Social Network Analysis

A crash course @ UPF

Dino Pedreschi

ISTI-CNR & Università di Pisa

http://kdd.isti.cnr.it

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNIVERSITÀ DI PISA

Complex (Social) Networks

- Big graph data and social, information, biological and technological networks
- The architecture of complexity and how real networks differ from random networks:
  - node degree and long tails,
  - social distance and small worlds,
  - clustering and triadic closure.
- Comparing real networks and random graphs.
- The main models of network science: small world and preferential attachment.

Complex (Social) Networks

- Strong and weak ties, community structure and long-range bridges.
- Robustness of networks to failures and attacks.
- Cascades and spreading. Network models for diffusion and epidemics. The strength of weak ties for the diffusion of information. The strength of strong ties for the diffusion of innovation.
- Practical network analytics with Cytoscape and Gephi.
- Simulation of network processes with NetLogo.

Complex (Social) Networks

- Textbooks
  - Albert-Laszlo Barabasi. *Network Science* (2016)
  - http://barabasi.com/book/network-science
  - David Easley, Jon Kleinberg: *Networks, Crowds, and Markets* (2010)
  - http://www.cs.cornell.edu/home/kleinber/networks-book/
- Network Analytics Software (open):
  - Cytoscape: http://www.cytoscape.org/
  - Gephi: http://gephi.github.io/
- Network Data Repository
  - http://networkrepository.com/
- Simulation of network models: NetLogo

Part 2

- Small-world & Preferential attachment recap
- Measuring small-worlds with big data
- Strength of weak ties
- Centrality measures
- Strength of weak ties, centrality and mobility
- Community discovery
- Link prediction
- Multi-dimensional network analysis

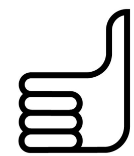
ARE REAL NETWORKS LIKE RANDOM GRAPHS?

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:

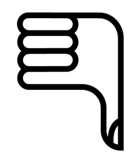
Average path length:

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



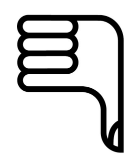
Clustering Coefficient:

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



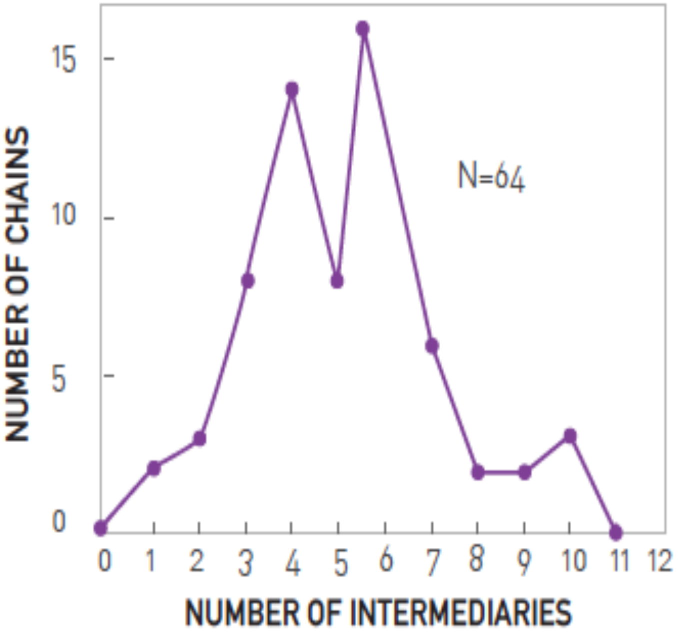
Degree Distribution:

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

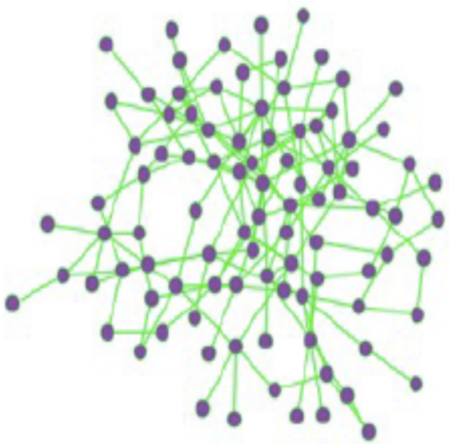


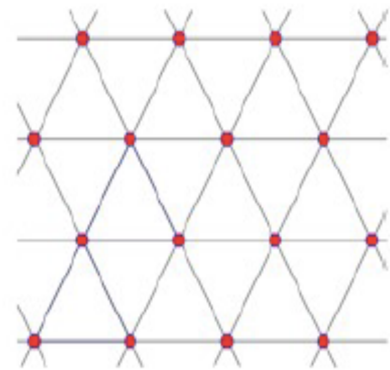
The small-world model

Milgram experiment



Real networks are between random networks and lattices



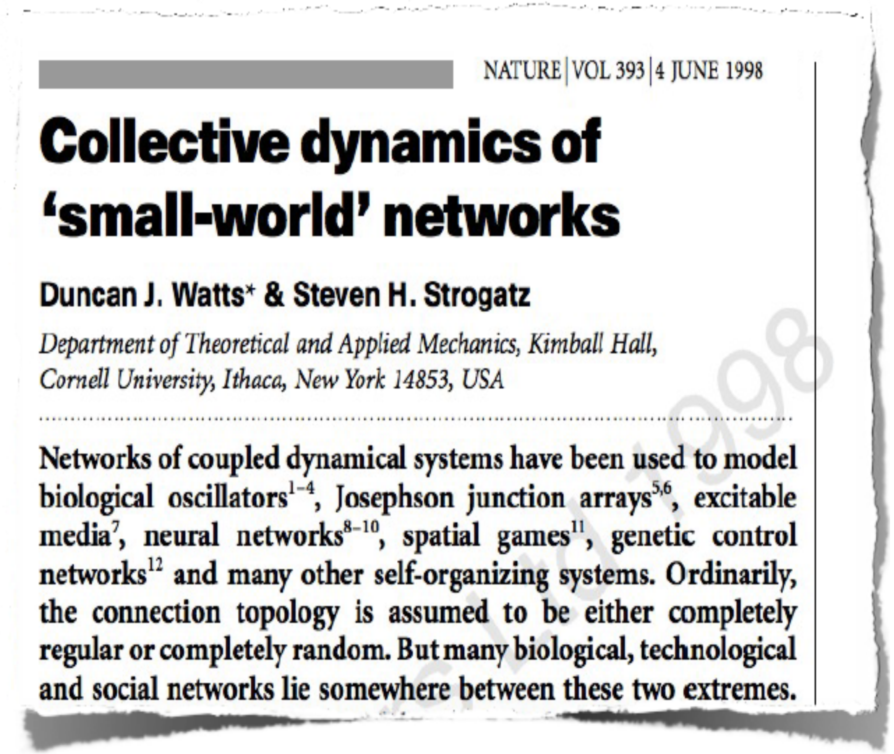


Real networks are somewhere here

Watts-Strogatz model

Duncan Watts

Steve Strogatz



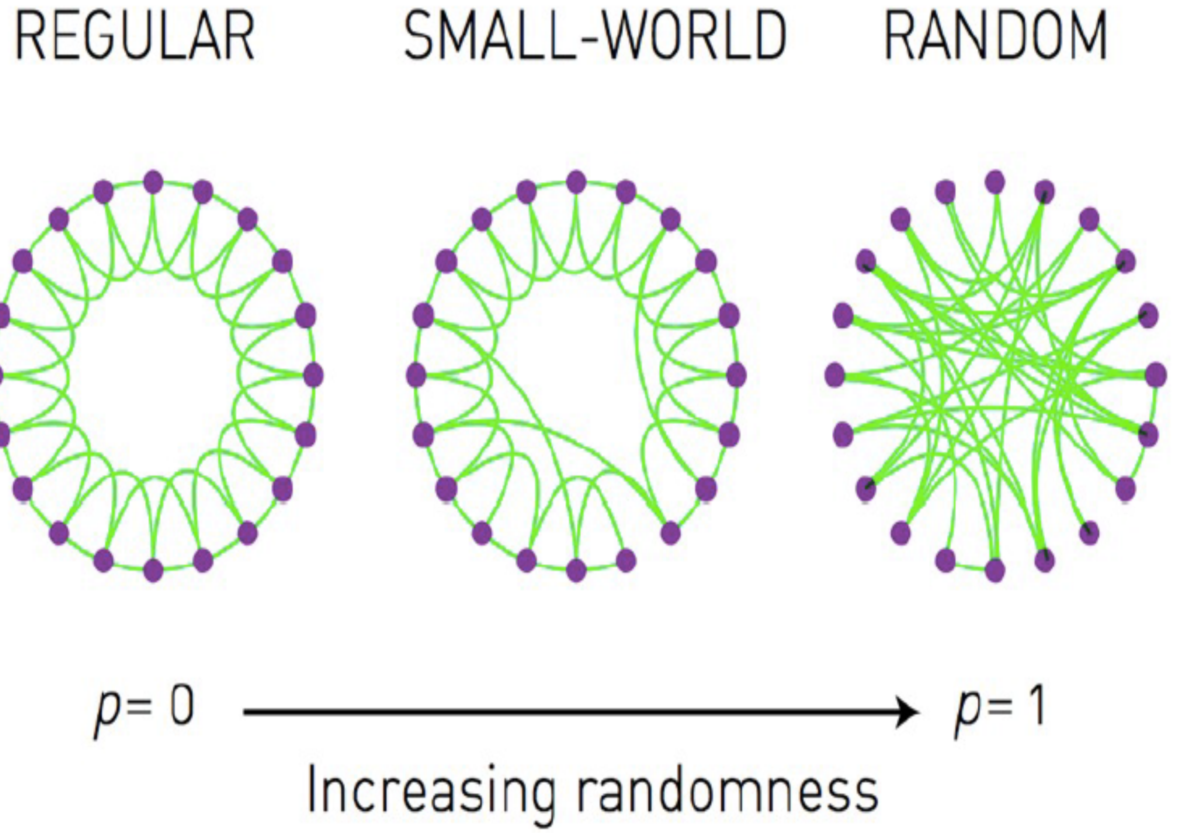
NATURE | VOL 393 | 4 JUNE 1998

Collective dynamics of 'small-world' networks

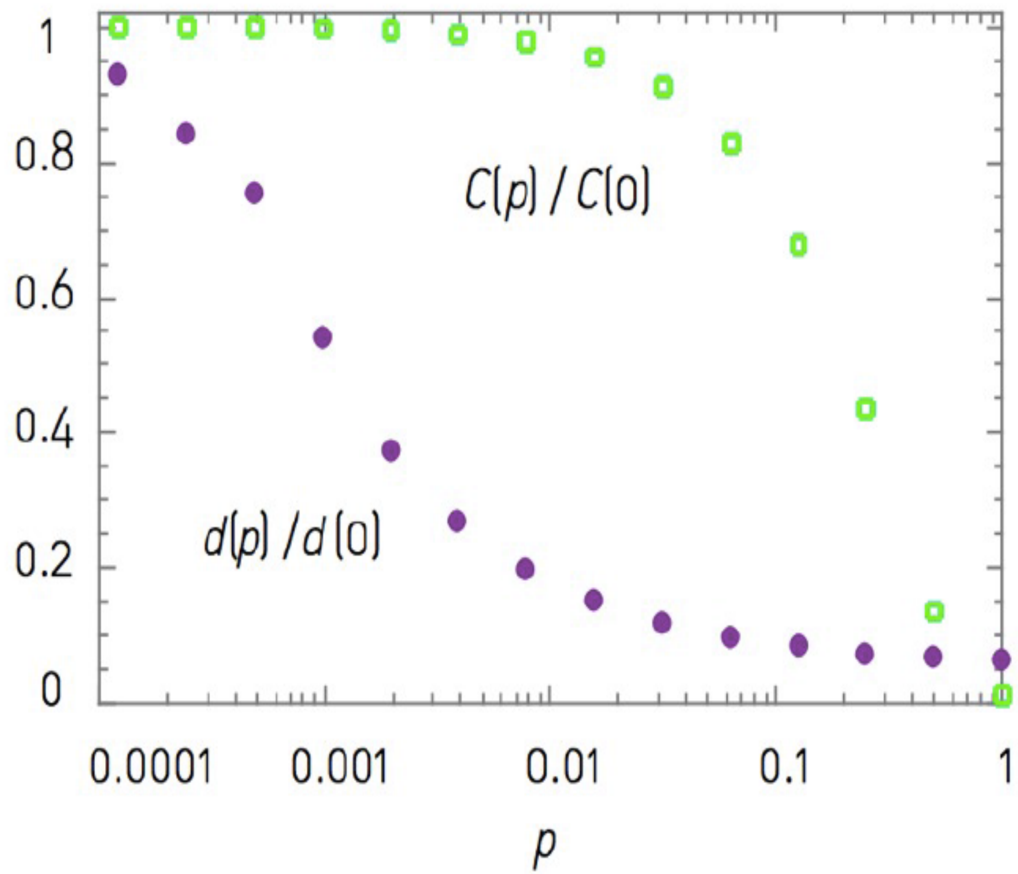
Duncan J. Watts* & Steven H. Strogatz

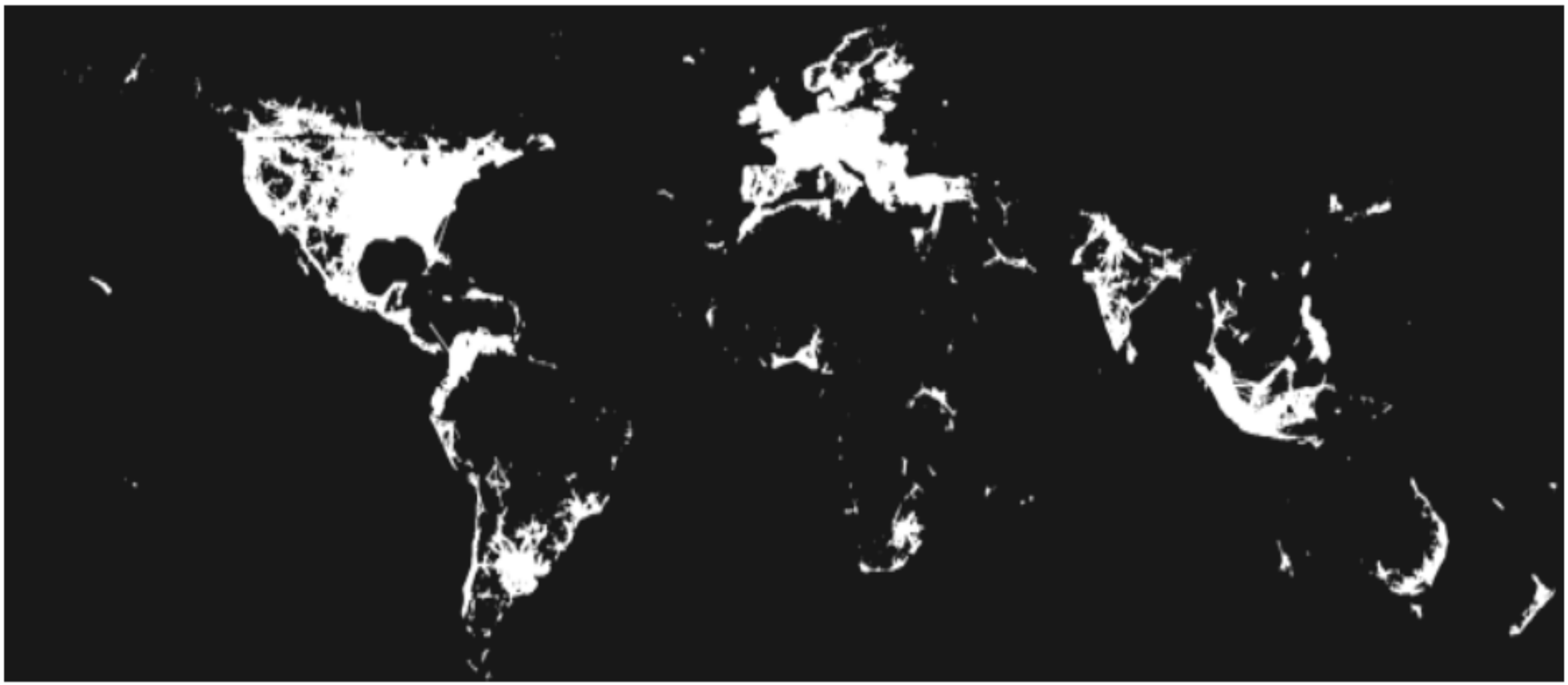
Department of Theoretical and Applied Mechanics, Kimball Hall, Cornell University, Ithaca, New York 14853, USA

Networks of coupled dynamical systems have been used to model biological oscillators[1–4], Josephson junction arrays[5,6], excitable media[7], neural networks[8–10], spatial games[11], genetic control networks[12] and many other self-organizing systems. Ordinarily, the connection topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes.


REGULAR
SMALL-WORLD
RANDOM
p= 0
p= 1
Increasing randomness

Average path length vs. clustering coefficient





Hubs represent the most striking difference between a random and a scale-free network. Their emergence in many real systems raises several fundamental questions:

- Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?

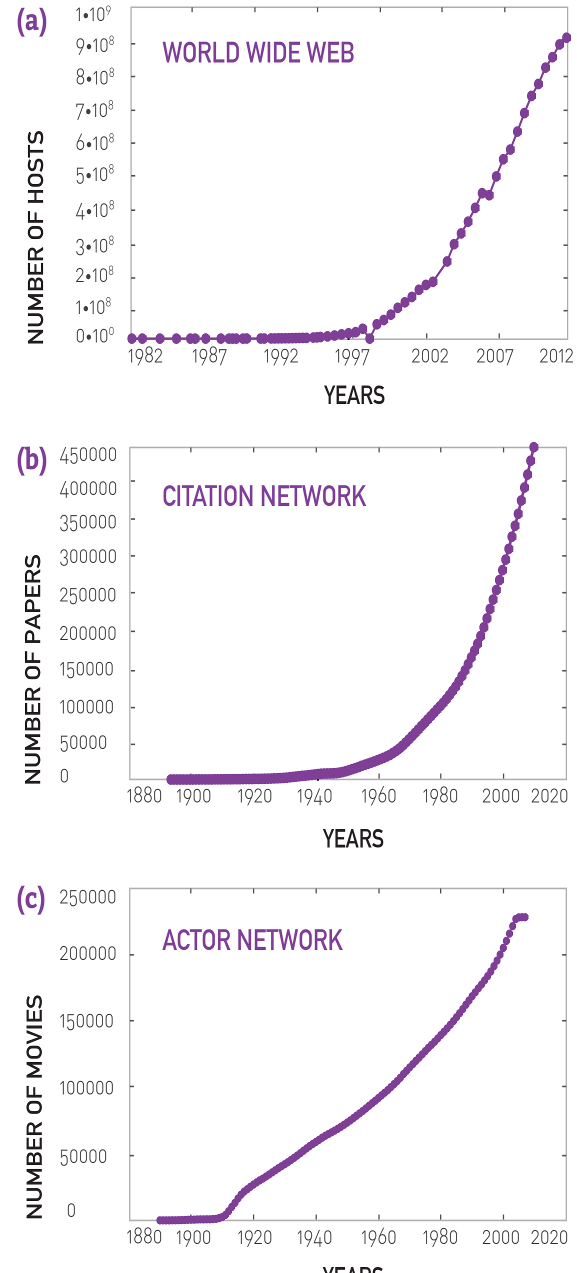
- Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

Growth and preferential attachment

BA MODEL: Growth

ER model:
the number of nodes, N, is fixed (static models)

networks expand through the addition of new nodes



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

BA MODEL: Preferential attachment

ER model: links are added randomly to the network

New nodes prefer to connect to the more connected nodes

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

Growth and Preferential Sttachment

The random network model differs from real networks in two important characteristics:

Growth: While the random network model assumes that the number of nodes is fixed (time invariant), real networks are the result of a growth process that continuously increases.

Preferential Attachment: While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.

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

The Barabási-Albert model

Origin of SF networks: Growth and preferential attachment

(1) Networks continuously expand by the addition of new nodes

WWW : addition of new documents

(2) New nodes prefer to link to highly connected nodes.

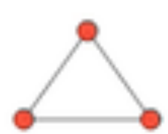
WWW : linking to well known sites

GROWTH:

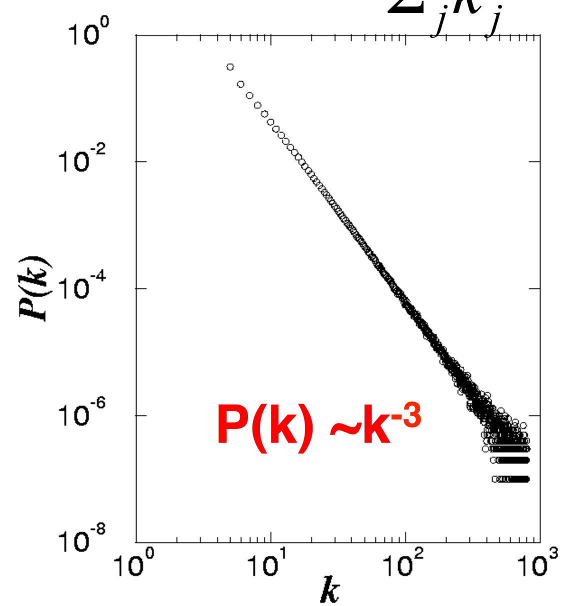
add a new node with m links

PREFERENTIAL ATTACHMENT:

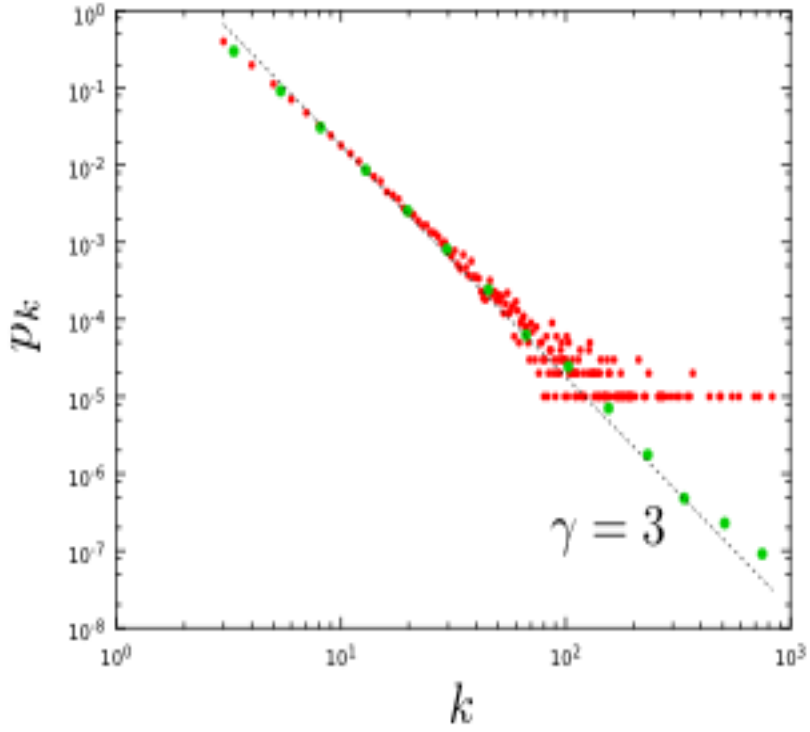
the probability that a node connects to a node with *k* links is proportional to *k*.



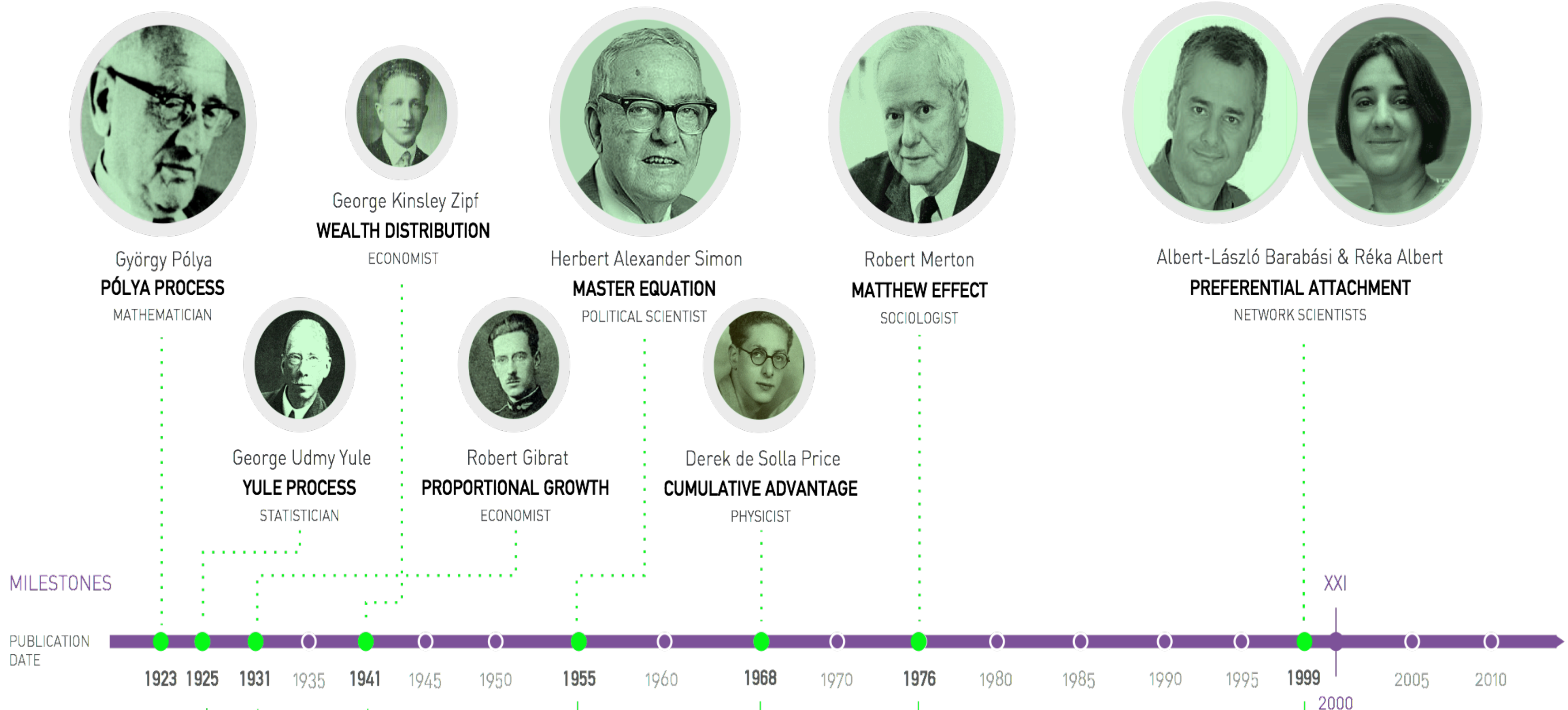
$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$



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



10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
p_k
10^0
10^1
10^2
10^3
k
$\gamma = 3$



György Pólya (1887-1985)
Preferential attachment made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya [2]. Hence, in mathematics preferential attachment is often called a **Pólya process**.

Robert Gibrat (1904-1980)
proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called **proportional growth**, this is a form of preferential attachment.

Herbert Alexander Simon (1916-2001)
used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].

Robert Merton (1910-2003)
In sociology preferential attachment is often called the **Matthew effect**, named by Merton [8] after a passage in the Gospel of Matthew.

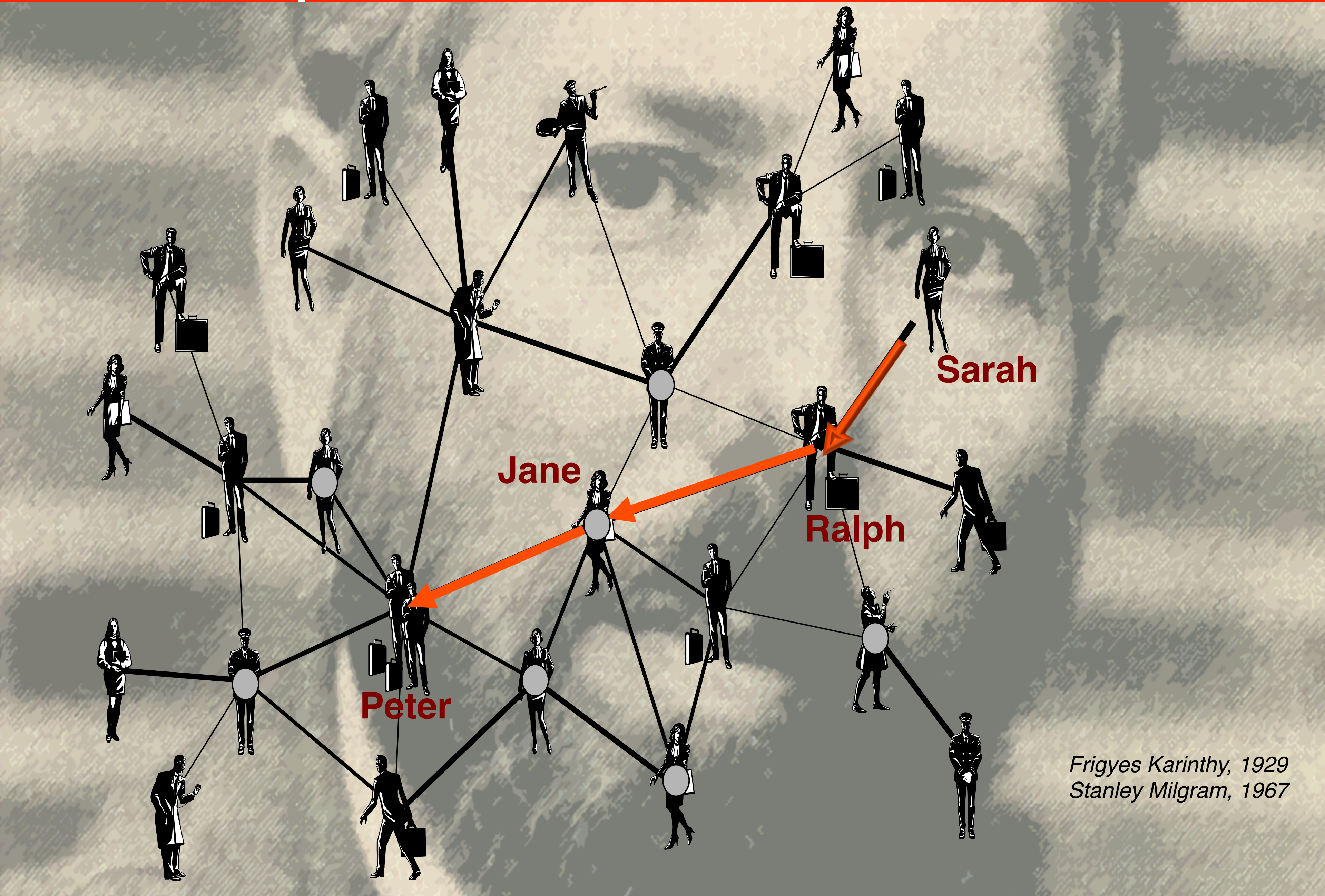
George Udmy Yule (1871-1951)
used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a **Yule process**.

George Kinsley Zipf (1902-1950)
used preferential attachment to explain the fat tailed distribution of wealth in the society [5].

Derek de Solla Price (1922-1983)
used preferential attachment to explain the citation statistics of scientific publications, referring to it as **cumulative advantage** [7].

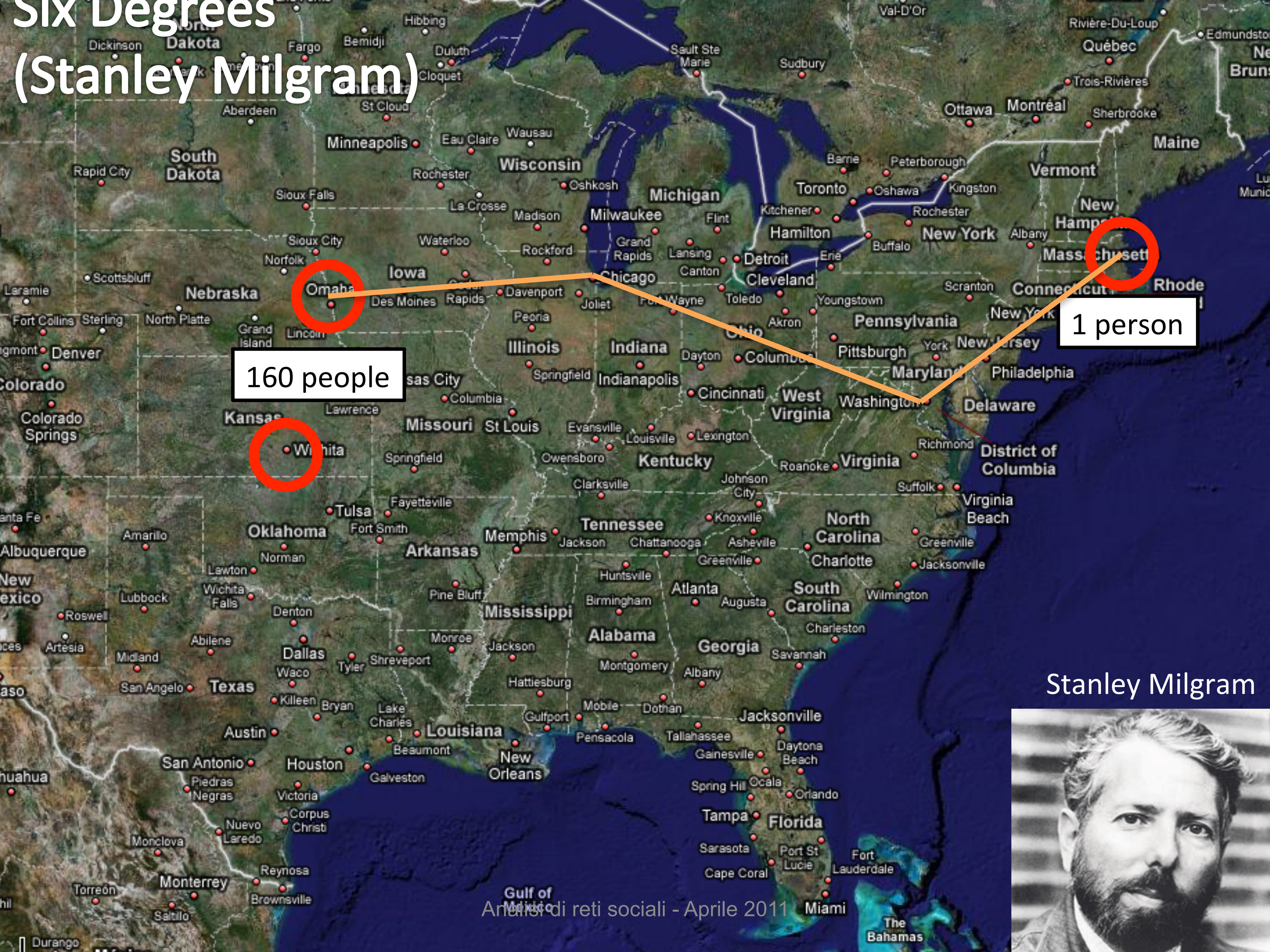
Barabási (1967) & **Albert** (1972)
introduce the term **preferential attachment** in the context of networks [1] to explain the origin of their power-law degree distribution.

