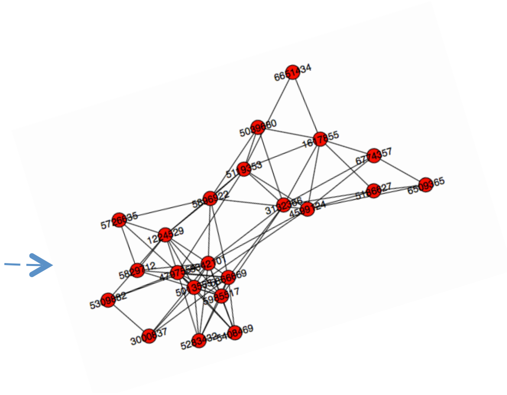
Tiles: evolutionary community discovery

Giulio Rossetti[1,2], Luca Pappalardo[1,2], Fosca Giannotti[2], Dino Pedreschi[1,2]

[1] Computer Science Dep., University of Pisa, Italy {rossetti,pedre}@di.unipi.it

[2] ISTI - CNR KDDLab, Pisa, Italy {fosca.giannotti, giulio.rossetti}@isti.cnr.it



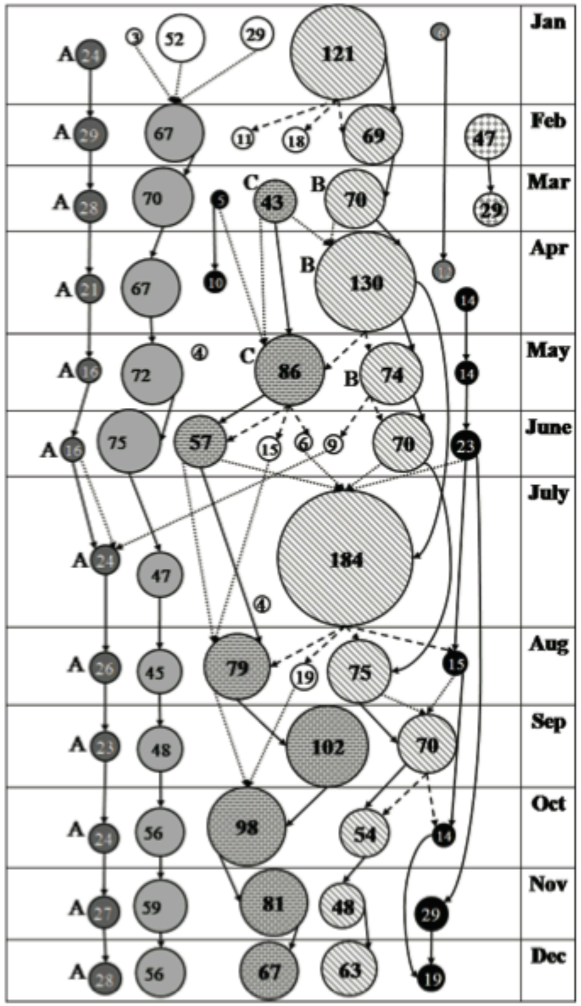


Dynamic Networks

- The majority of data mining problems on network have been formulated to fit static scenarios
  - Community Discovery, Link Prediction, Frequent Pattern Mining

- Evolution has been analyzed almost only through *temporal discretization*…
  - Separate analysis of chronologically ordered snapshot of the same network

- … and\or through *temporal "aggregation"*
  - i.e. producing a single weighted graph
    (edge weighted w.r.t. their number of presence, frequency…)



Are we missing something?

Real world networks evolve quickly:

– Social interactions
– Buyer-seller
– Stock-exchanges
– …

In these scenarios a QSSA (Quasi Steady State Assumption) rarely holds:

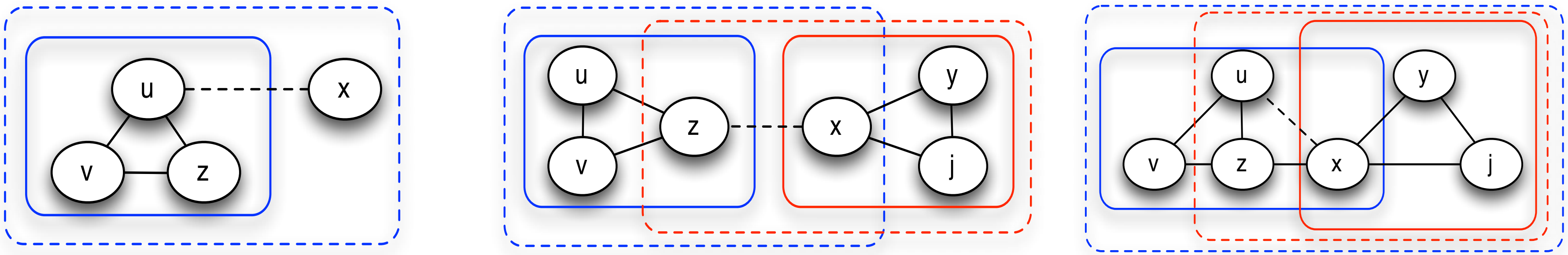
– Network cannot be *"frozen in time"*
  - Nodes and edges rise and fall producing perturbation on the whole topology
– The reduction to static scenarios trough temporal discretization is not always a good idea
  - How can we chose the temporal threshold?
  - To what extent can we trust the obtained results?

The Idea… TILES

Temporal Interaction a Local Edge Strategy

- Imagined for social "interaction" networks
  - Multiple time stamped interactions between the same couple of nodes

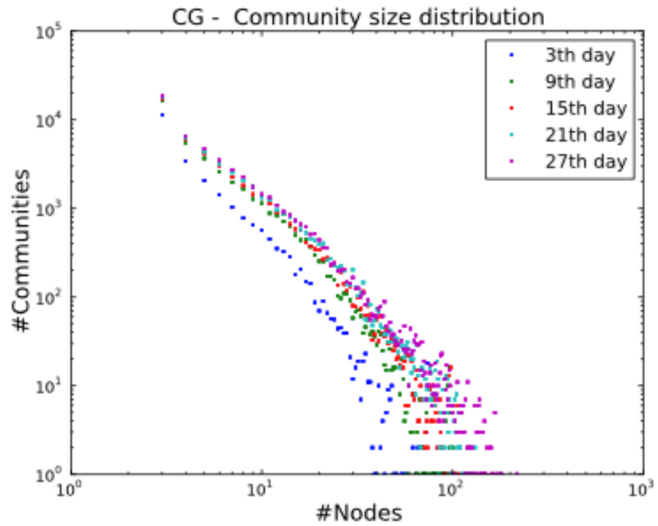
- Domino Effect
  - TILES *incrementally updates* community memberships when a new interaction take place (it operates on an interaction stream)
  - A single parameter: interaction time to live (TTL) that regulates interaction vanishing (non monotonic network growth)
- Output
  - Multiple time stamped *observation* of overlapping communities

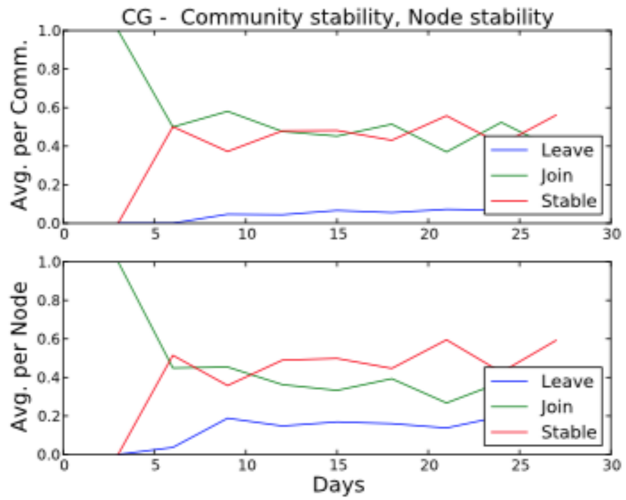


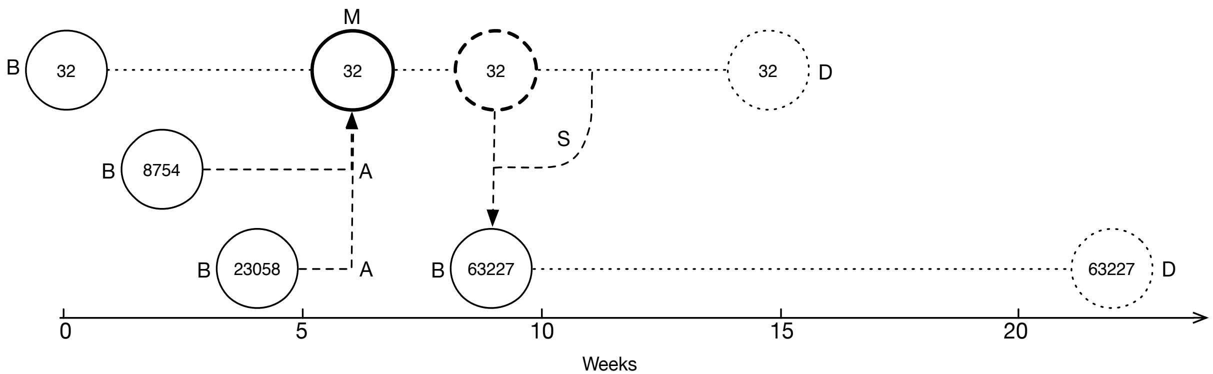
Tiles Community Insights

Experiments real interaction networks show that:

- Community size distribution and overlap distribution are long tailed
- Community stability vary w.r.t. its topology
- TTL affect community life-cycle
  (birth, split, merge, death events)
- Smaller and denser communities live longer than bigger and sparser ones







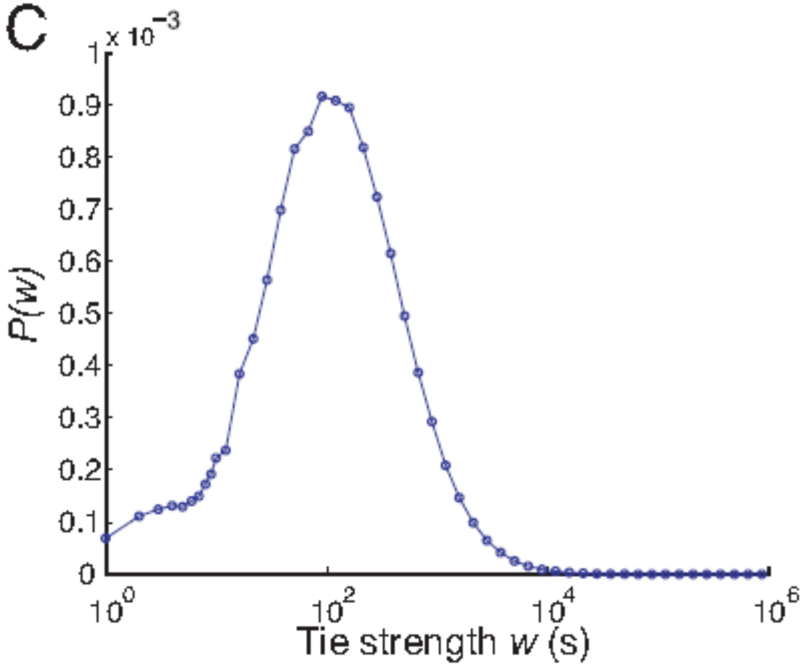
Community discovery

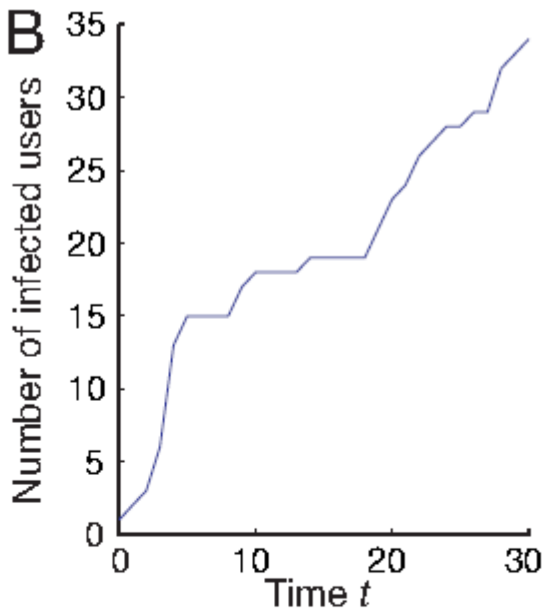
- Challenging task
- Many competing approaches
- Huge literature
- A recent survey:
  - Michele Coscia, Fosca Giannotti, Dino Pedreschi: A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining* 4(5): 512-546 (2011)

Diffusion and cascades

The strength of weak ties …

- For information **diffusion** (**spreading** of news and rumors on a social network)





