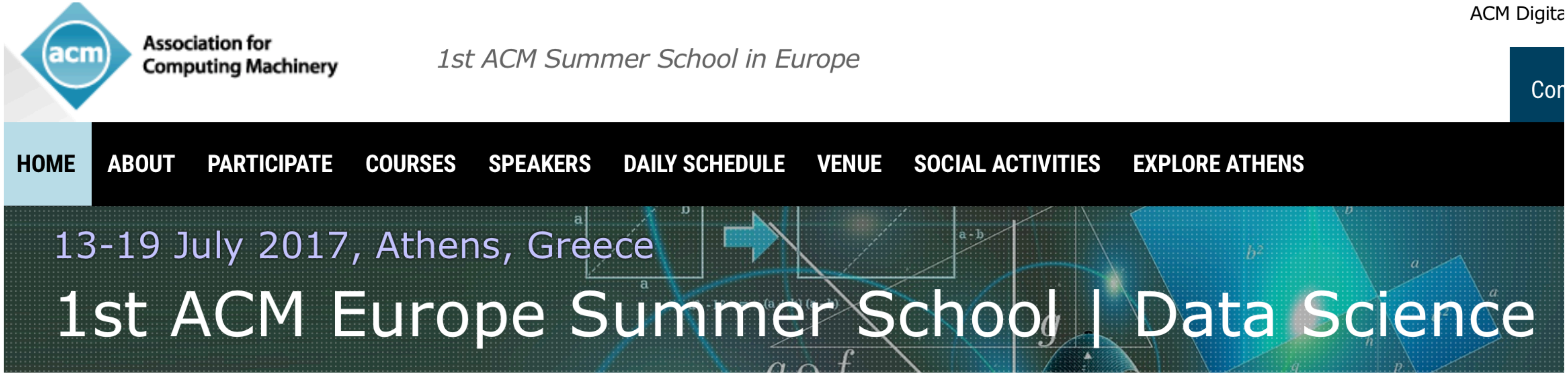


Social Network Analysis

Dino Pedreschi

KDD Lab

UNIVERSITÀ DI PISA

Università di Pisa & ISTI-CNR

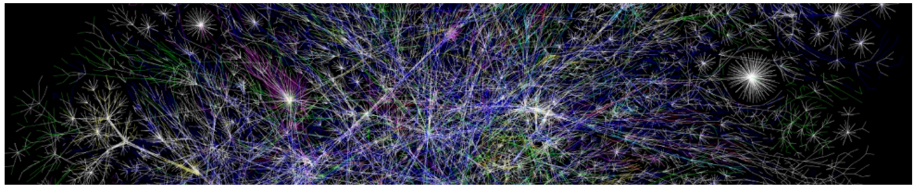
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

http://kdd.isti.cnr.it

KDD Lab

Home Research Projects People Resources Blog

Mobility Data Mining for Science of Cities



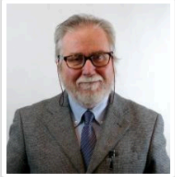
Social Network Analysis and Visual Analytics

Ethical Data Mining

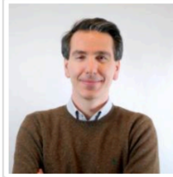
Analytical Platforms and Infrastructures for Social Mining