Measuring the small-world property



Frigyes Karinthy, 1929
Stanley Milgram, 1967

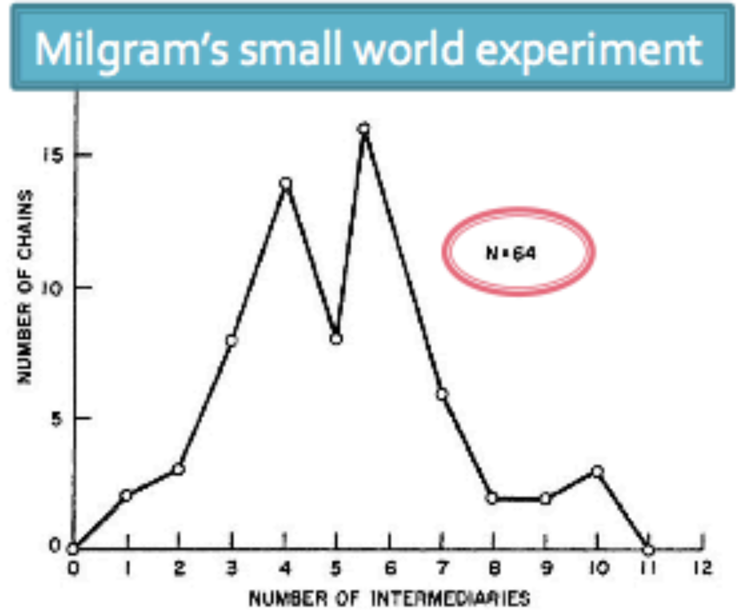
Six Degrees (Stanley Milgram)



Stanley Milgram

The Small-world experiment

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



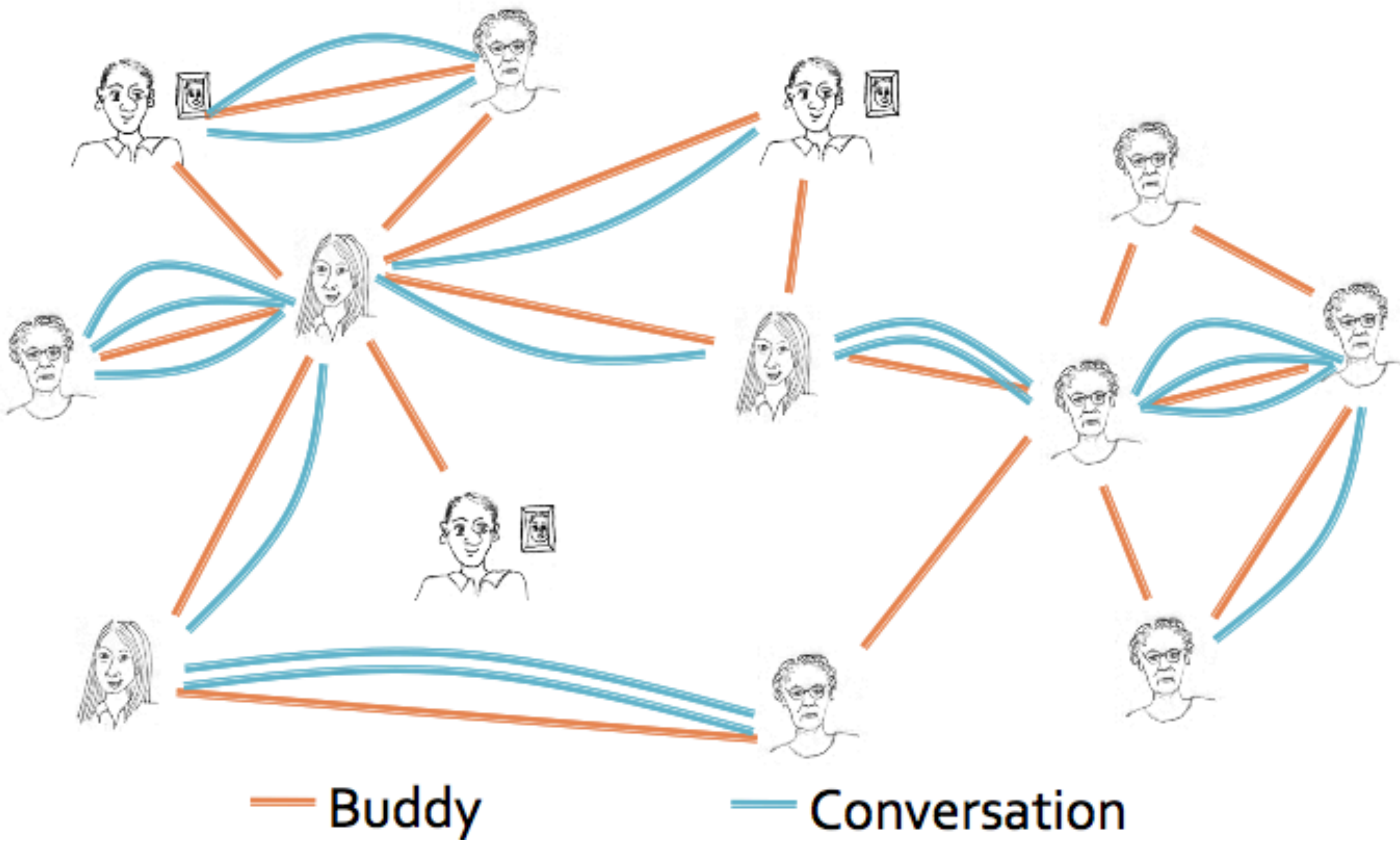
- Further observations:
  - People what owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
  - People from the Boston area have even closer paths: 4.4

Planetary-Scale Views on an Instant-Messaging Network

Jure Leskovec & Eric Horvitz

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

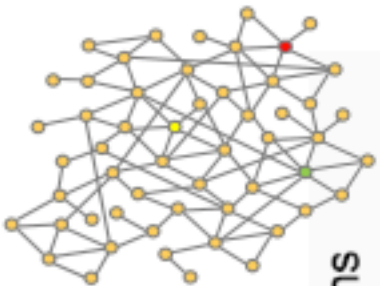
Messaging as a network

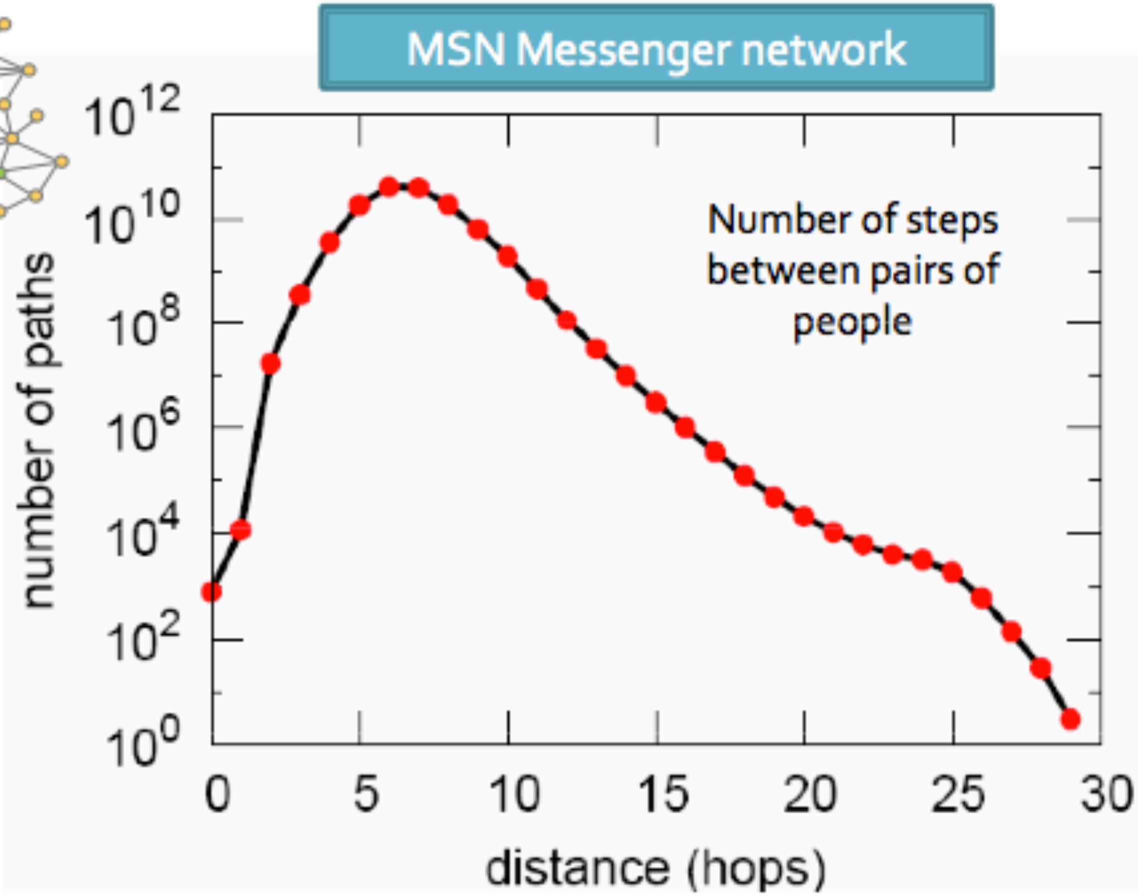


IM communication network

- **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



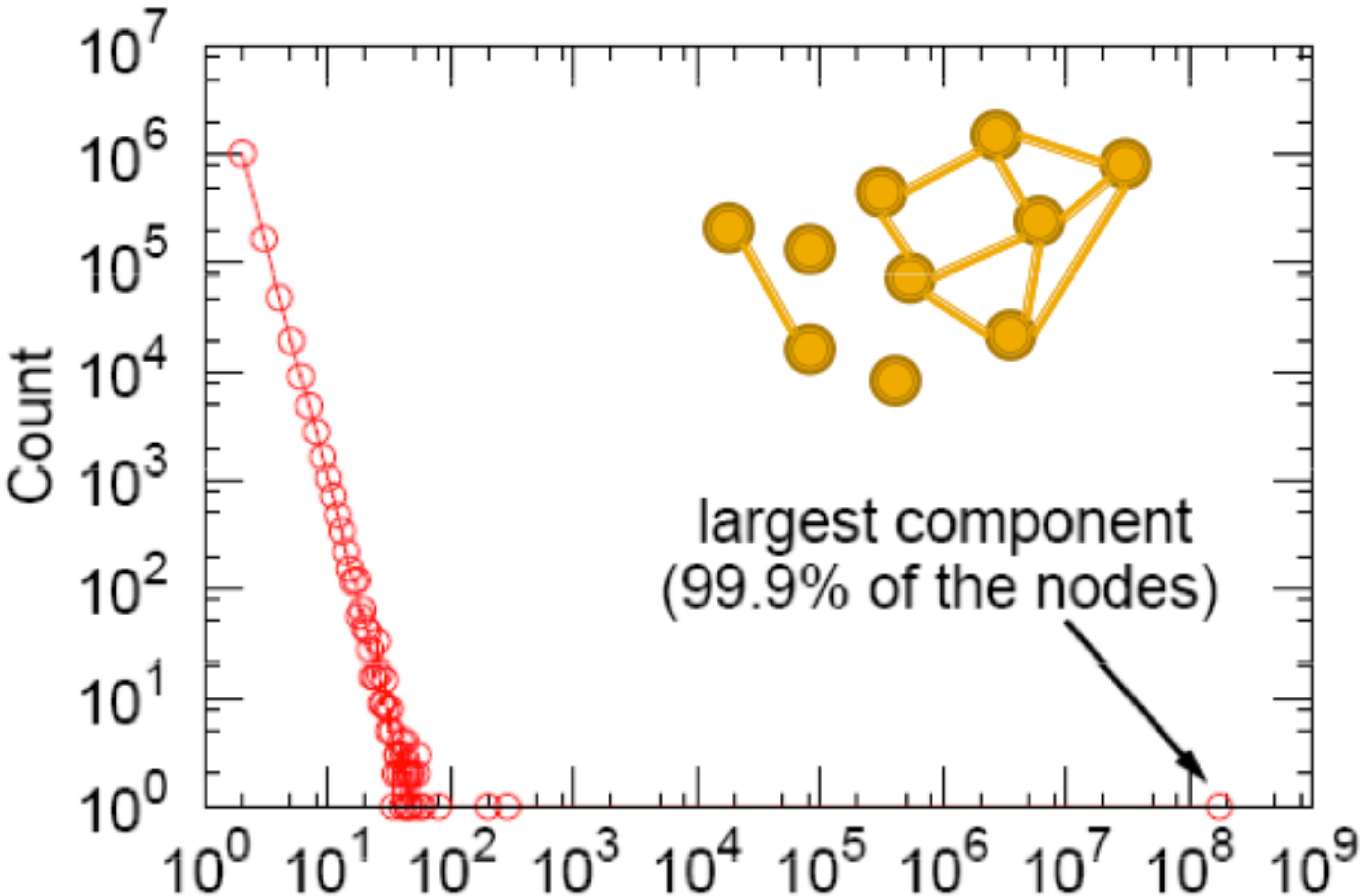


Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Avg. path length **6.6**

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

The giant connected component



WWW: 19 DEGREES OF SEPARATION

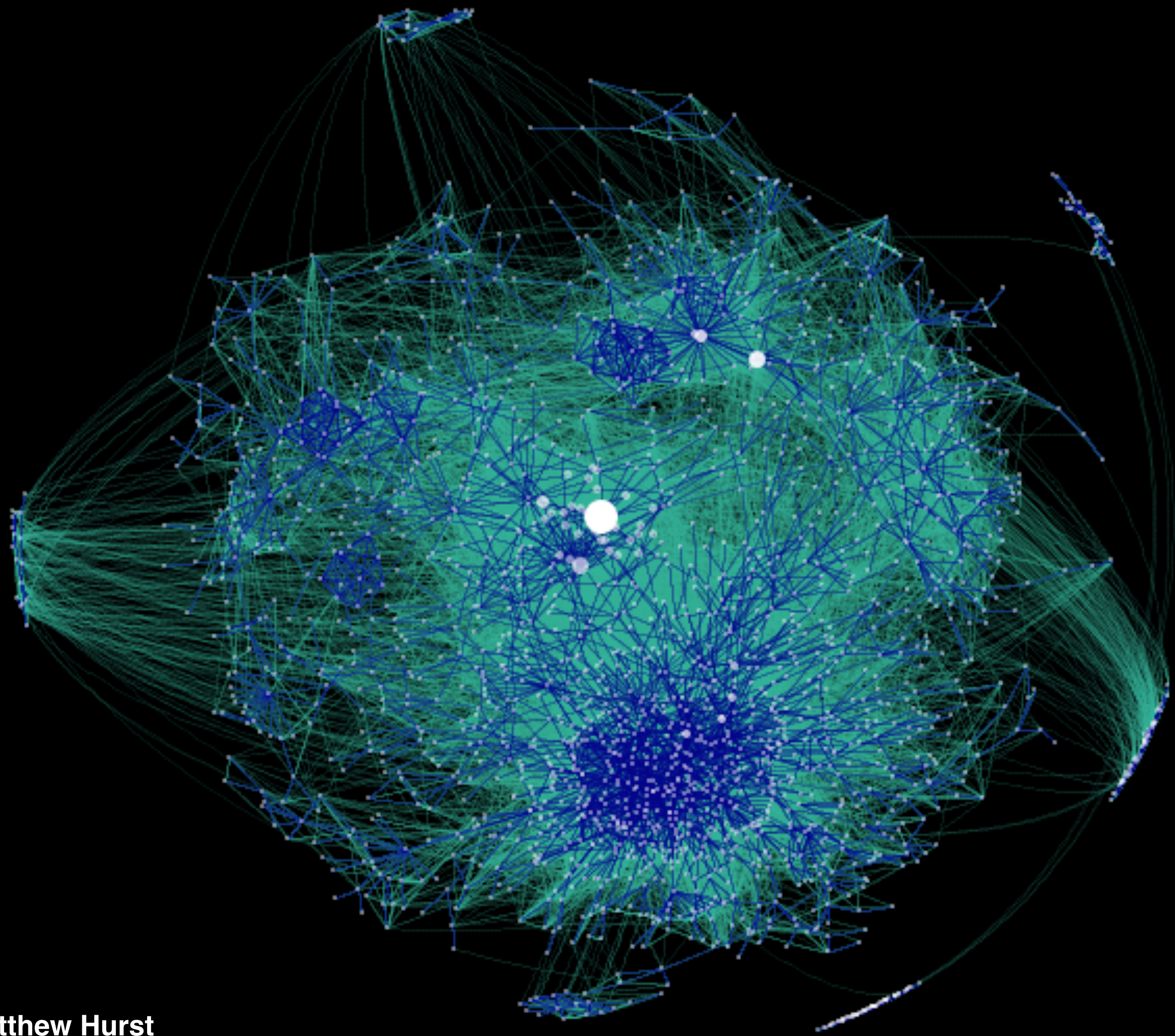


Image by **Matthew Hurst**
Blogosphere

The strength of weak ties

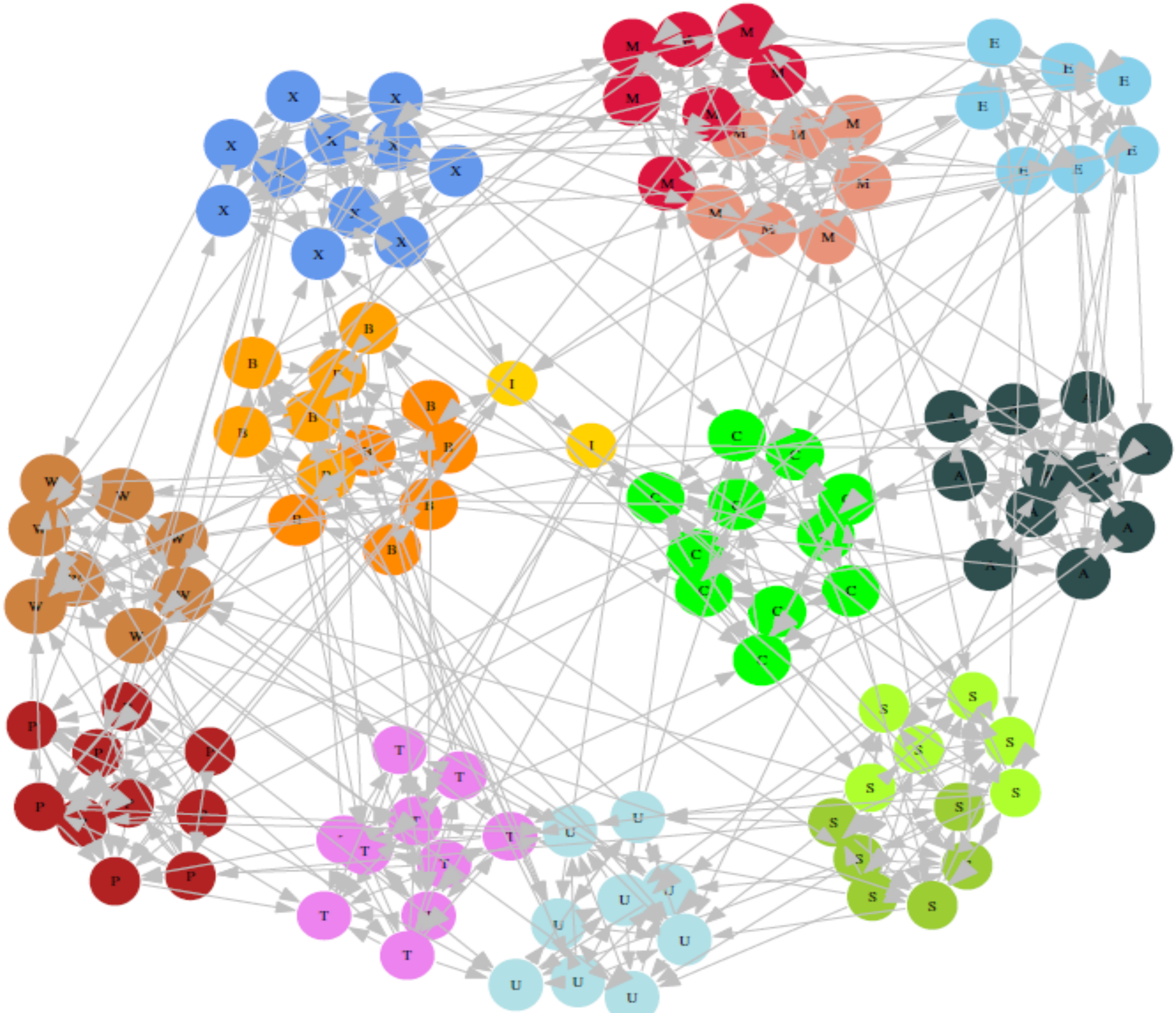
The strength of weak ties

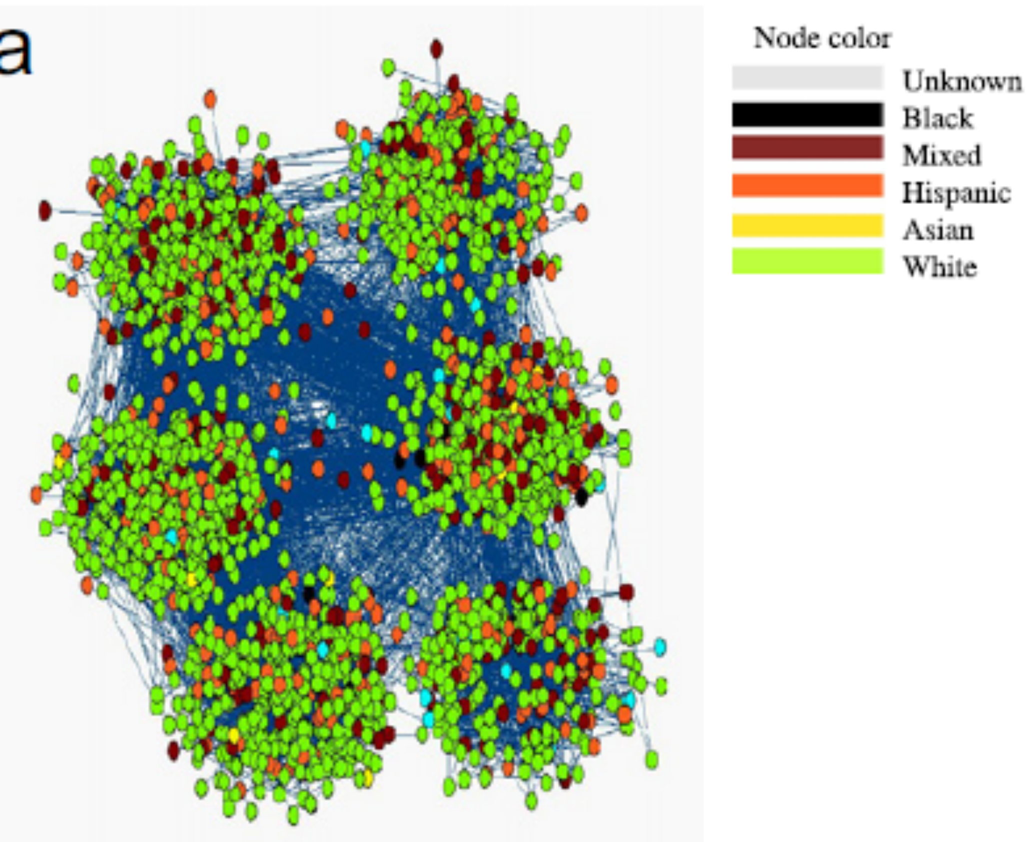
- Mark S. **Granovetter**, 1973
- His PhD thesis: how people get to know about new jobs?
- Through personal contacts
- Surprise: often acquaintances, **not** close friends
- Why?

The Strength of Weak Ties

Mark S. Granovetter

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





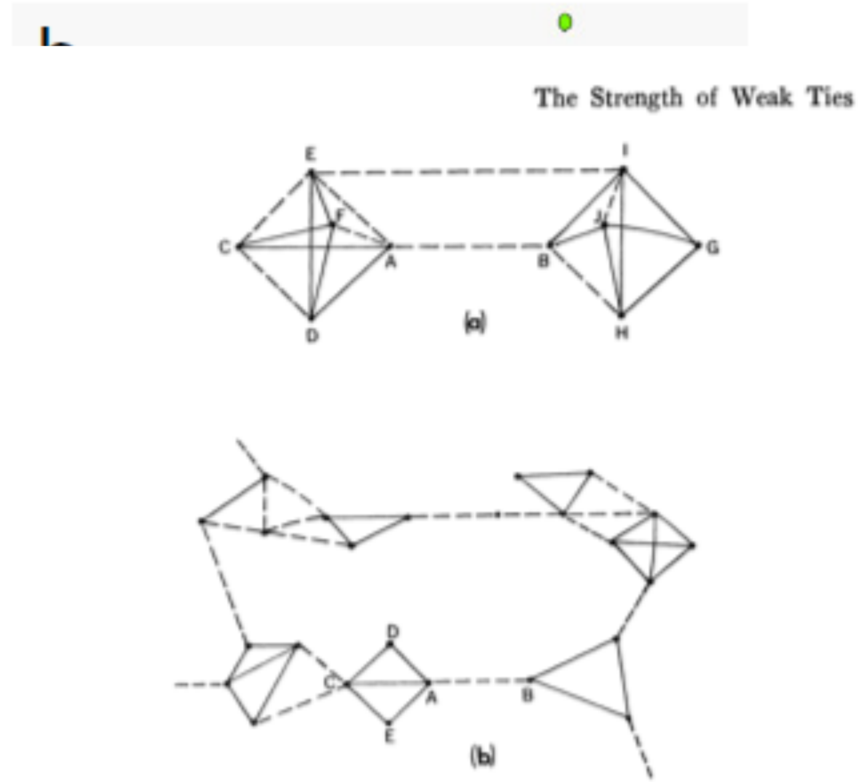
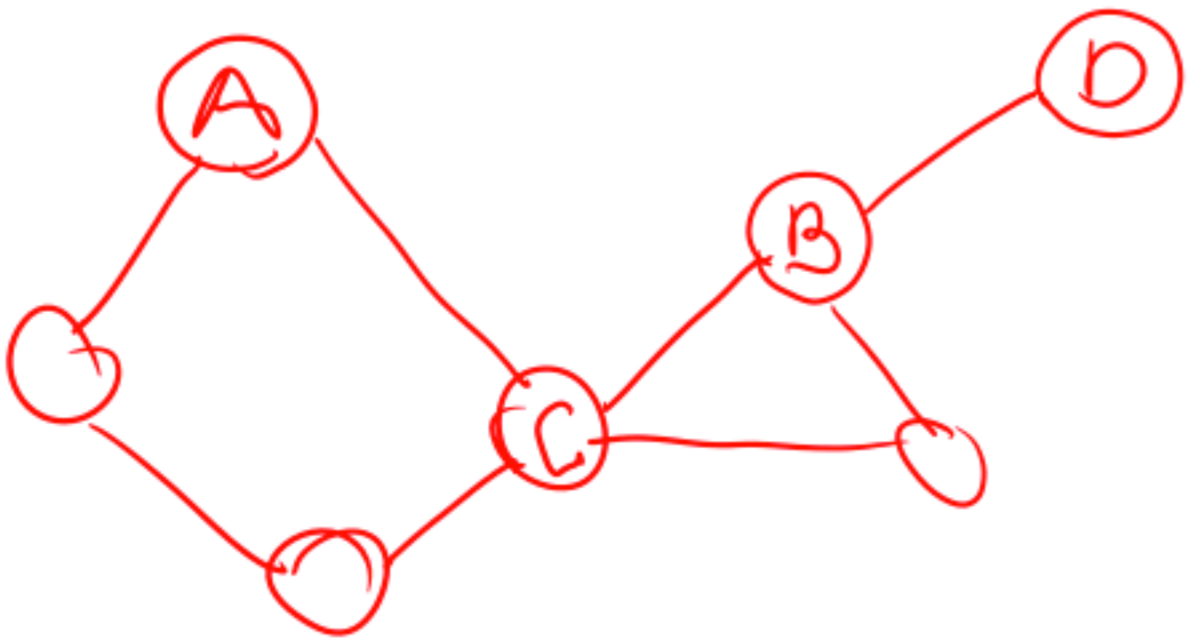


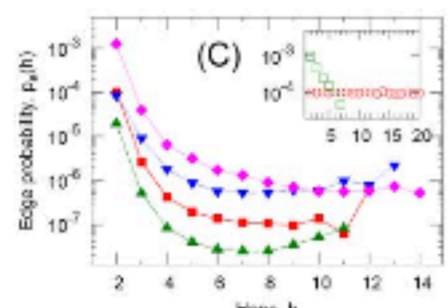
FIG. 2.—Local bridges. *a*, Degree 3; *b*, Degree 13. ——— = strong tie; – – – = weak tie.

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:

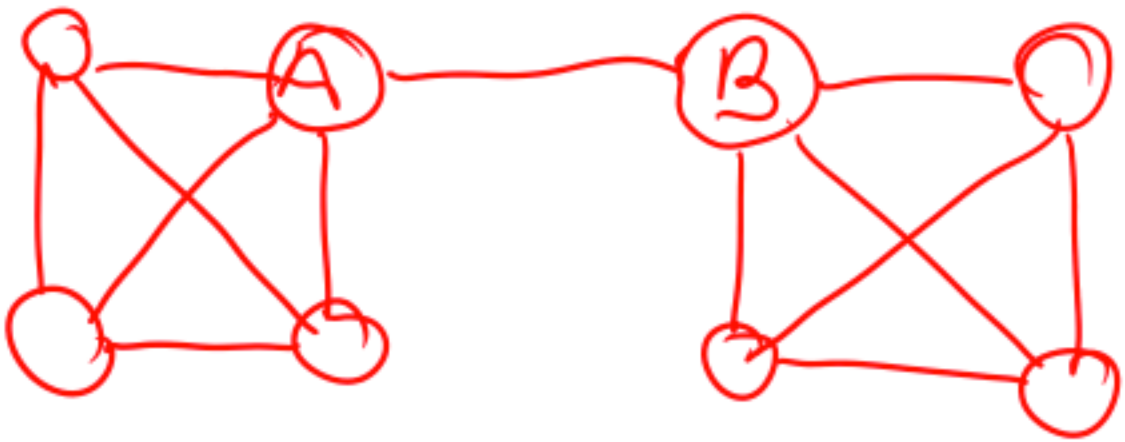
- If B and C have a friend A in common, then:
  - B is more likely to meet C
    - (since they both spend time with A)
  - B and C trust each other
    - (since they have a friend in common)
  - A has incentive to bring B and C together
    - (as it is hard for A to maintain two disjoint relationships)

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)

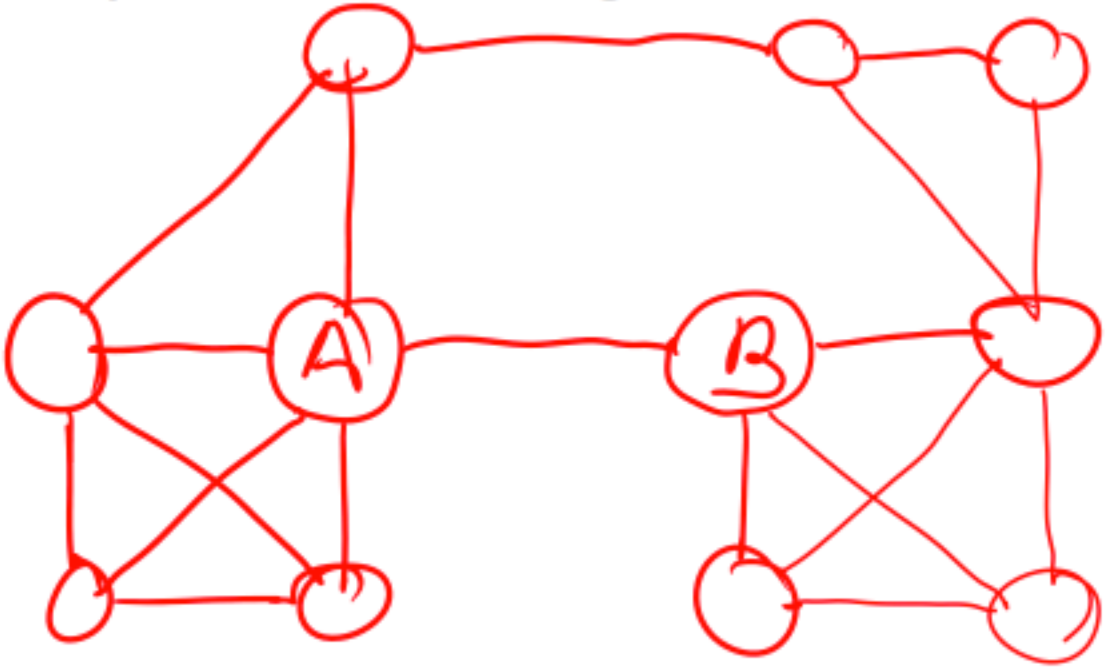
Bridges and Local Bridges

- 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)

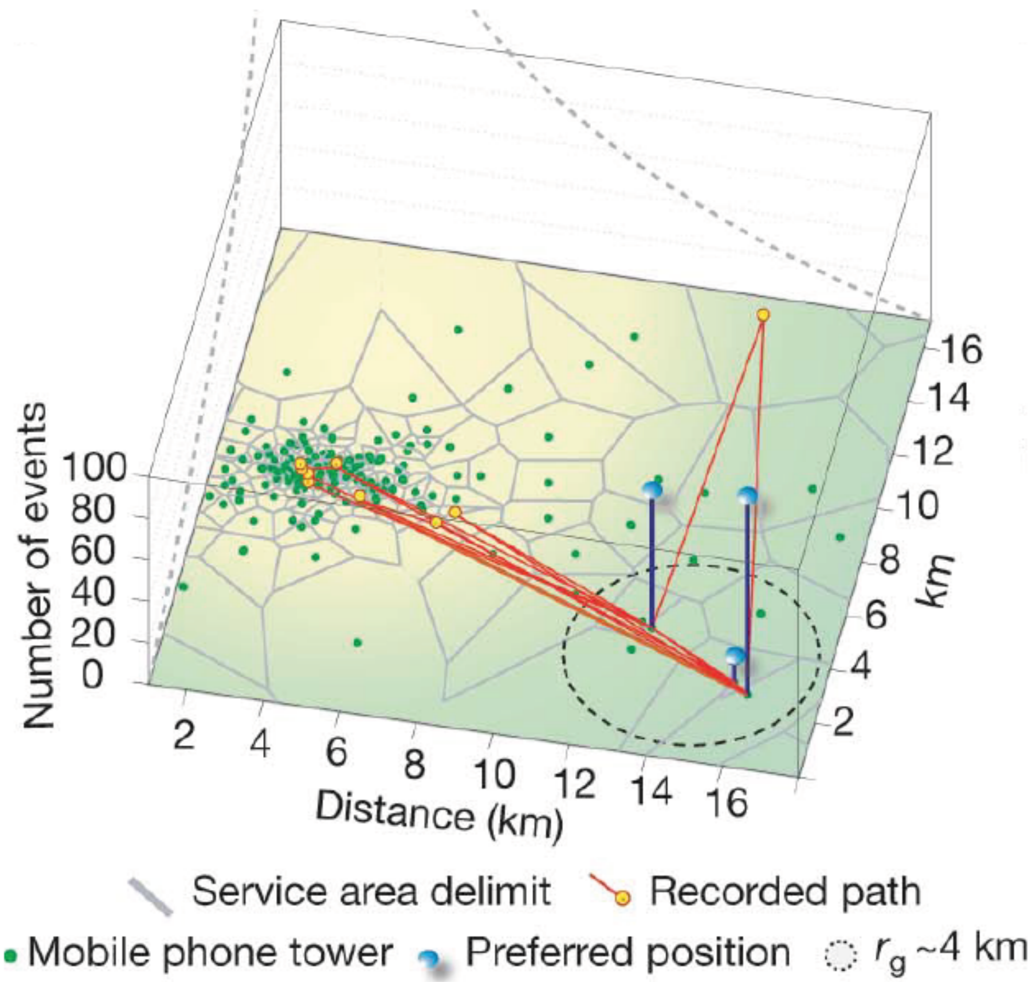
Local Bridges and Weak ties

- 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

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

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).