The weakness of weak ties

- Diffusion of **innovation / adoption**

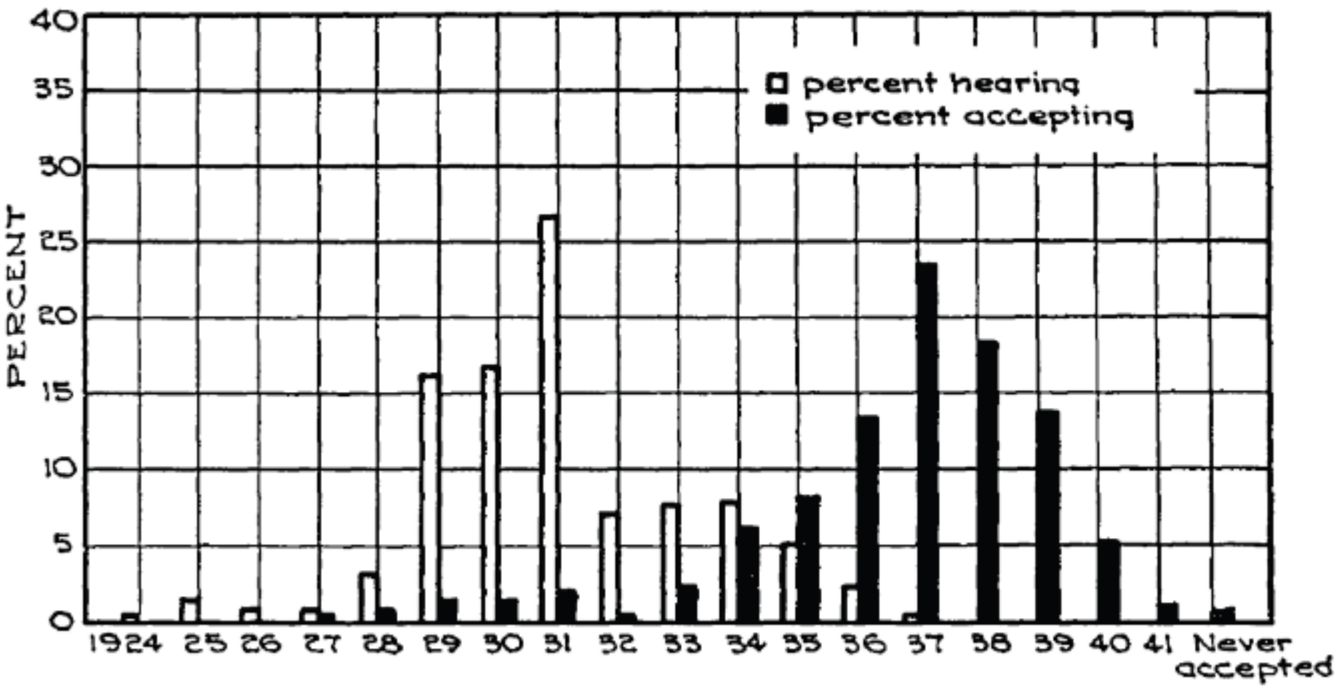
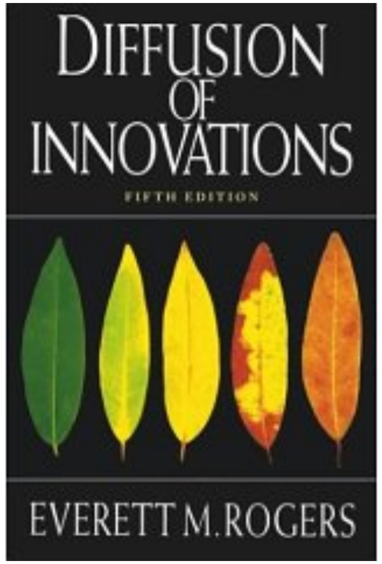
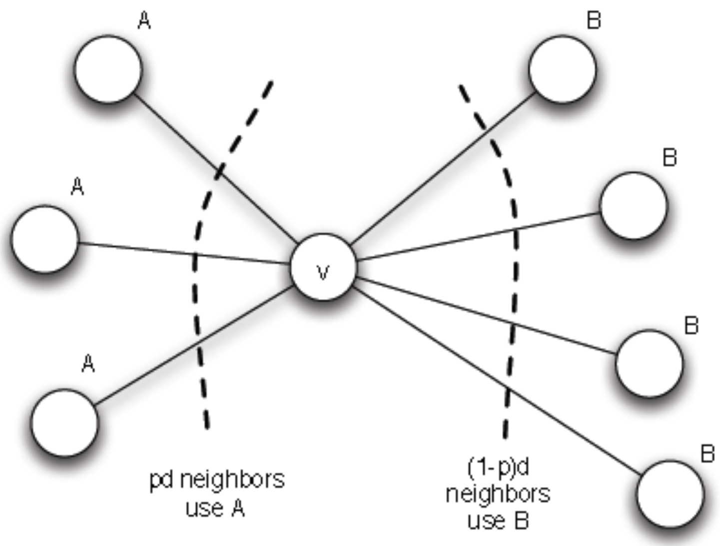
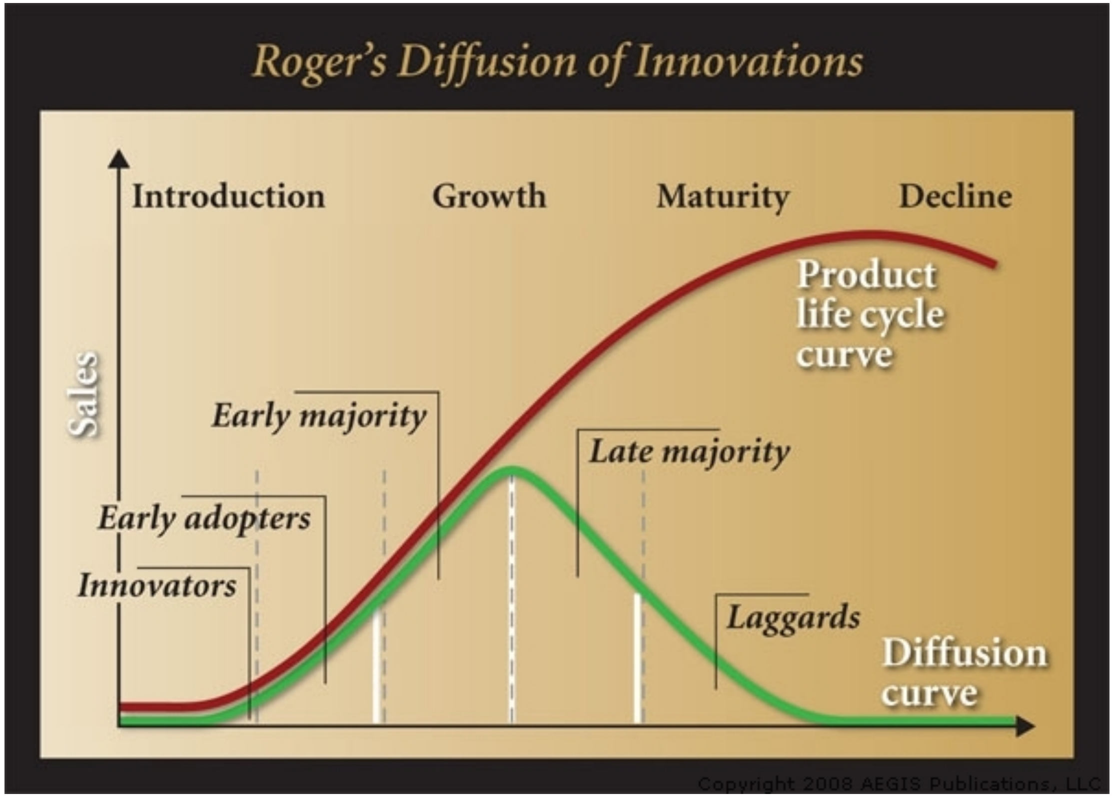


Figure 19.10: The years of first awareness and first adoption for hybrid seed corn in the Ryan-Gross study. (Image from [358].)