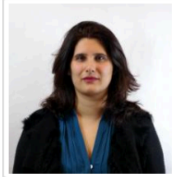
http://kdd.isti.cnr.it/people


Dino Pedreschi
Full Professor
pedre@di.unipi.it


Franco Turini
Full Professor
turini@di.unipi.it

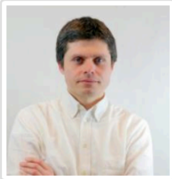

Fosca Giannotti
Director of Research
fosca.giannotti@isti.cnr.it


Salvatore Ruggieri
Full Professor
ruggieri@di.unipi.it


Anna Monreale
Assistant Professor
annam@di.unipi.it

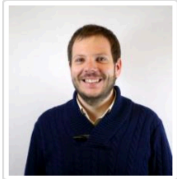

Alina Sîrbu
Assistant Professor
alina.sirbu@unipi.it

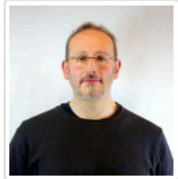

Barbara Furletti
Researcher
barbara.furletti@isti.cnr.it

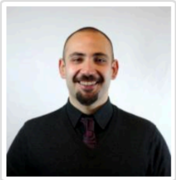

Mirco Nanni
Researcher
mirco.nanni@isti.cnr.it


Salvatore Rinzivillo
Researcher
rinzivillo@isti.cnr.it


Roberto Trasarti
Researcher
roberto.trasarti@isti.cnr.it


Paolo Cintia
Post Doc
paolo.cintia@isti.cnr.it

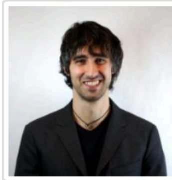

Valerio Grossi
Post Doc
vgrossi@di.unipi.it

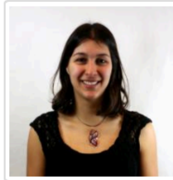

Luca Pappalardo
Post Doc
luca.pappalardo@isti.cnr.it


Giulio Rossetti
Post Doc
giulio.rossetti@isti.cnr.it

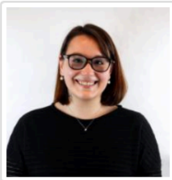

Alessandro Baroni
Ph.D. Student
baroni@di.unipi.it

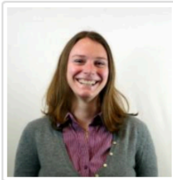

Lorenzo Gabrielli
Ph.D. Student
lorenzo.gabrielli@isti.cnr.it


Riccardo Guidotti
Ph.D. Student
riccardo.guidotti@isti.cnr.it


Ioanna Miliou
Ph.D. Student
ioanna.miliou@for.unipi.it


Letizia Milli
Ph.D. Student
letizia.milli@isti.cnr.it


Laura Pollacci
Ph.D. Student
laura.pollacci@di.unipi.it


Francesca Pratesi
Ph.D. Student
francesca.pratesi@isti.cnr.it


Farzad Vaziri
Ph.D. Student
vaziri@di.unipi.it


Viola Bachini
Collaborator
viola.bachini@isti.cnr.it


Daniele Fadda
Collaborator
daniele.fadda@isti.cnr.it

Knowledge Discovery & Data Mining Lab
http://kdd.isti.cnr.it

Complex (Social) Networks

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

Wiki of the course

- http://didawiki.di.unipi.it/doku.php/wma/acm-athens-july2017

- Special thanks to
  - Fosca Giannotti, ISTI-CNR Pisa
  - Albert-Laszlo Barabasi, Northeastern Univ. Boston
  - Giulio Rossetti, University of Pisa
  - Jure Leskovec, Stanford Univ.