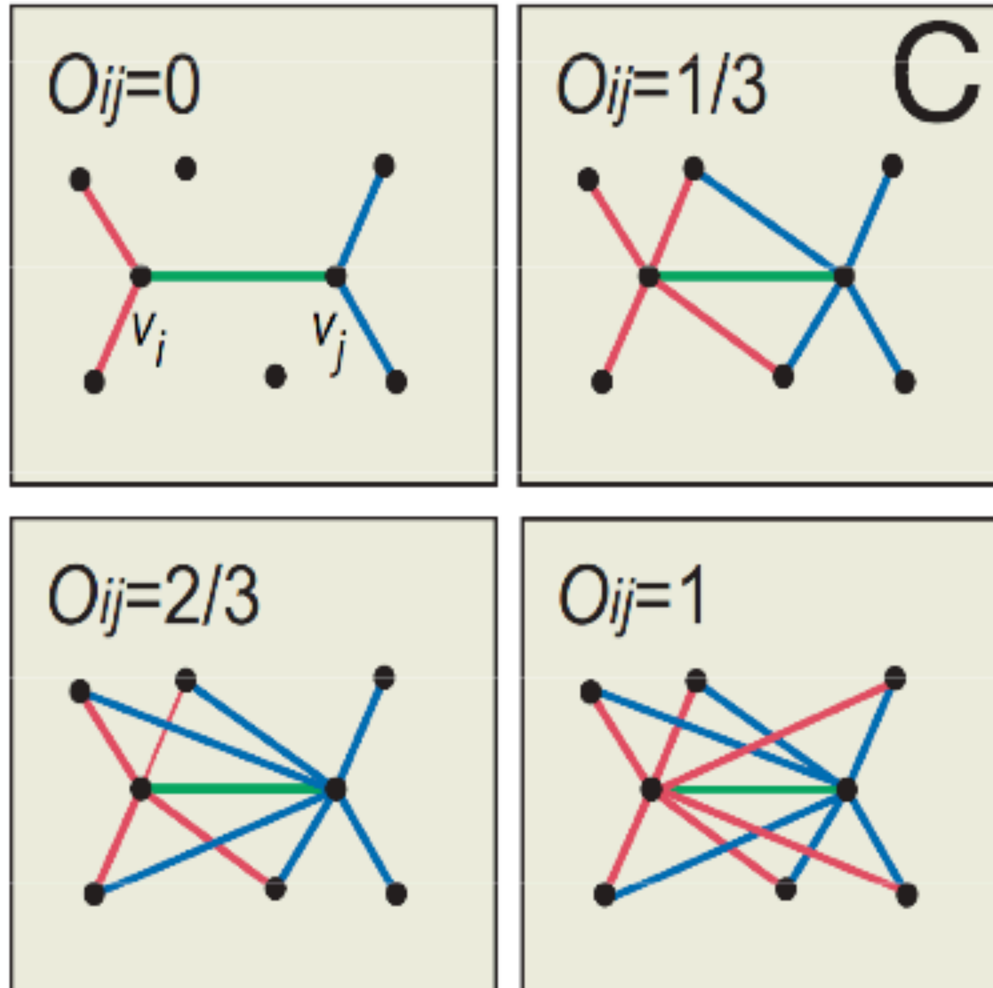
Neighborhood Overlap

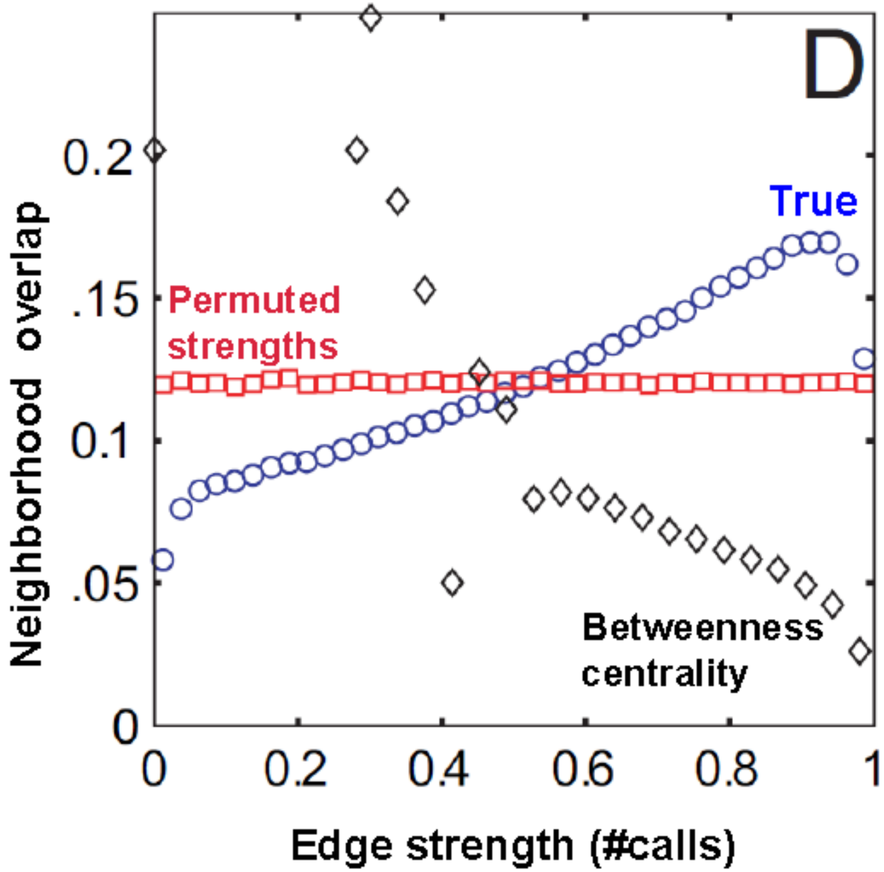
- Overlap:

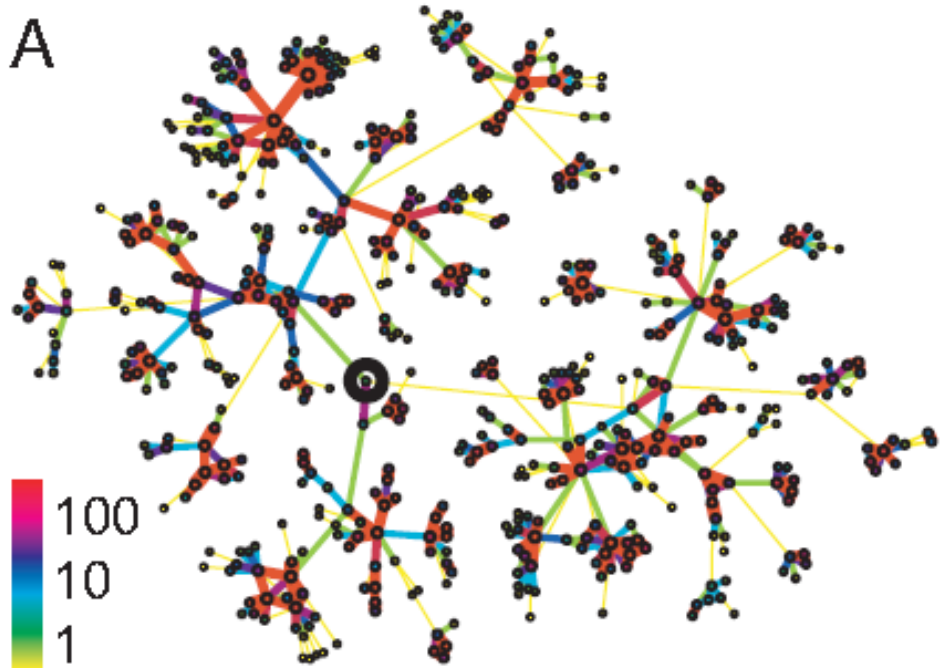
  $$O_{ij} = \frac{n(i) \cap n(j)}{n(i) \cup n(j)}$$

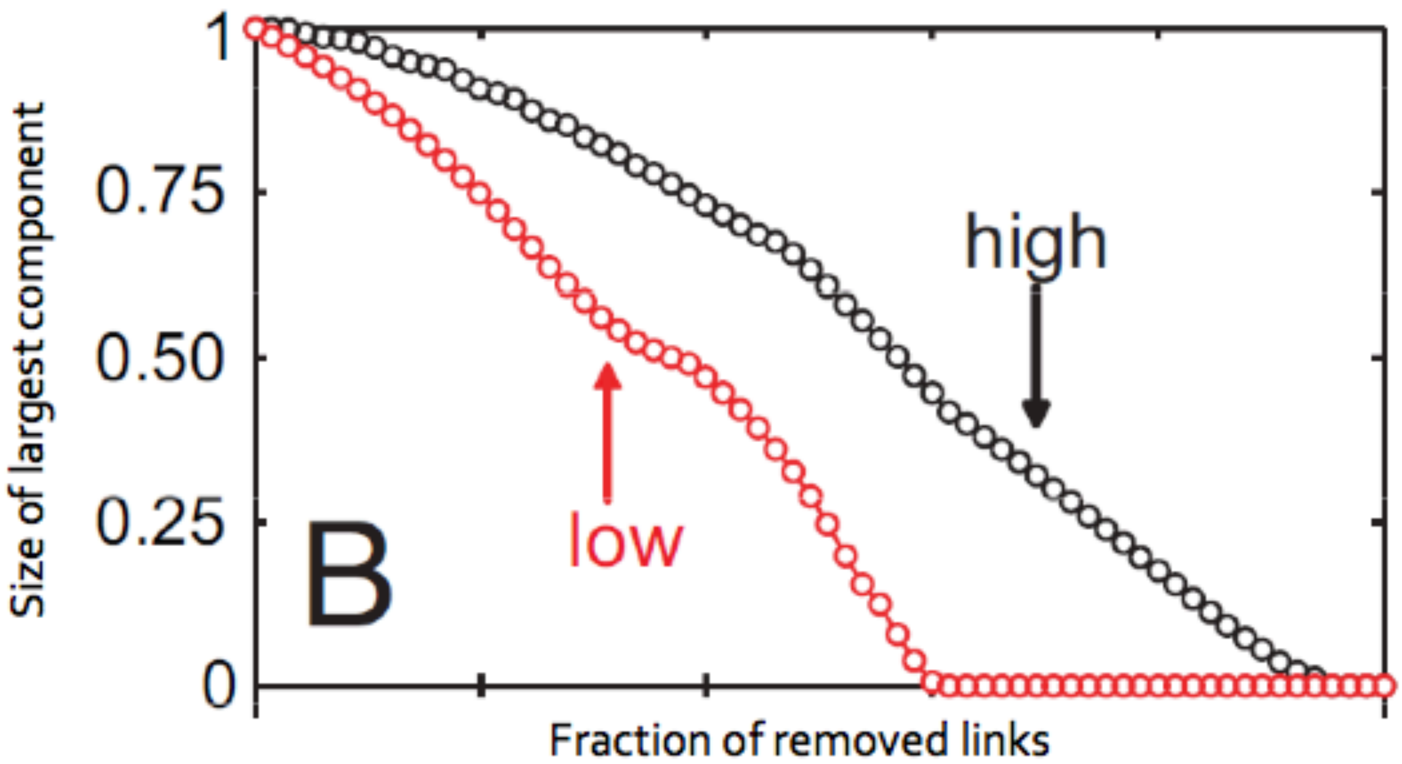
  - n(i) … set of neighbors of A

- Overlap = 0 when an edge is a local bridge




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


A
100
10
1

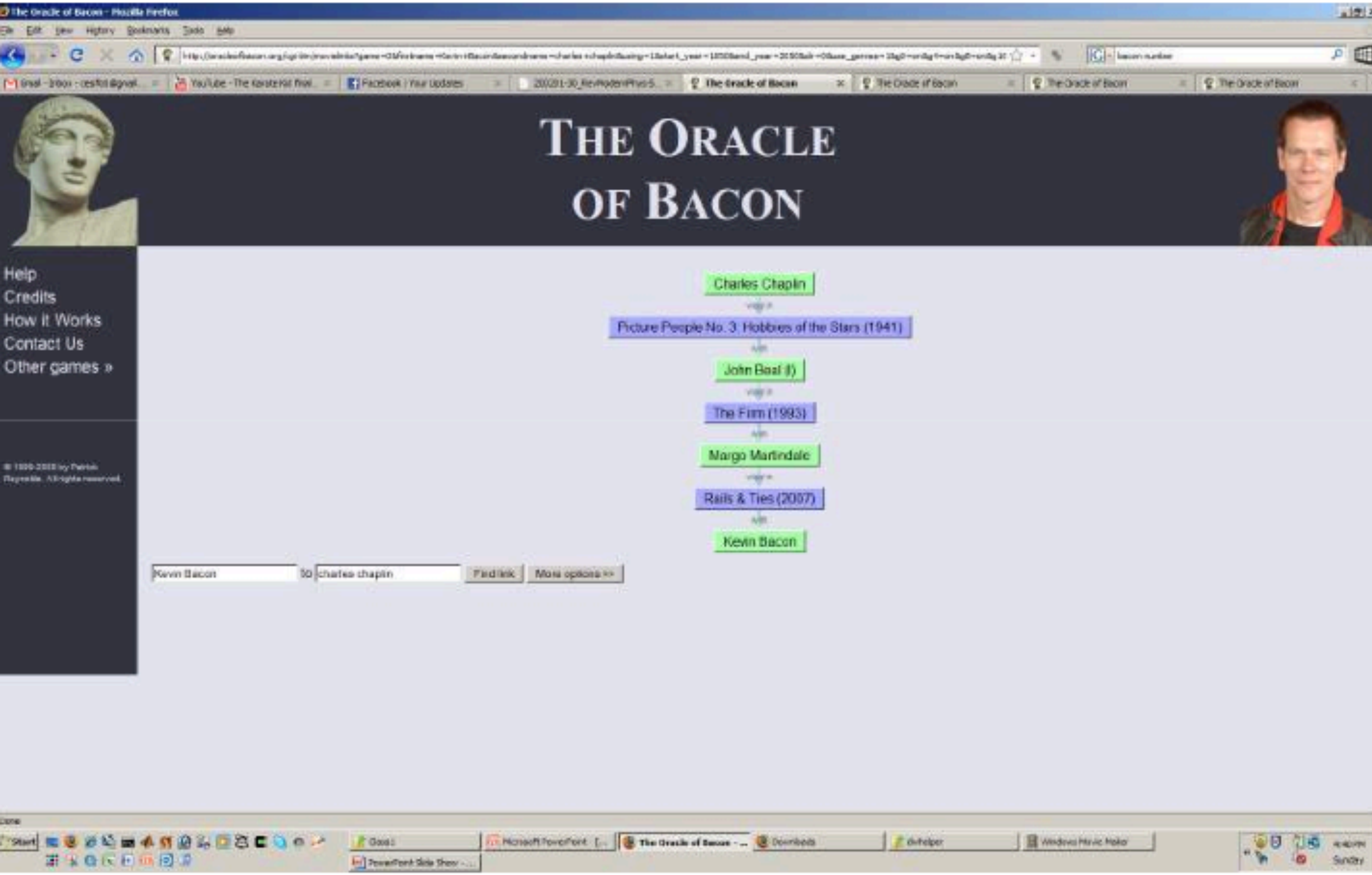


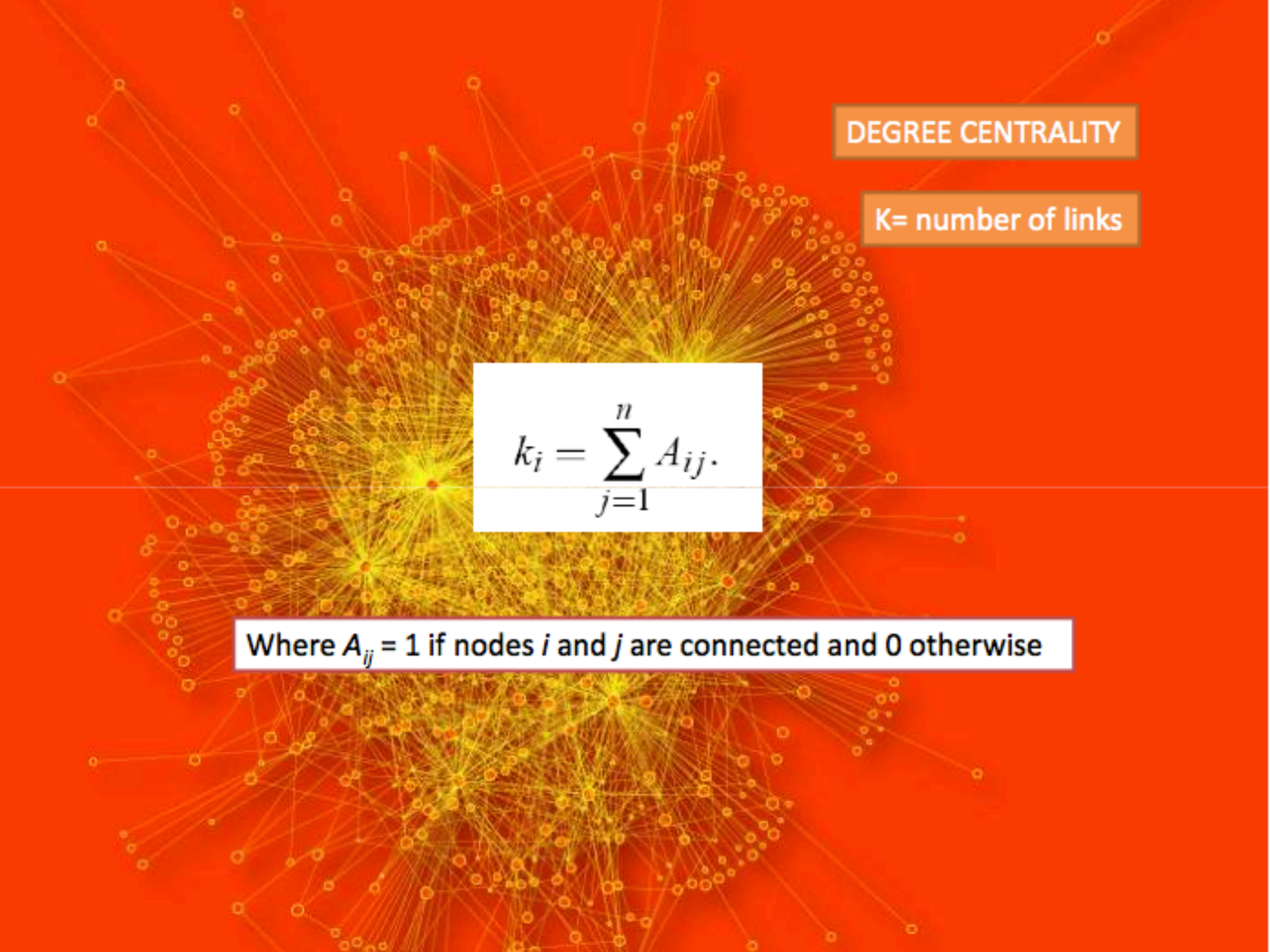
- Removing links based on **overlap**
  - Low to high
  - High to low

Seminar 4

Centrality

How important is 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 >>



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

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

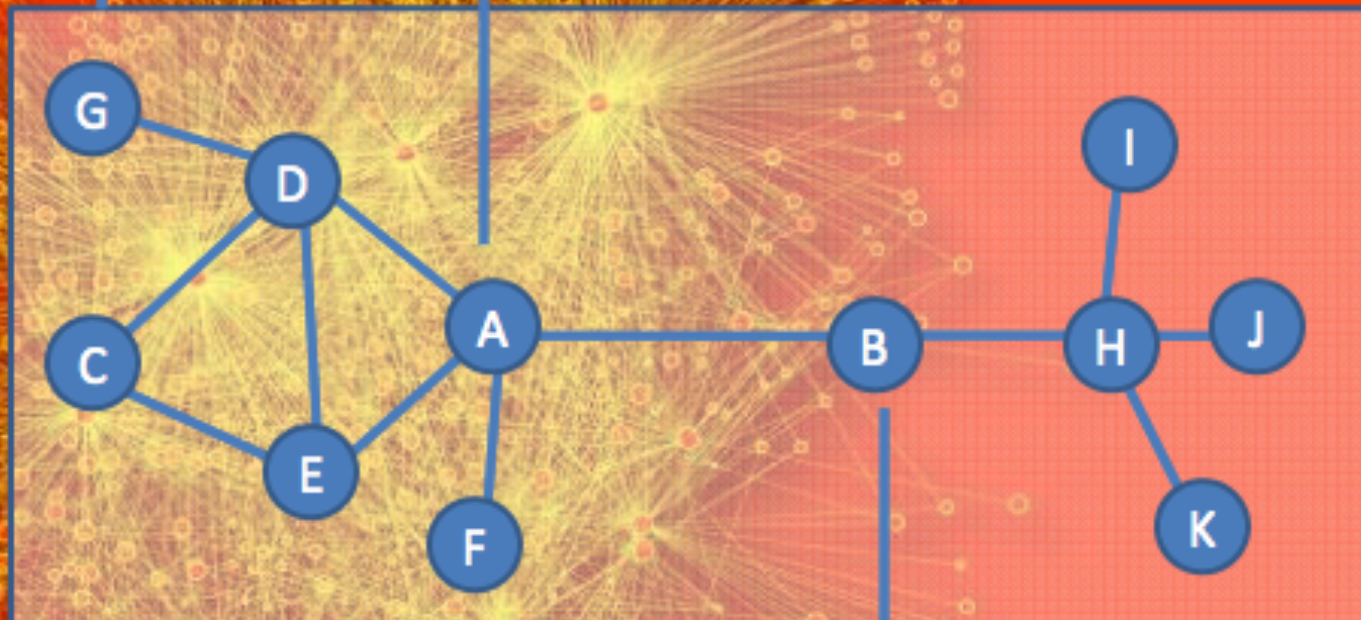
A-L Barabasi, "Linked", 2002

CLOSENESS CENTRALITY

C= Average Distance to neighbors

C(G)=1/10(1+2*3+2*3+4+3*5)
C(G)=3.2

C(A)=1/10(4+2*3+3*3)
C(A)=1.9

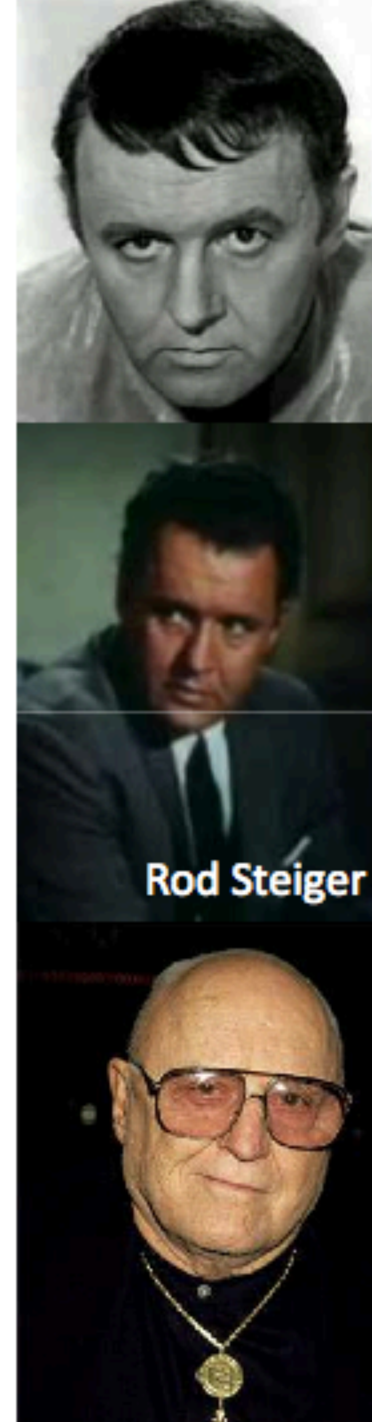


C(B)=1/10(2+2*6+2*3)
C(B)=2

N=11

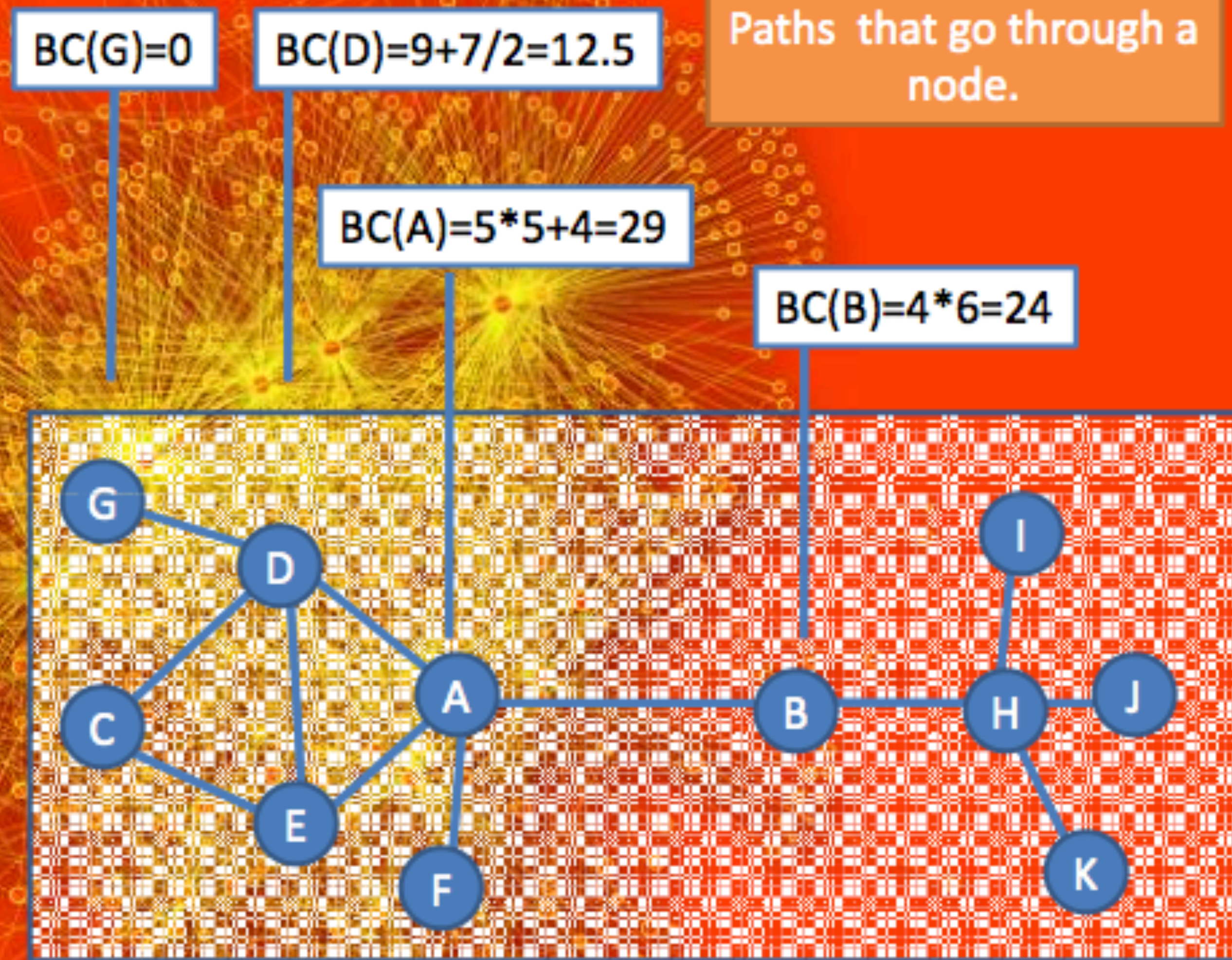
Hollywood Revolves Around

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)



BETWENNESS CENTRALITY

BC= number of shortest Paths that go through a node.



N=11

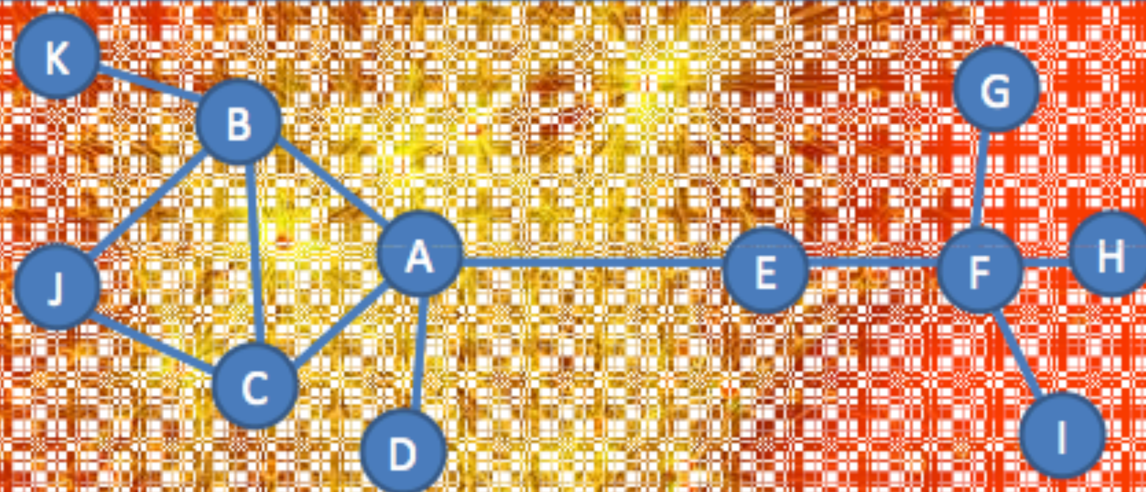
A set of measures of centrality based on betweenness

LC Freeman - Sociometry, 1977 - jstor.org

PAGE RANK

PR=Probability that a random walker with interspersed Jumps would visit that node.

PR=Each page votes for its neighbors.



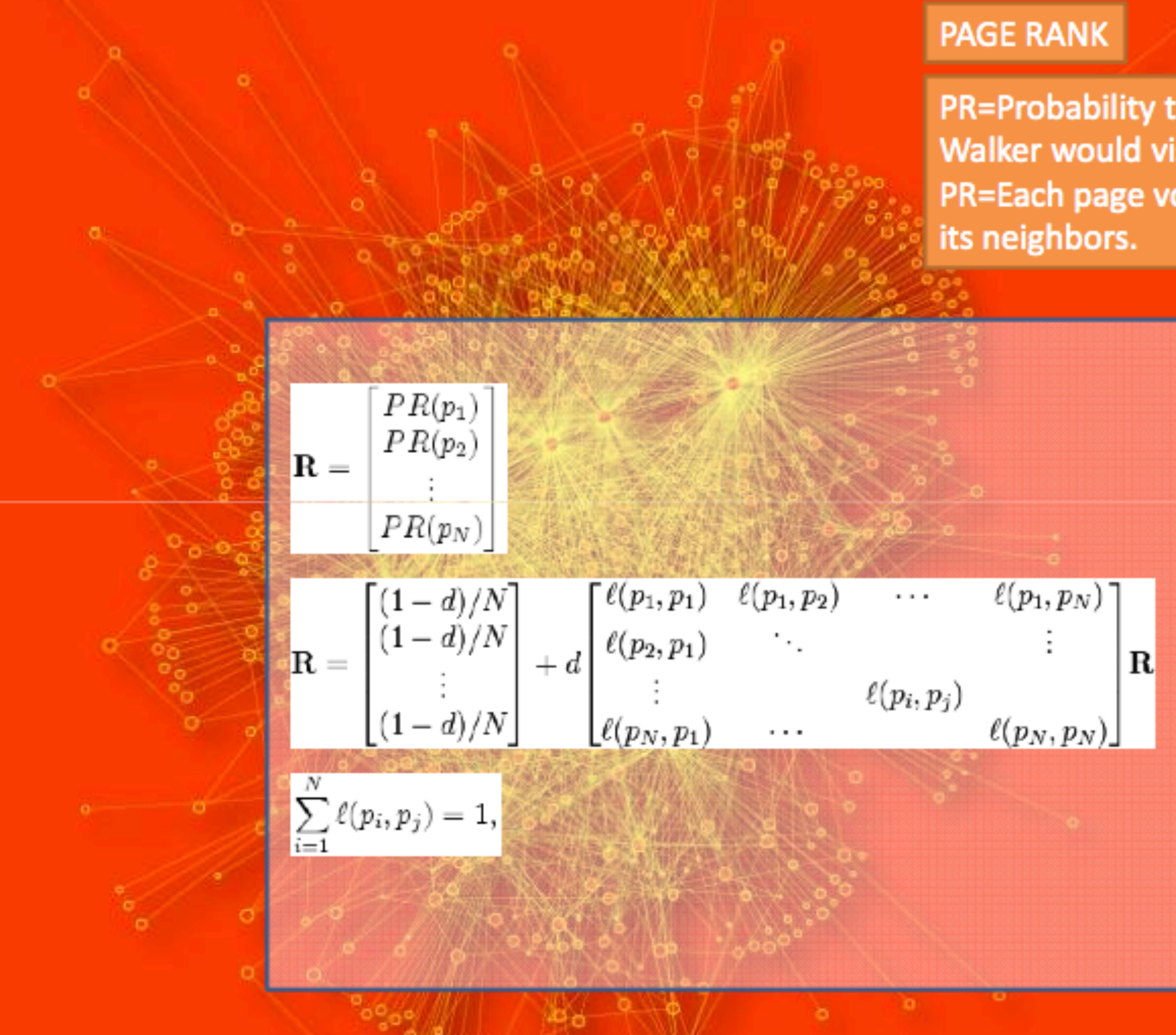
PR(A)=PR(B)/4 + PR(C)/3 + PR(D)+PR(E)/2

A random surfer eventually stops clicking

PR(X)=(1-d)/N + d(ΣPR(y)/k(y))