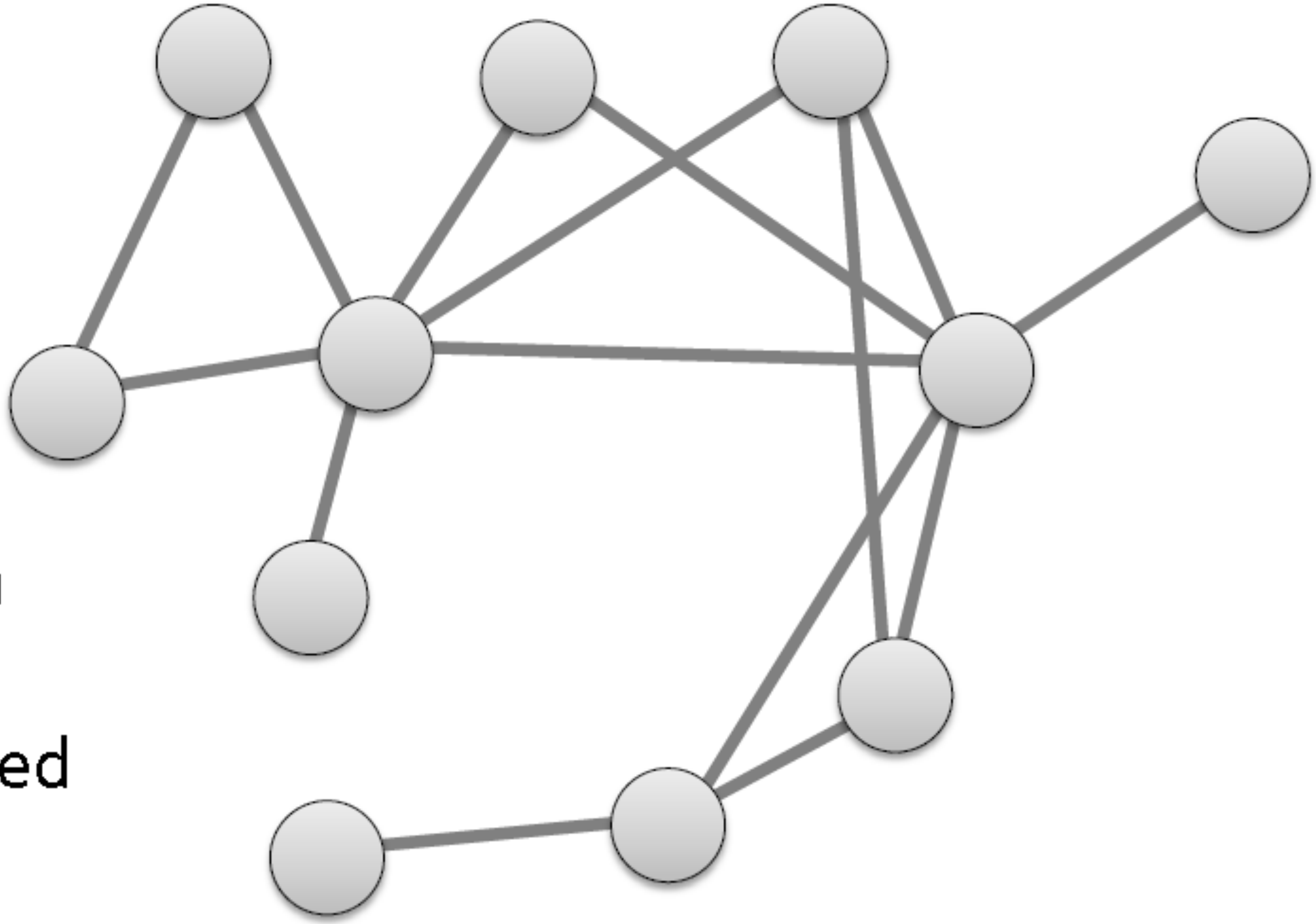
The strength of the strong ties for the





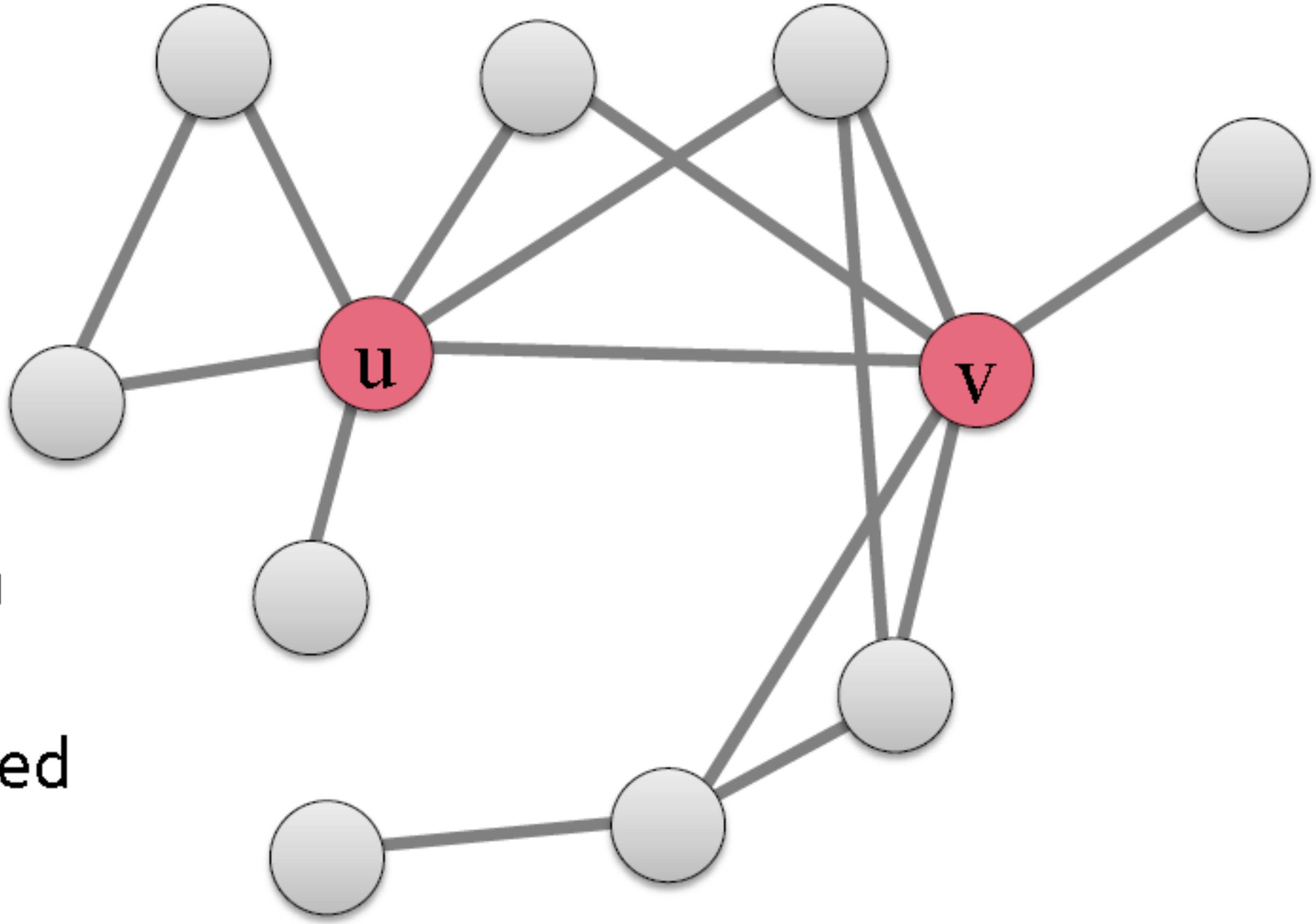


$$S = \{u, v\}$$



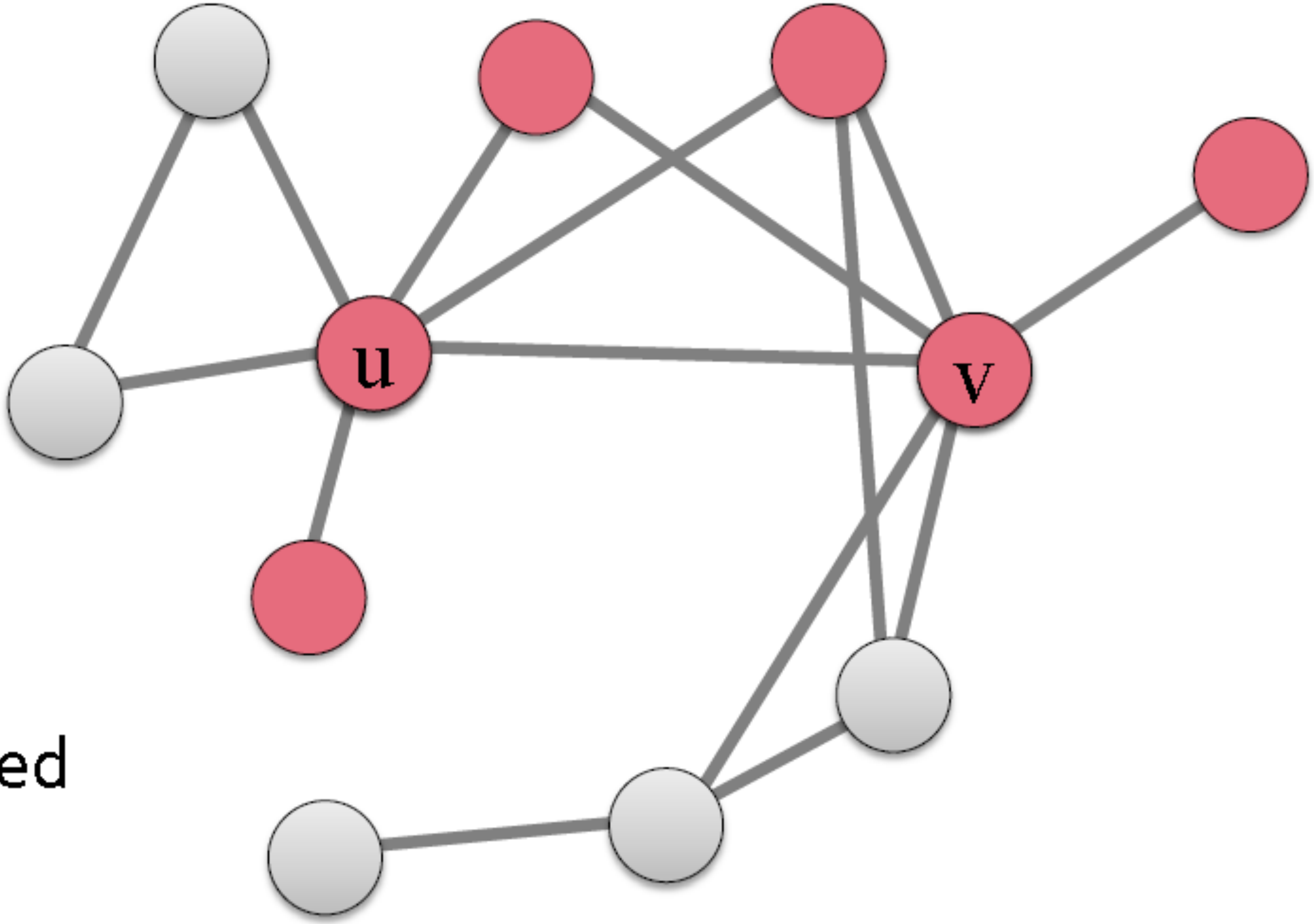
If more than 50% of my friends are red I'll be red

$$S = \{u, v\}$$



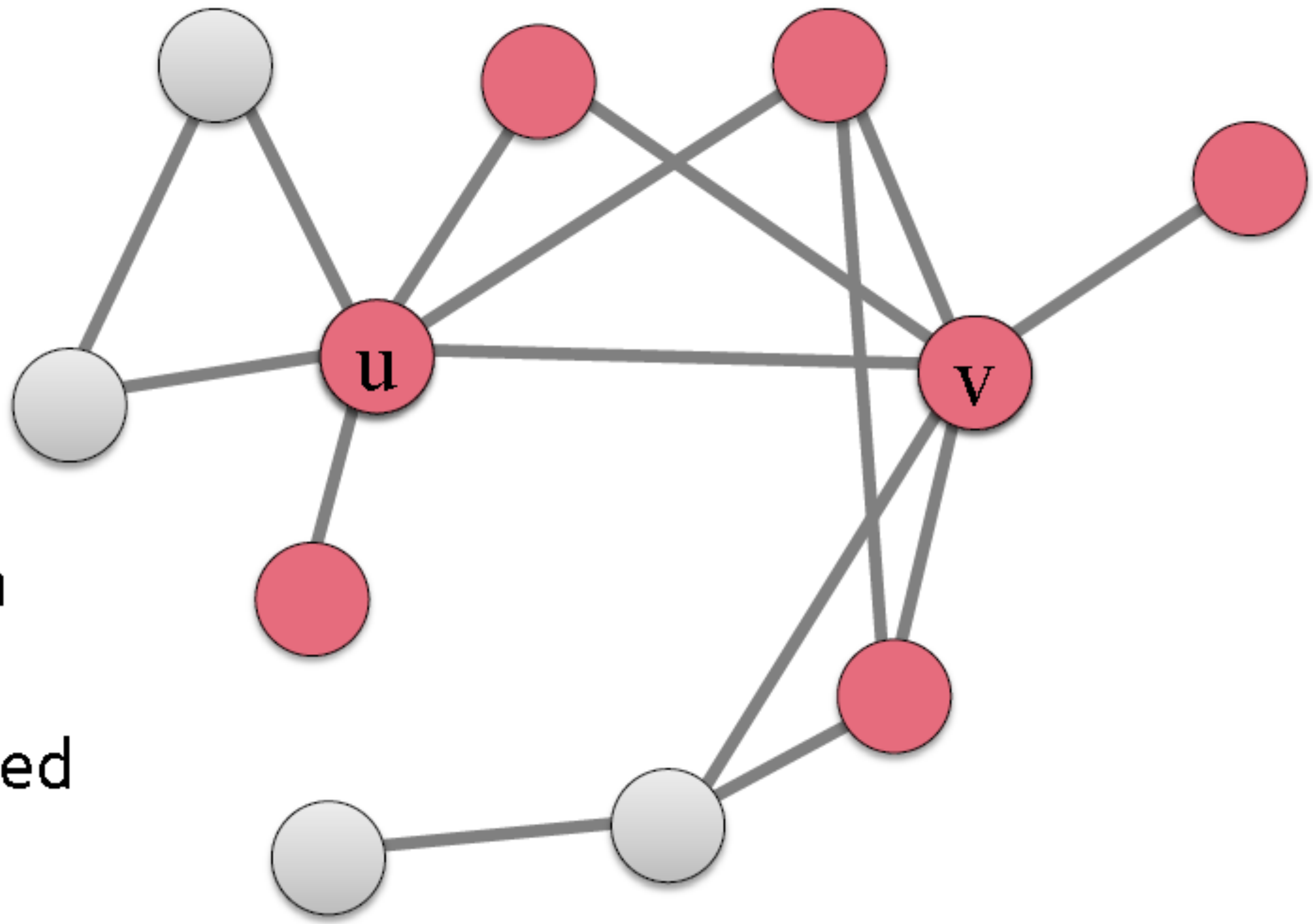
If more than 50% of my friends are red I'll be red

$$S = \{u, v\}$$



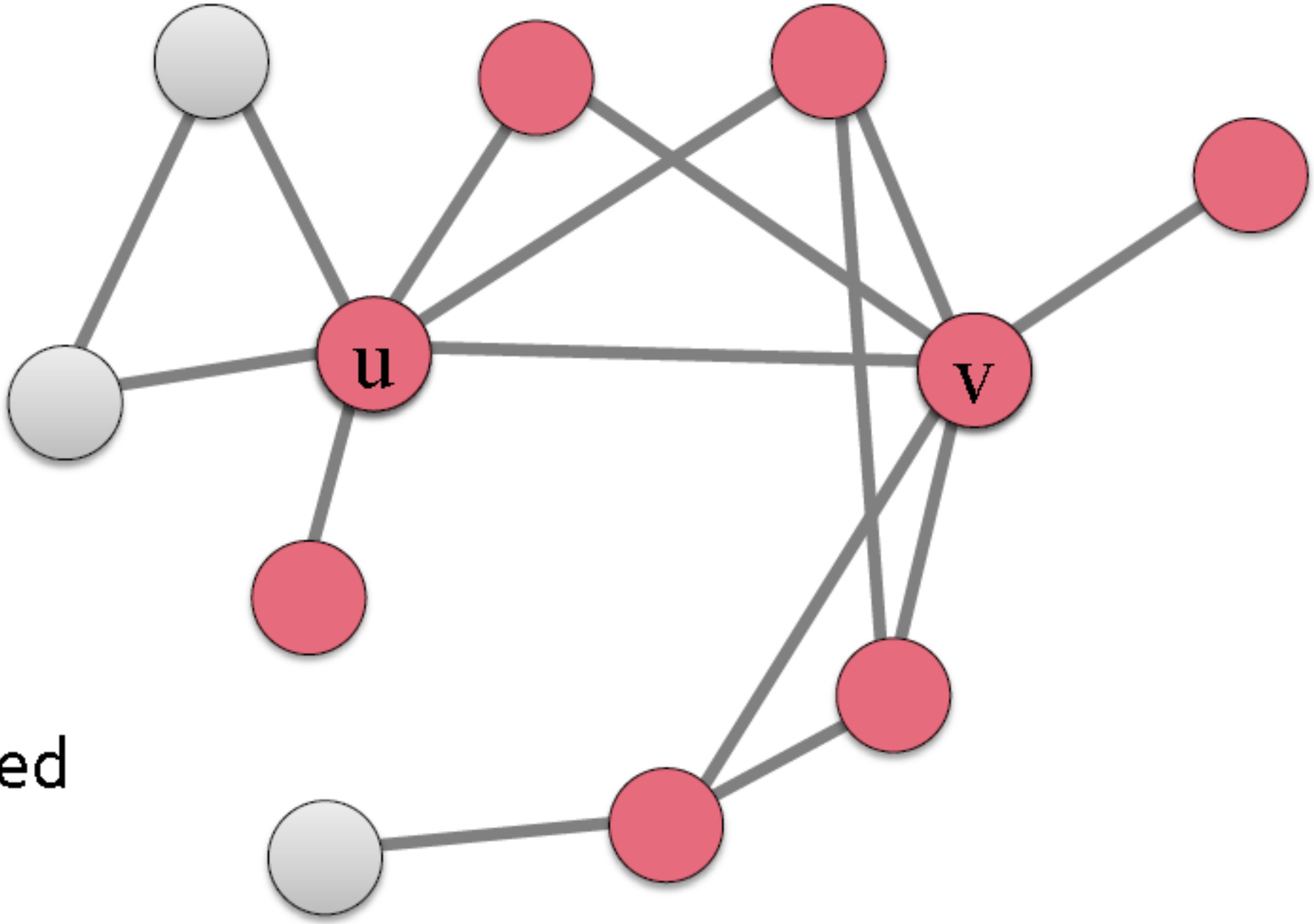
If more than 50% of my friends are red I'll be red