The power of complex networks

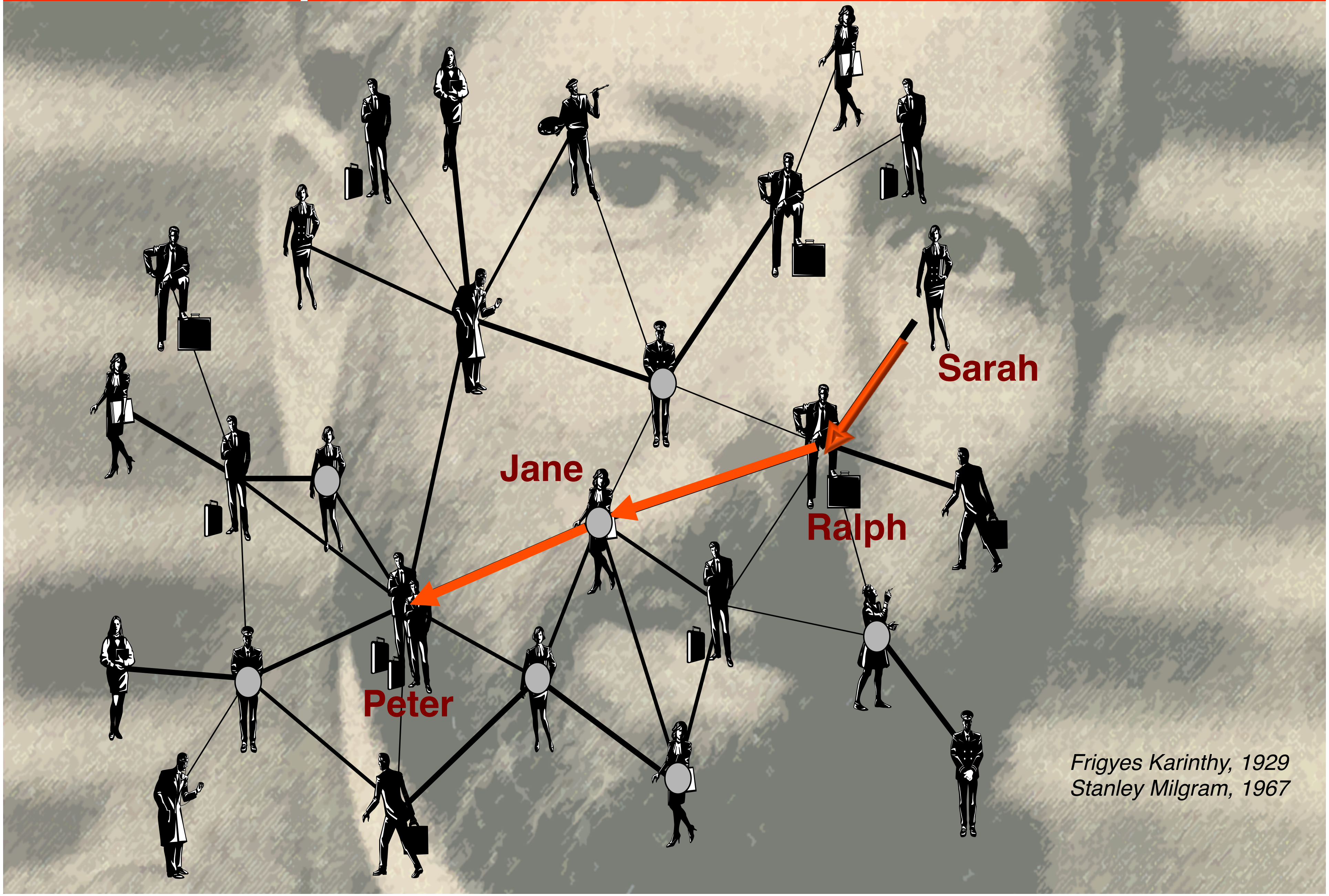
Lecture 2

Part 2

- Measuring small-worlds with big data
- Strength of weak ties
- Centrality measures
- Community structure
- Link prediction
- Robustness
- Cascades
- Epidemic spreading