PAGE RANK

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

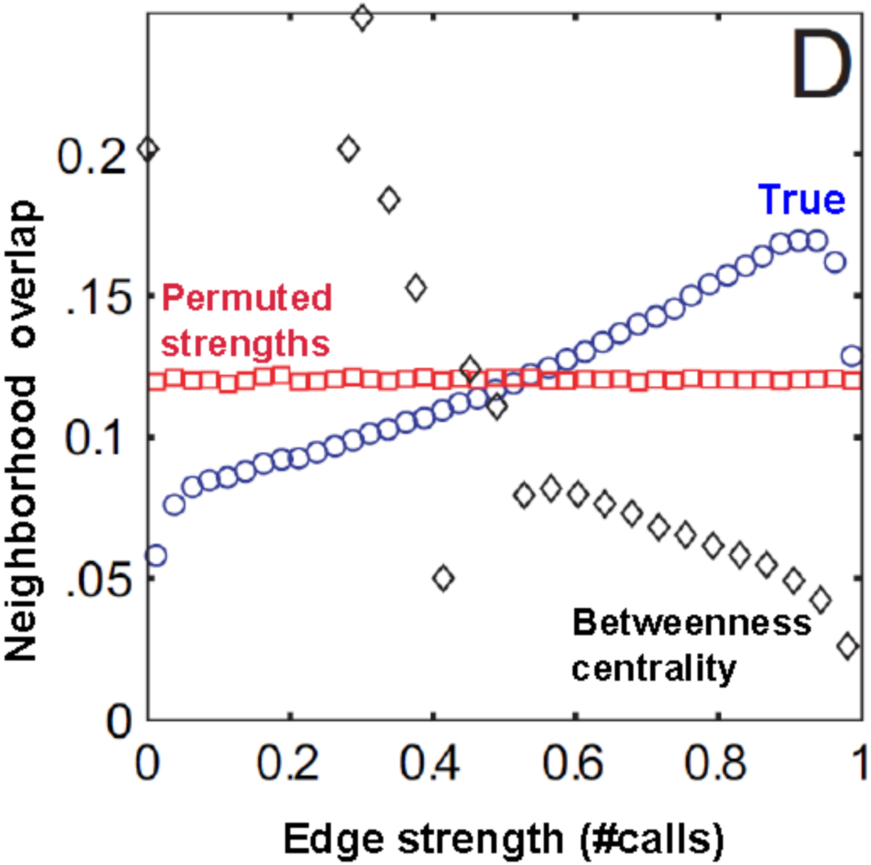


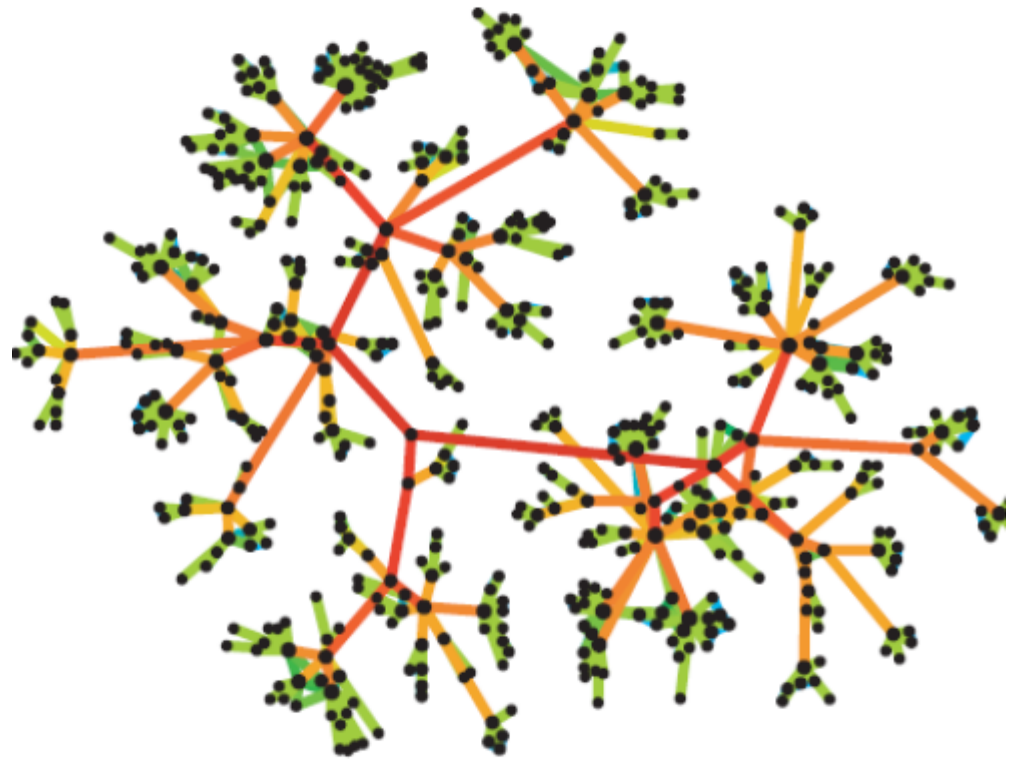
$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

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

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

Back to Granovetter





Human mobility, social ties and link prediction

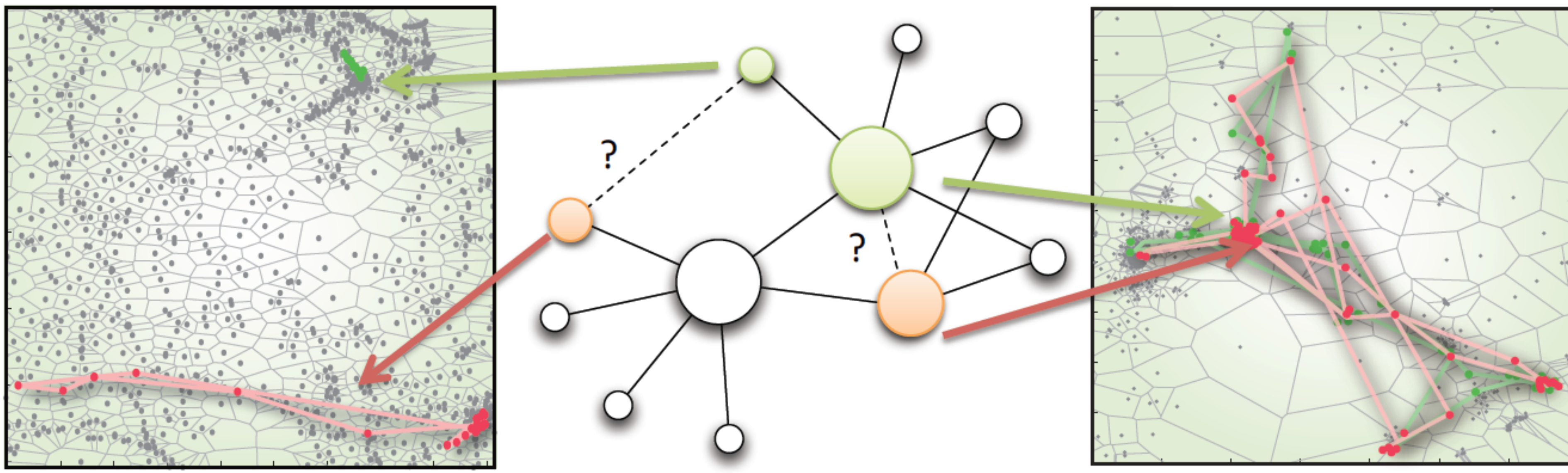
Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti,
Albert-Lászlo Barabási

SIGKDD Int. Conf. on Knowledge Discovery and Data Mining – KDD 2011

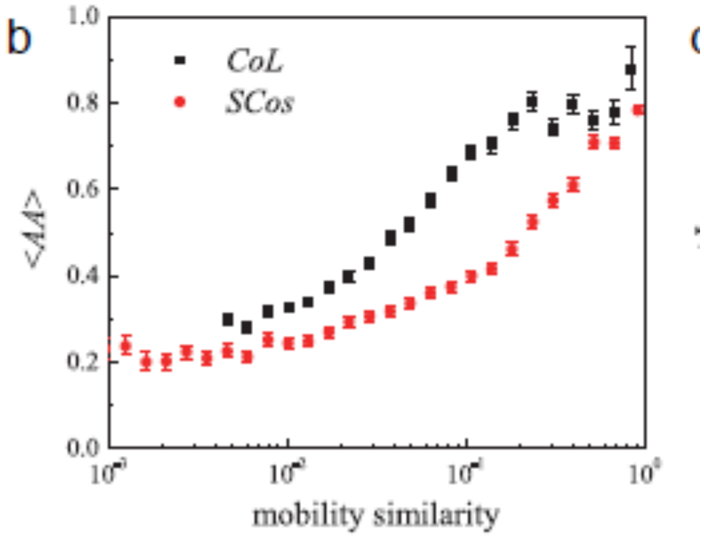
Colocation, social proximity, tie strength

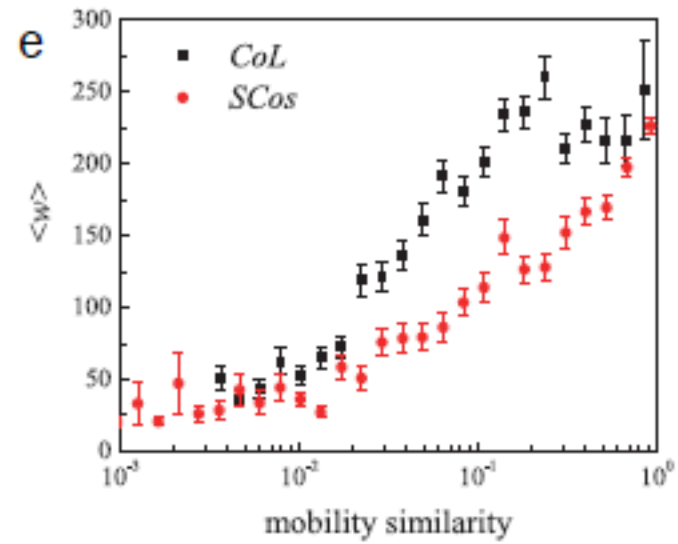
- How similar is the movement of users u and v
  - Various **co-location measures**, quantifying the similarity between the movement routines of u and v (mobile homophily)
- How connected are u and v in the social network.
  - Various well-established **measures of network proximity**, based on the common 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**

Network proximity vs. mobile homophily

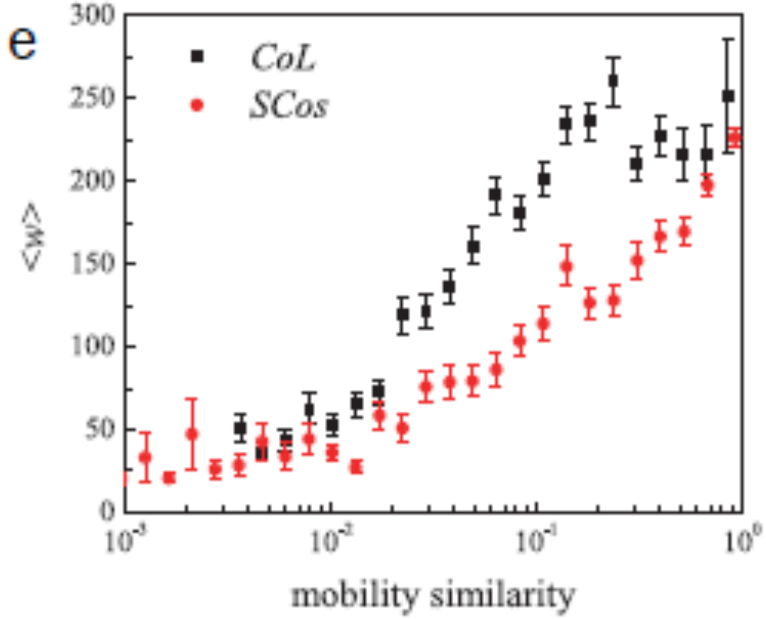


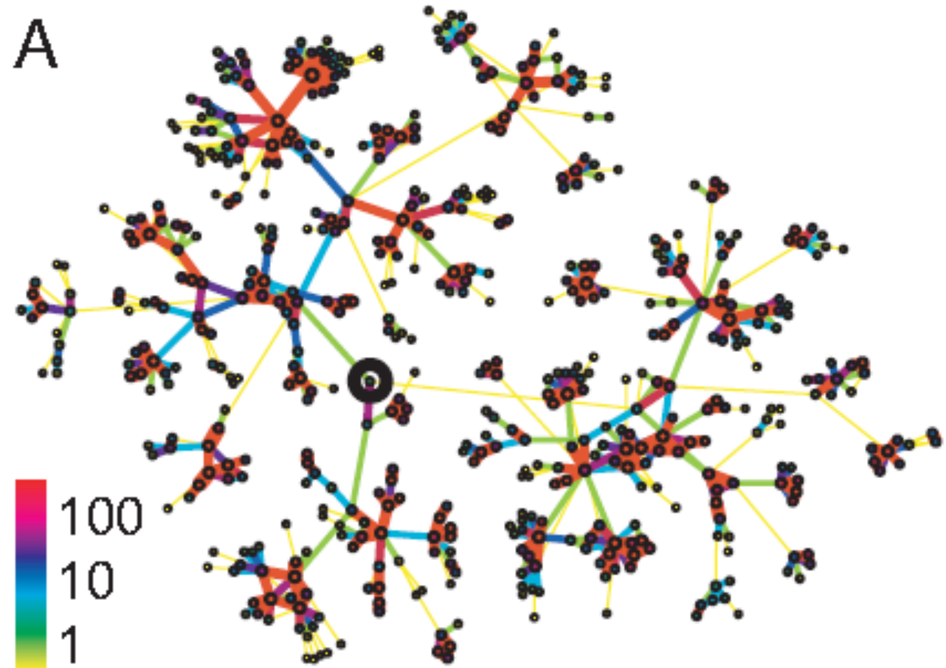
mobility dimension of the "strength of weak ties"





- co-location, network proximity and tie strength strongly correlate with each other
- measured on 3 months of calls, 6 Million users, nation-wide (large European country)

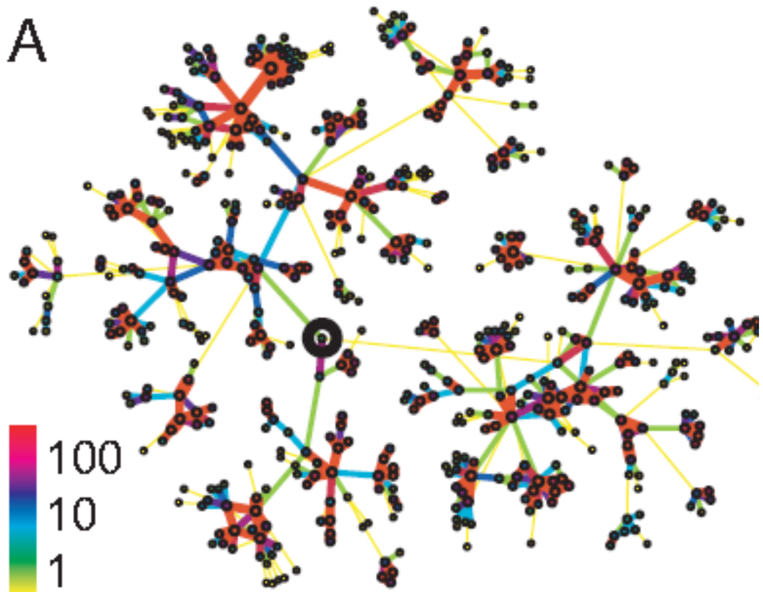

e
CoL
SCos
<w>
mobility similarity
300
250
200
150
100
50
0
10^{-3}
10^{-2}
10^{-1}
10^{0}

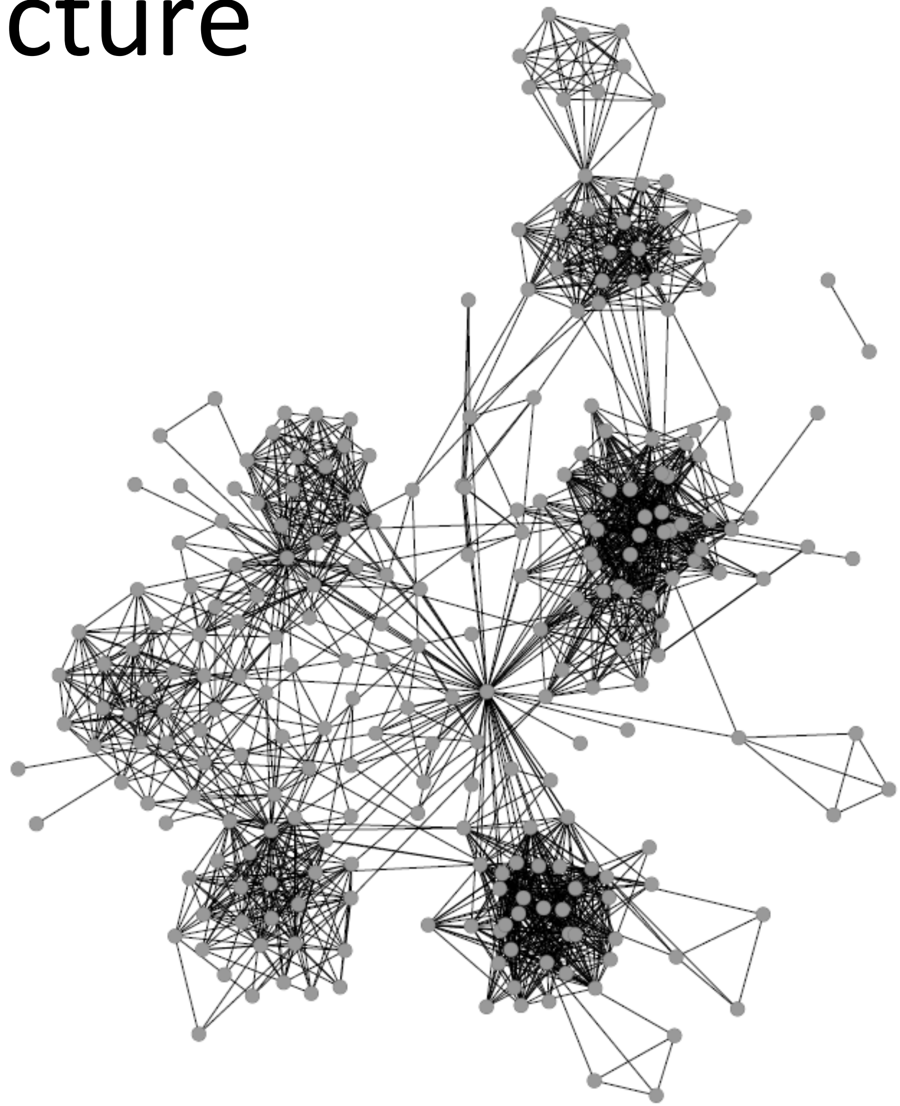

A
100
10
1

Community discovery

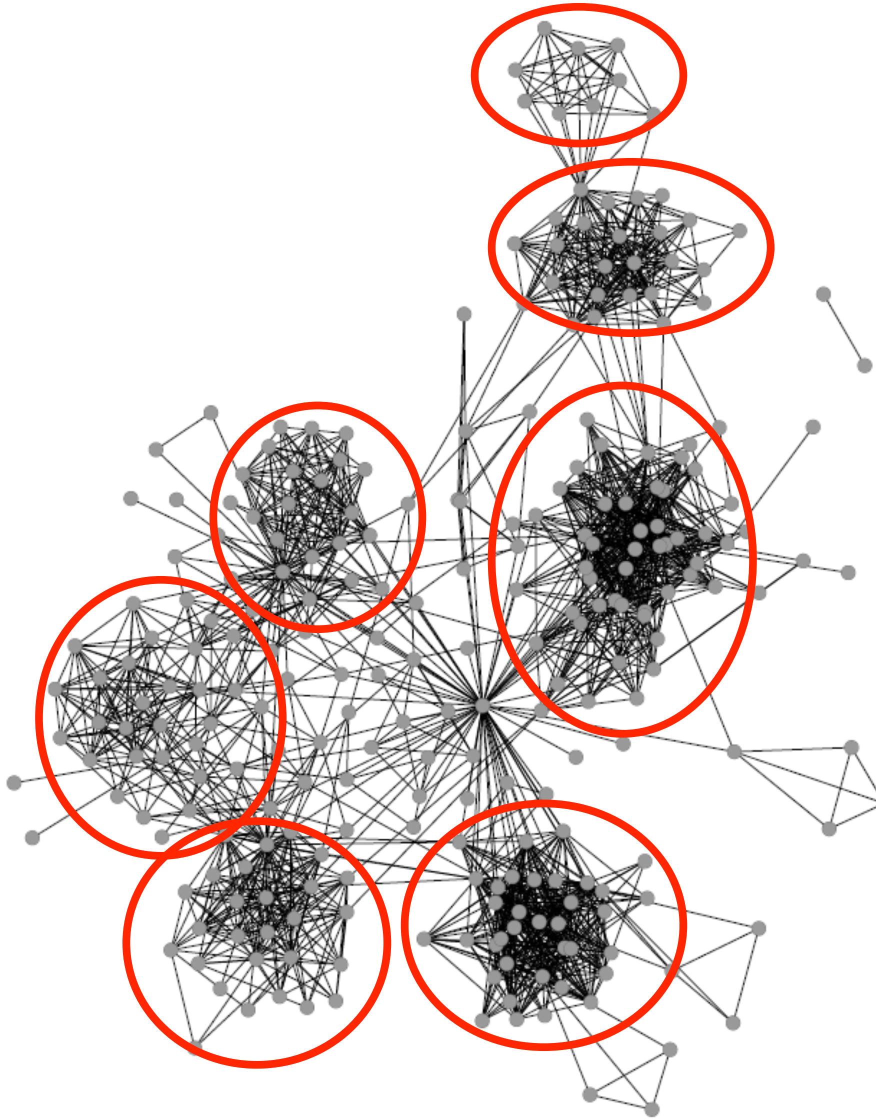
How to highlight the modular structure of a network?

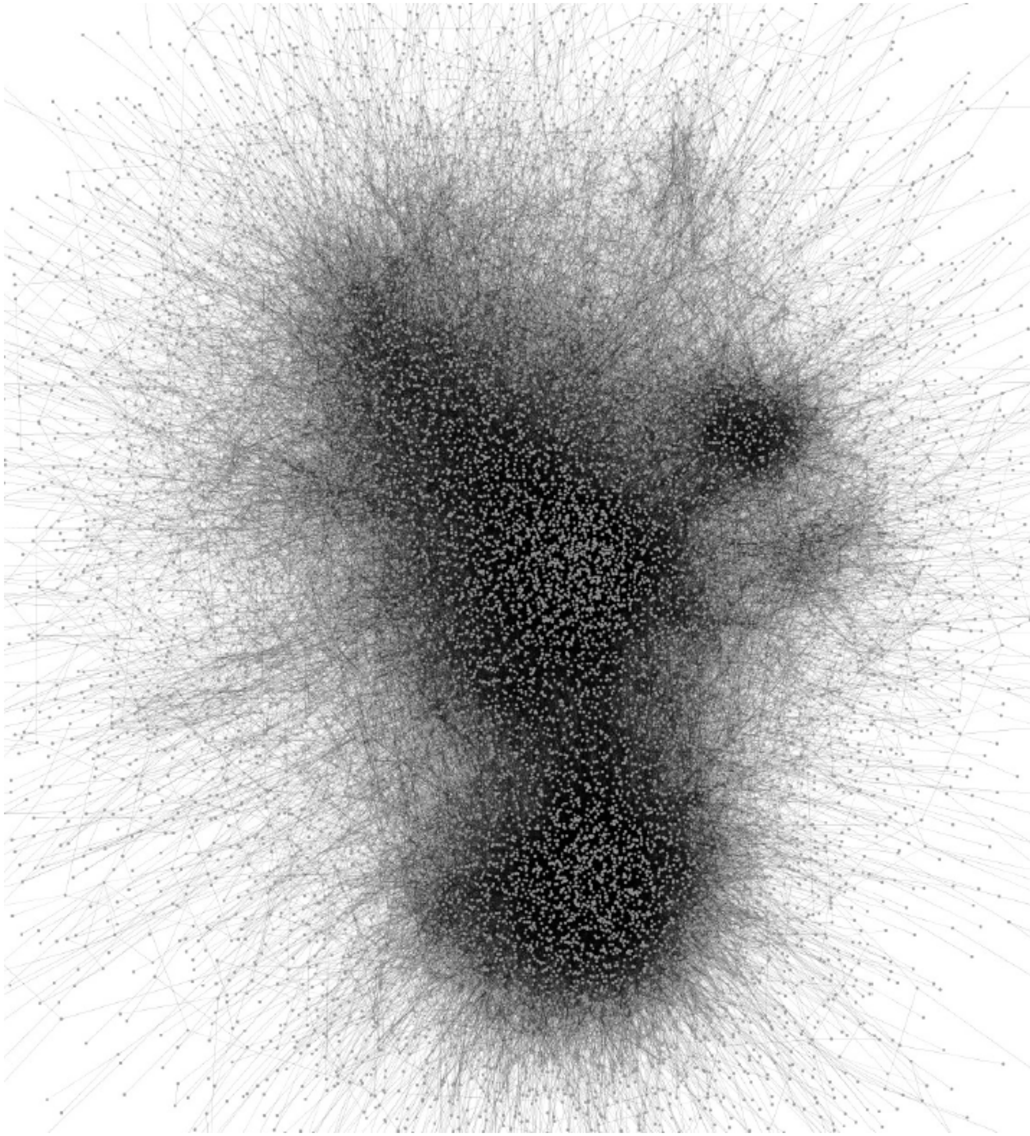
Community structure

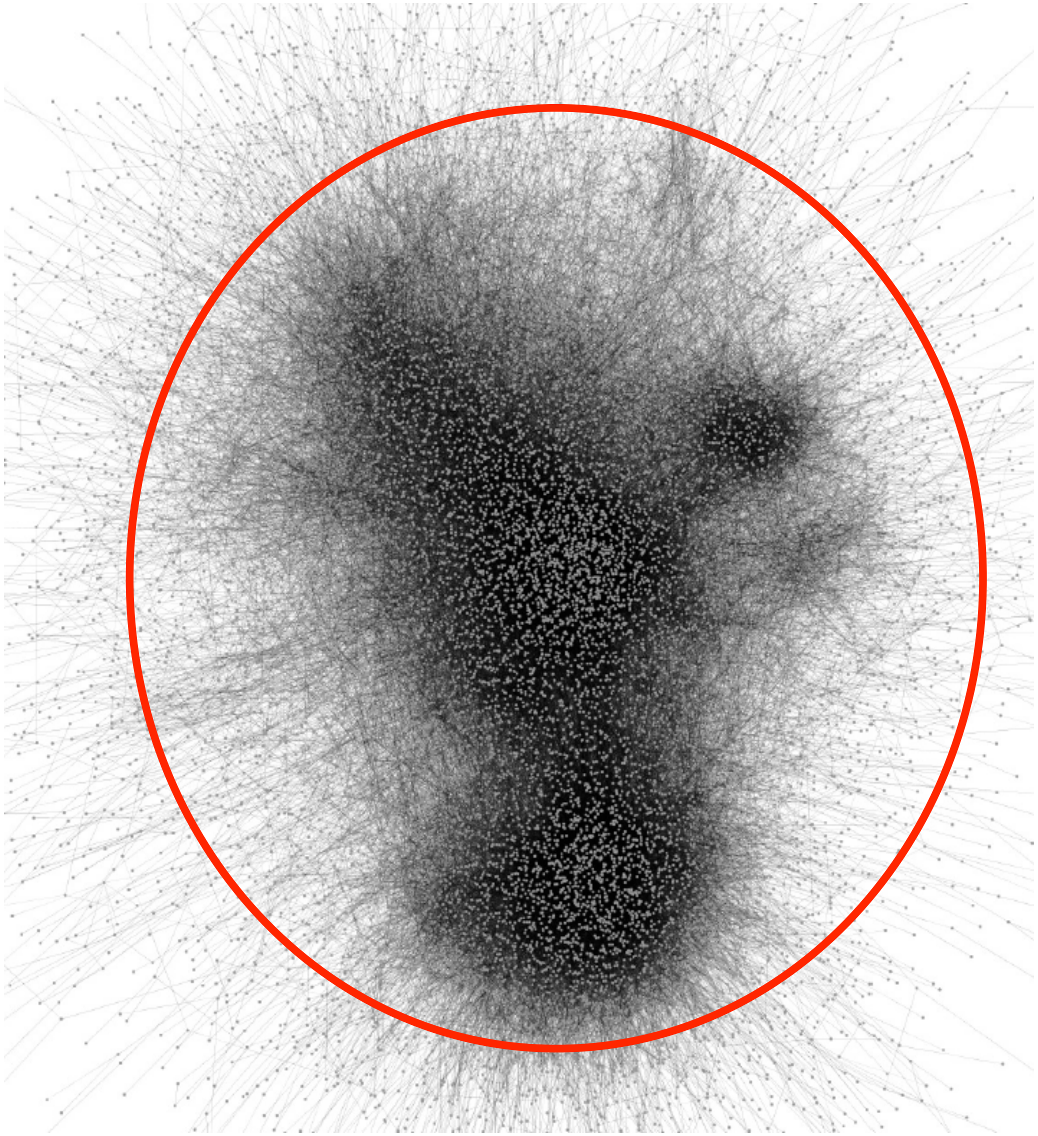




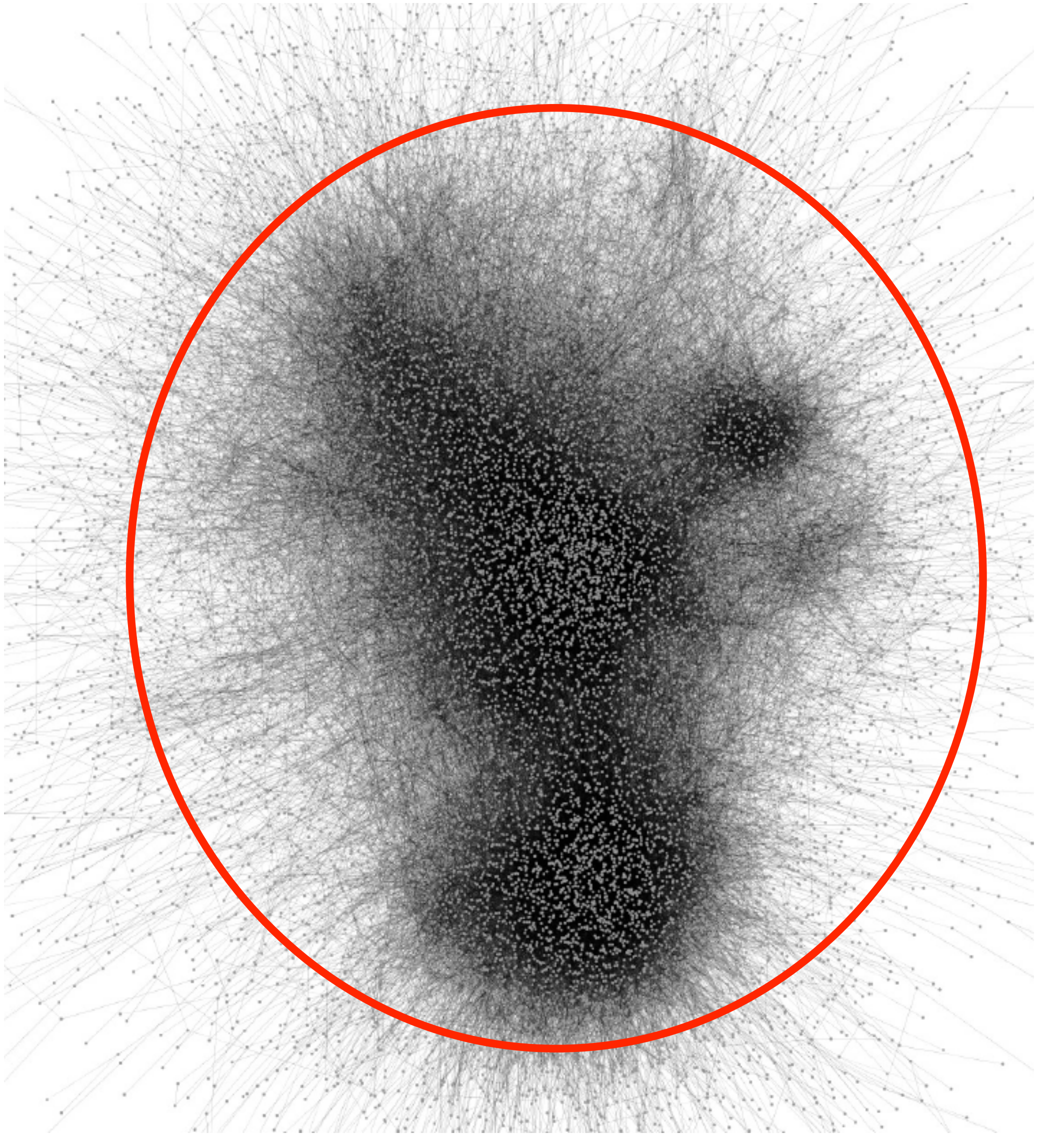
Communities



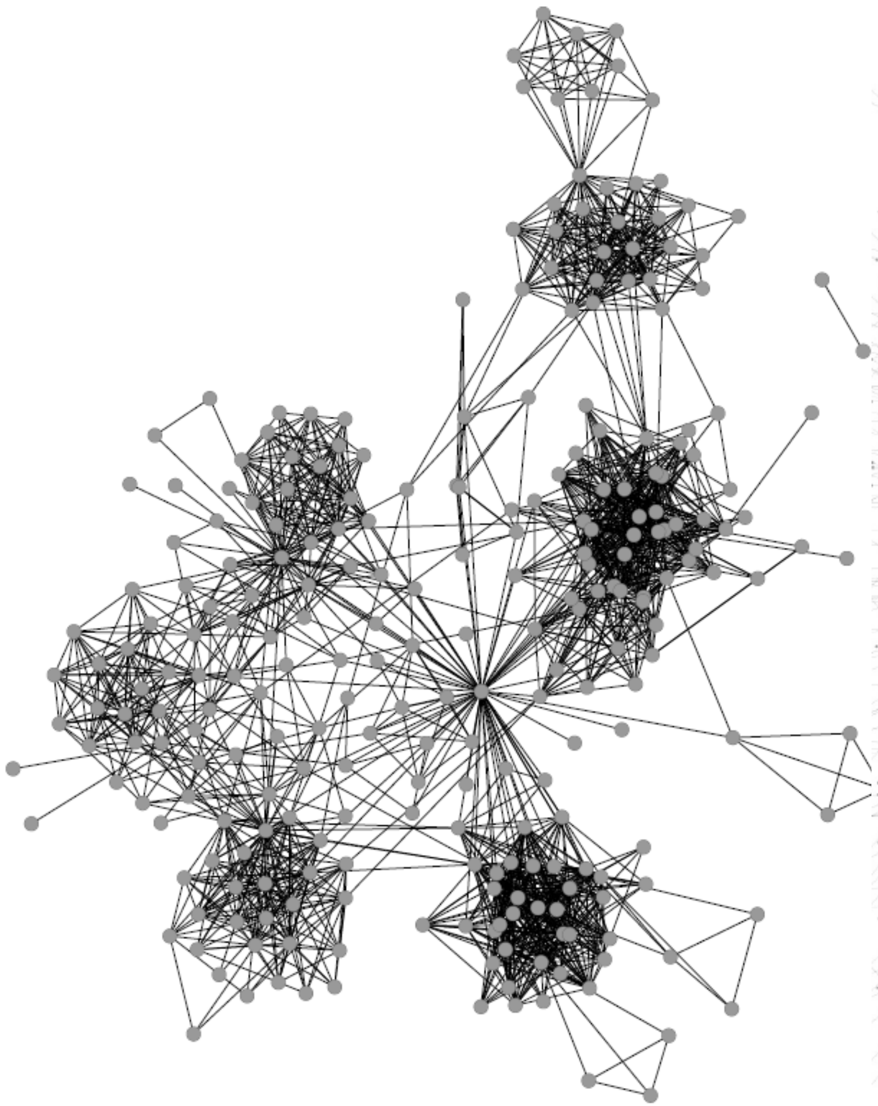


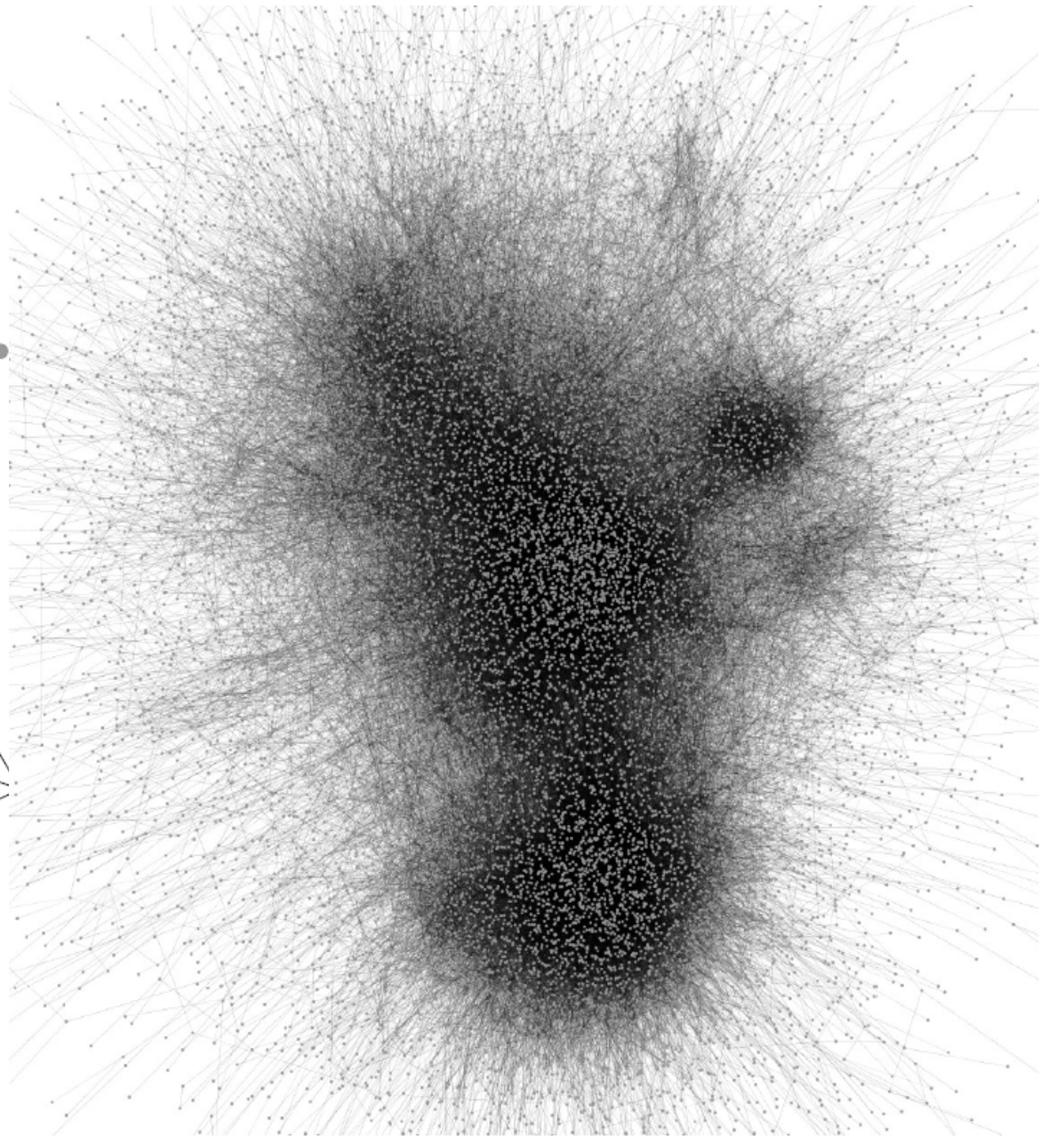


?