$$S = \{u, v\}$$



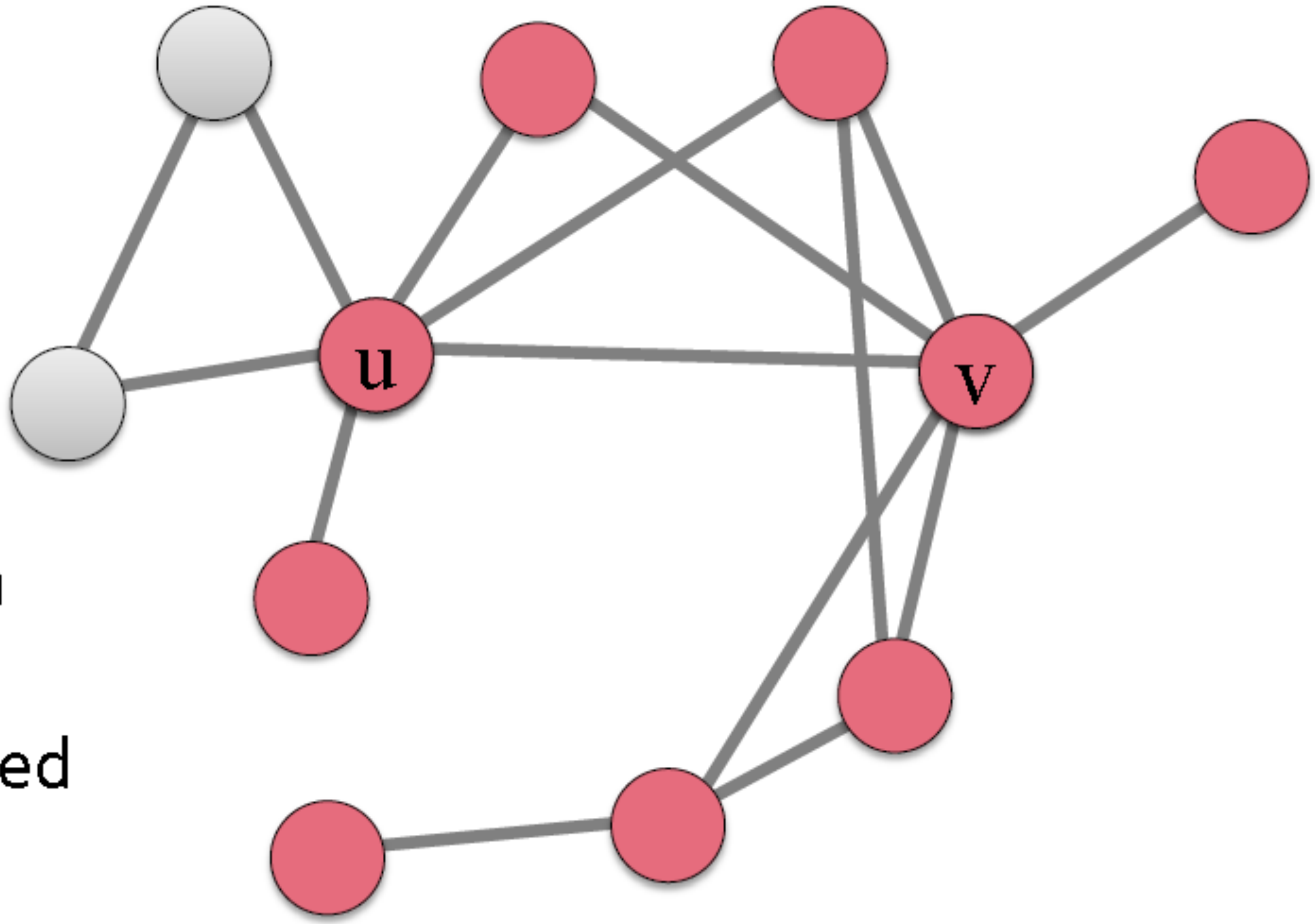
If more than 50% of my friends are red I'll be red

$S = \{u, v\}$



If more than 50% of my friends are red I'll be red

$$S = \{u, v\}$$

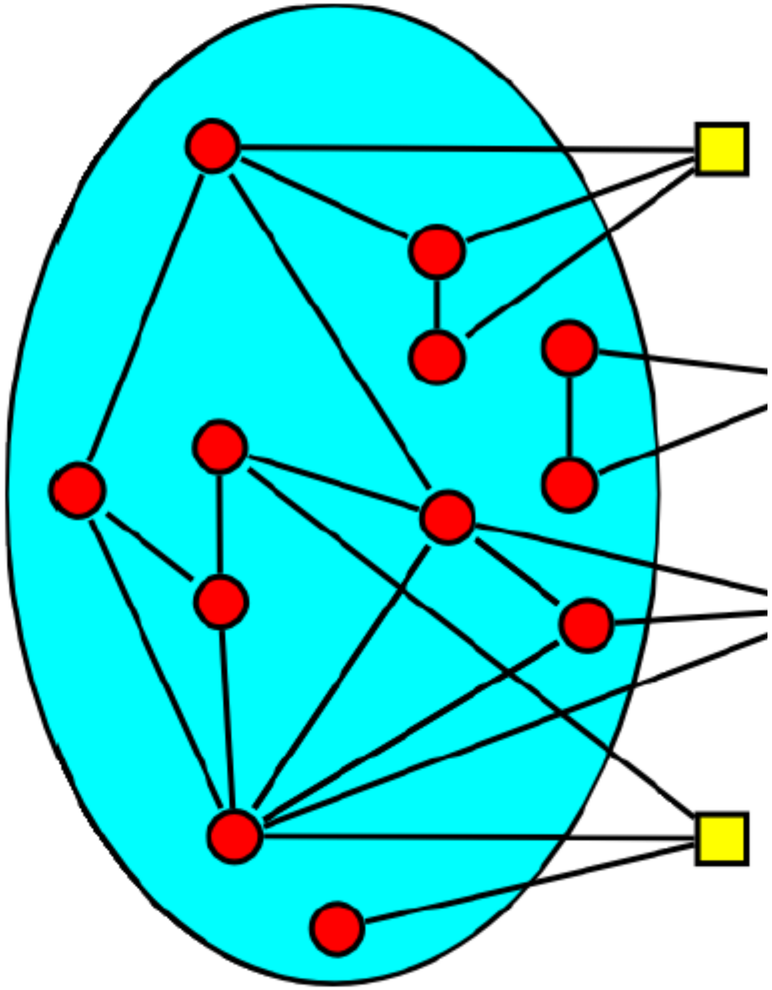


If more than 50% of my friends are red I'll be red

[Backstrom et al. KD

Adoption Curve: LiveJournal

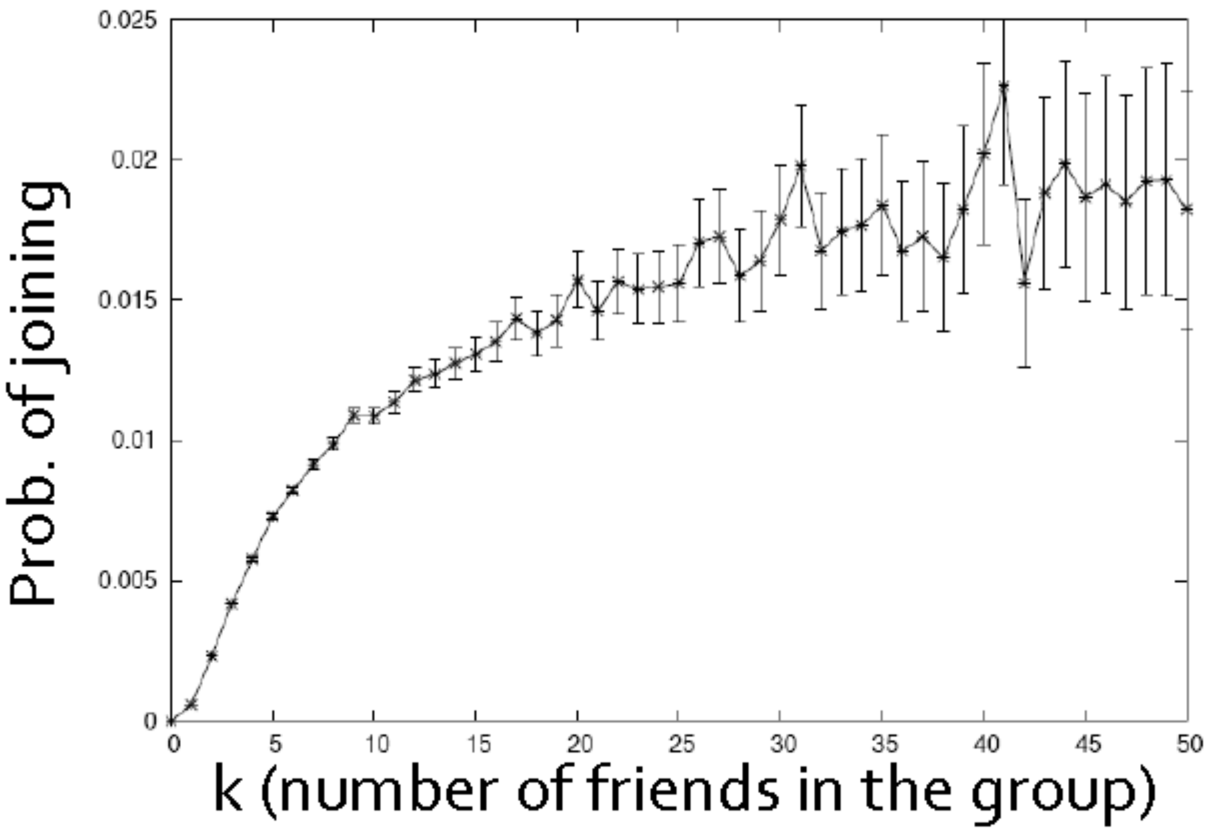
- **Group memberships spread over the network:**
  - Red circles represent existing group members
  - Yellow squares may join
- **Question:**
  - How does prob. of joining a group depend on the number of friends already in the group?



Adoption Curve: LiveJournal

[Backstrom et al., KDD '0

- **LiveJournal group membership**



[Leskovec et al., TWEB '07]

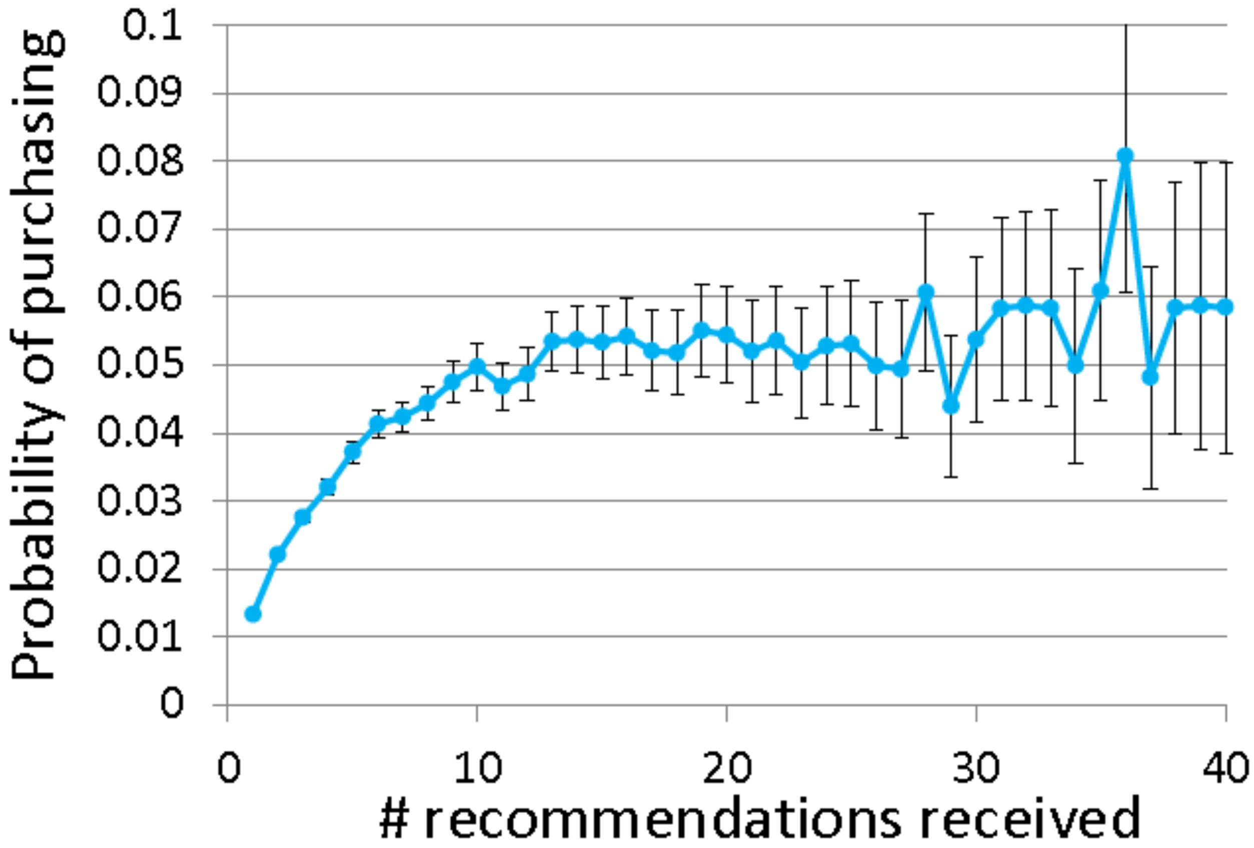
Diffusion in Viral Marketing

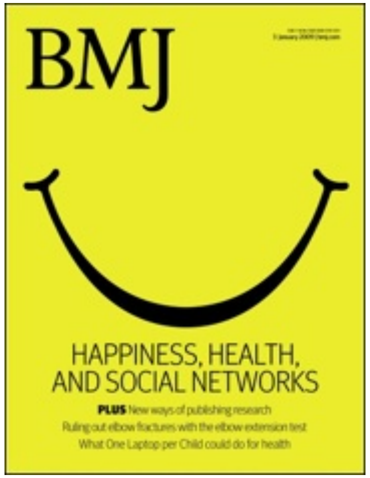
- **Senders and followers of recommendations receive discounts on products**