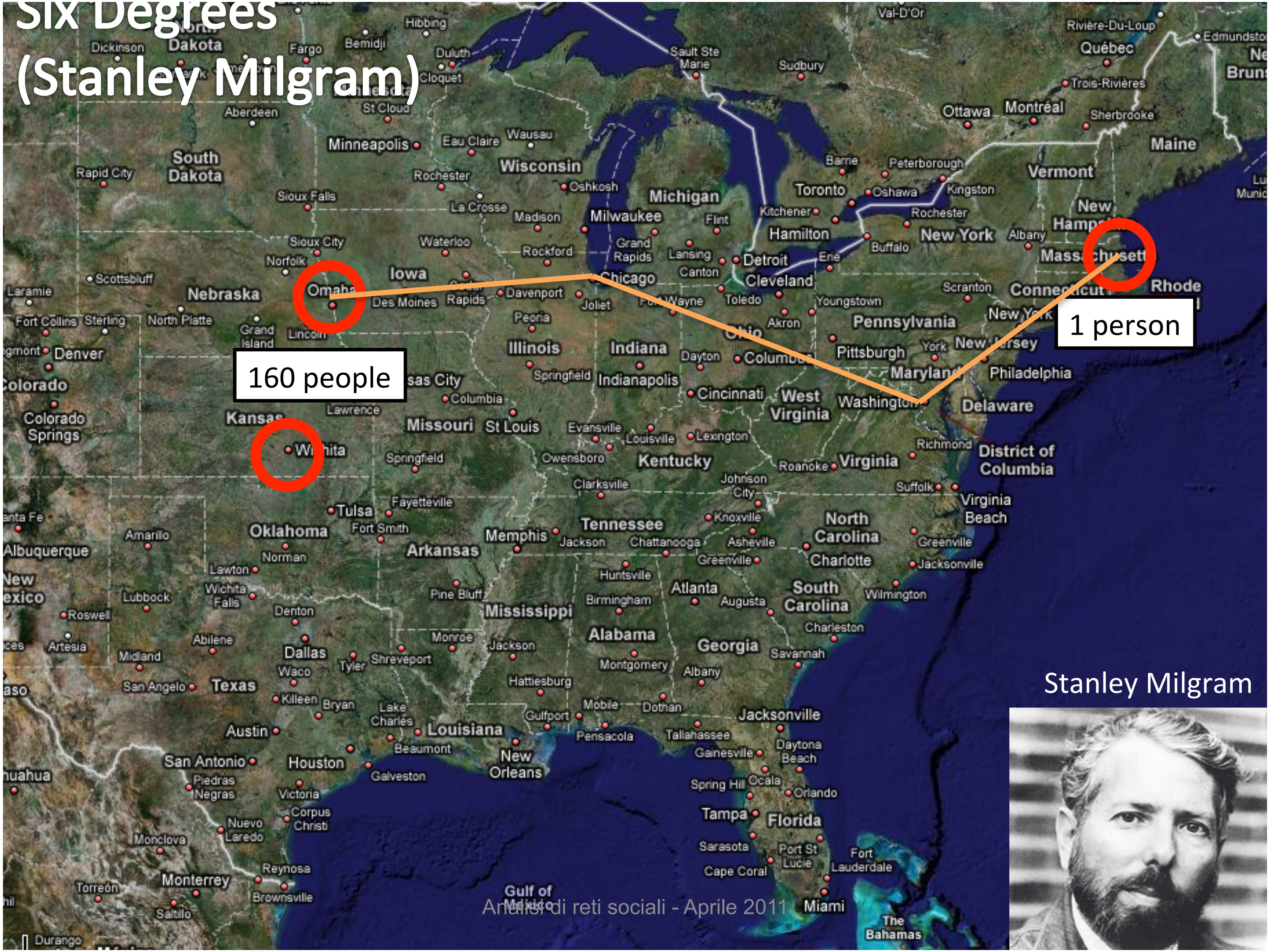
Measuring the small-world property

SIX DEGREES | small worlds



Frigyes Karinthy, 1929
Stanley Milgram, 1967

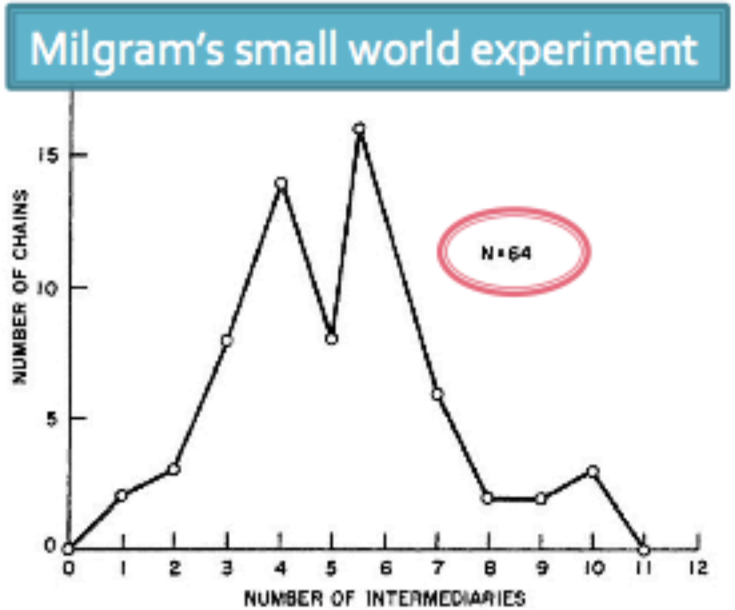
Six Degrees (Stanley Milgram)



Stanley Milgram

The Small-world experiment