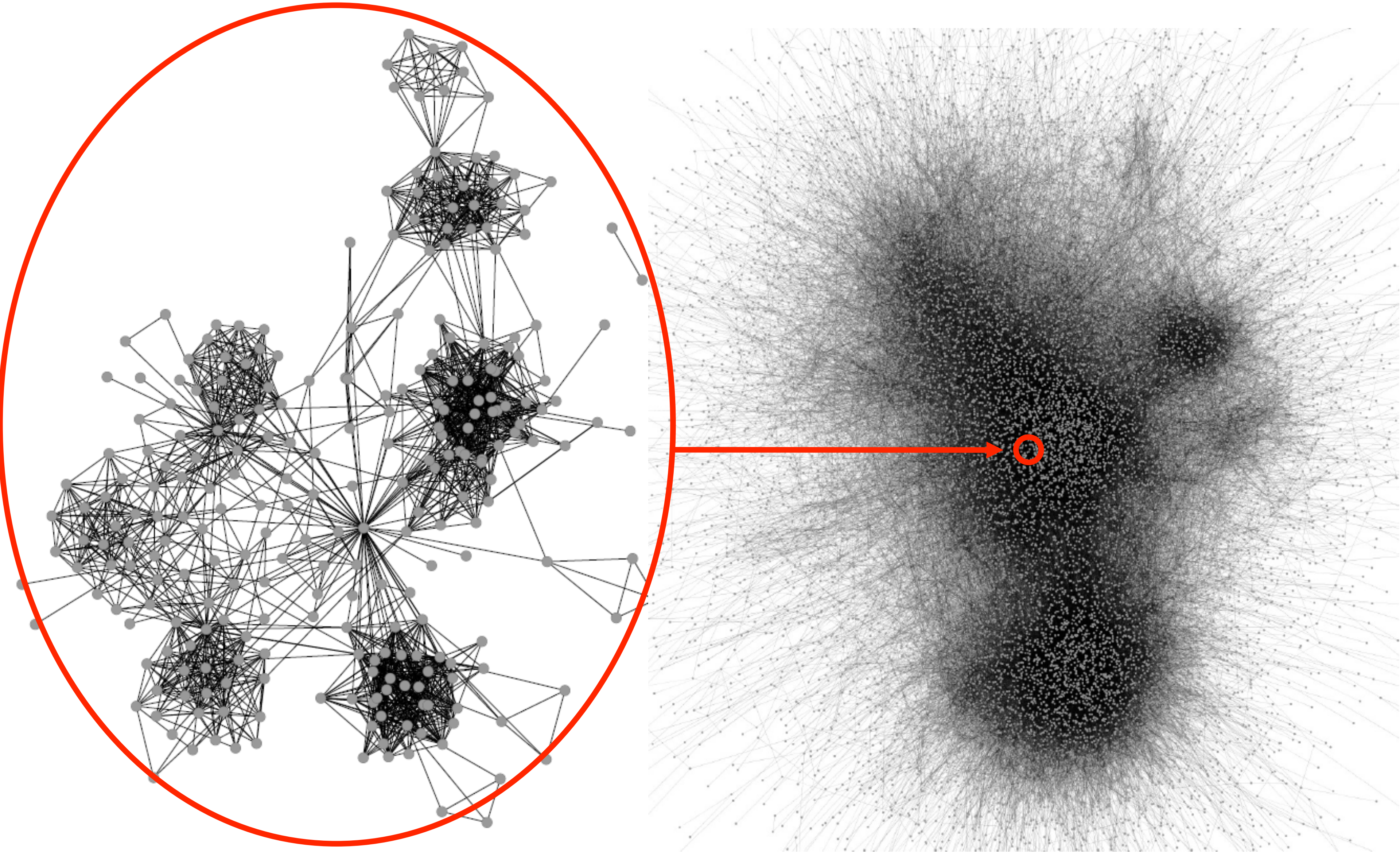


Are these two different networks?





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

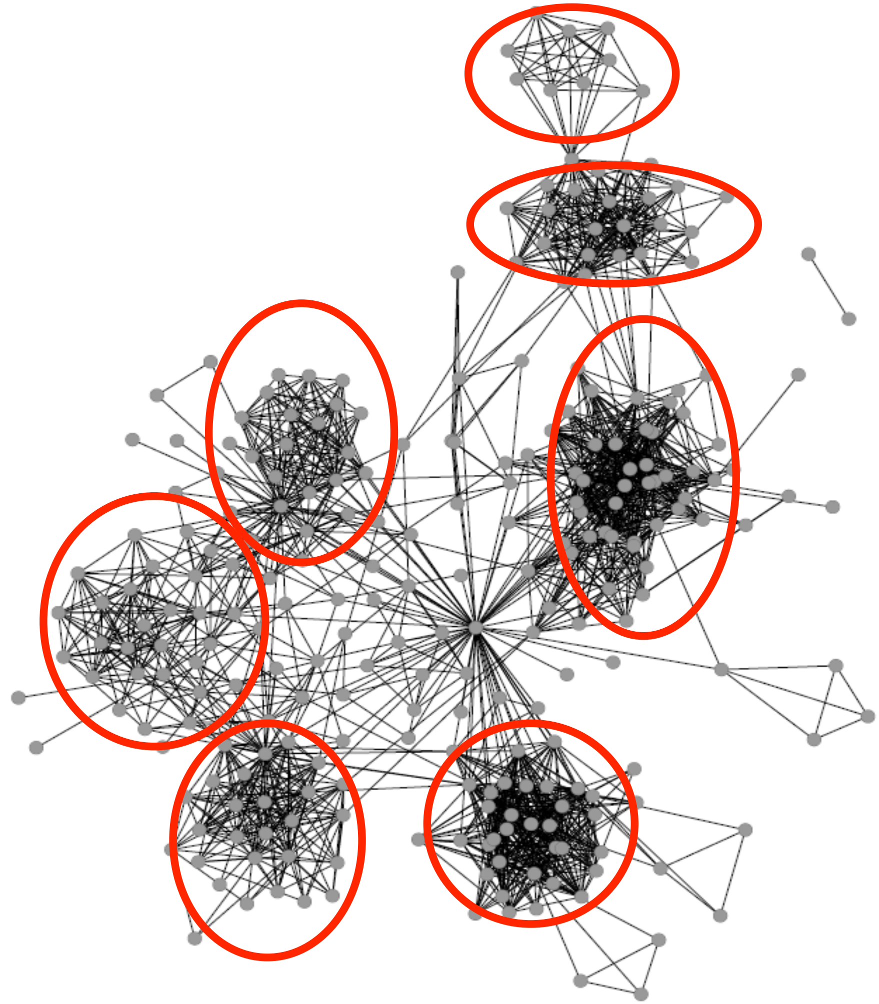
DEMON

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..



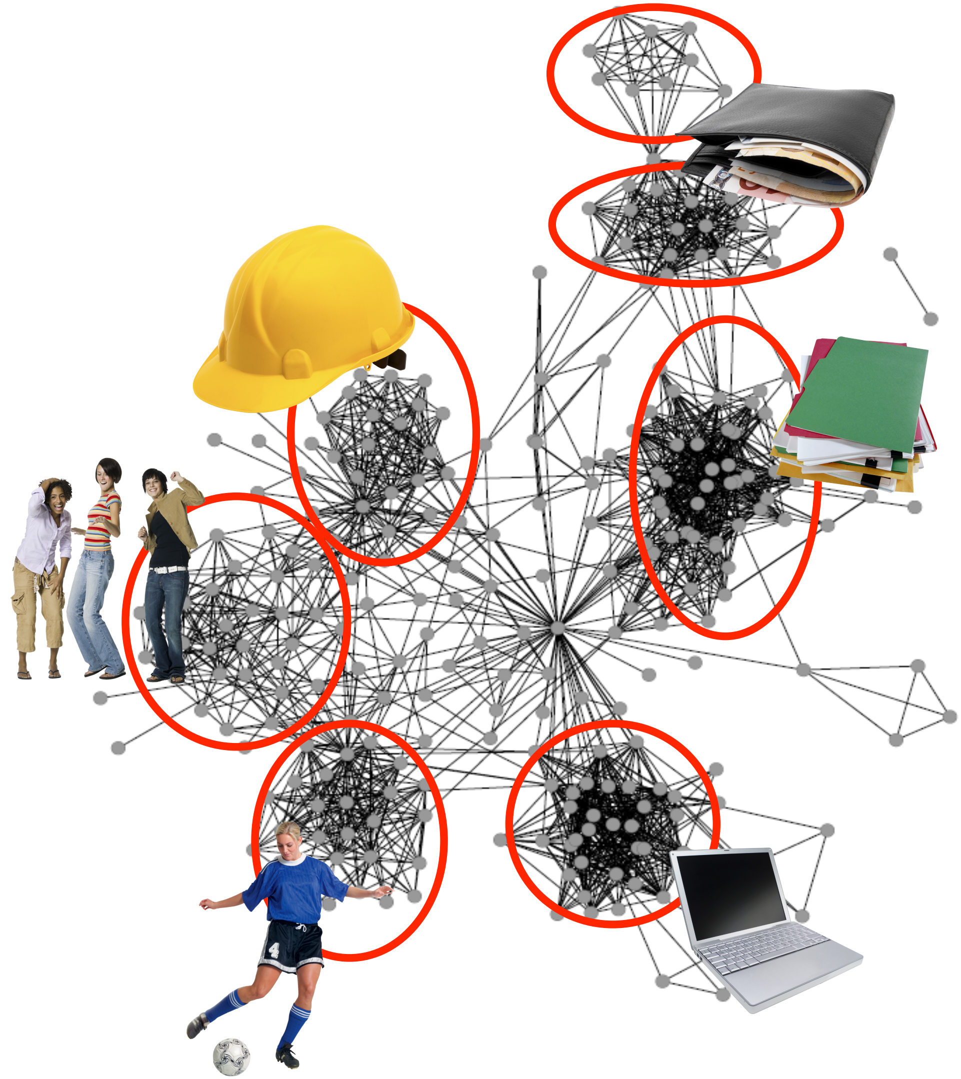
Reducing the complexity

Real Networks are Complex Objects

Can we make them “simpler”?

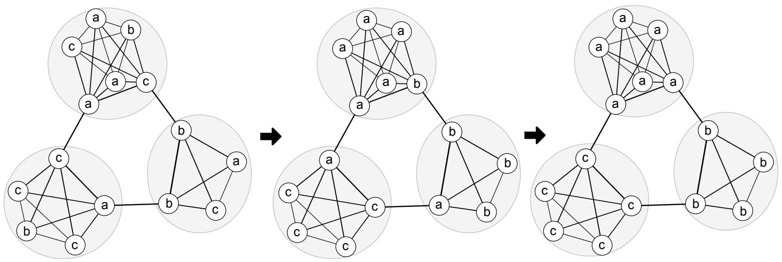
Ego-Networks

networks built upon a focal node , the "*ego*”, and the nodes to whom *ego* is directly connected to, including 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

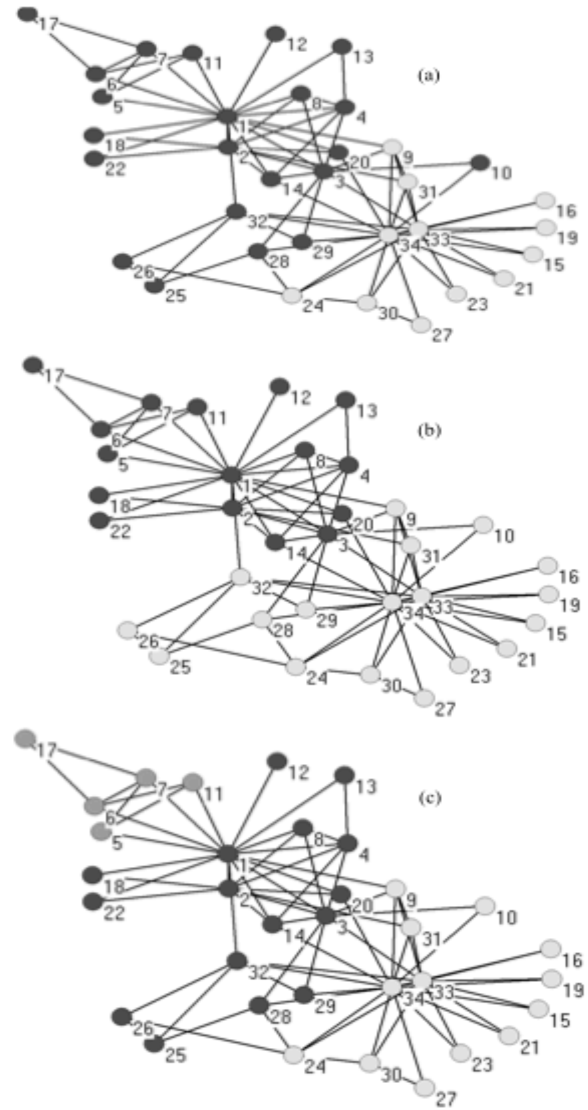


- For each raw community c in C
  1. Merge with "similar" ones in the set (given a threshold)
     (i.e. merge iff at most the ε% of the smaller one is not included in the bigger one)

[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

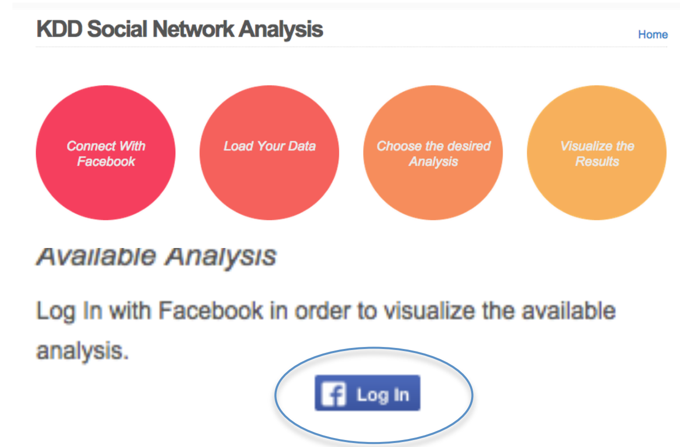
Product Networks

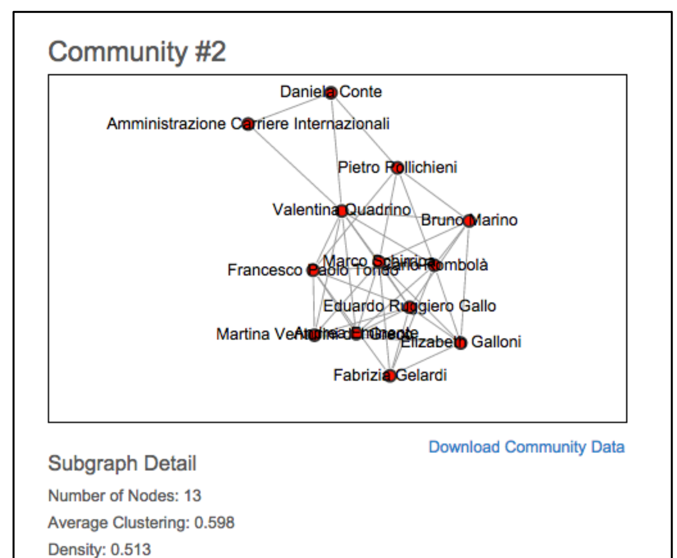
- Amazon

DEMON@Work

Personal Facebook Communities

1. Log out from Facebook and clean your browser cookies
2. Visit: **kddsna.isti.cnr.it:8080**
3. Log In with Facebook
4. Select one of the two options:
   1. “Visualize your network”
   2. “Demon Communities”
5. **Wait** for the data to be collected and displayed
6. Zoom-in/out and drag communities with your mouse





DEMON Biblio

- Michele Coscia, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi:
  DEMON: a local-first discovery method for overlapping communities.
  The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012: 615-623

- Michele Coscia, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi:
  Uncovering Hierarchical and Overlapping Communities with a Local-First Approach.
  ACM Transactions on Knowledge Discovery from Data TKDD 9(1): 6 (2014)

Bottom-up (local) vs top-down (global) community detection

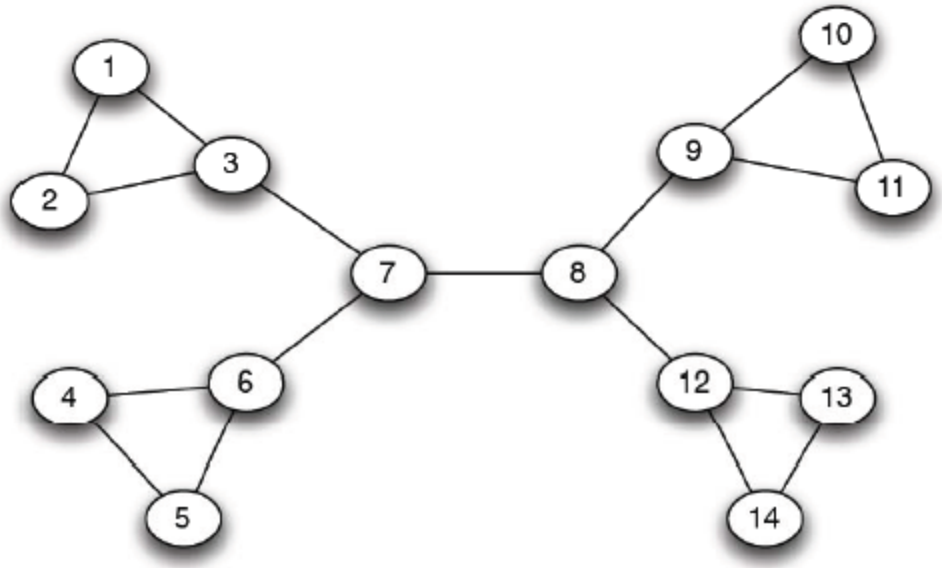
[Girvan-Newman PNAS '02]

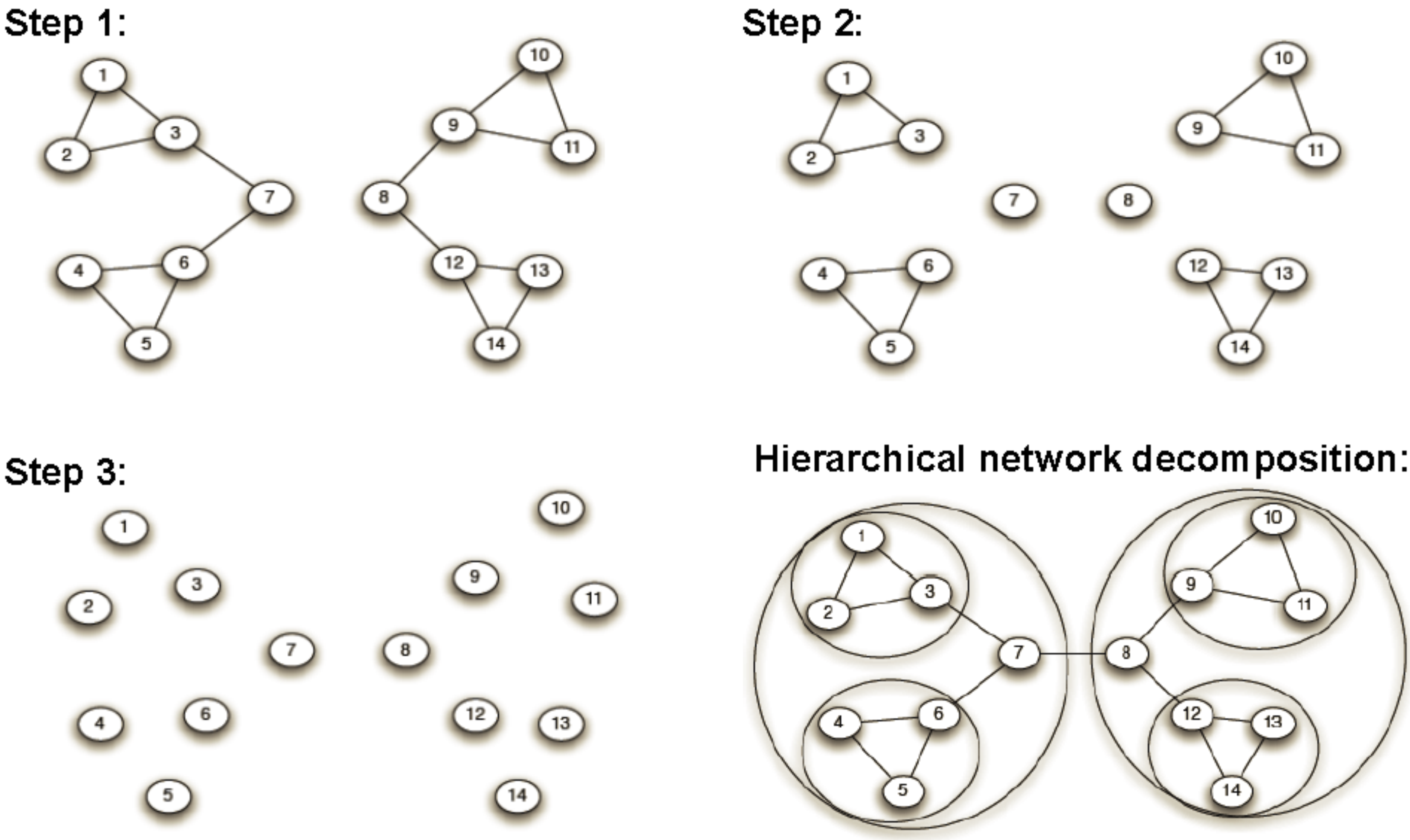
Method 1: Girvan-Newman

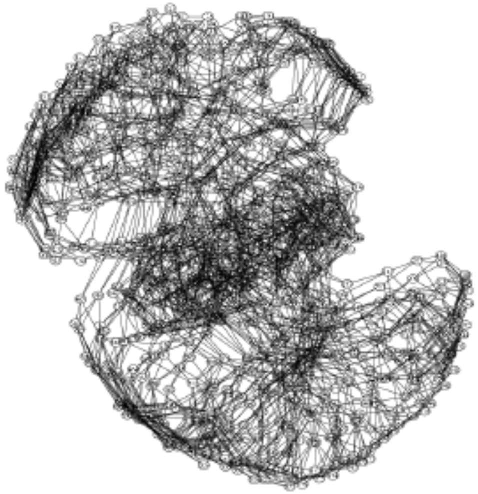
- Divisive hierarchical clustering based on the notion of edge betweenness:

  Number of shortest paths passing through the edge

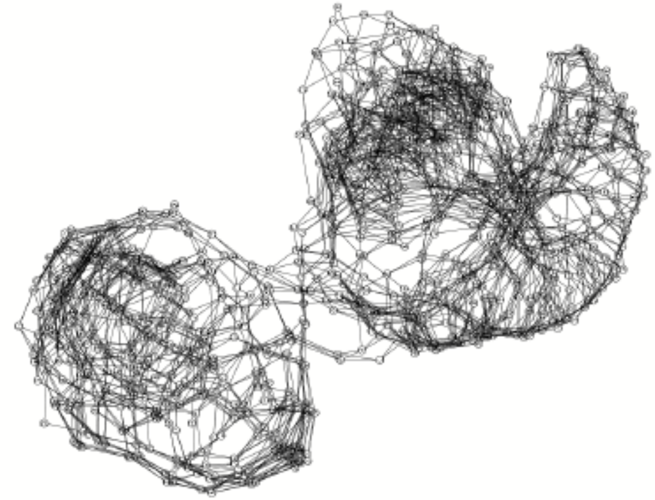
- Remove edges in decreasing betweenness
- Example:



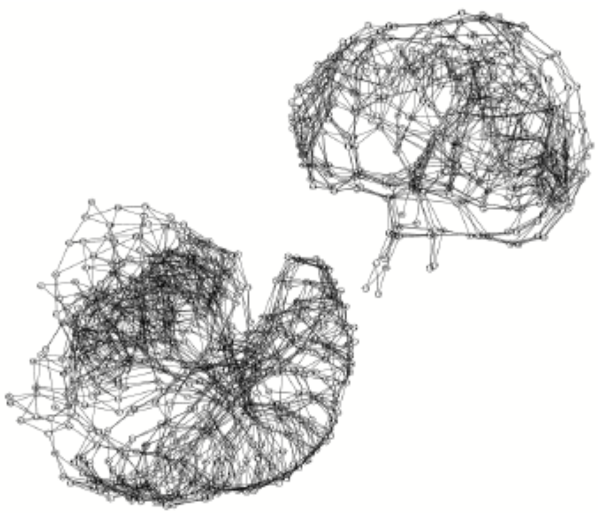

Step 1:
Step 2:
Step 3:
Hierarchical network decomposition:



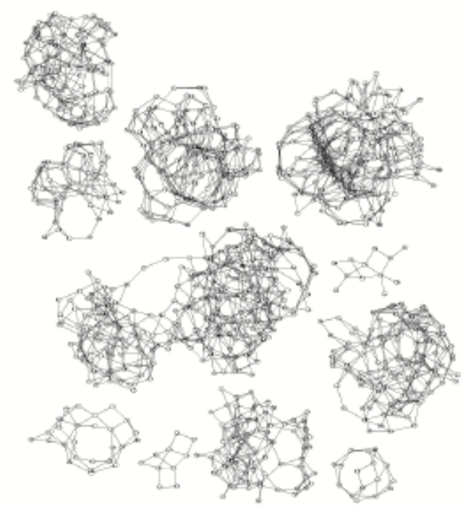
0 cuts



100 cuts



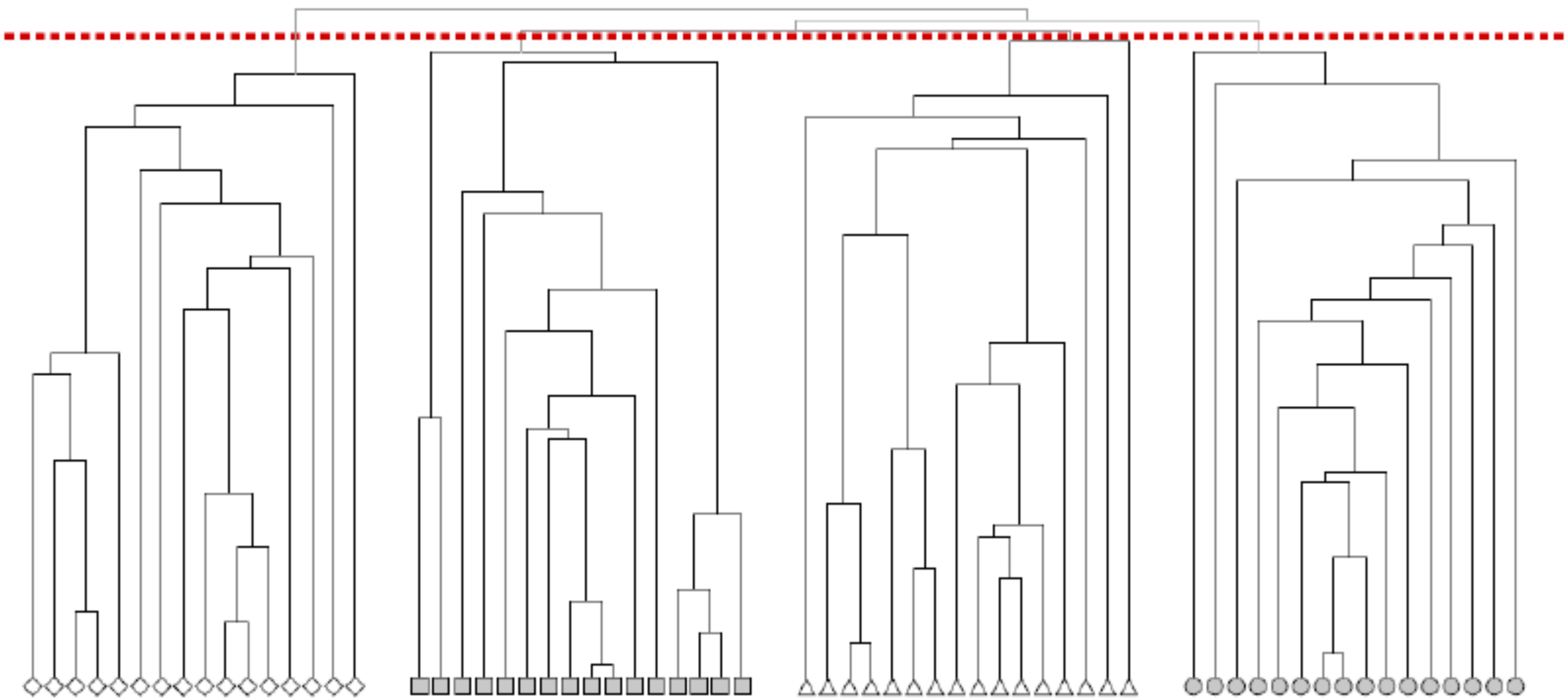
120 cuts



500 cuts

Hierarchical decomposition

- How to select the number of clusters/communities?



How to evaluate the quality of a network partition into communities?