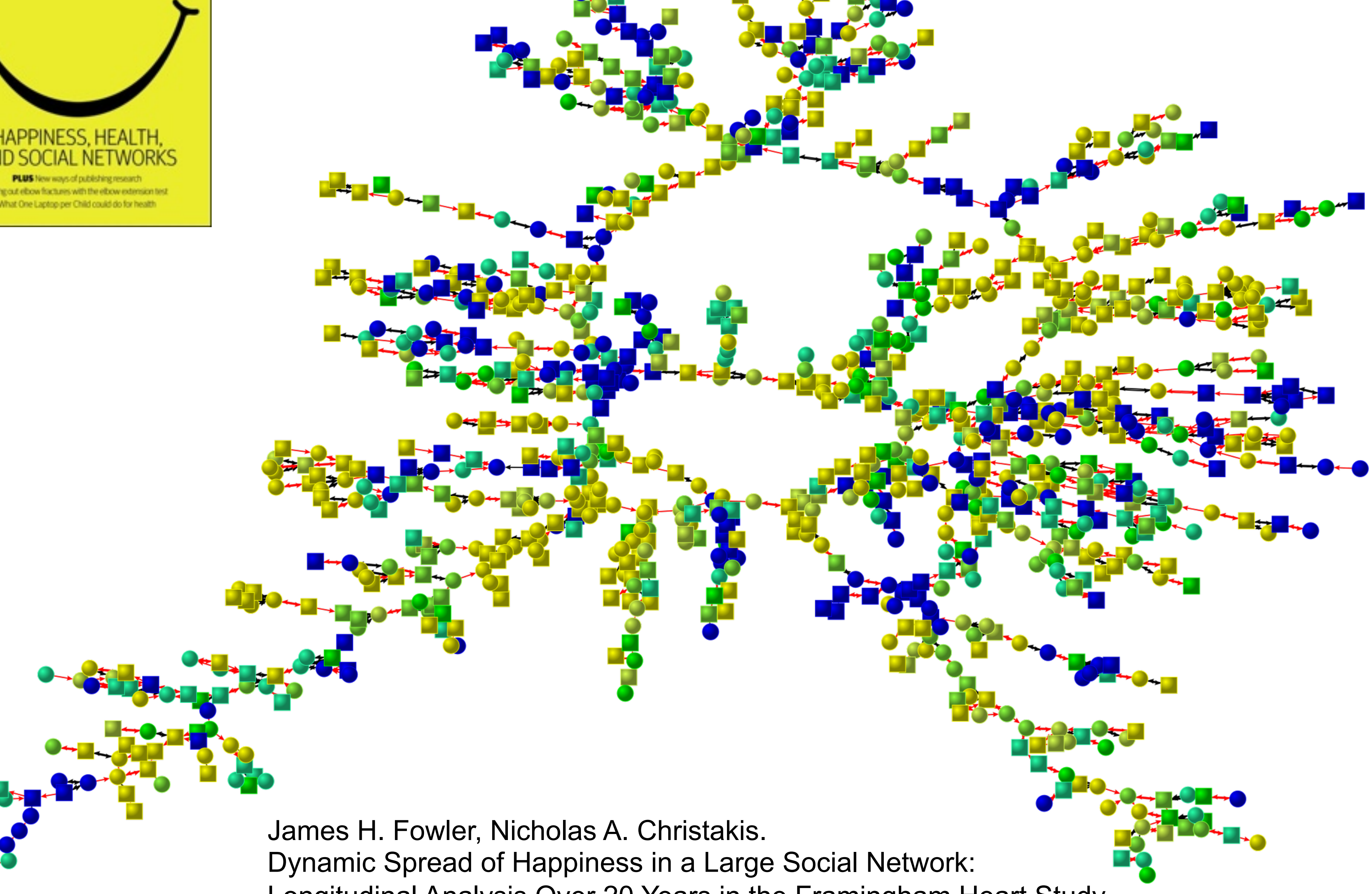
- **Data: Incentivized Viral Marketing program**
  - 16 million recommendations
  - 4 million people, 500k products

[Leskovec et al., TWEB '07]

Adoption Curve: Validation







James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

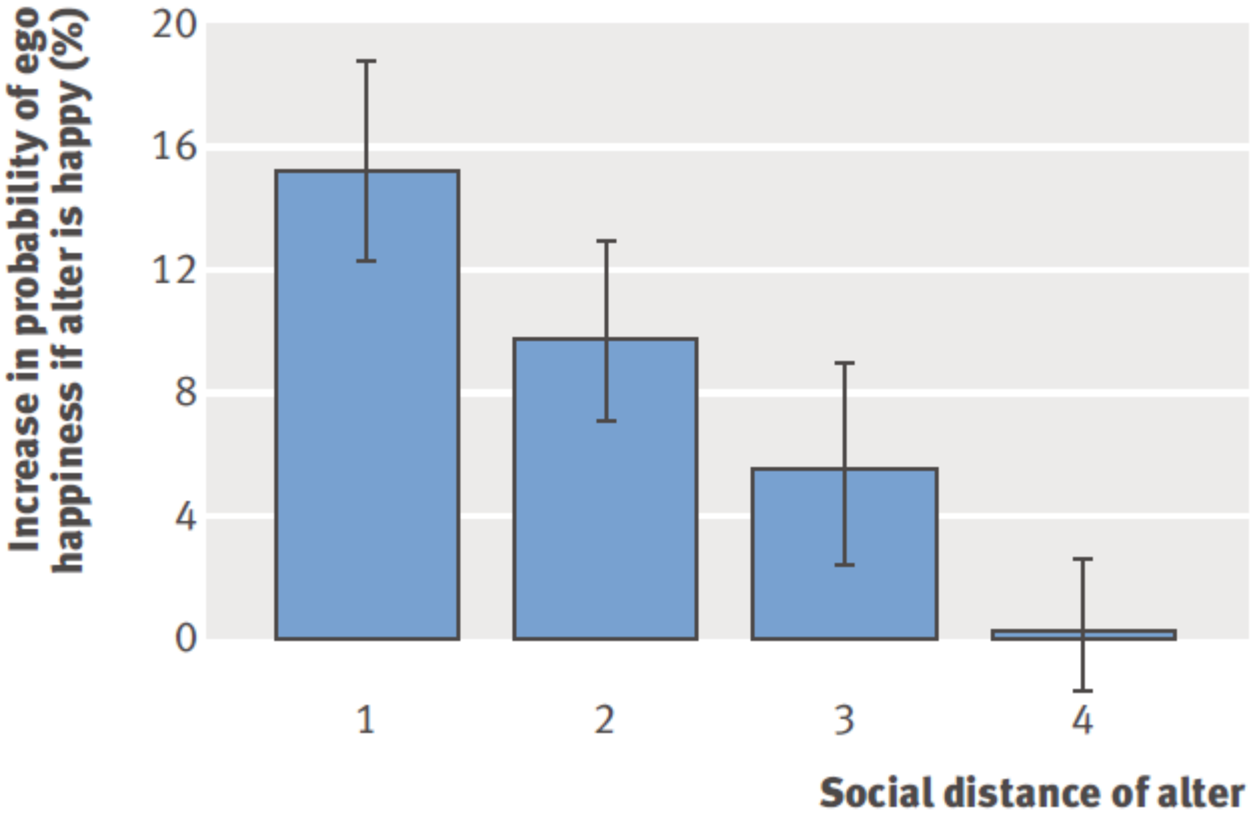
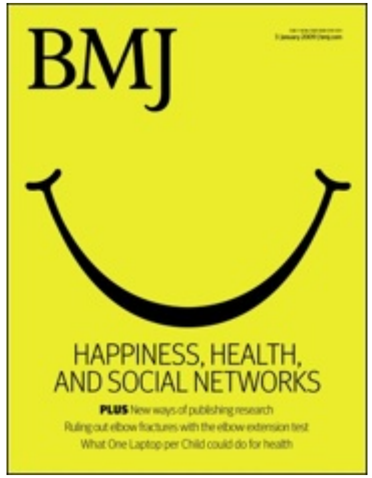


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >0 at social distances up to three degrees of separation, meaning that a person's happiness is associated with happiness of people up to three degrees removed from them in the network. Values derived by comparing conditional



Social influence
or
homophily?

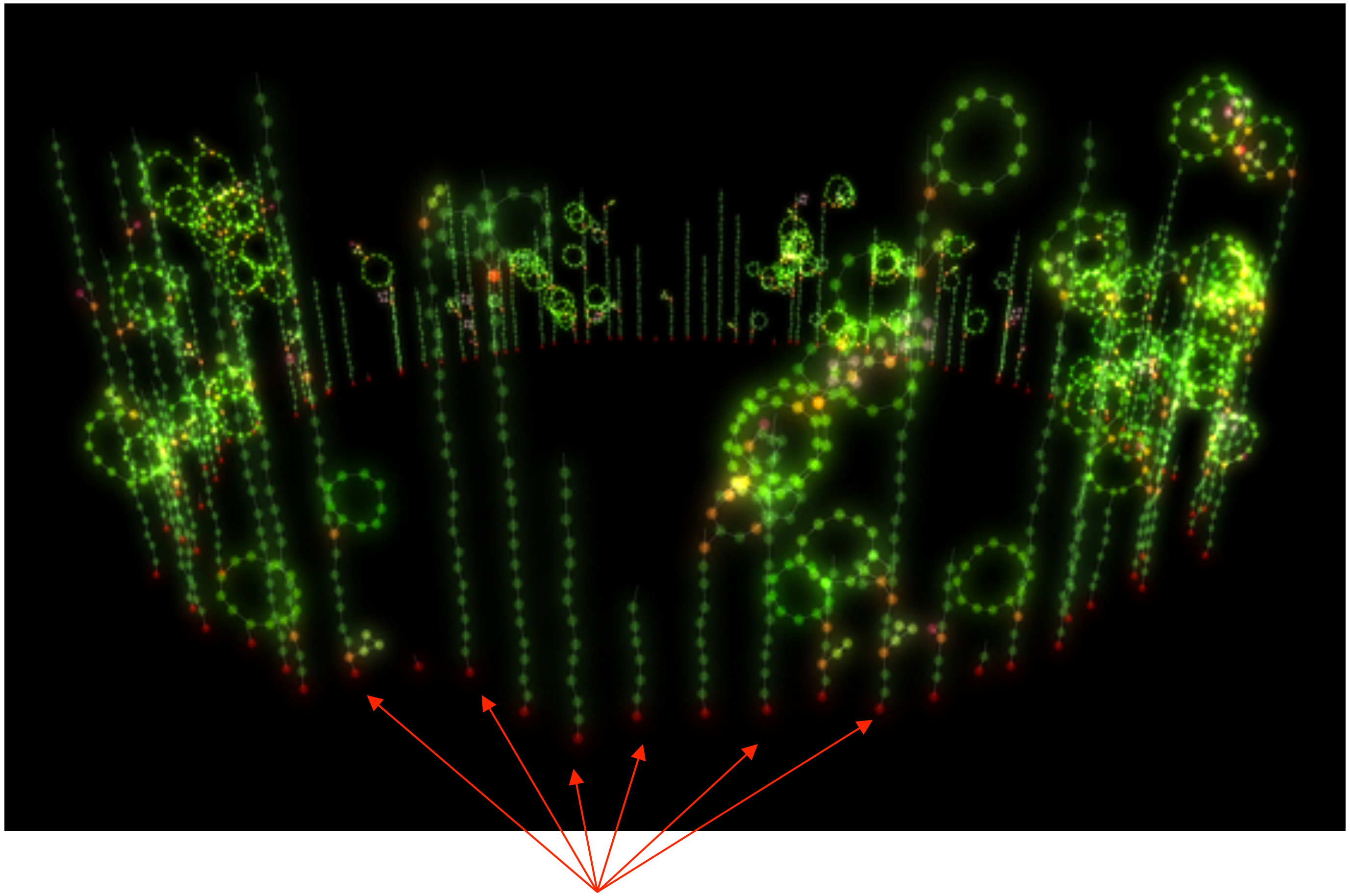
The Three Dimensions of Social Prominence

Diego Pennacchioli[2,3], Giulio Rossetti[1,2], Luca Pappalardo[1,2], Fosca Giannotti[2], Dino Pedreschi[1,2]