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



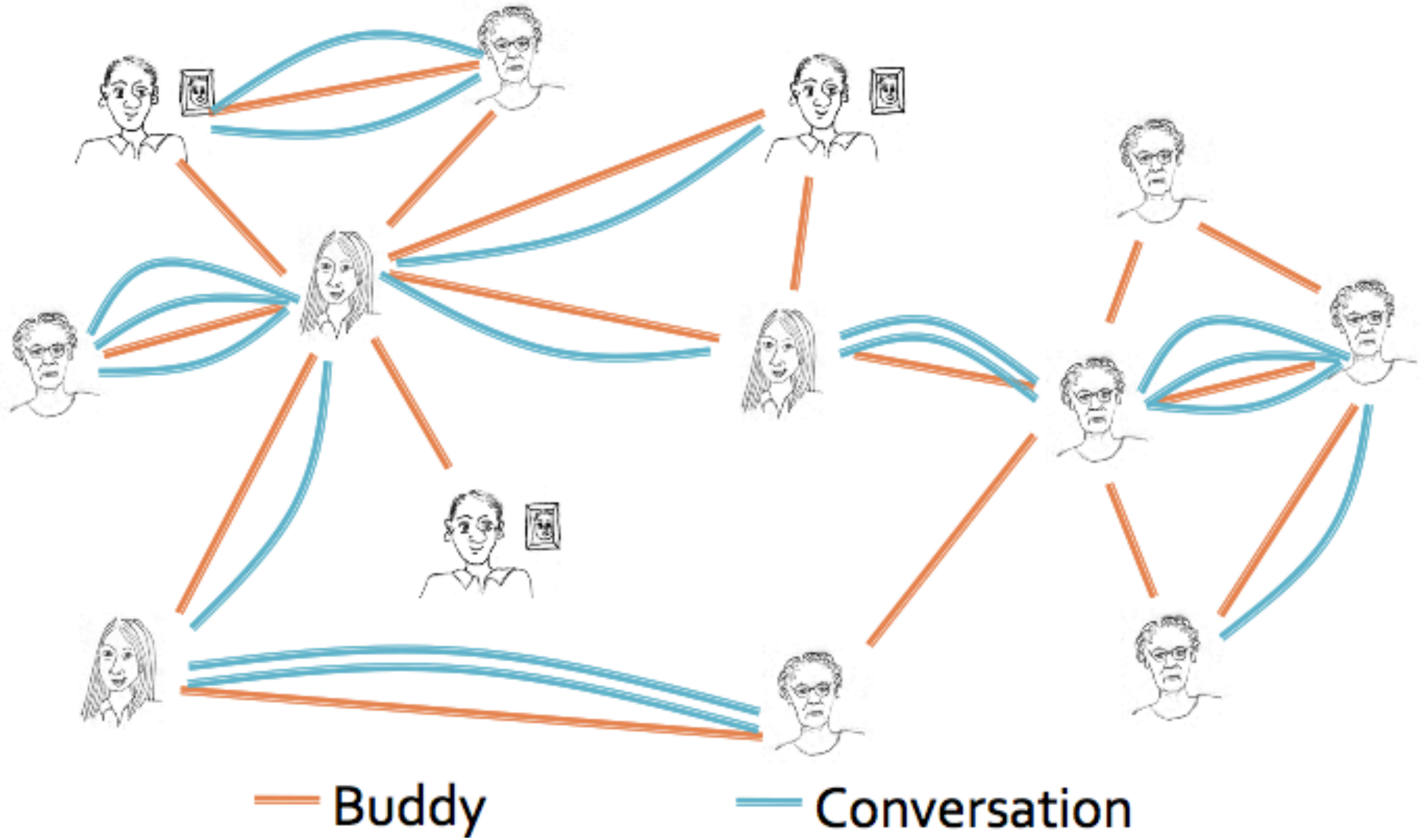
- Further observations:
  - People what owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
  - People from the Boston area have even closer paths: 4.4