Modularity

Q = (number of edges within groups) – (expected number within groups)

Actual number of edges between i and j is

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge } (i,j), \\ 0 & \text{otherwise.} \end{cases}$$

Expected number of edges between *i* and *j* is

$$\text{Expected number} = \frac{k_i k_j}{2m}.$$

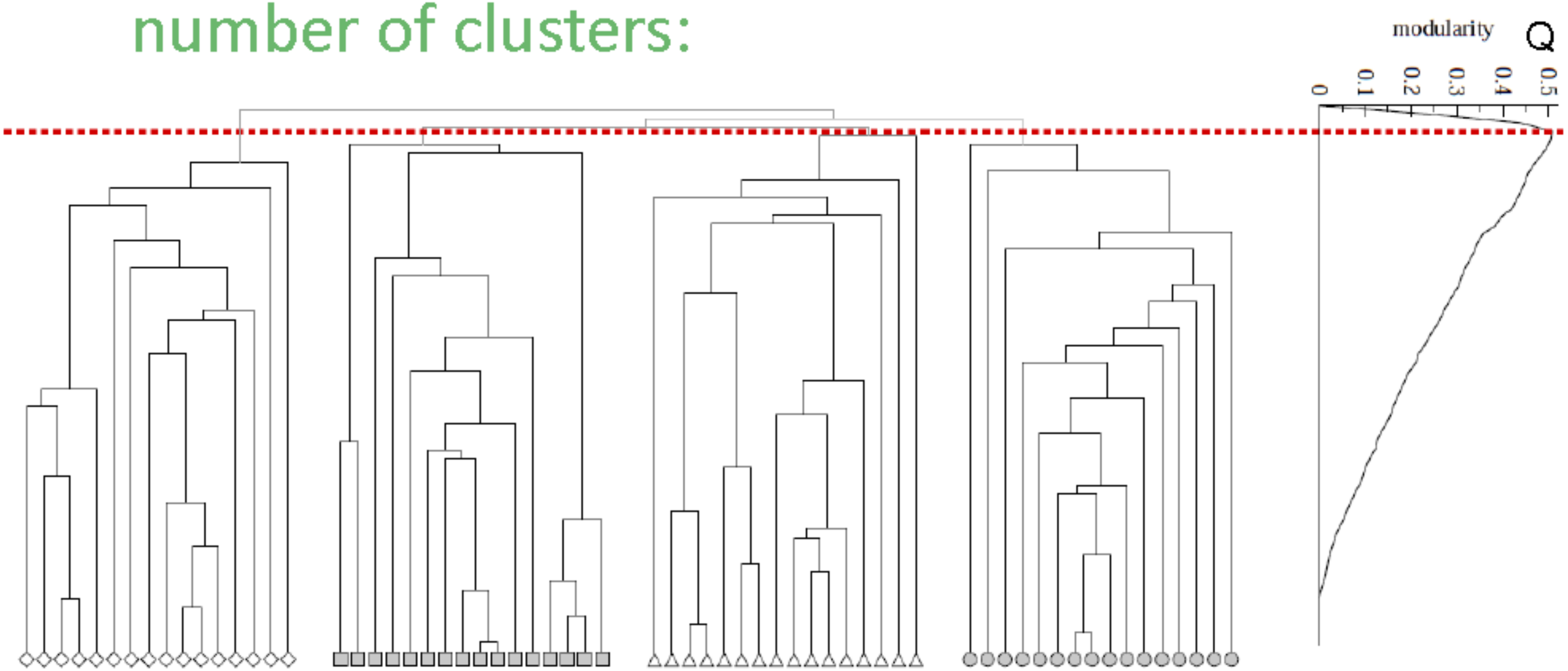
Modularity

- Q = (number of edges within groups) – (expected number within groups)
- Then:

$$Q = \frac{1}{4m}\left[\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)\right]$$

m … number of edges
A_{ij} … 1 if (i,j) is edge, else 0
k_i … degree of node i
c_i … group id of node i
$\delta(a, b)$ … 1 if a=b, else 0

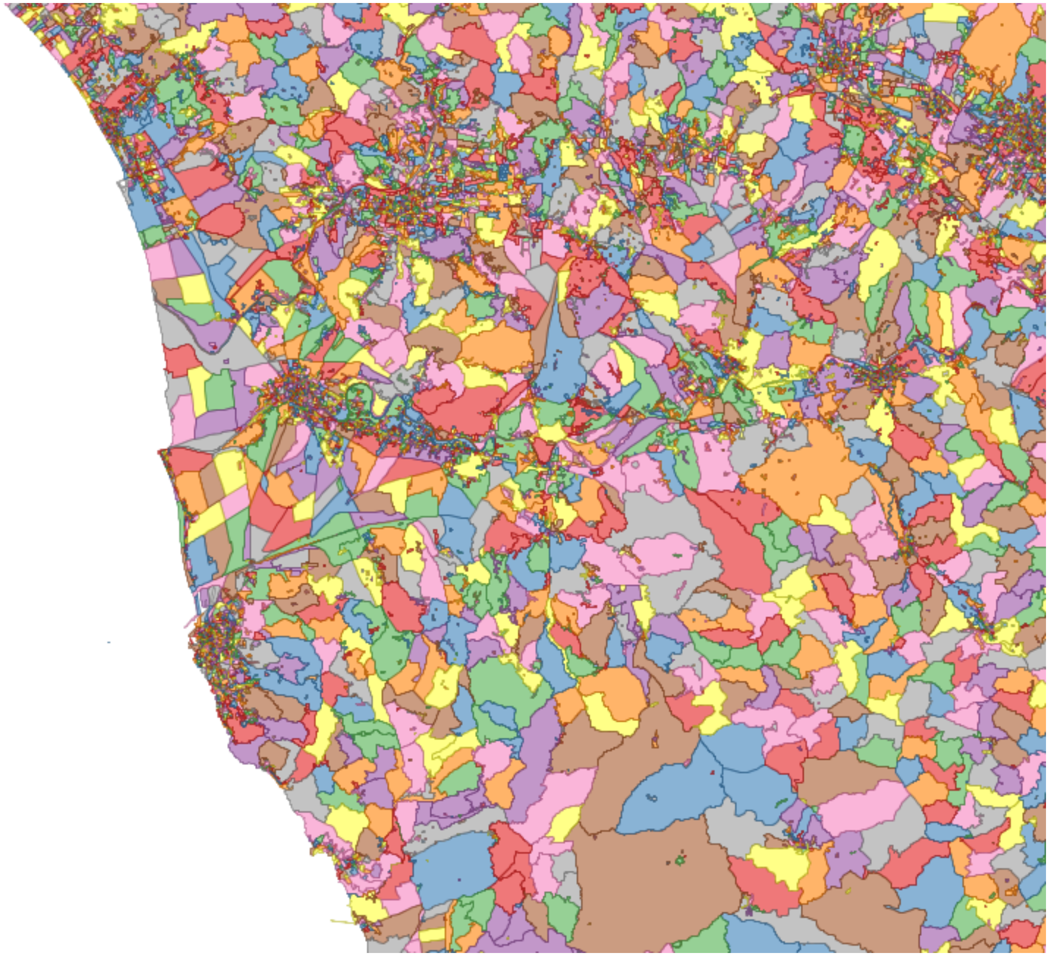
- Modularity is useful for selecting the number of clusters:



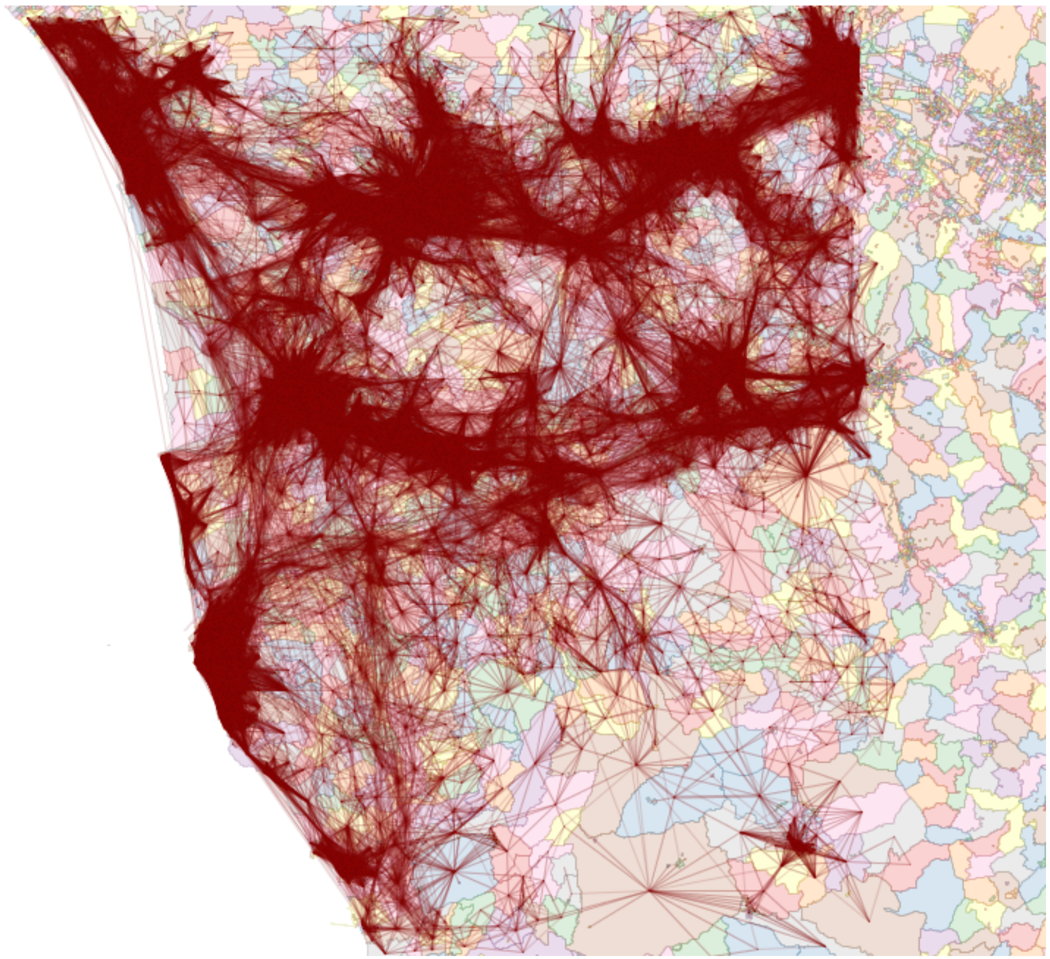
Community discovery

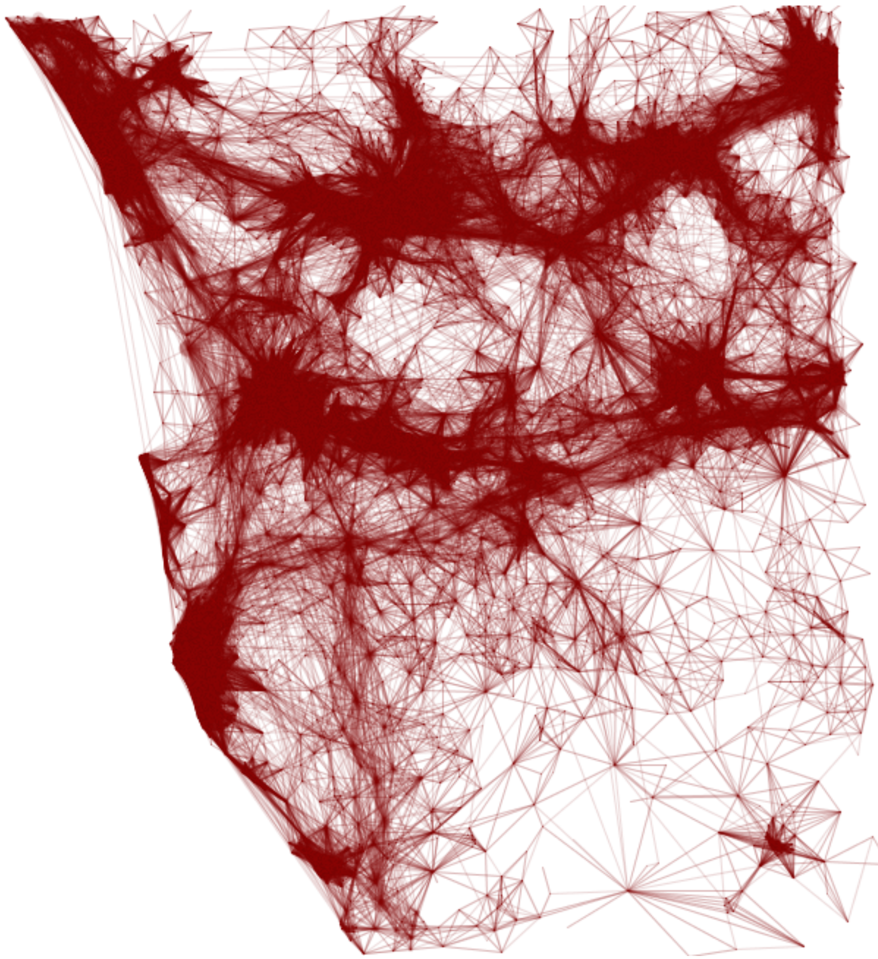
- Challenging task
- Many competing approaches
- Huge literature
- Recent surveys:
  - Michele Coscia, Fosca Giannotti, Dino Pedreschi: A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining* 4(5): 512-546 (2011)
  - Santo Fortunato: Community detection in graphs *Physics Reports* 486 (3), 75-174 (2010)

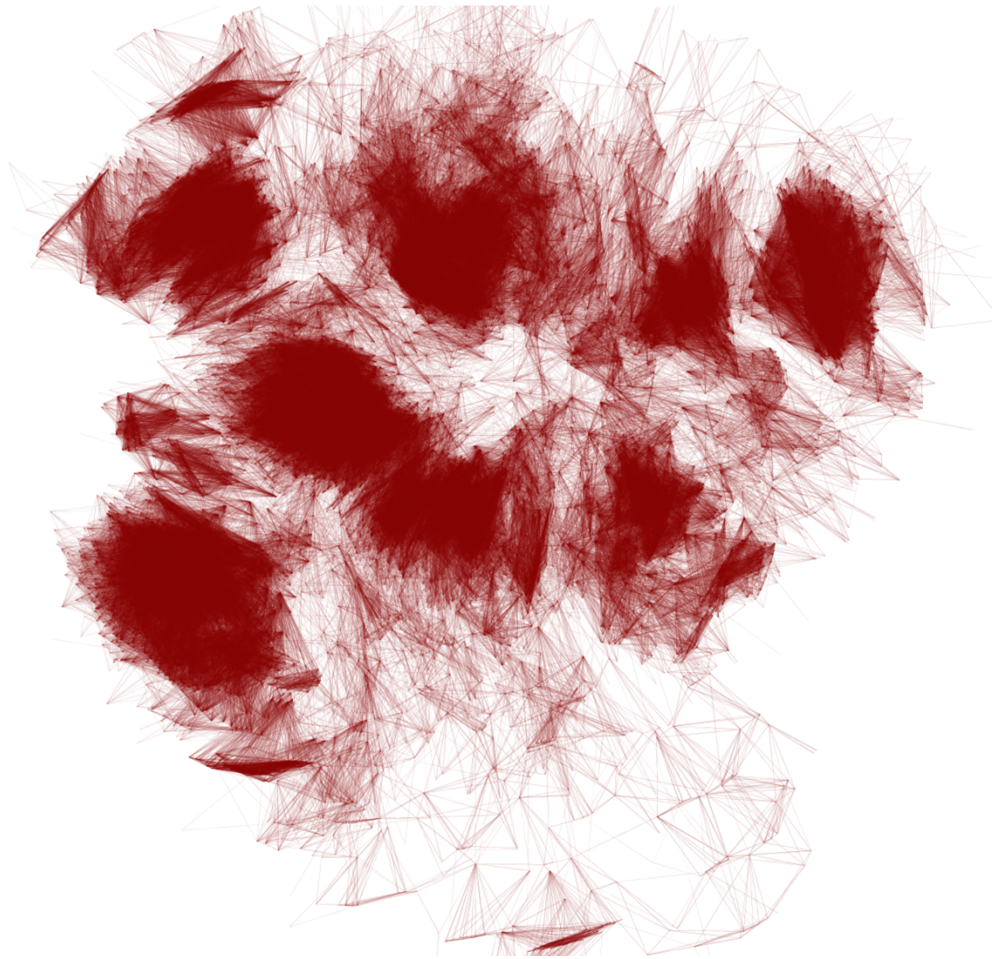
Discover the borders of mobility

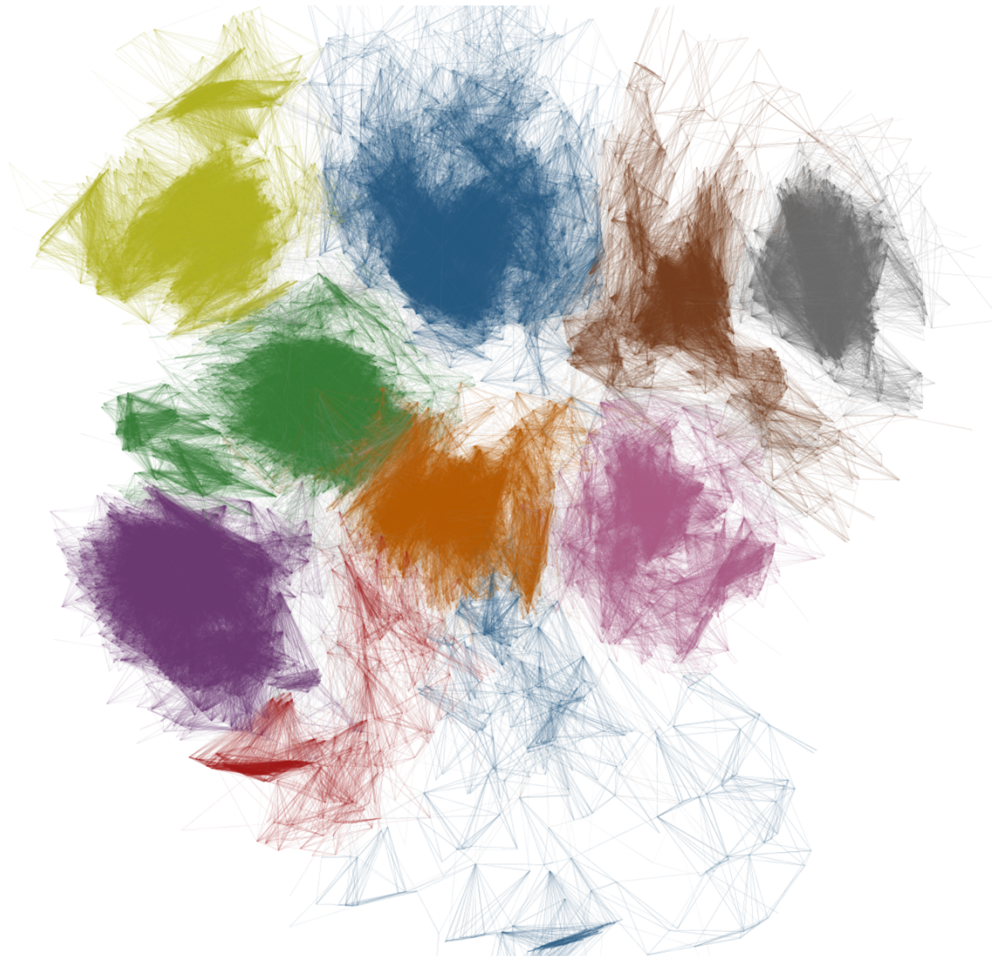


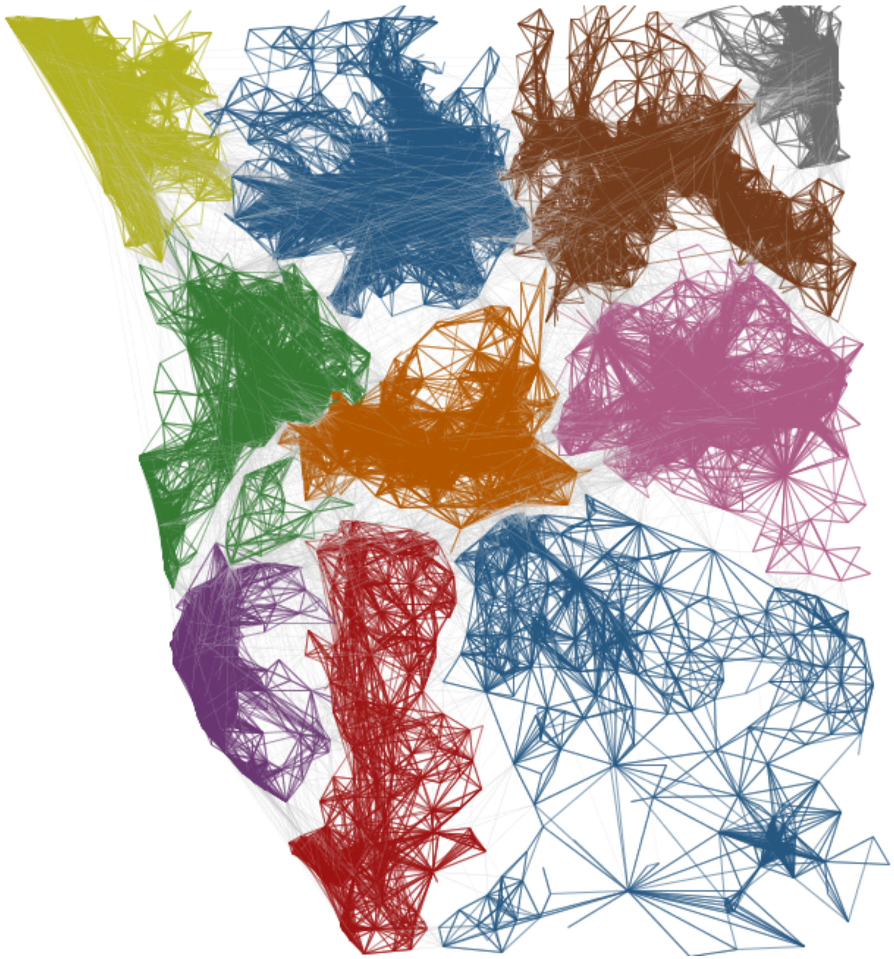
Salvatore Rinzivillo, Mainardi, Pezzoni, Michele Coscia, Dino Pedreschi, Fosca Giannotti: Discovering the Geographical Borders of Human Mobility. KI 26(3): 253-260 (2012)