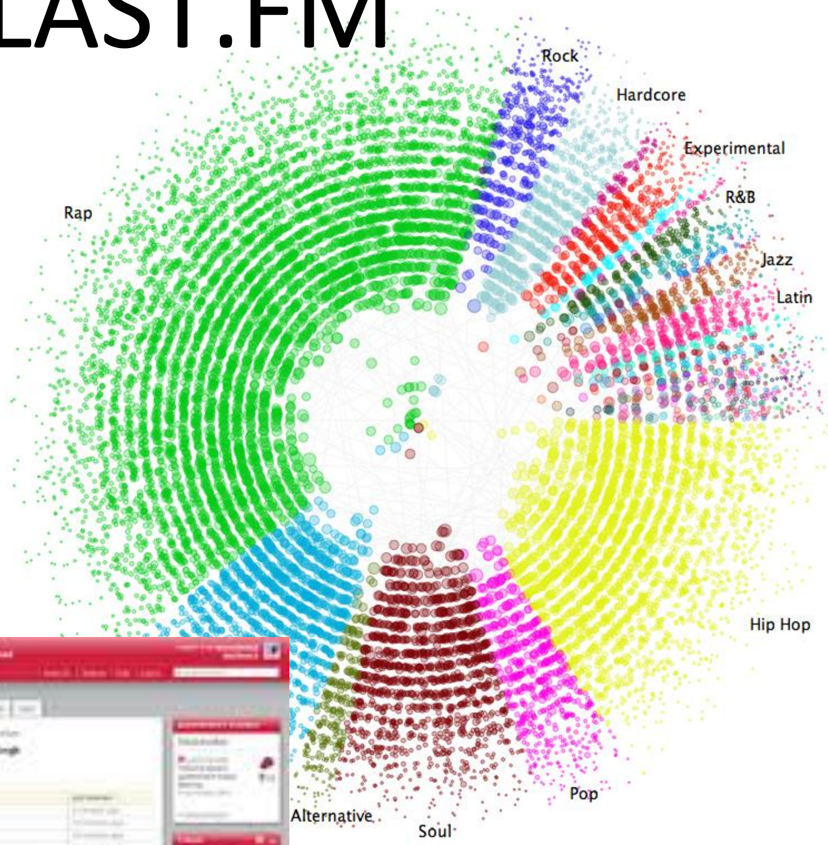
[1] Computer Science Dep., University of Pisa, Italy {rossetti,pedre}@di.unipi.it
[2] ISTI - CNR KDDLab, Pisa, Italy {fosca.giannotti, giulio.rossetti}@isti.cnr.it
[3]IMT Lucca, Italy

last.fm™
the social music revolution

Social Influence: Leaders

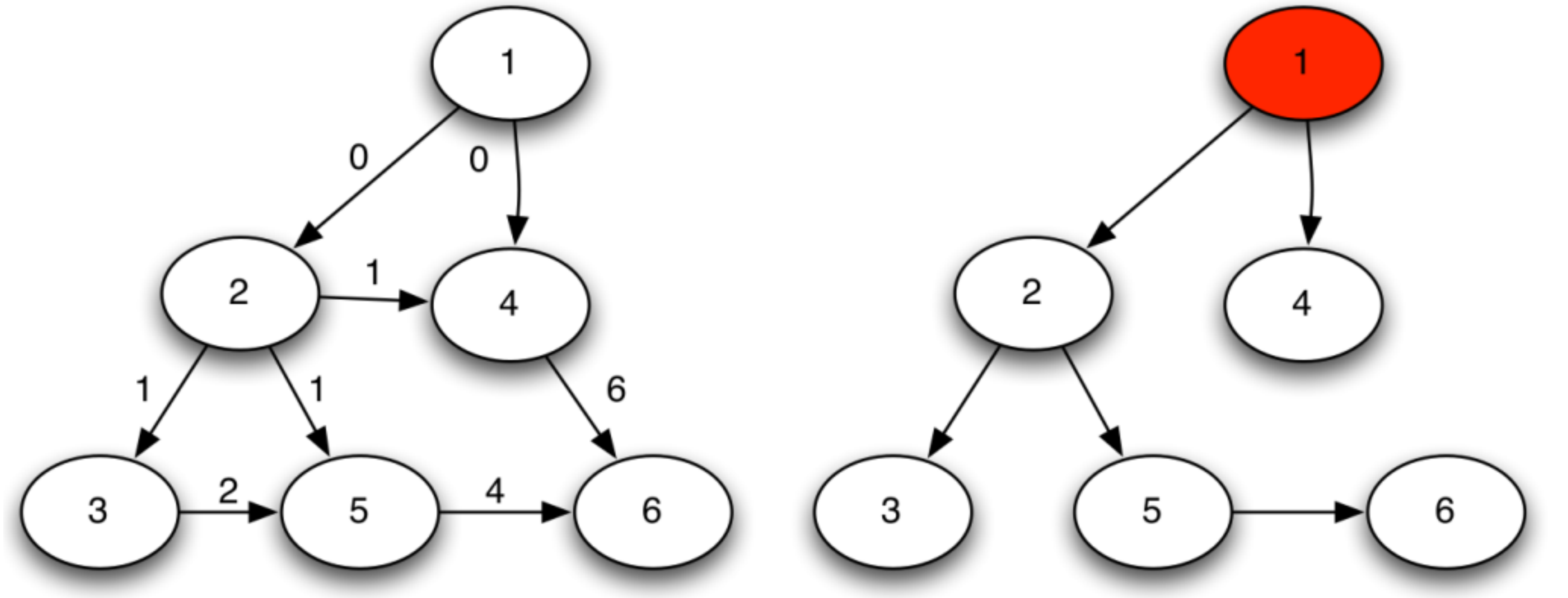


Chiediamo a LAST.FM

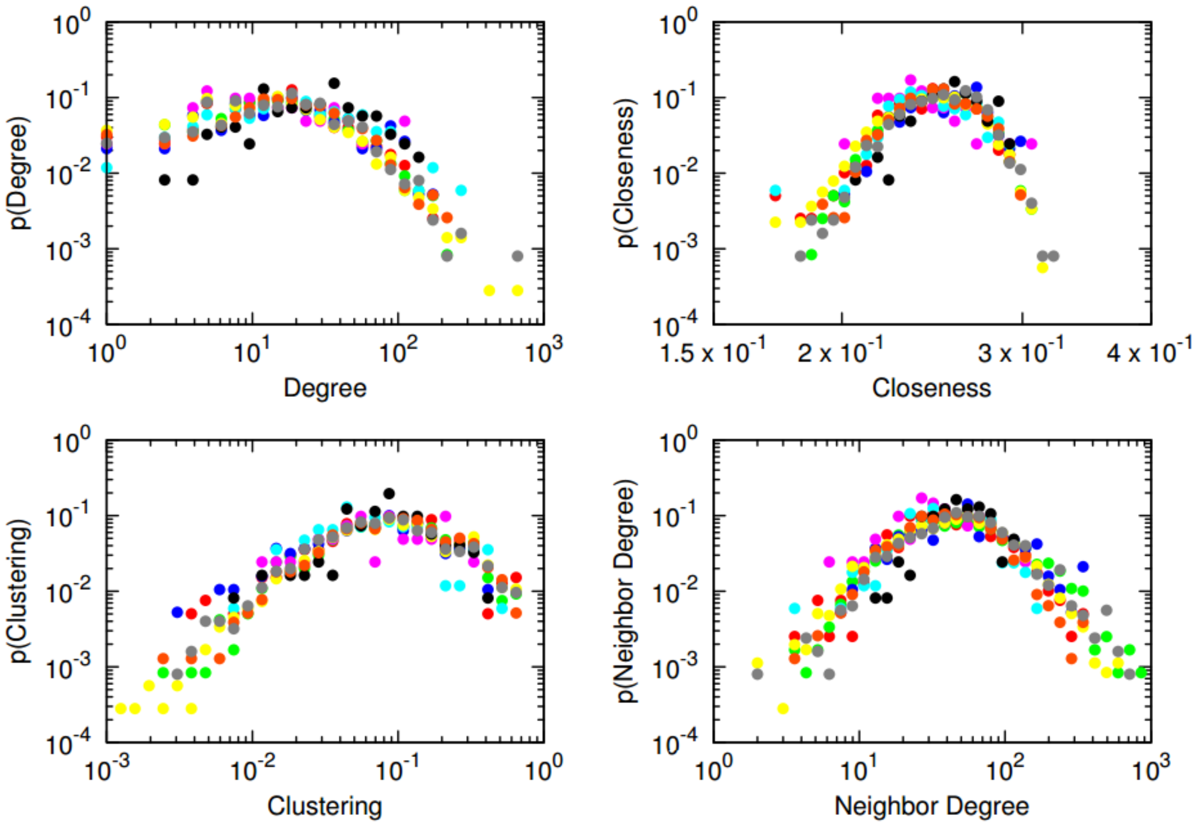


80.000utenti, 4000.000 connessioni

Leader finding



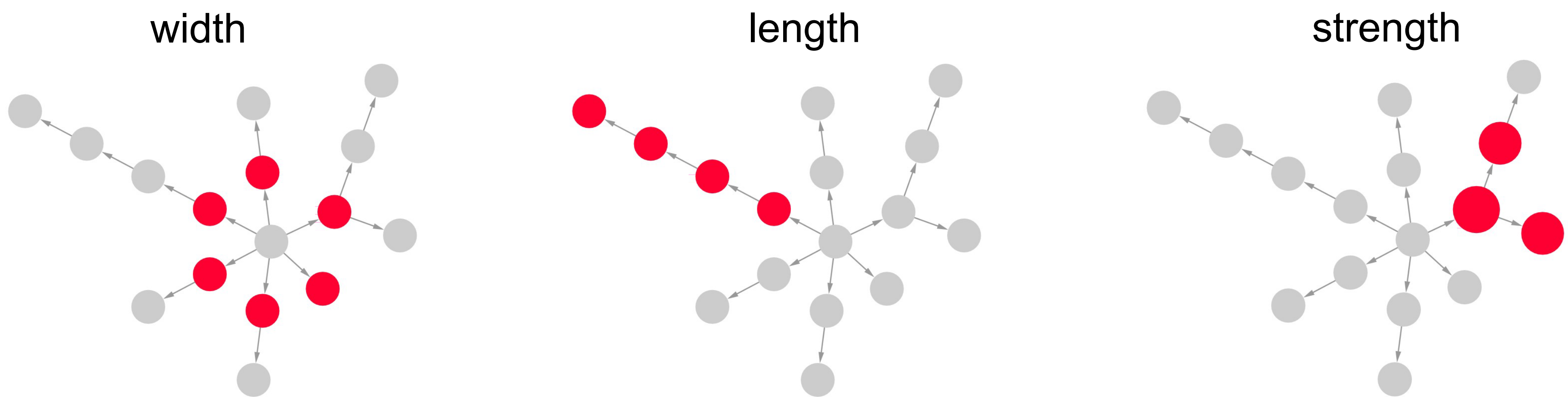
dai BigData...i veri influenzer non sono i leaders



… abbiamo scoperto che i leader teorici, quelli che avrebbero in teoria il potere di influenzare la rete sociale, non hanno una grande influenza pratica sulla rete.

What is Social Prominence?

- It has been observed that a small set of users in a Social Network is able to anticipate (or influence) the behavior of the entire network
- We detected 3 possible scenarios:

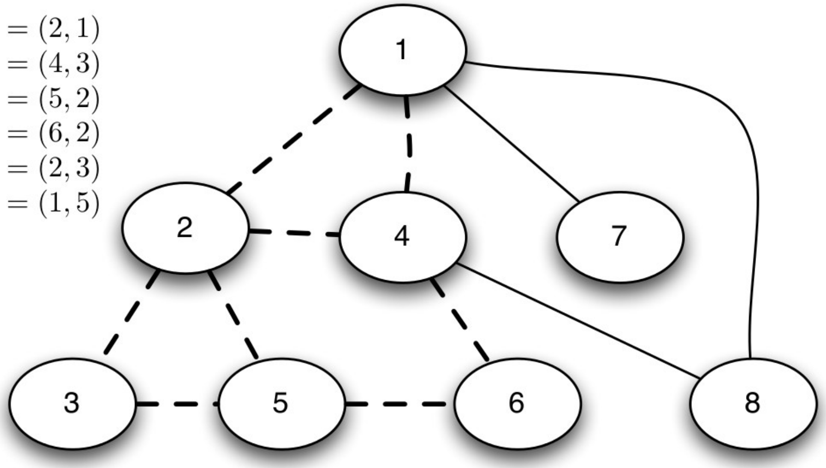


The Idea

- Define what a “leader” is
- Identify three measures of social prominence (width, depth and strength)
- Analyze their relationship with the topological characteristic of prominent actors in a network
- Look for patterns distinguishing different objects spreading in a social network

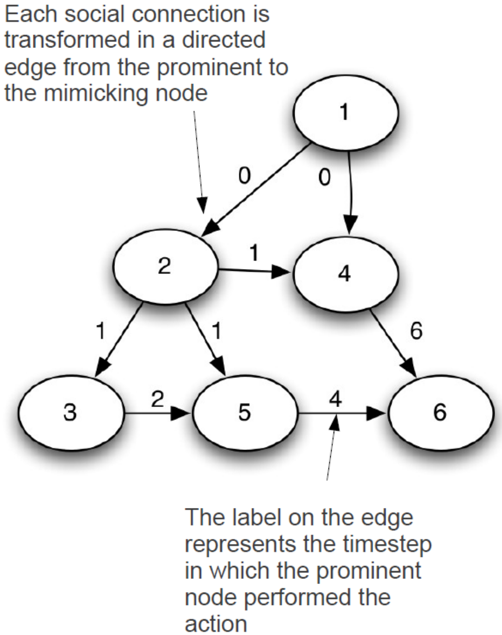
Leaders and structure

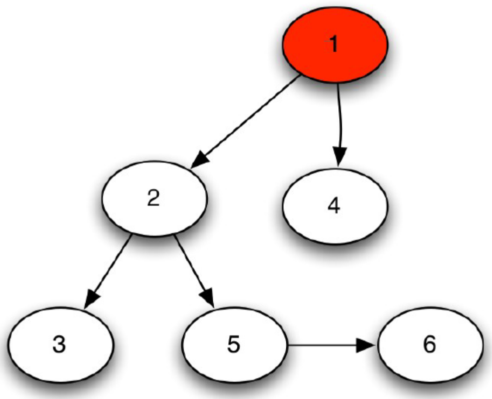
$$
\begin{array}{ll}
a_{1,x} = (1,0) & a_{4,y} = (2,1) \\
a_{2,x} = (2,1) & a_{7,y} = (4,3) \\
a_{3,x} = (1,2) & a_{8,y} = (5,2) \\
a_{4,x} = (4,6) & a_{6,y} = (6,2) \\
a_{5,x} = (1,4) & a_{1,y} = (2,3) \\
a_{6,x} = (6,7) & a_{2,y} = (1,5)
\end{array}
$$