Planetary-Scale Views on an Instant-Messaging Network

Jure Leskovec & Eric Horvitz

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

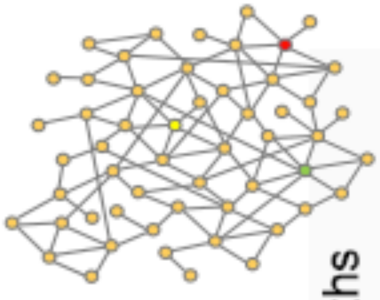
Messaging as a network

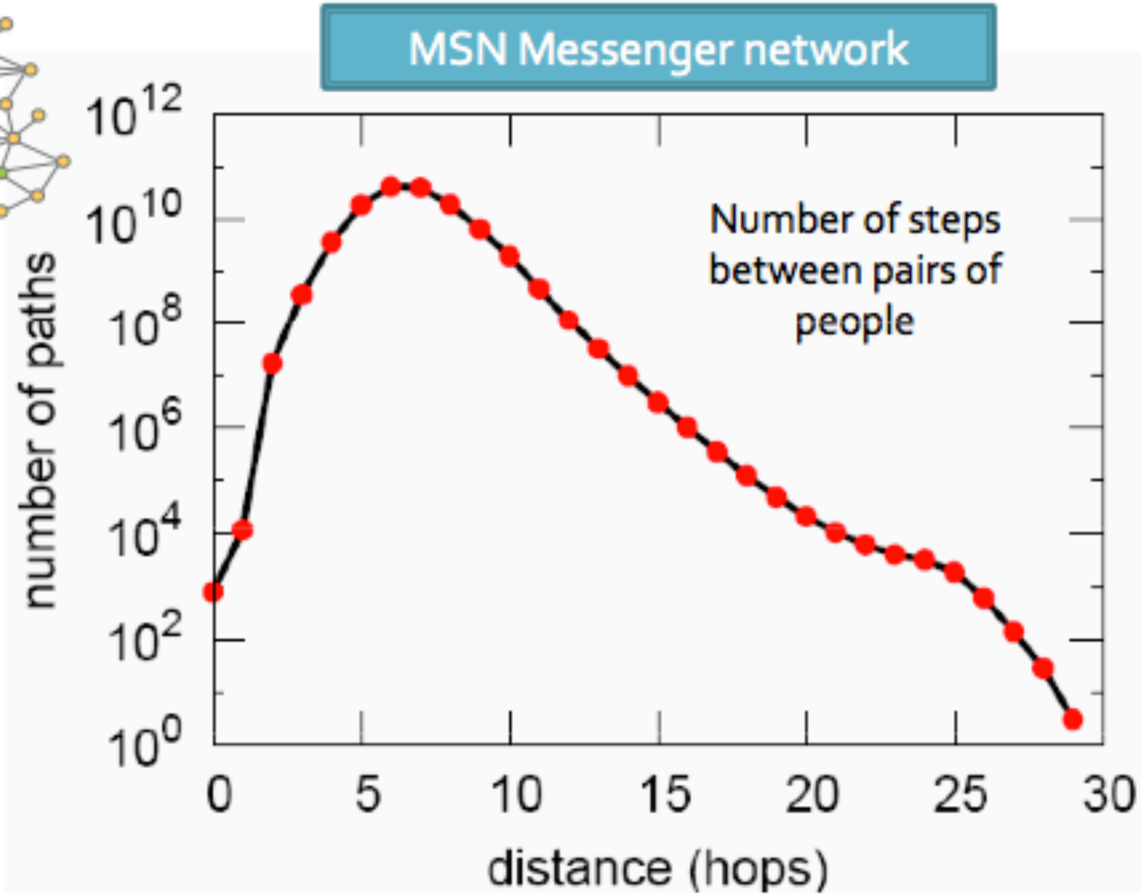


IM communication network

- **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



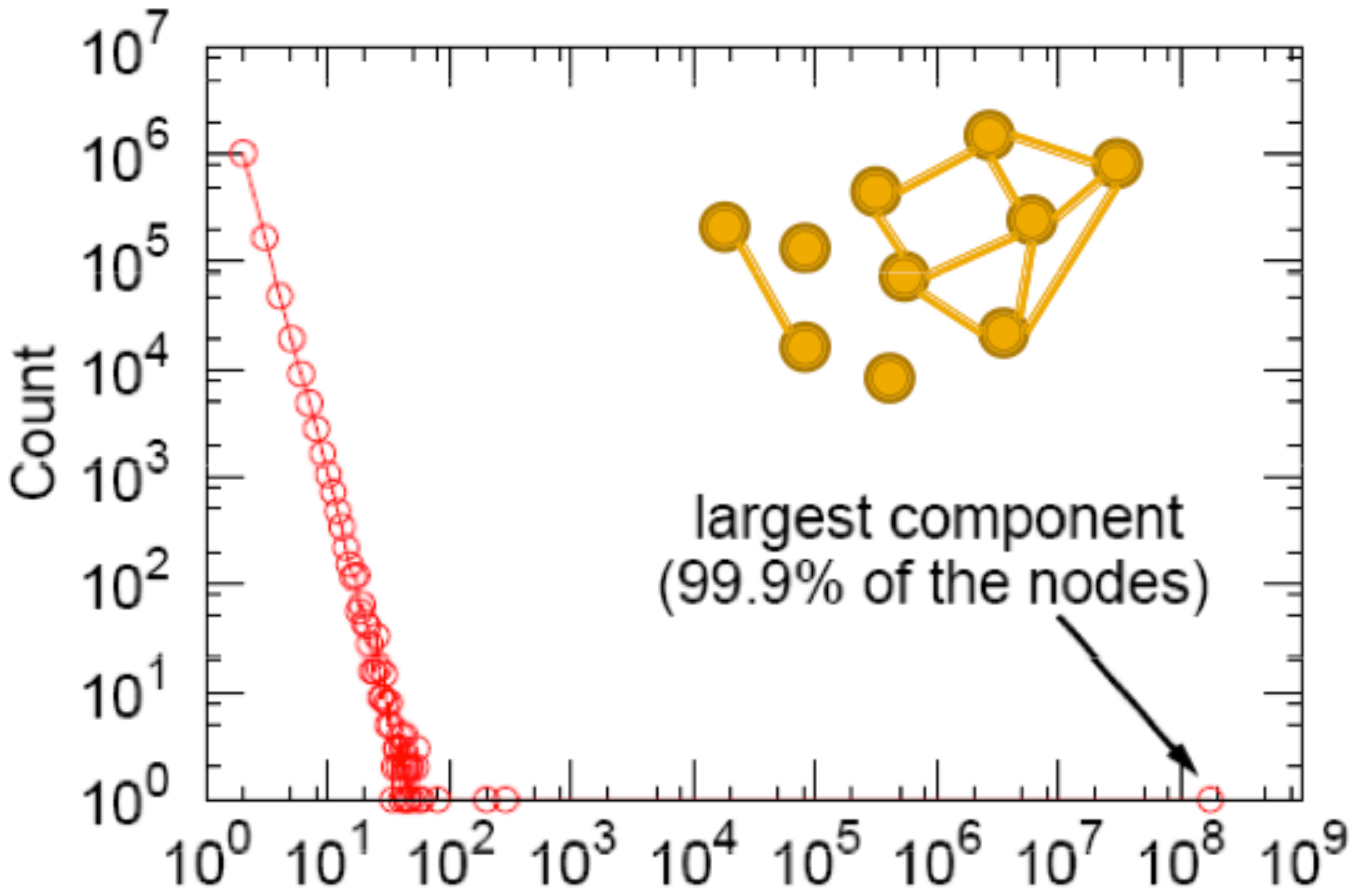


Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Avg. path length **6.6**

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

The giant connected component



The strength of weak ties

The strength of weak ties

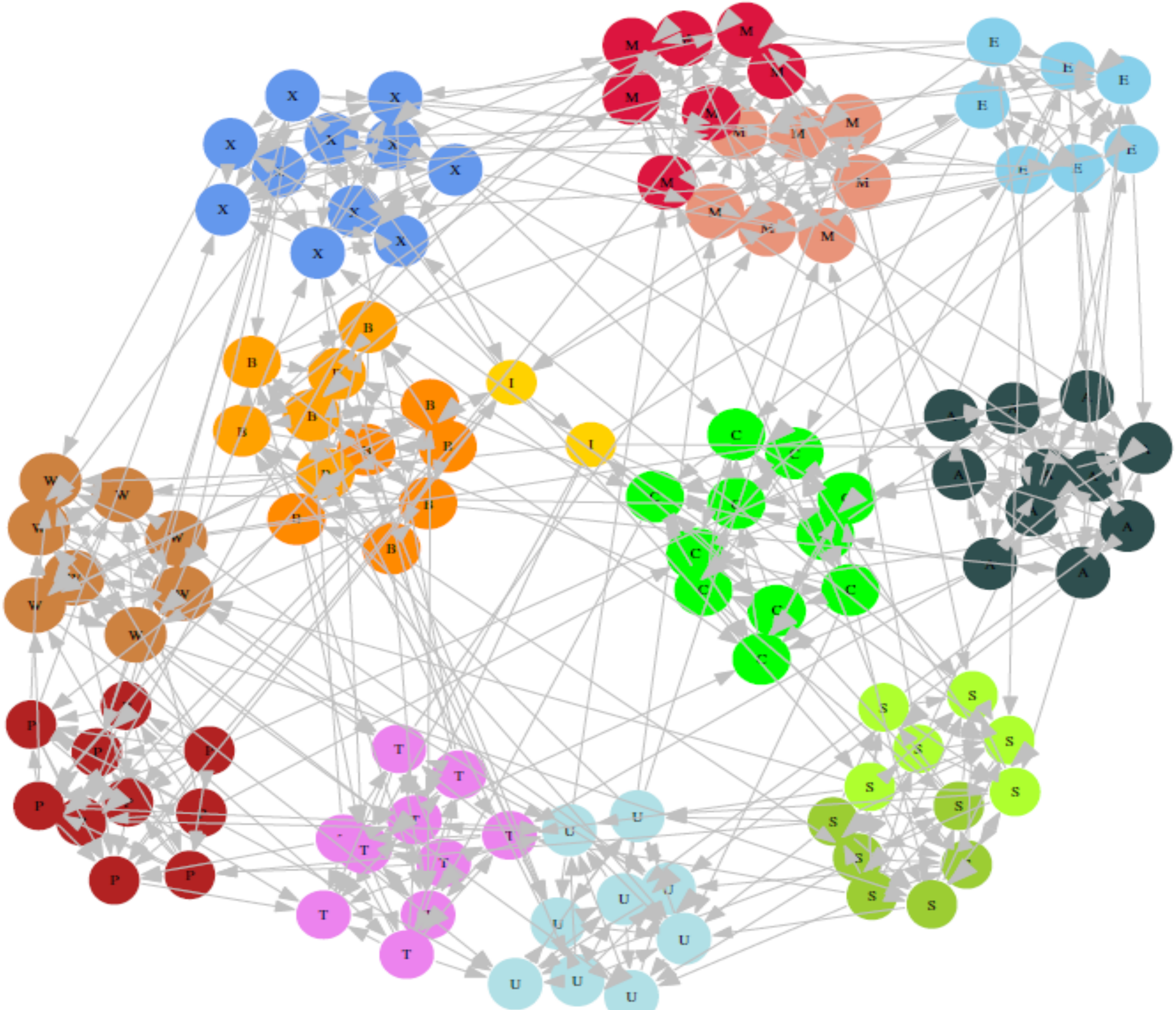
- Mark S. **Granovetter**, 1973
- His PhD thesis: how people get to know about new jobs?
- Through personal contacts
- Surprise: often acquaintances, **not** close friends
- Why?