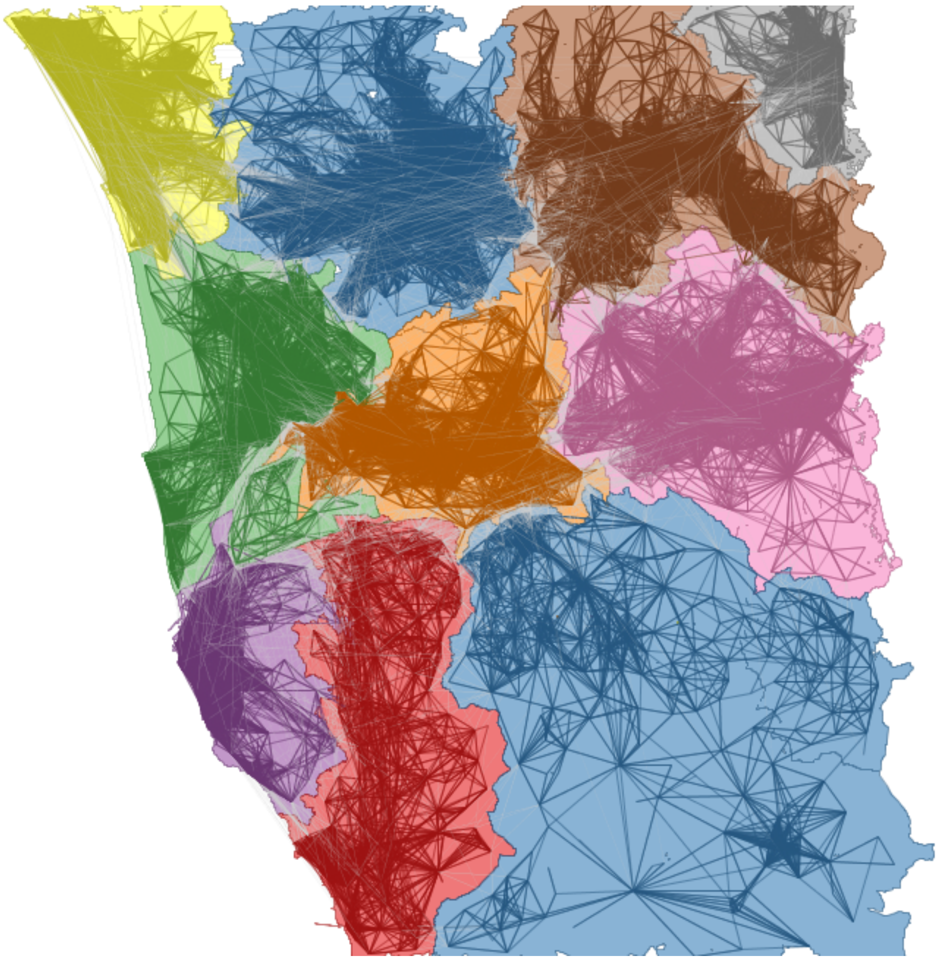


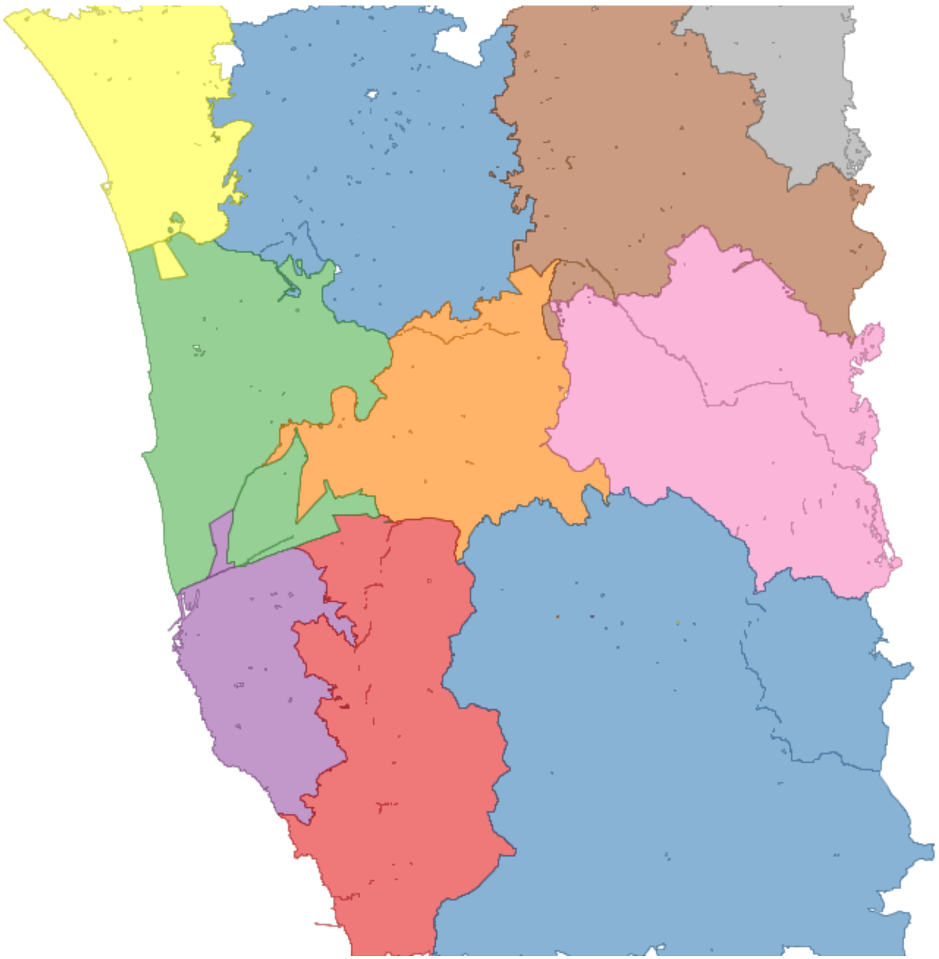


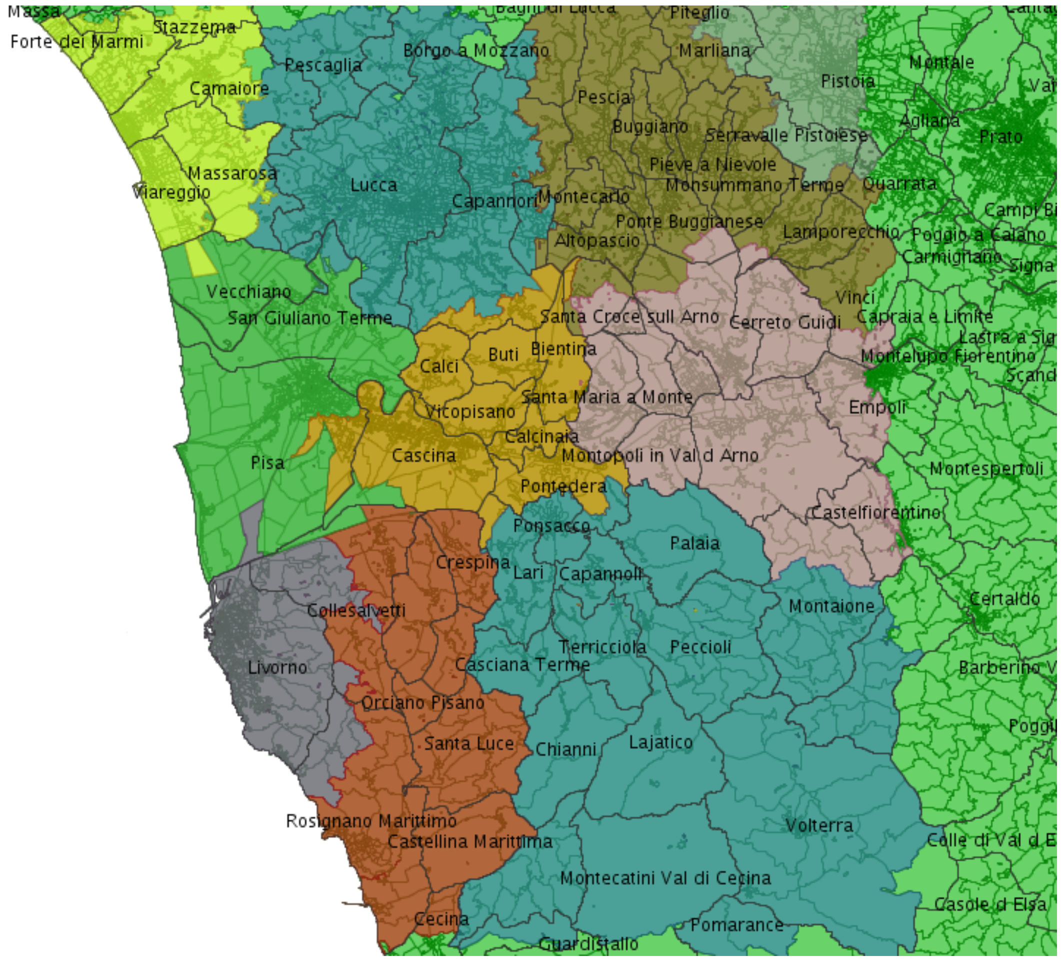


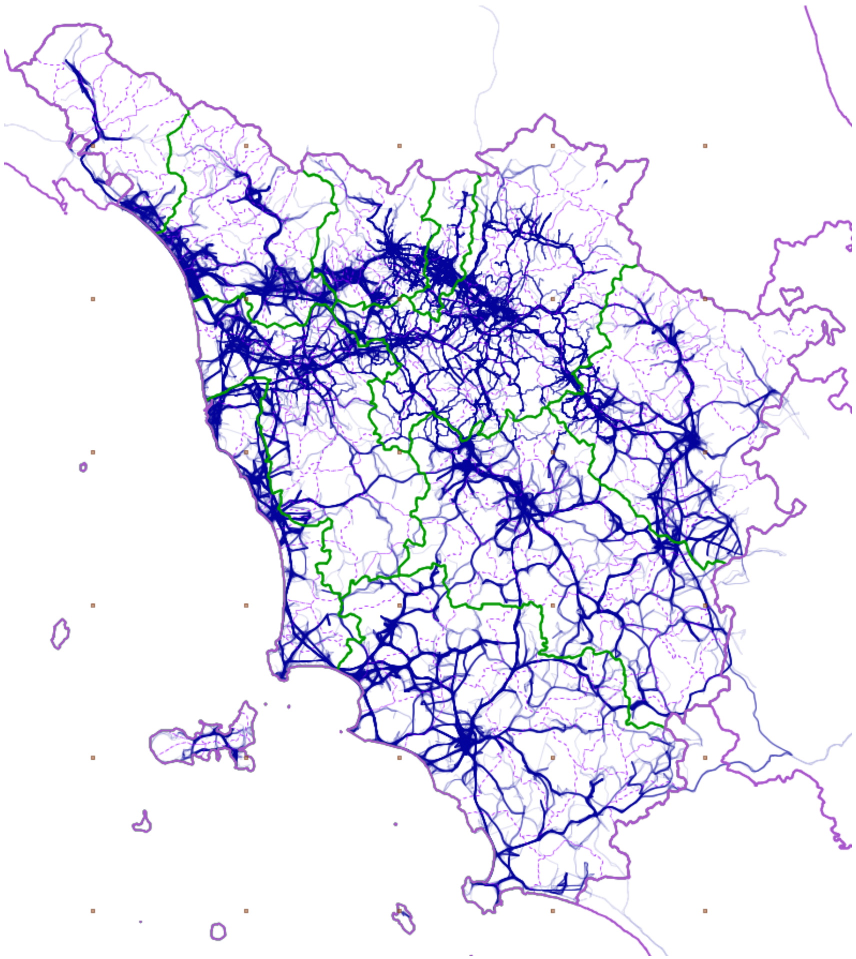


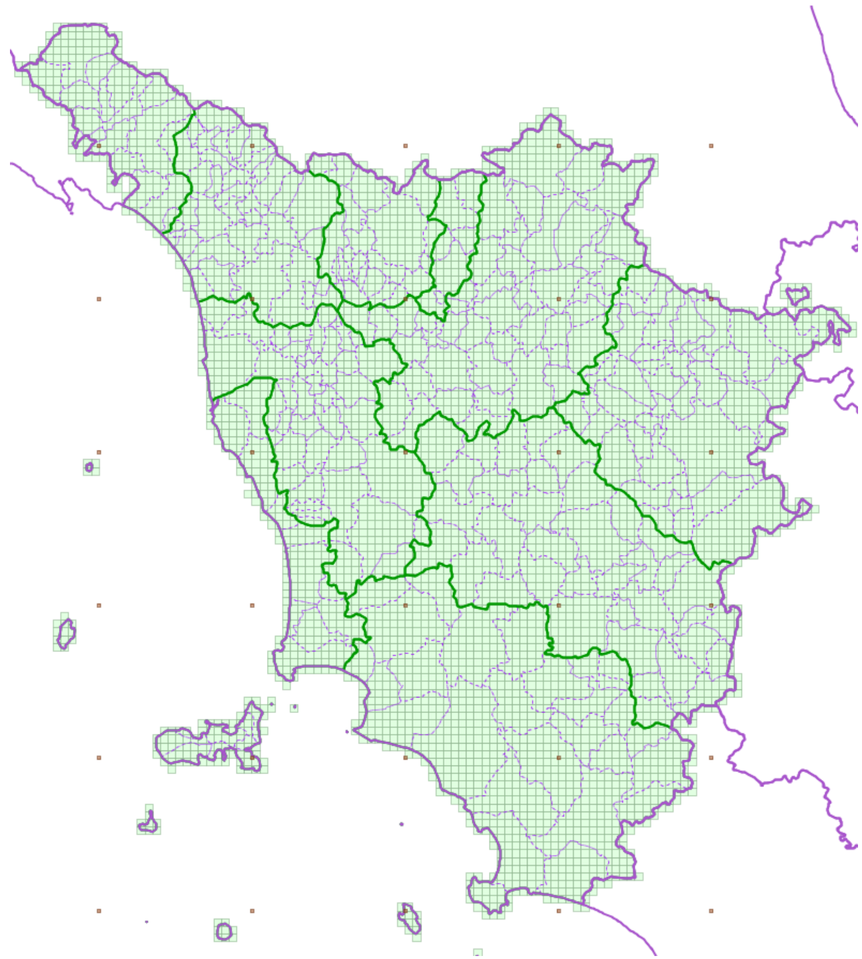


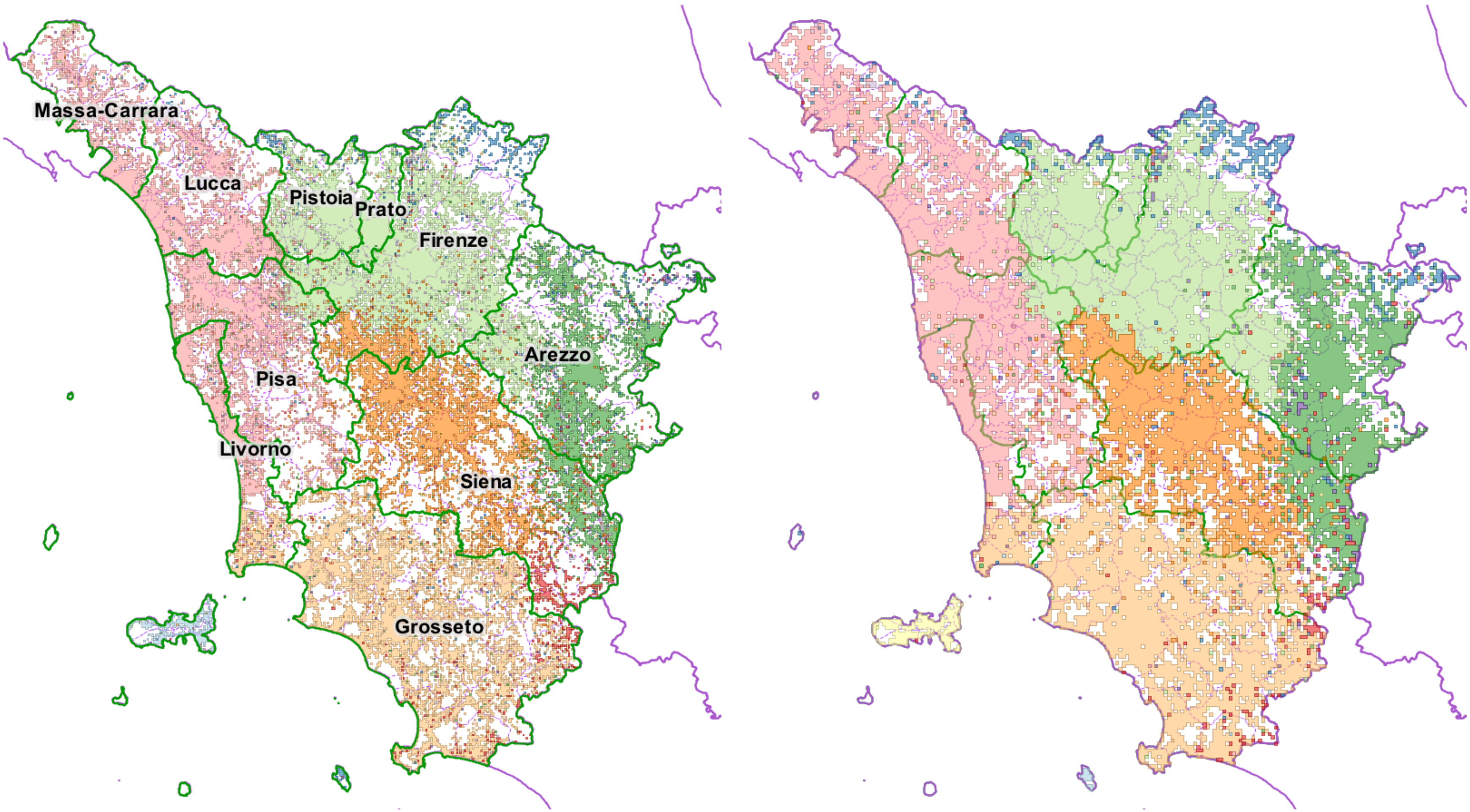


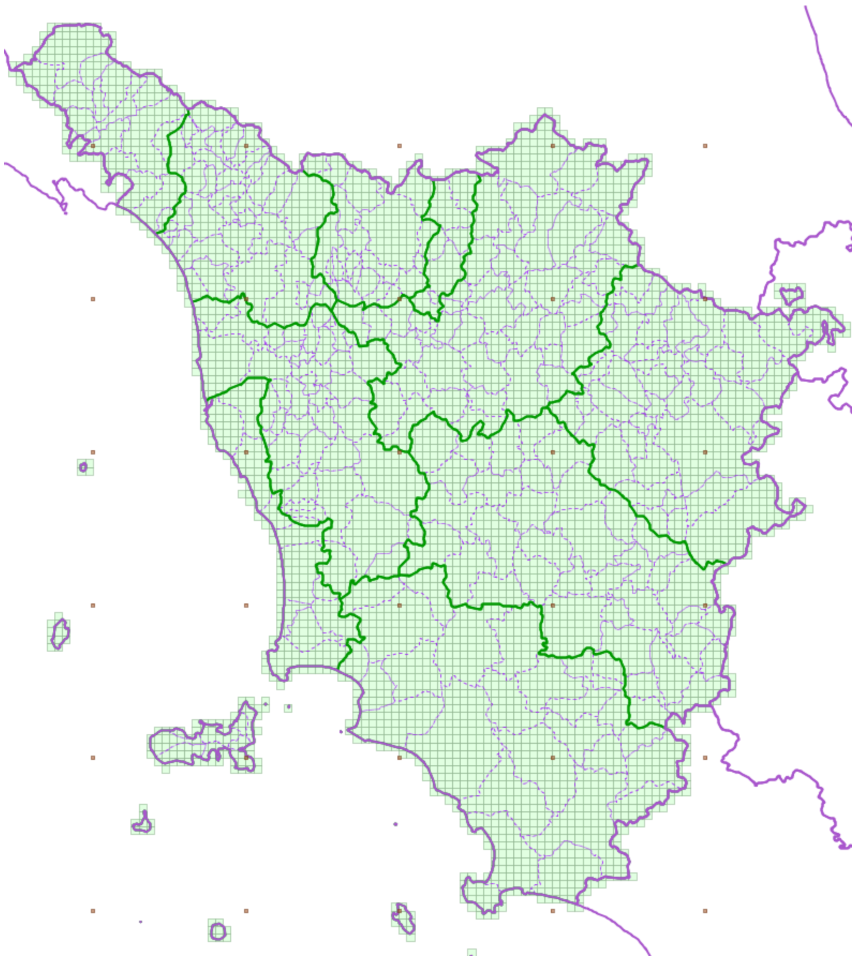


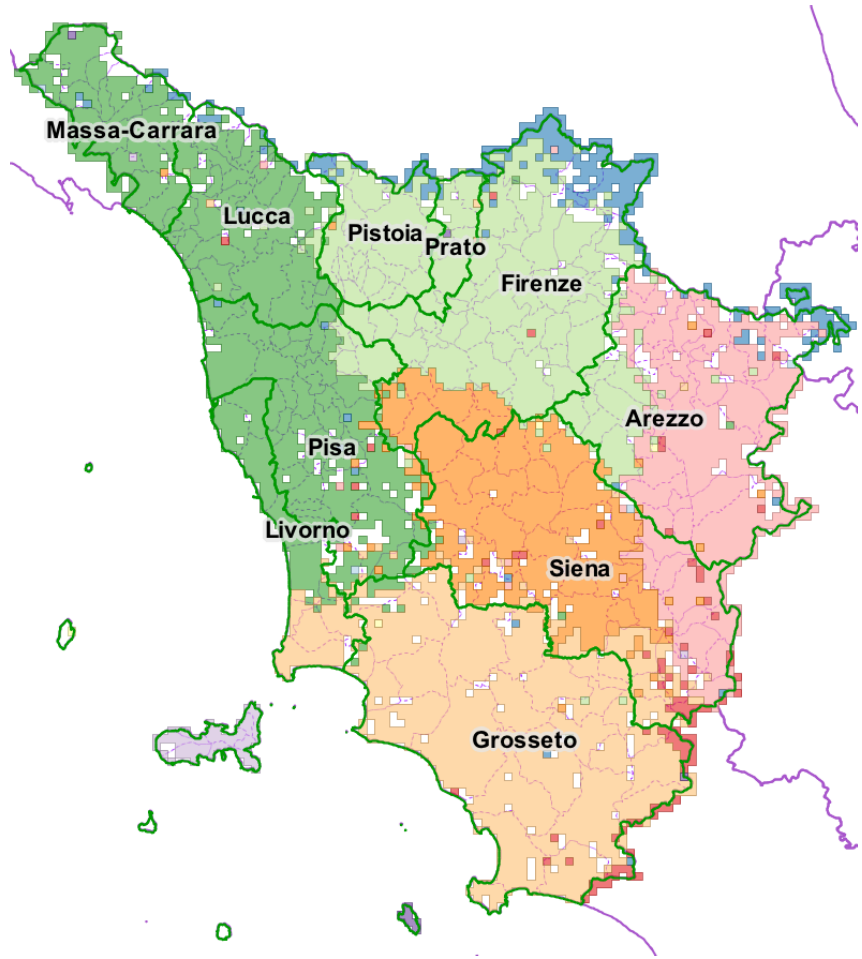

Massa
Stazzema
Forte dei Marmi
Pescaglia
Borgo a Mozzano
Pieve Fosciana
Pitegliо
Marliana
Pistoia
Montale
Camaiore
Pescia
Buggiano
Serravalle Pistoiese
Agliana
Prato
Massarosa
Viareggio
Lucca
Capannori
Montecarlo
Pieve a Nievole
Monsummano Terme
Quarrata
Ponte Buggianese
Altopascio
Lamporecchio
Poggio a Caiano
Carmignano
Signa
Vecchiano
San Giuliano Terme
Santa Croce sull Arno
Cerreto Guidi
Vinci
Capraia e Limite
Buti
Bientina
Calci
Montelupo Fiorentino
Santa Maria a Monte
Vicopisano
Empoli
Calcinaia
Cascina
Pisa
Montopoli in Val d Arno
Montespertoli
Pontedera
Castelfiorentino
Ponsacco
Palaia
Crespina
Lari
Capannoli
Montaione
Certaldo
Collesalvetti
Livorno
Terricciola
Peccioli
Casciana Terme
Orciano Pisano
Santa Luce
Chianni
Lajatico
Rosignano Marittimo
Castellina Marittima
Volterra
Montecatini Val di Cecina
Cecina
Pomarance
Guardistallo
Casole d Elsa






Massa-Carrara
Lucca
Pistoia
Prato
Firenze
Arezzo
Pisa
Livorno
Siena
Grosseto

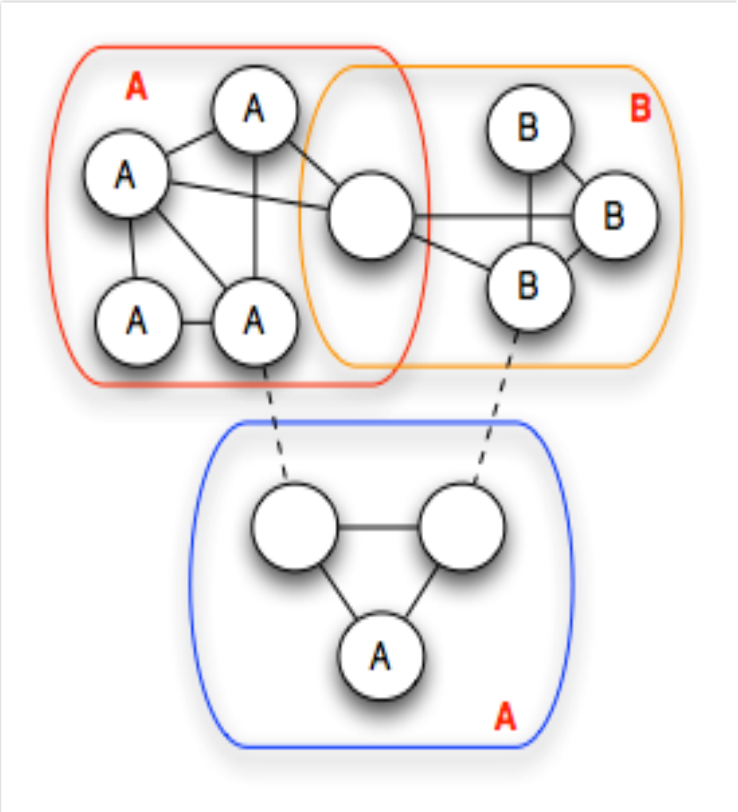



Massa-Carrara
Lucca
Pistoia
Prato
Firenze
Arezzo
Pisa
Livorno
Siena
Grosseto

Demon communities

- Overlapping
- Microscopic
- High homophily

People belonging to the same
e social context often show some degree of homopily:
(i.e. same age, level of education)



- Application: classification
- E.g. user engagement

Skype Network Data

Semantic rich dataset:

- **Social Graph**
  (built upon users contact lists
  ~billions of nodes)

- **Users Geographic presence**
  (city, nation…)

- **Users Monthly Activity**
  (individual's days of Audio\Video\Chat
  products usage)

Problem: Service Usage

Given an online platform we often we need to *estimate* how its services (i.e., Skype Audio\Video call) are used by the registered users.
In particular we can be asked to answer the following questions:

Q1: Can Service Usage be described as a function of the Network Data?

Q2: If so, at which scale should we analyze the network in order to perform a descriptive analysis?

Classifier features

For each network partition obtained, we built classifier and trained it to discriminate between High and Low active communities.

STRUCTURAL FEATURES

N	number of nodes
M	number of edges
D	density
CC	global clustering
CC_{avg}	average clustering
A_{deg}	degree assortativity
deg^{C}_{max}	max degree (community links)
deg^{C}_{avg}	avg degree (community links)
deg^{all}_{max}	max degree (all links)
deg^{all}_{avg}	avg degree (all links)
T	closed triads
T_{open}	open triads
O_v	neighborhood nodes
O_e	outgoing edges
E_{dist}	num. edges with distance
d	approx. diameter
r	approx. radius
g	conductance

COMMUNITY FORMATION FEATURES

T_f	first user arrival time
IT_{avg}	avg user inter-arrival time
IT_{std}	std of user inter-arrival time
$IT_{l,f}$	last-first inter-arrival time

GEOGRAPHIC FEATURES

N_s	number of countries
E_s	country entropy
S_{max}	percentage of most represented country
N_t	number of cities
E_t	city entropy
$dist_{avg}$	avg geographic distance
$dist_{max}$	max geographic distance

ACTIVITY FEATURES

Video	mean number of days of video
Chat	mean number of days of chat

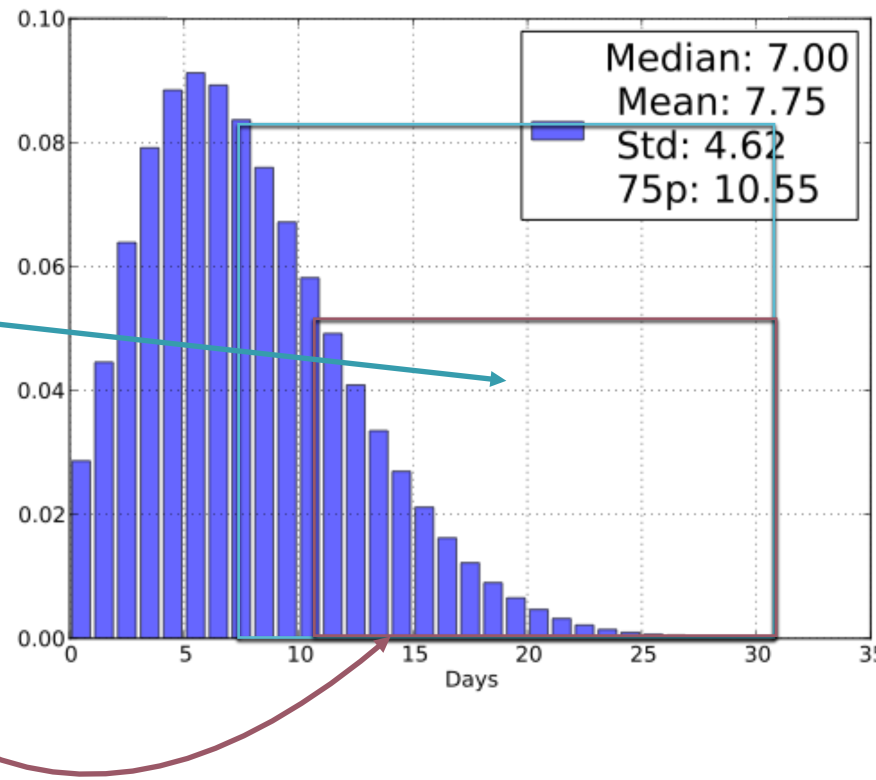
Target Class (for each service)

The target class identify the Service Activity Level (High/Low)

Two scenarios:

1. Low/High activity is identified by the median of the distribution (i.e., an highly active community have and avg activity > than the median of the overall activity distribution)

2. High activity communities are the one above the 75th percentile

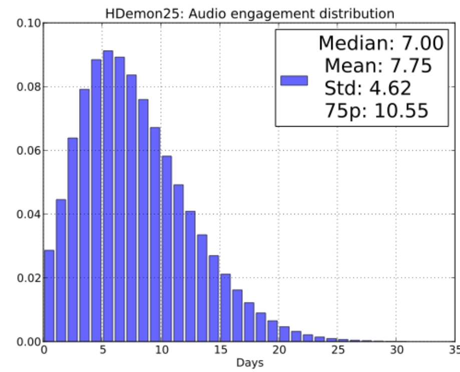


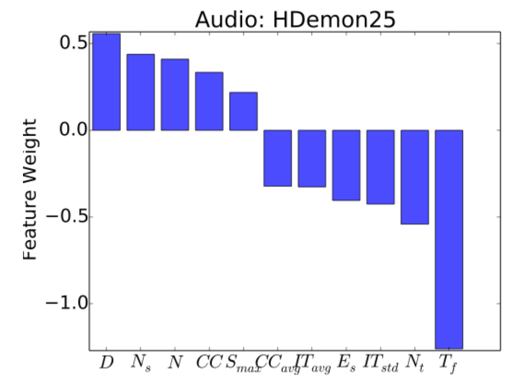
“Social Engagement” : Skype social graph

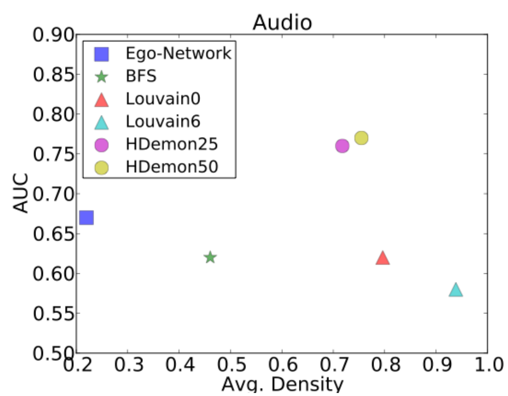
- **Problem:**
  Given the Skype social graph and its user information (i.e., location…) predict average level of community activity for the Audio \Video services.

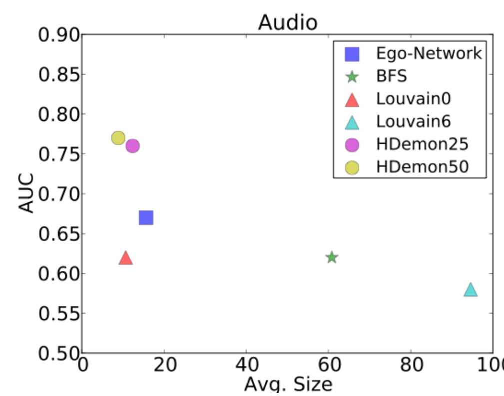
- **Question:**
  The CD method chosen will affect the classification results?

- **Main Results:**
  - The smaller and denser communities are the better
  - Demon outperforms Louvain, Ego-Nets and BFS
  - Topological, Temporal and Geographical features of communities are valuable activity level predictors

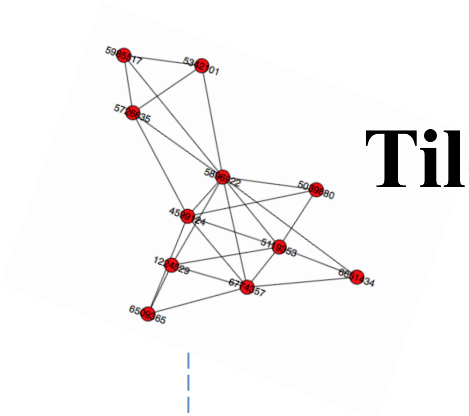








G. Rossetti, L. Pappalardo, R. Kikas, F. Giannotti, D. Pedreschi, M. Dumas *Community-centric analysis of service en- gagement in Skype social networks*. IEEE ASONAM 2015, France (Accepted)

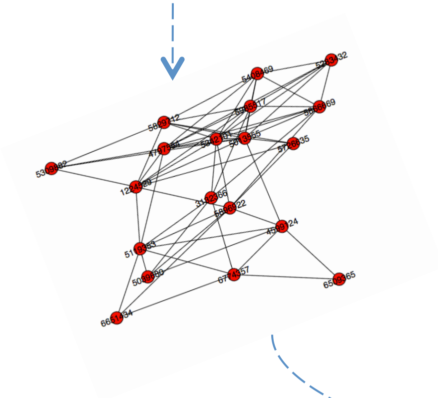


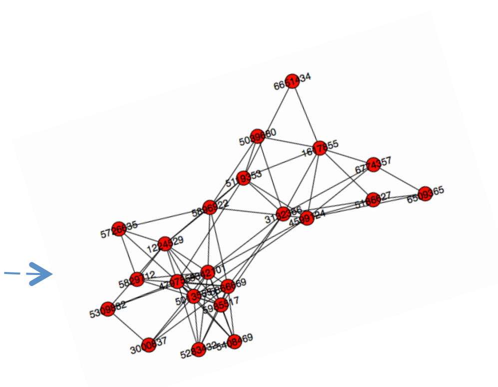
Tiles: evolutionary community discovery

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



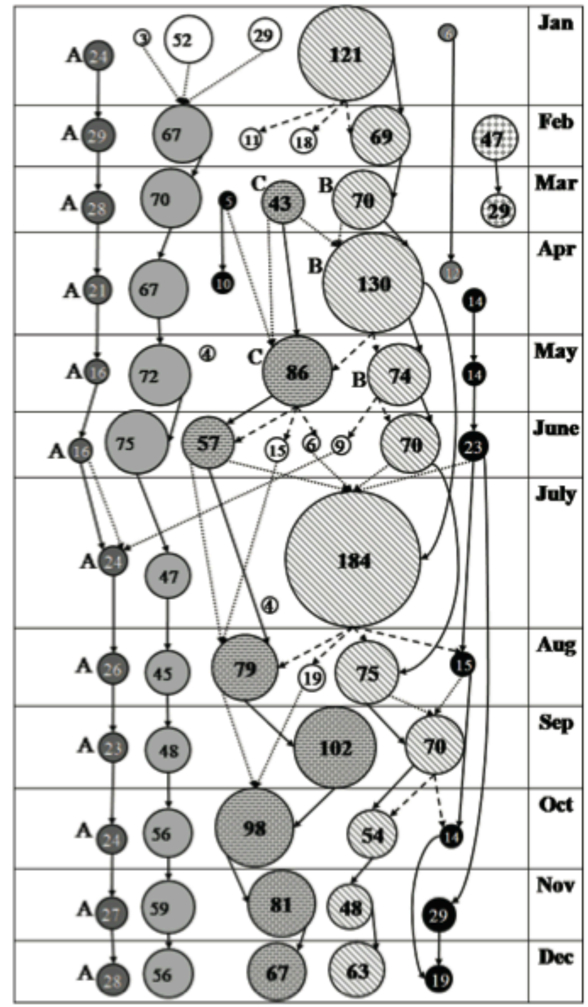


Dynamic Networks

- 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
- …

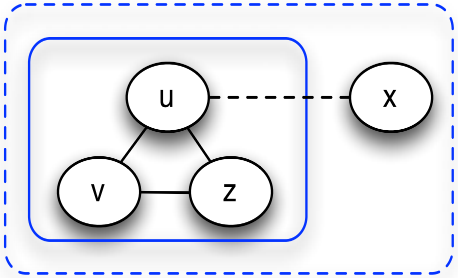
In these scenarios a QSSA (Quasi Steady State Assumption) rarely holds:

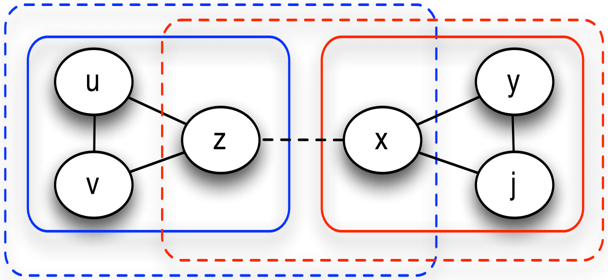
- Network cannot be *“frozen in time”*
  - Nodes and edges rise and fall producing perturbation on the whole topology
- The reduction to static scenarios trough temporal discretization is not always a good idea
  - How can we chose the temporal threshold?
  - To what extent can we trust the obtained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

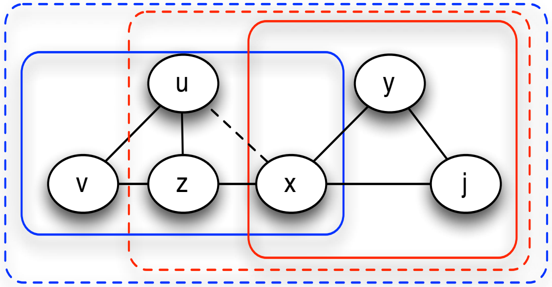
The Idea… TILES

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



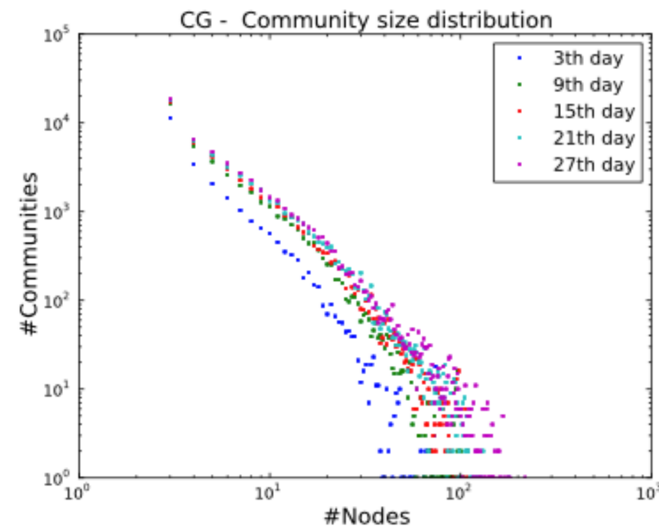


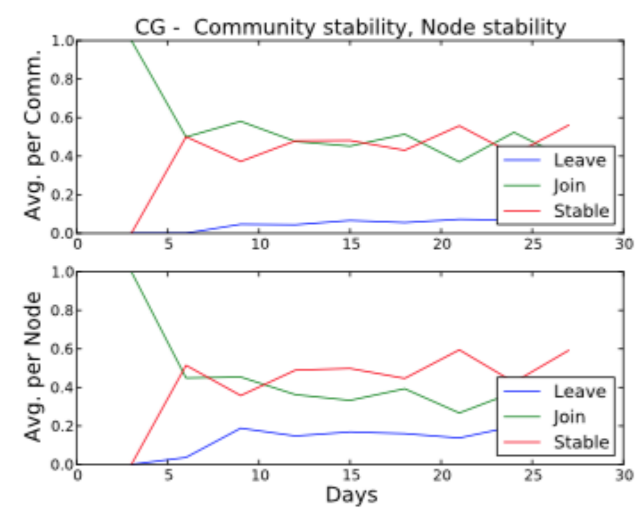


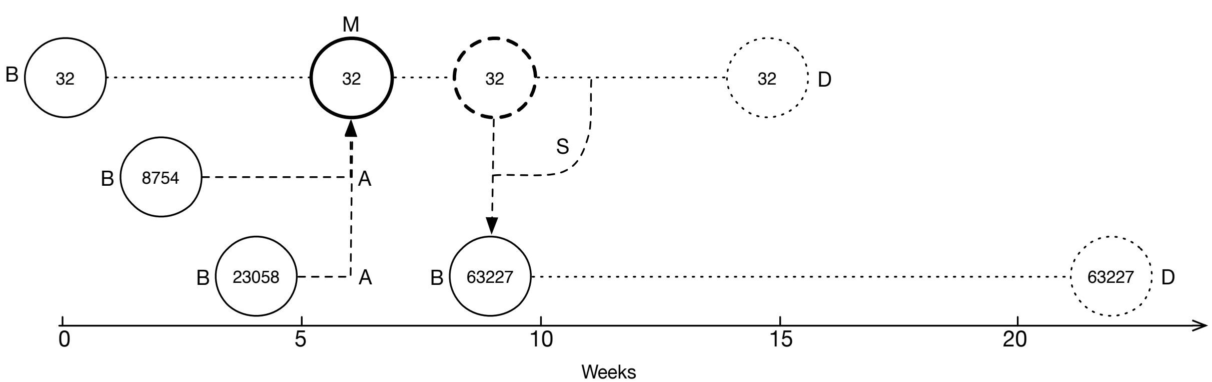
Tiles Community Insights

Experiments real interaction networks show that:

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



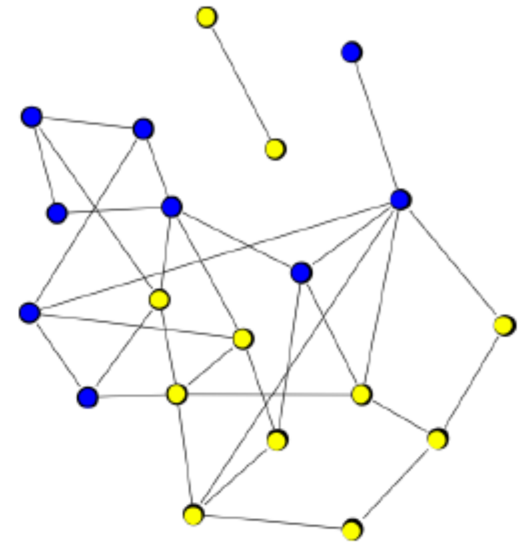


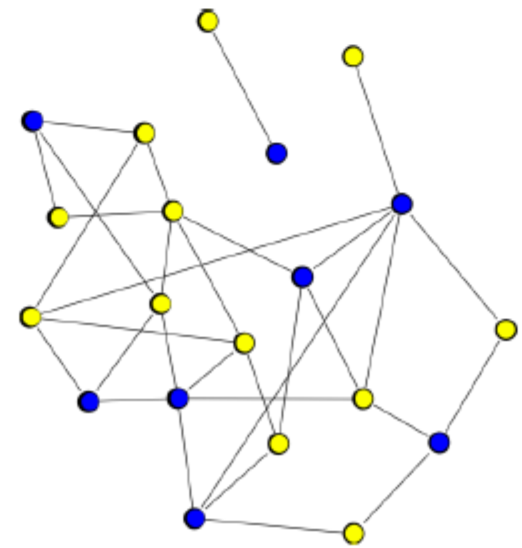


Group formation dynamics

Group formation in networks

- In a social network **nodes explicitly declare group membership:**
  - Facebook groups, Publication venue
- Can think of groups as **node colors**
- Gives **insights into social dynamics:**
  - Recruits friends? Memberships spread along edges
  - Doesn't recruit? Spread randomly
- **What factors influence a person's decision to join a group?**

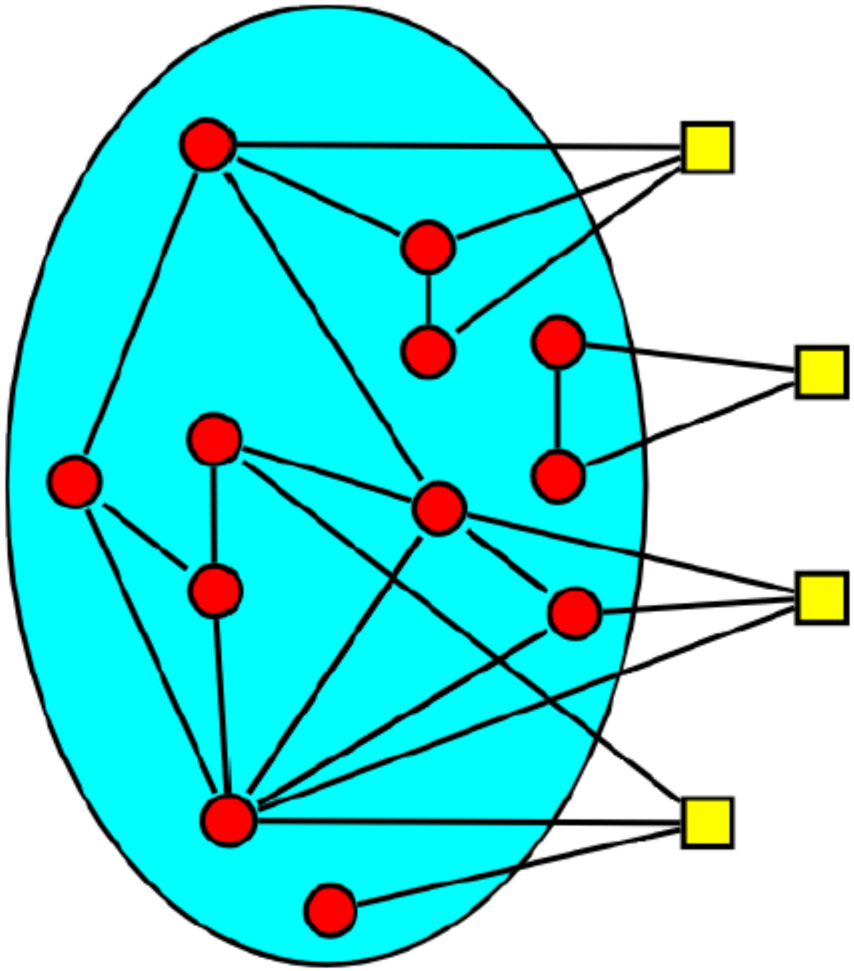


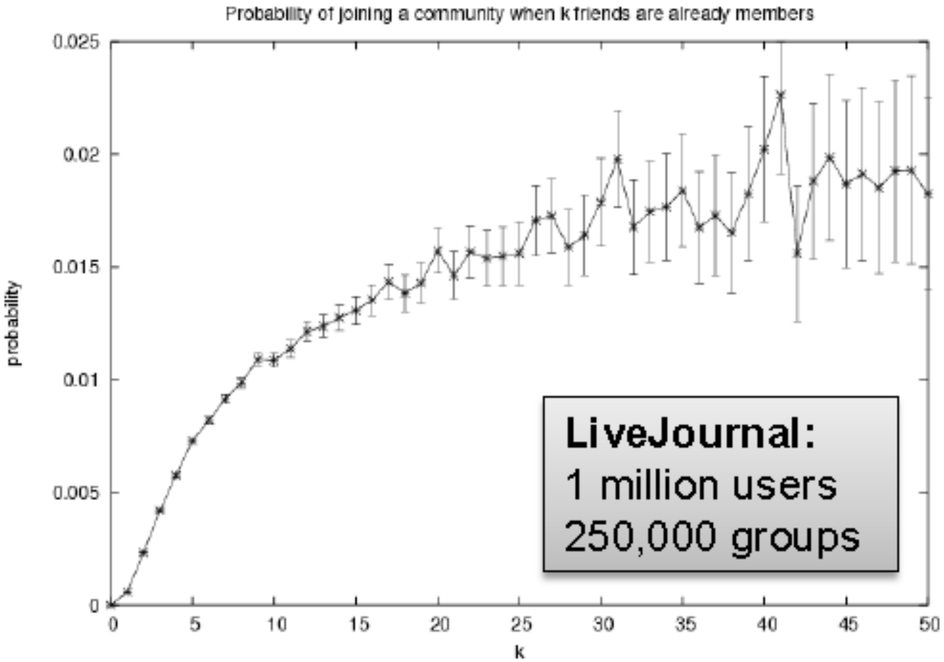


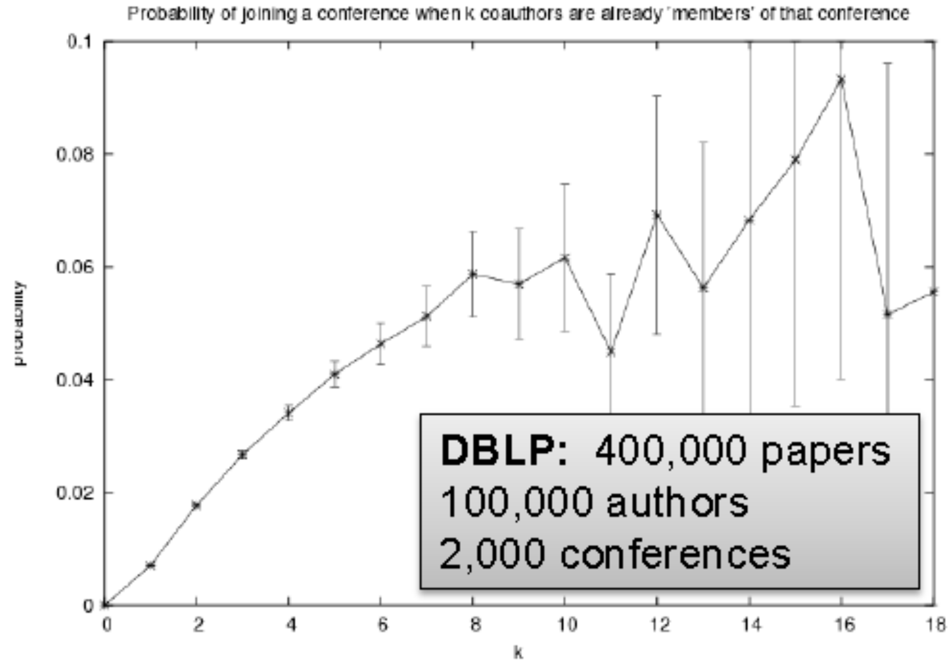
Group memberships spread over the network:

- 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?





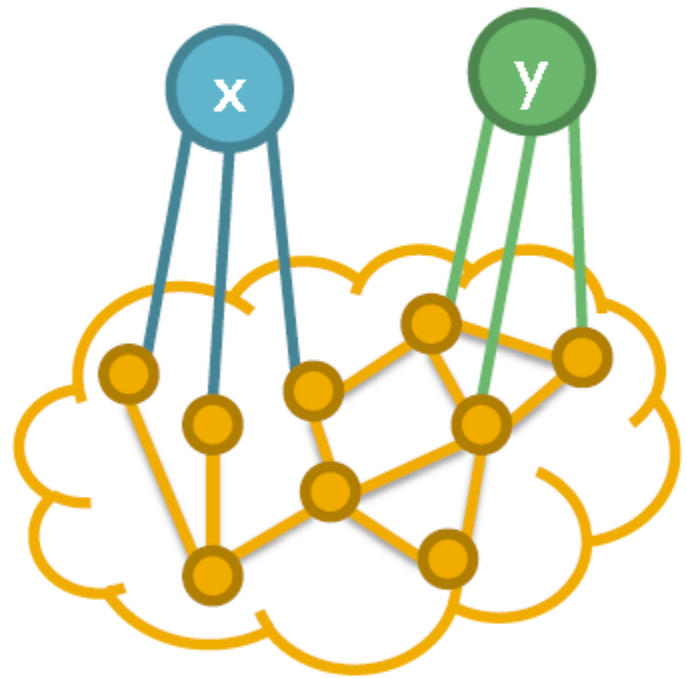


- **Diminishing returns:**
  - Probability of joining increases with the number of friends in the group
  - But increases get smaller and smaller

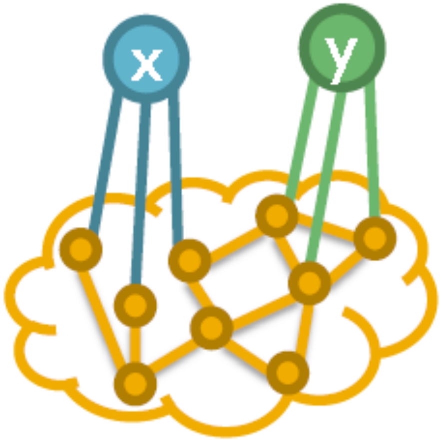
Connectedness of friends and group membership

- *x* and *y* have three friends in the group
- *x*'s friends are independent
- *y*'s friends are all connected

Who is more likely to join?