For each Artist we extract the induced temporal subgraph of its Listeners

We define Leader all those nodes that are the first, in their neighborhood
to adopt the given artist





The Minimum Diffusion Tree (MDT) is then the minimum spanning tree

Data, experiments and results

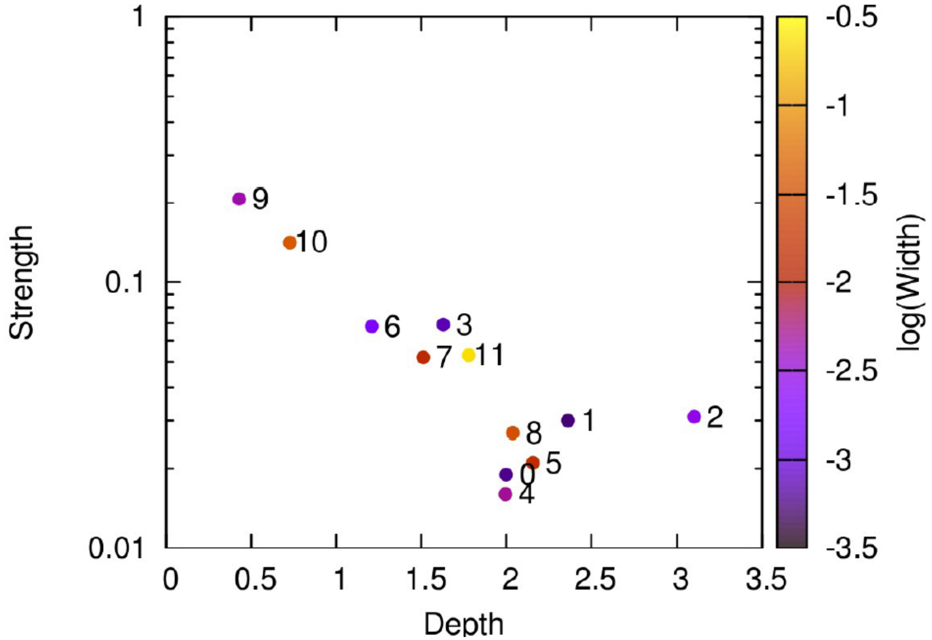
Data gently provided by

	Width	Strength	Degree	Clustering	Neigh Deg	Bet Centr	Clo Centr
AVG Depth	-0.03	**-0.23**	-0.08	0.05	-0.08	-0.02	**-0.13**
Width	-	0.01	**-0.31**	**0.13**	0.05	-0.07	**-0.59**
Strength	-	-	0.02	-0.02	0.03	0.00	0.04
Degree	-	-	-	**-0.16**	-0.02	**0.77**	**0.56**
Clustering	-	-	-	-	-0.05	-0.06	**-0.32**
Neigh Deg	-	-	-	-	-	-0.00	**0.39**
Bet Centr	-	-	-	-	-	-	**0.22**

Central nodes are characterized by low Depth & Width
High Width are usually reached only by nodes in tightly knit communities
There is a trade-off between Depth and Strength (not between D and W nor between S a

Data, experiments and results

Cluster	size	dance	ele	folk	jazz	met	pop	punk	rap	rock
0	1822	1.25	1.13	**1.54**	1.37	1.50	0.76	1.31	1.13	1.10
1	136	1.28	1.55	1.28	**2.35**	0.78	0.73	0.64	1.35	0.70
2	664	0.59	0.87	0.98	0.48	0.95	0.97	**1.50**	1.20	1.19
3	482	1.26	1.16	1.09	1.12	0.91	0.80	**2.48**	1.24	0.89
4	973	1.14	1.20	1.15	**1.41**	0.80	0.91	0.66	0.97	0.97
5	512	**1.29**	0.96	0.95	1.09	1.10	0.97	0.33	1.06	1.01
6	682	0.89	0.79	0.61	0.64	**1.13**	1.08	1.07	1.08	1.01
7	124	0.75	**1.45**	0.35	0.64	0	1.09	0	1.02	0.62
8	524	0.93	1.01	1.12	0.91	**1.15**	1.07	0.43	0.95	0.87
9	937	0.40	0.46	0.19	0.23	0.45	**1.56**	0.13	0.37	1.06
10	232	0.72	0.57	0.27	0.99	0.38	**1.44**	0.38	0.46	1.00
11	612	0.74	0.94	0.71	0.40	0.70	**1.27**	0.07	0.68	0.83



Jazz:
1 lowest width
4 lowest strength
Not easy to be prominent

Pop:
9, 10, 11
Lowest depth, highest strength
Leaders for pop artists are embedded in groups of users very engaged with the new artist, but not prominent among their friends

Punk:
2 high depth, low width and strength
Long cascades, exactly the opposite of the pop genre, similar to folk!

Dance:
5 high depth, high width, low strength

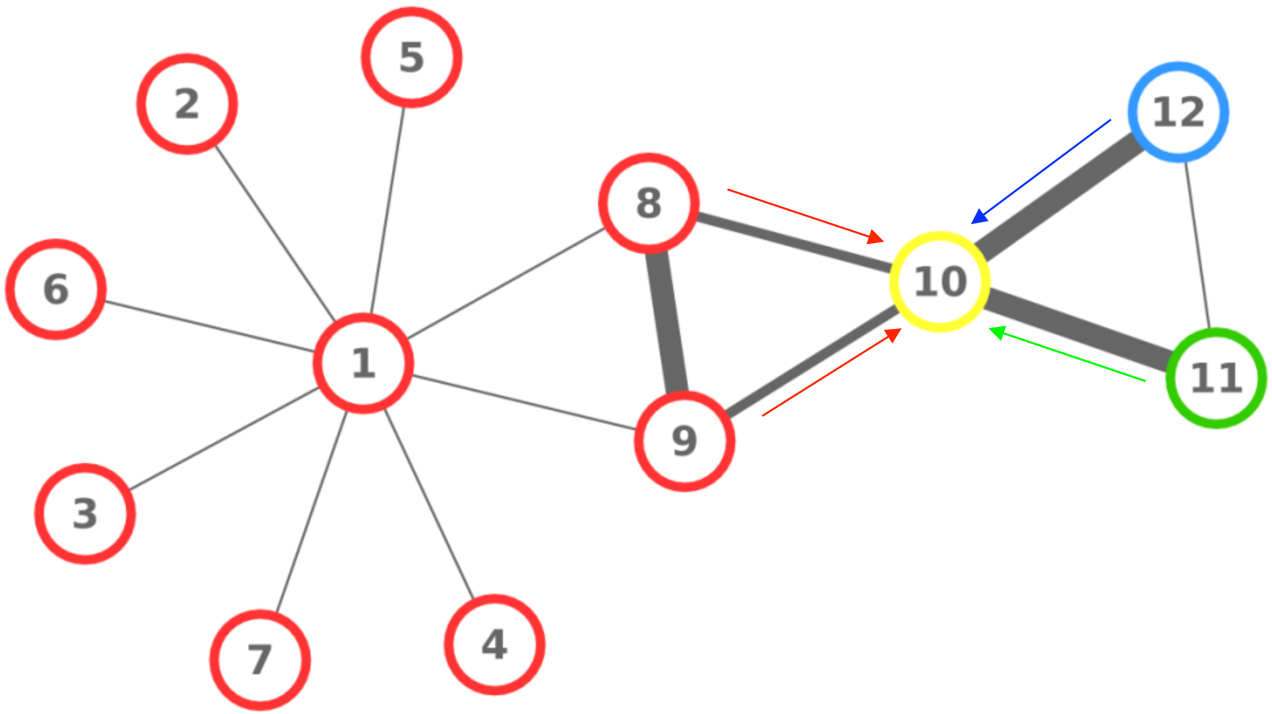
Social knowledge: **U** know **B**ecause **I** **K**now”

La Conoscenza Sociale

Quello che possiamo
Contenere nel nostro cervello
E' una minima frazione
Della conoscenza umana

Ma le nostre connessioni sociali
Contengono ognuna un'altra parte
E la loro somma puo' essere significat
Come valutare, quindi, una persona?

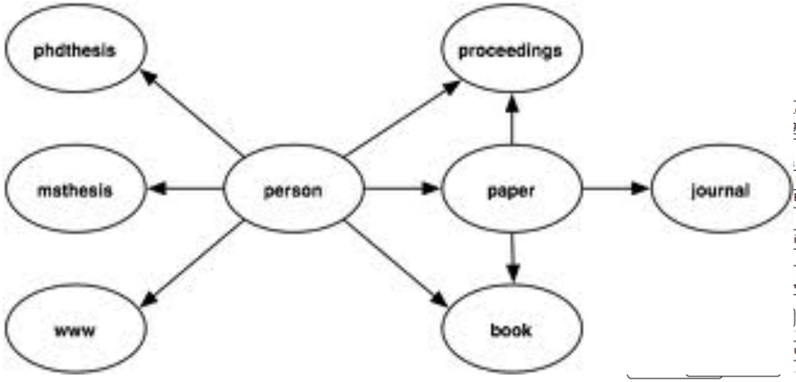
Calcolare la propria conoscenza sociale

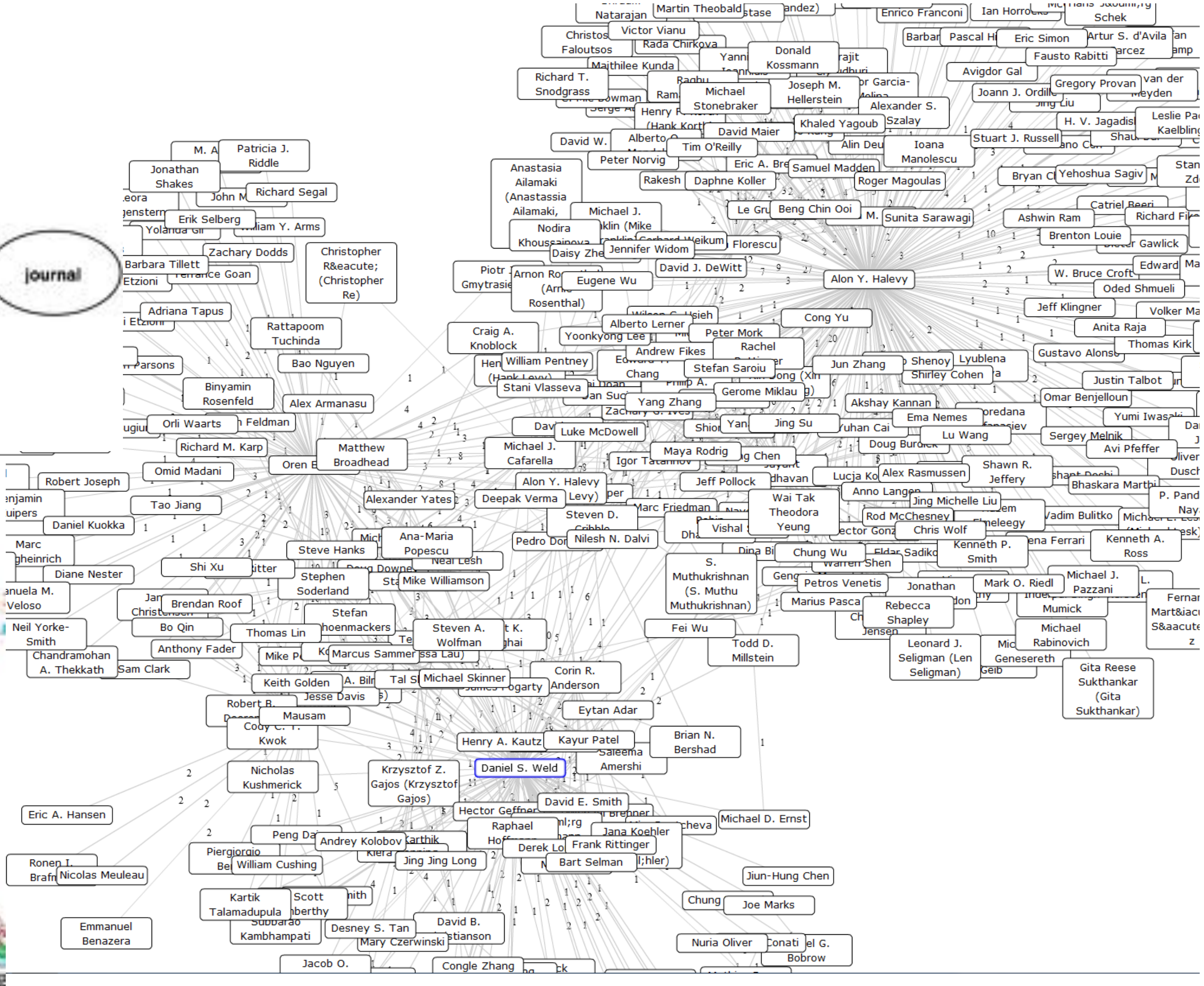


Con processi di diffusione su reti possiamo quantificare l'ammontare di “skill” che ogni connessione ci permette di accedere

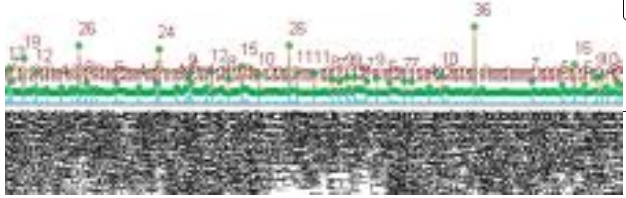
dblp
computer science bibliography

Le pubblicazioni di 40.000 ricercatori in DataBase & DataMining per 30 anni



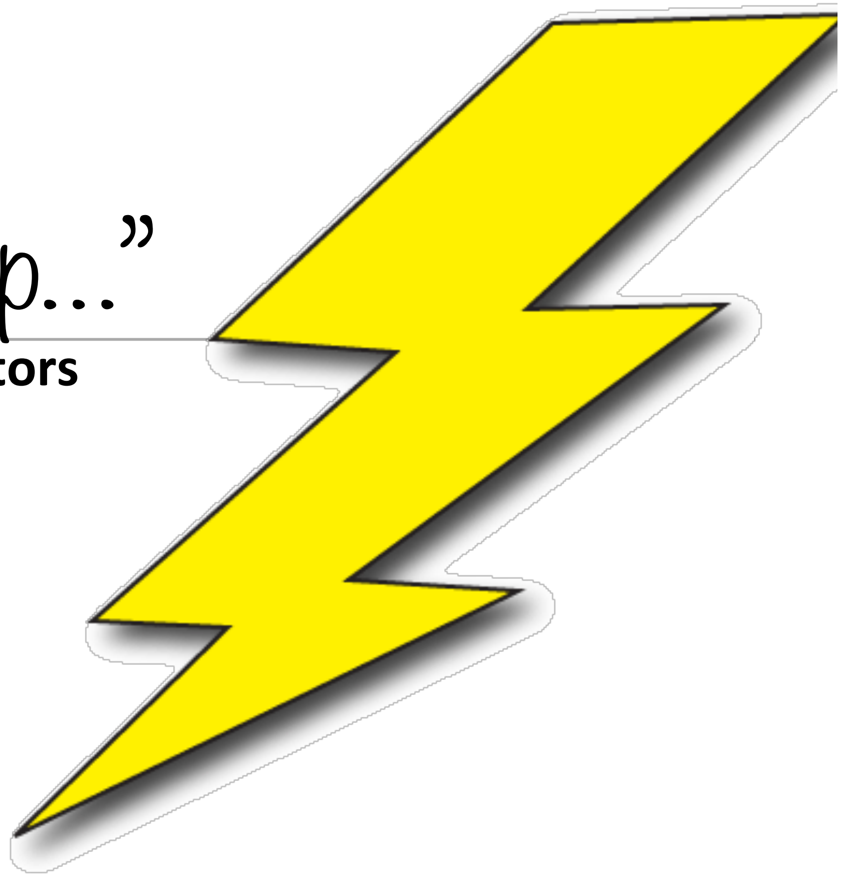


Co-author Graph



“It’s a long way to the top…”

Predicting **Success** via **Innovators**

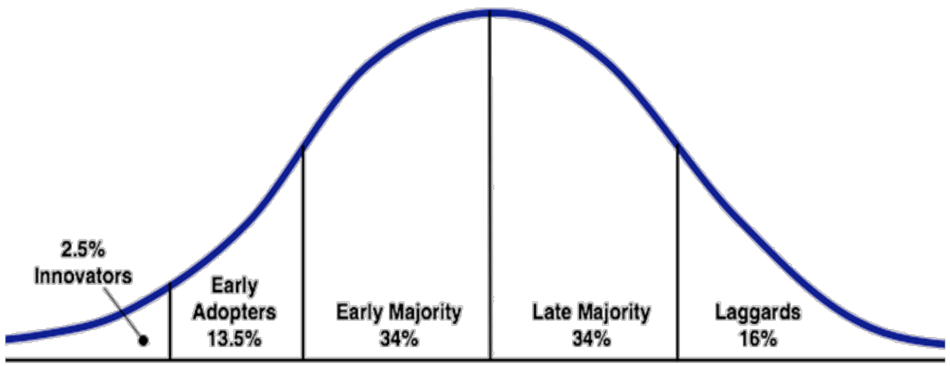


G. Rossetti, D. Pennacchioli, L. Milli, D. Pedreschi and F. Giannotti - 2015

Adopters: Innovators

- **Diffusion of Innovations**
  [Rogers 1962]

- Five “category” of **Adopters** based on the time of first adoptions:
  - Each one has its own semantics;
  - Temporal distribution
    Assumed to be a Gaussian;
  - Categories proportion is univocally determined
    (i.e. Innovators are <u>always</u> the first 2.5%)



Goods: Hits and Flops

- Retail market products,
- Music Artists,
- Business and stores…

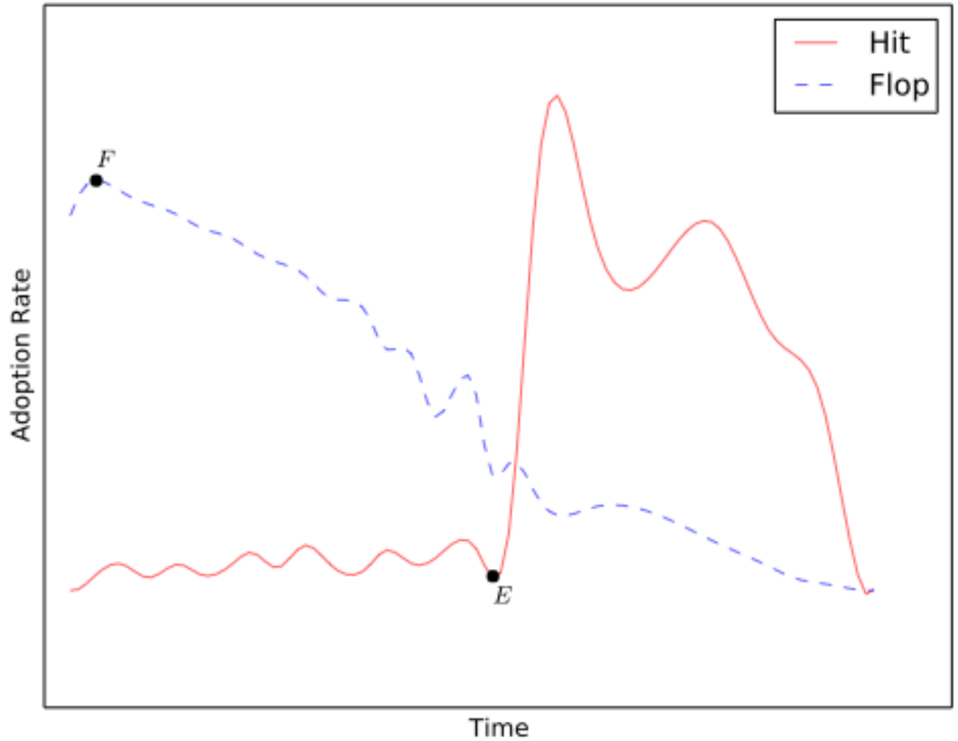
What is a successful (Hit) good?

And an unsuccessful (Flop) one?

Hits and Flops share the same set of adopters or not?

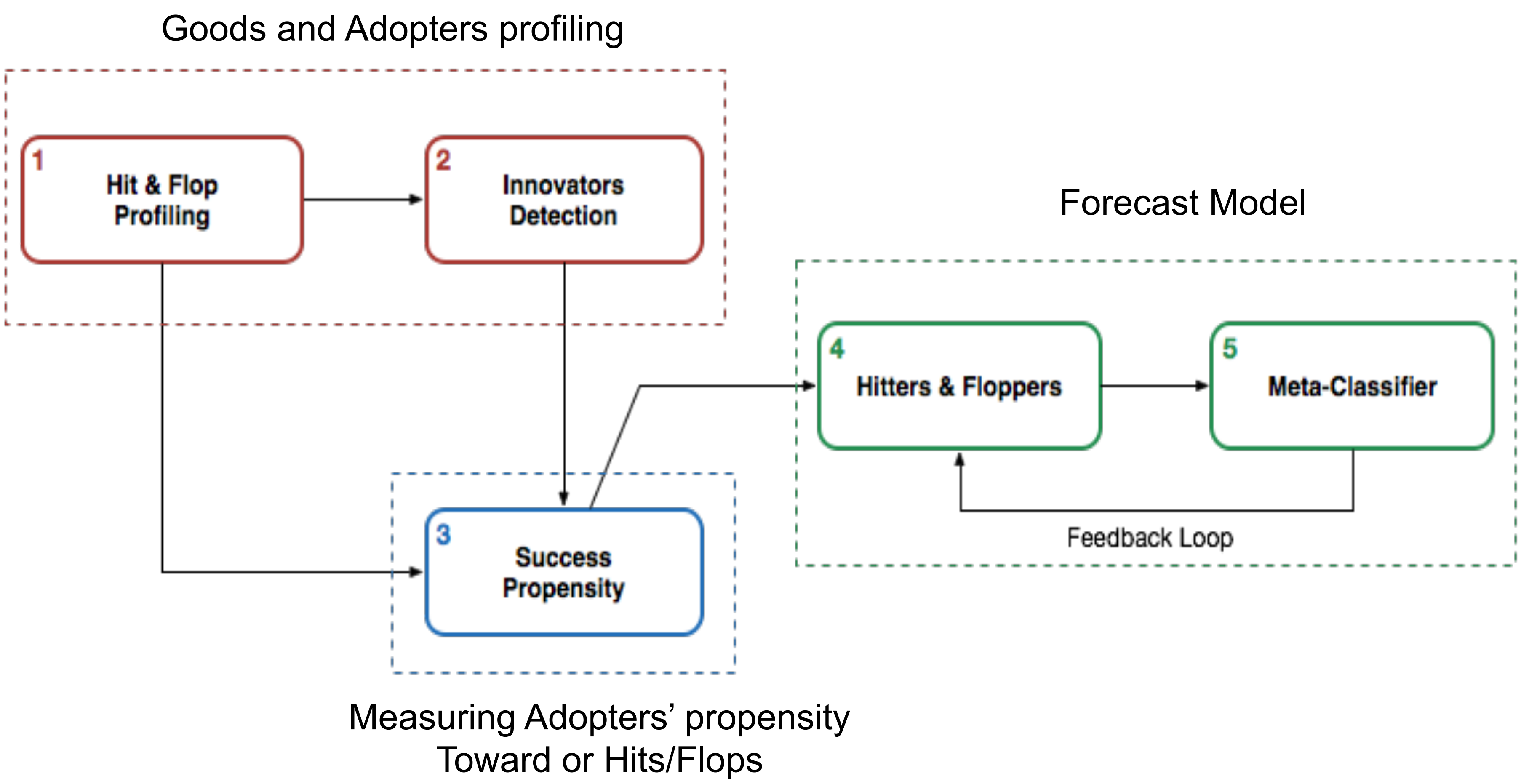
Hits & Flops: qualitative definitions

- Hit
  - A good whose trend slowly increases trough time until reaching an explosion point that marks the start of a sharp rising of its adoptions.
- Flop
  - A good whose adoption trend does not increase considerably over time or even reaches an early maximum only to sharply decrease.



Given a **partial observation** of the **adoptions** of a **novel good** can we decide if it will became an **Hit** or a **Flop**?

Hit&Flop: Workflow



Forecast Evaluation*

COOP	H&F	ER-H&F	ER	NM
PPV	**.781**(.09)	.825(.21)	0(0)	.547(.01)
NPV	.316(.12)	.384(.06)	.292(0)	.05(.03)
Recall	**.586**(.29)	.03(.01)	0(0)	.818(.04)
Specificity	.522(.38)	.982(.02)	1(0)	.361(.02)

Last.fm	H&F	ER-H&F	ER	NM
PPV	**.766**(.03)	.290(.37)	0(0)	.644(0)
NPV	**.471**(.04)	.047(.39)	.351(0)	.026(.04)
Recall	.520(.04)	.006(.01)	0(0)	.990(.02)
Specificity	.727(.06)	.970(.02)	1(0)	.007(.01)

Yelp	H&F	ER-H&F	ER	NM
PPV	**.990**(.01)	1(0)	0(0)	.488(.04)
NPV	**.631**(.17)	.341(.11)	.306(0)	.099(.08)
Recall	**.897**(.09)	.654(.11)	0(0)	.933(.01)
Specificity	**.906**(.10)	1(0)	1(0)	.007(.01)

Datasets

Dataset	Goods	Adopters	Adoptions	Period	Obs. window
COOP	5605	620026	11204984	1 year	4 weeks
Last.fm	1806	50837	882845	2 years	2 months
Yelp	2499	141936	427894	10 years	30 months

Competitors

H&F: Hits&Flops
ER-H&F: Hits&Flops with Roger's Innovators
ER: Rogers's Innovators
NM: Hits&Flops on Null Model (avg. 100 models)

Results in a nutshell

- H&F guarantee the most stable predictive performances in terms of PPV and Recall
- ER is not able to provide useful classification (2.5% fixed innovator threshold)
- ER–H&F suffer the constrains imposed by ER

*Results after a 10-fold cross validati

Knowledge Discovery & Data Mining Laboratory | kdd.isti.cnr.it

Textbooks & reading

- David Easley, Jon Kleinberg: *Networks, Crowds, and Markets*. http://www.cs.cornell.edu/home/kleinber/networks-book/
- Albert-Laszlo Barabasi. *Network Science Book Project* (2013, ongoing) http://barabasilab.neu.edu/networksciencebook/
- A.-L. Barabasi. *Linked*. Plume, 2002

Courses

- Pedreschi + Giannotti @ University of Pisa
  - http://didawiki.cli.di.unipi.it/doku.php/wma/start
- Barabasi @ Northeastern University
  - http://barabasilab.neu.edu/courses/phys5116/
- Leskovec @ Stanford University
  - http://www.stanford.edu/class/cs224w/handouts.html
- Slides from this course are freely adapted from those of Laszlo Barabasi, Jure Leskovec, Fosca Giannotti, besides my own. Thanks!