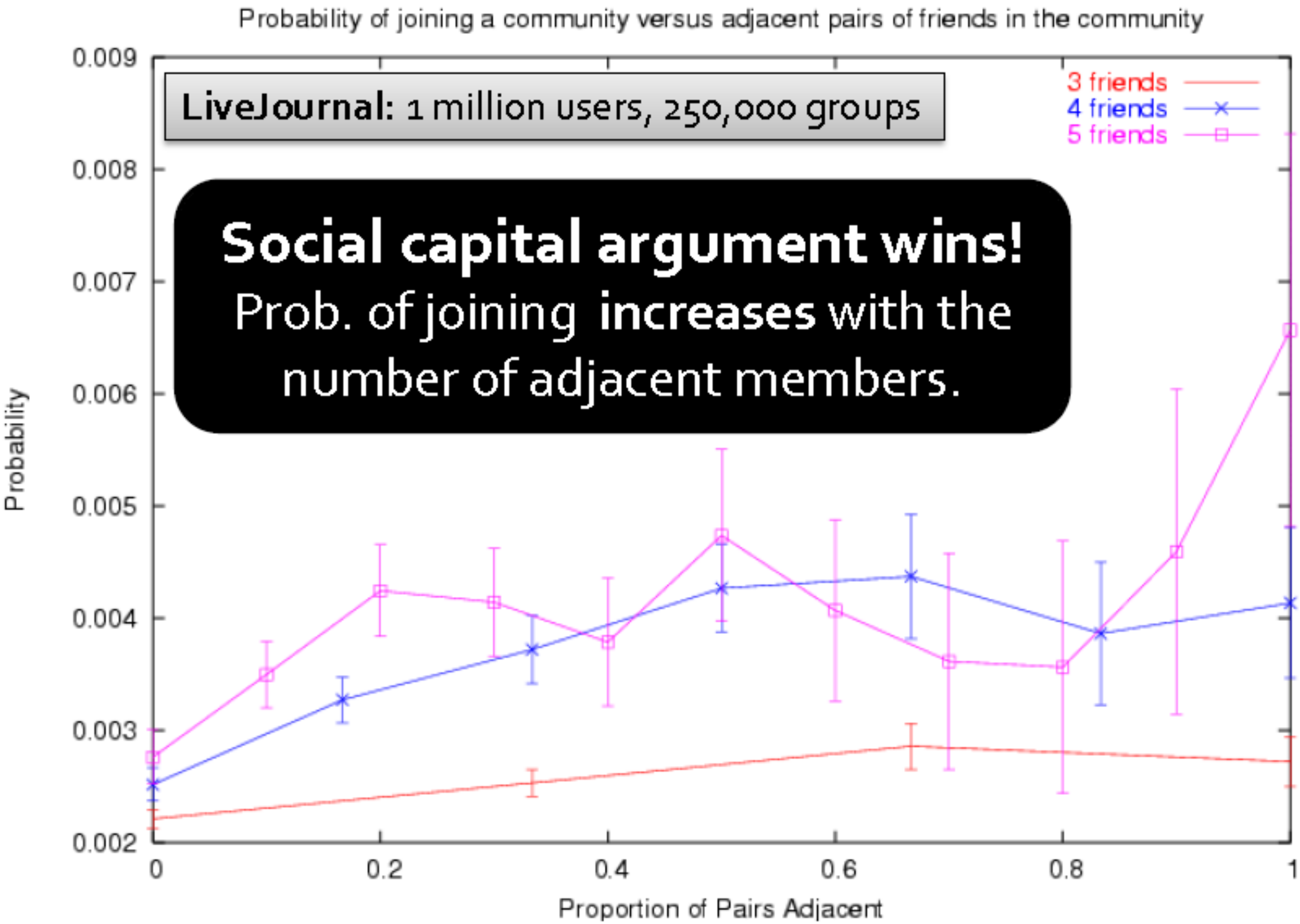
- **Competing sociological theories:**
  - Information argument [Granovetter '73]
  - Social capital argument [Coleman '88]



- **Information argument:**
  - Unconnected friends give independent support
- **Social capital argument:**
  - Safety/trust advantage in having friends who know each other

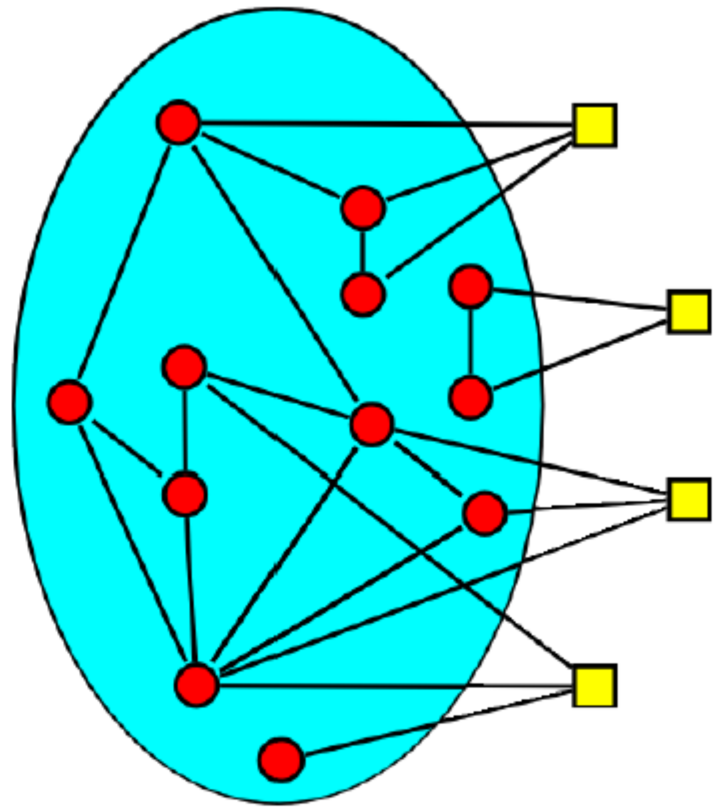
… and the winner is …

[Backstrom et al., KDD 2006]



The strength of **strong** ties

- **A person is more likely to join a group if**
  - she has more friends who are already in the group
  - friends have more connections between themselves
- **So, groups form clusters of tightly connected nodes**

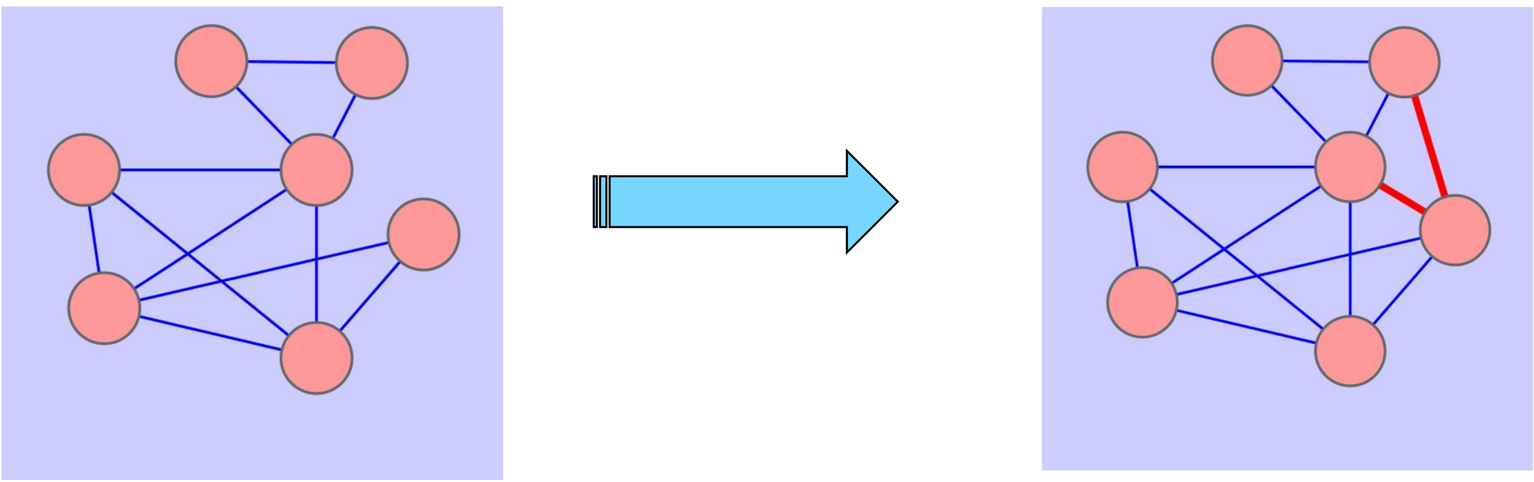


Link prediction

Which new links will appear in the social network?

Link prediction in social networks

- Can new social links be predicted?



Link prediction in social networks

- Social networks are very sparse
- Disproportion between possible links and links that actually form in the network.
- From a machine learning perspective, link prediction is a binary classification problem over an extremely unbalanced dataset, where positive cases are overwhelmed by negative cases.

The link prediction challenge

- In a phone call graph with 10^6 users, the average degree is around 4, so we have $4*10^6$ links, vs. the number of potential links in the order of 10^{12}
  - One new link every one million possibilities!
- Therefore, the trivial "**no-link**" classifier that always predicts the absence of any links has an extremely low classification error around 10^{-6}, i.e. an amazing accuracy of 99.999999 %!
- The challenge is in improving the **classification accuracy on the positive cases (precision).**

- Previous results seem to imply that new links form more likely WITHIN communtites rather than ACROSS communities

Unsupervised vs. Supervised methods

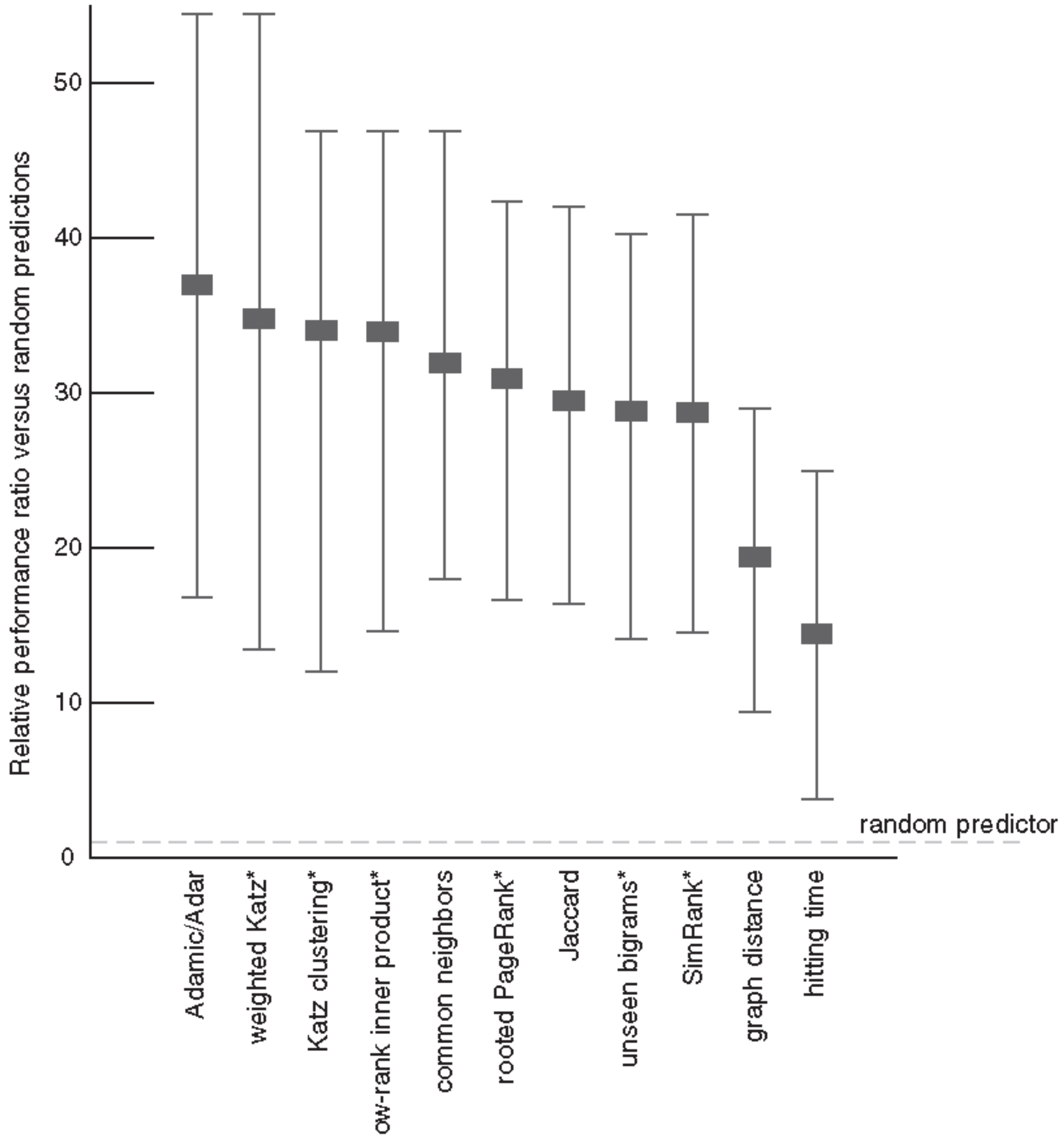
- **Unsupervised** link prediction, based on scores of topology measures such as common neighbors, Jaccard coefficient, Adamic/Adar measure, Katz
    - D. Liben-Nowell, J. Kleinberg. The link prediction problem for social networks. *J. of Am. Soc. for Information Science and Technology*, 58(7):1019-1031, 2007.
- **Supervised classification**, based on techniques for handling the disproportion of the negative cases of various machine learning/data mining methods
    - R. N. Lichtenwalter, J. T. Lussier, N. V. Chawla. New perspectives and methods in link prediction. ACM SIGKDD – Int. Conf on Knowledge Discovery in Databases. 2010.

How likely two nodes x and y belong to the same community?

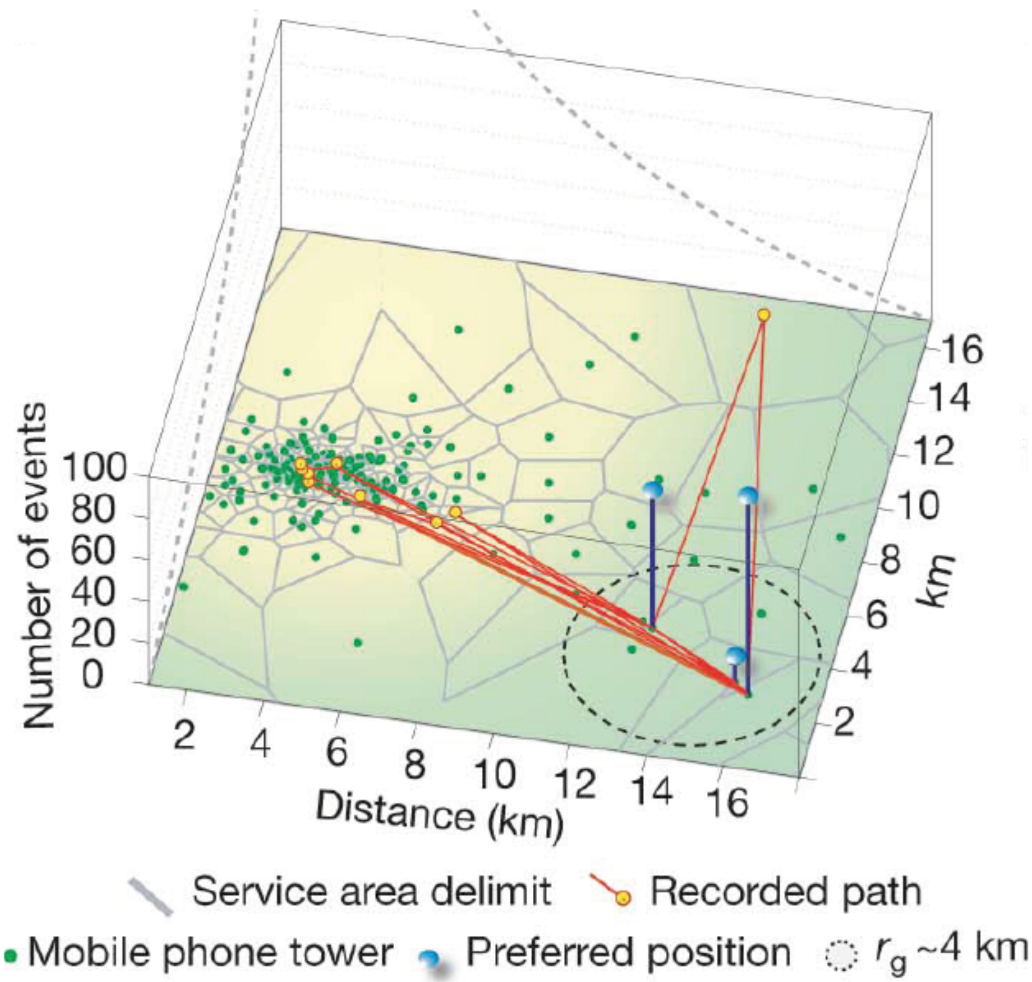
- [Liben-Nowell and Kleinberg 2006]

common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
Adamic/Adar	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \lvert\Gamma(z)\rvert}$
preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Katz_β	$\sum_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\text{paths}_{x,y}^{\langle\ell\rangle}\rvert$
	where $\text{paths}_{x,y}^{\langle\ell\rangle} := \{\text{paths of length exactly } \ell \text{ from } x \text{ to } y\}$

Performance of predictors (wrt random)



Country-wide tele-communication data



when you call

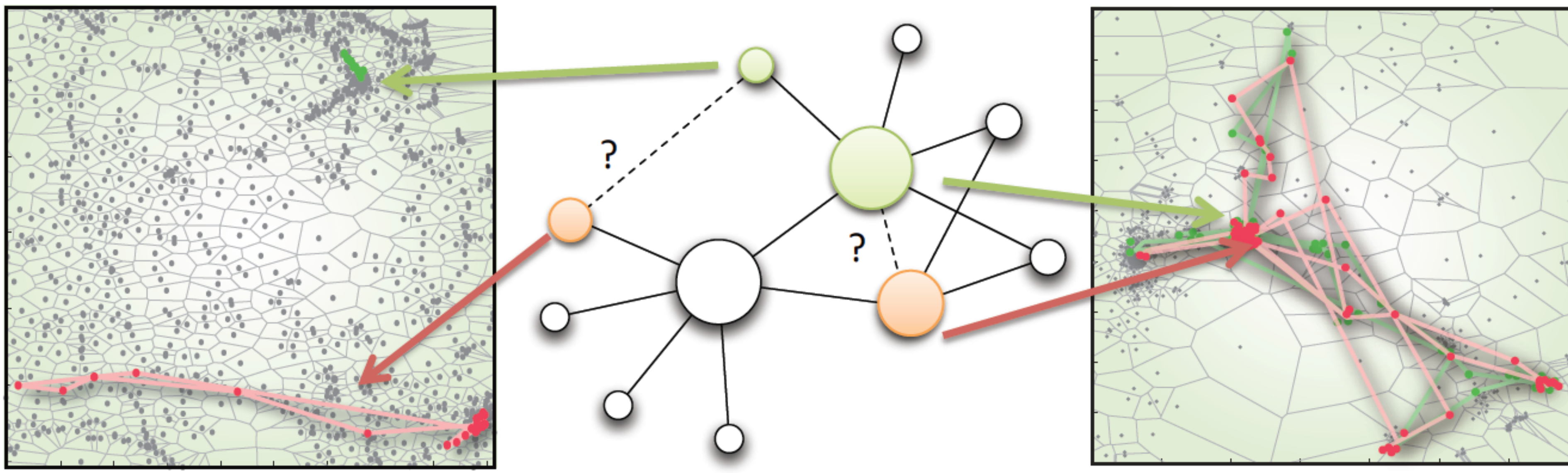
where you call

who you call

Link prediction in **mobile** social networks

- In mobile call records we have also location/ mobility in space and time as a further dimension, besides topology
- Is mobility a good predictor for future links?
- Can we build high-precision link predictors using combined topology/mobility features?

Link prediction in geo-social networks

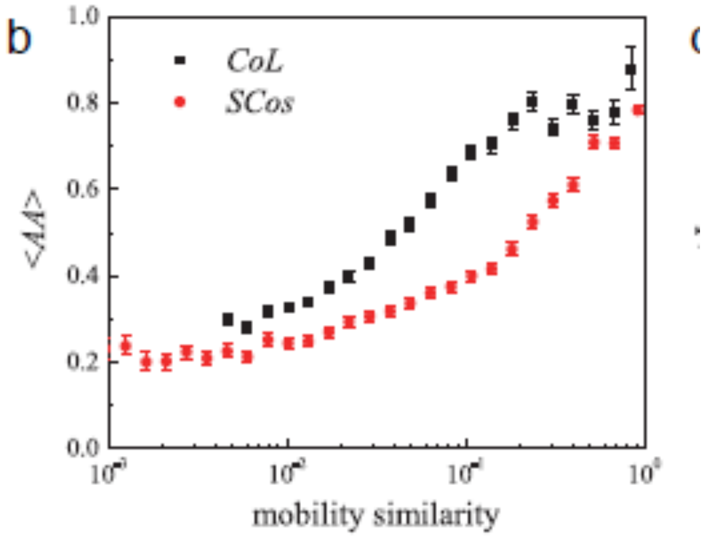


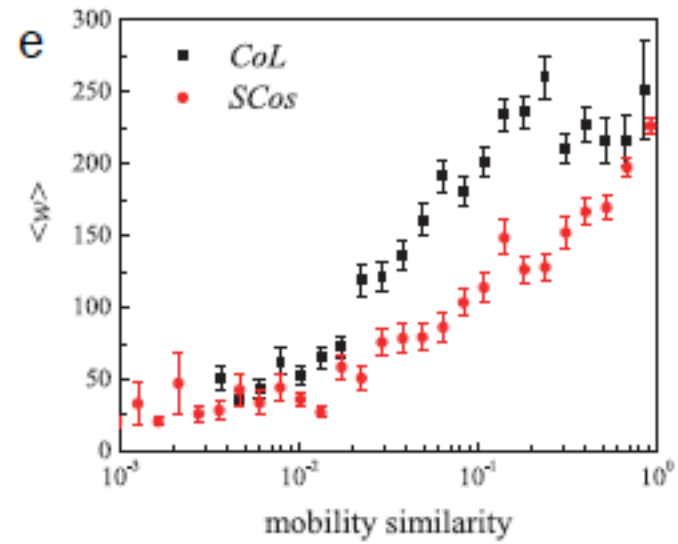
Correlation: Colocation, social proximity, tie strength

Table: Pearson Coefficients

	CoL	*SCos*	*J*	*CN*	*AA*	*K*	*w*
CoL	1	0.76	0.25	0.19	0.23	0.19	0.15
SCos	0.76	1	0.31	0.26	0.29	0.25	0.14
J	0.25	0.31	1	0.82	0.88	0.81	0.11
CN	0.19	0.26	0.82	1	0.94	0.99	0.06
AA	0.23	0.29	0.88	0.94	1	0.94	0.09
K	0.19	0.25	0.81	0.99	0.94	1	0.05
w	0.15	0.14	0.11	0.06	0.09	0.05	1

Human mobility and social ties

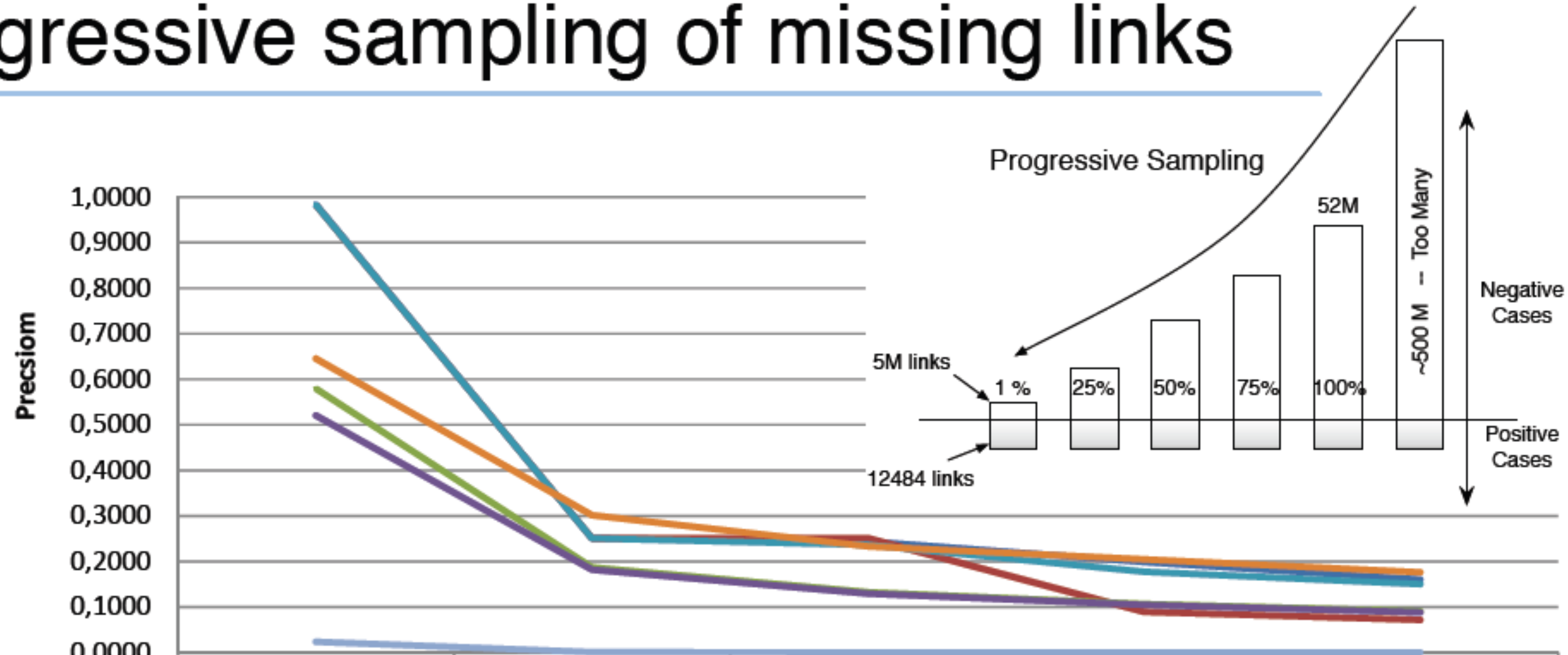




- co-location, network proximity and tie strength strongly correlate with each other
- measured on 3 months of calls, 6 Million users, nation-wide (large European country)
- **mobility dimension of the “strength of weak ties”**

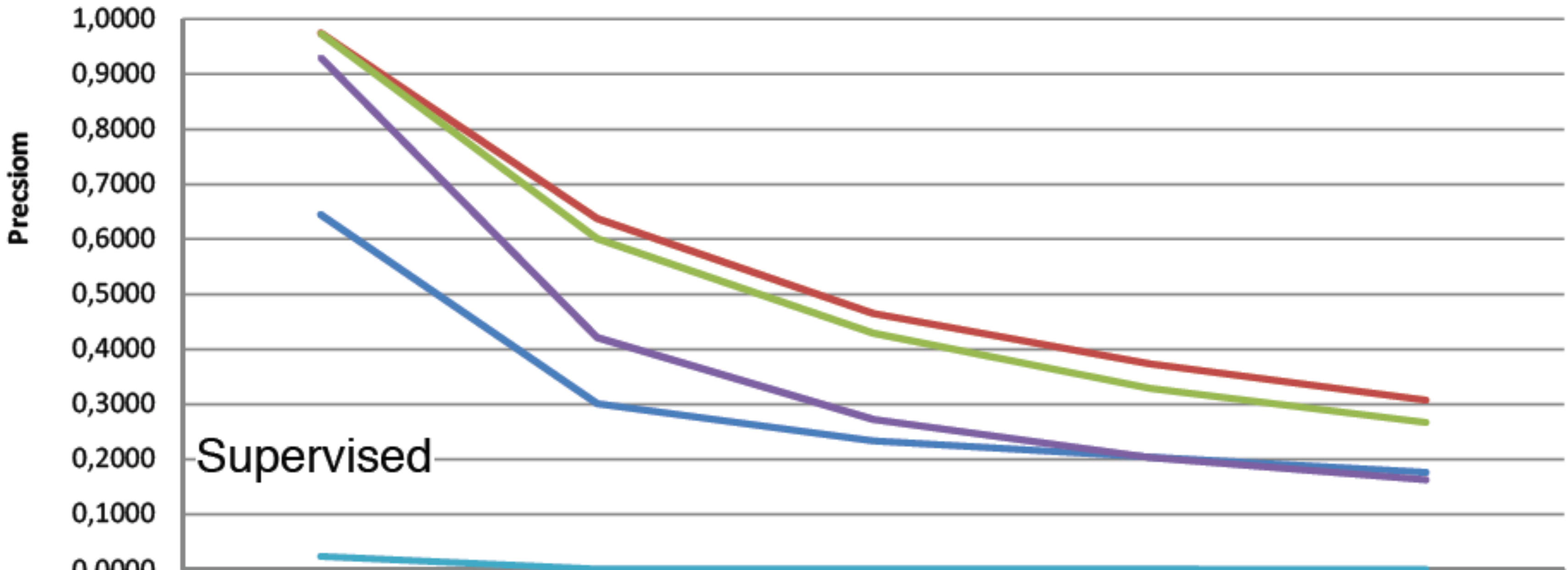
Unsupervised link prediction

Progressive sampling of missing links



	1%	25%	50%	75%	100%
Adamic Adar	0,9841	0,2507	0,2441	0,1988	0,1602
Common Neighbors	0,9829	0,2507	0,2507	0,0895	0,0715
Cosine Colocation	0,5794	0,1871	0,1325	0,1069	0,0906
ST Colocation	0,5203	0,1817	0,1295	0,1049	0,0884
Jaccard	0,9833	0,2507	0,2363	0,1777	0,1505
Katz	0,6451	0,3014	0,2333	0,2047	0,1762
Random	0,0237	0,0010	0,0005	0,0003	0,0002

Supervised link prediction



	1%	25%	50%	75%	100%
Katz (unsupervised)	0,6451	0,3014	0,2333	0,2047	0,1762
Topology & Mobility	0,9746	0,6378	0,4654	0,3740	0,3076
Topology	0,9741	0,6008	0,4294	0,3295	0,2668
Mobility	0,9306	0,4214	0,2724	0,2036	0,1629
Random	0,0237	0,0010	0,0005	0,0003	0,0002

Potential links with common neighbors

Unsupervised precision

Katz	9.1%
Adamic-Adar	7.8%
SCos	5.6%
Weighted *SCos*	5.6%
Extra-role *CoL*	5.1%
Weighted *CoL*	5.1%
CN	5.1%
CoL	5.0%
Jaccard	3.0%

Classification

	Pred. class=0	Pred. class=1
actual class=0	6,627	82
actual class=1	117	228

decision-tree: *AA*>0.5 and *SCoL*>0.7
73.5% precision and 66.1% recall

Combining topology and mobility measures is the key to achieving high precision and recall.

People is predictable!

- Probability of a new link between two (disconnected) random users:

  10^{-6}

- Best prediction accuracy using only social features:

  10%

- Best prediction accuracy using **social + mobility** features:

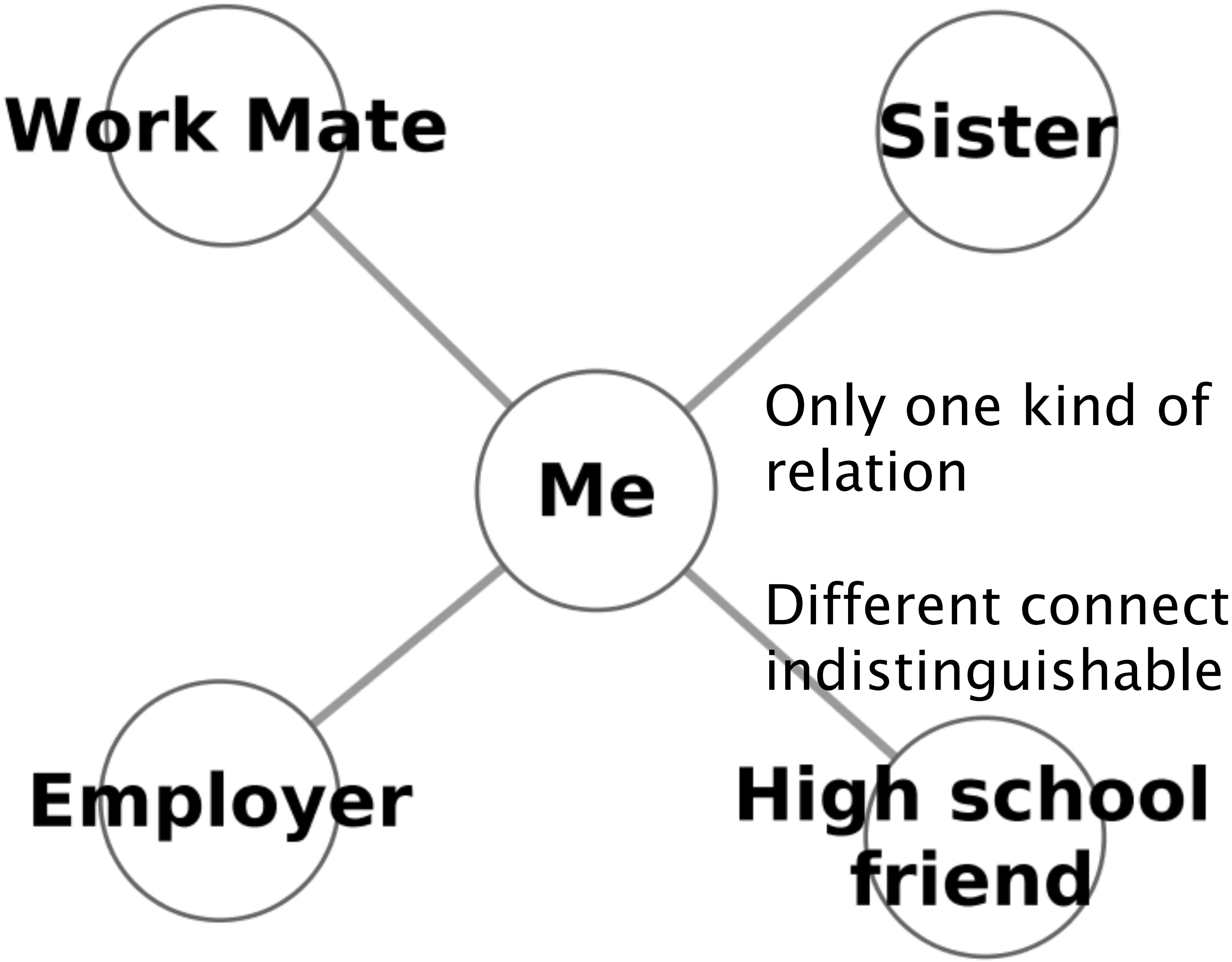
  75%

Multi-dimensional network analysis

M Berlingerio, M Coscia, F Giannotti, A Monreale, D Pedreschi. Multidimensional networks: foundations of structural analysis. *World Wide Web* 16 (5-6), 567-593 (2013)

Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, Dino Pedreschi: The pursuit of hubbiness: Analysis of hubs in large multidimensional networks. *Journal of Computational Science* 2(3): 223-237 (2011)

Classical Network Representation



Only one kind of relation

Different connections indistinguishable

Multigraphs as multidimensional networks