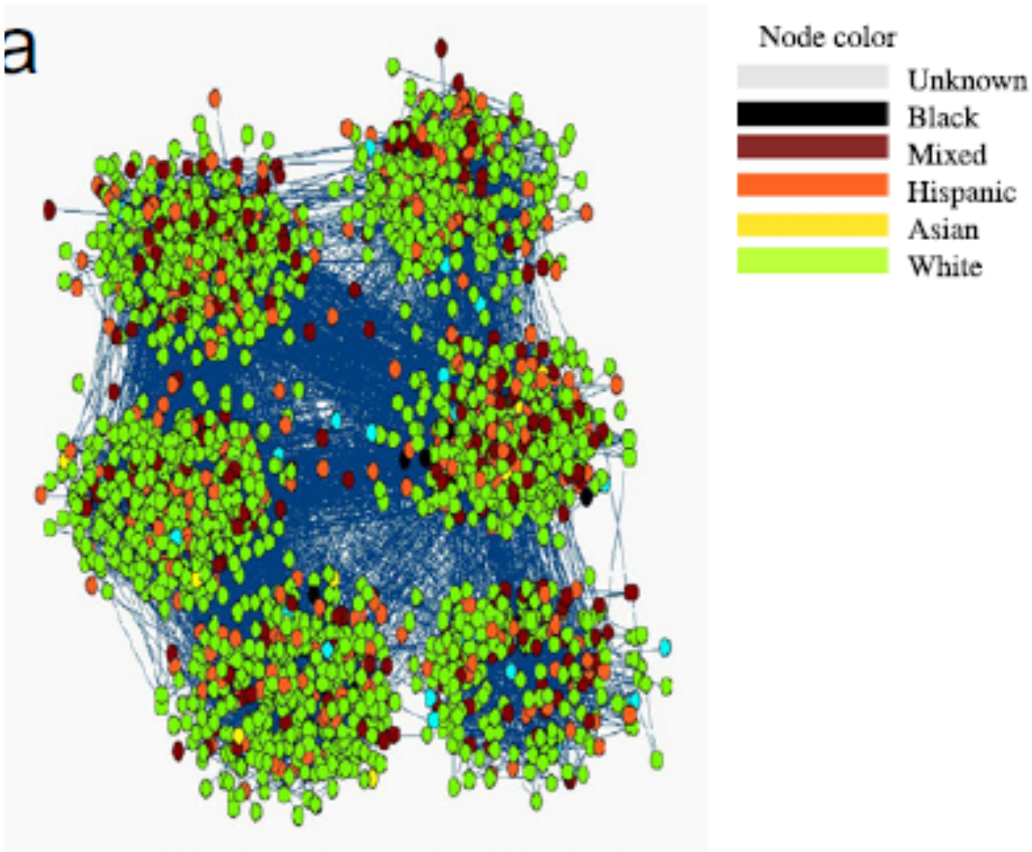
JSTOR

The Strength of Weak Ties

Mark S. Granovetter

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





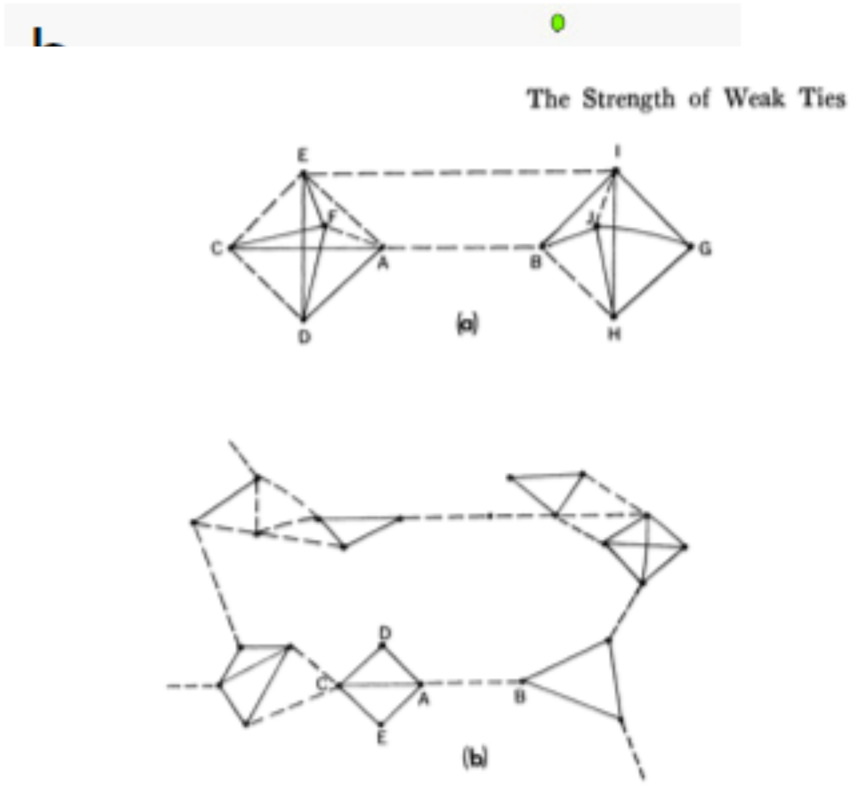
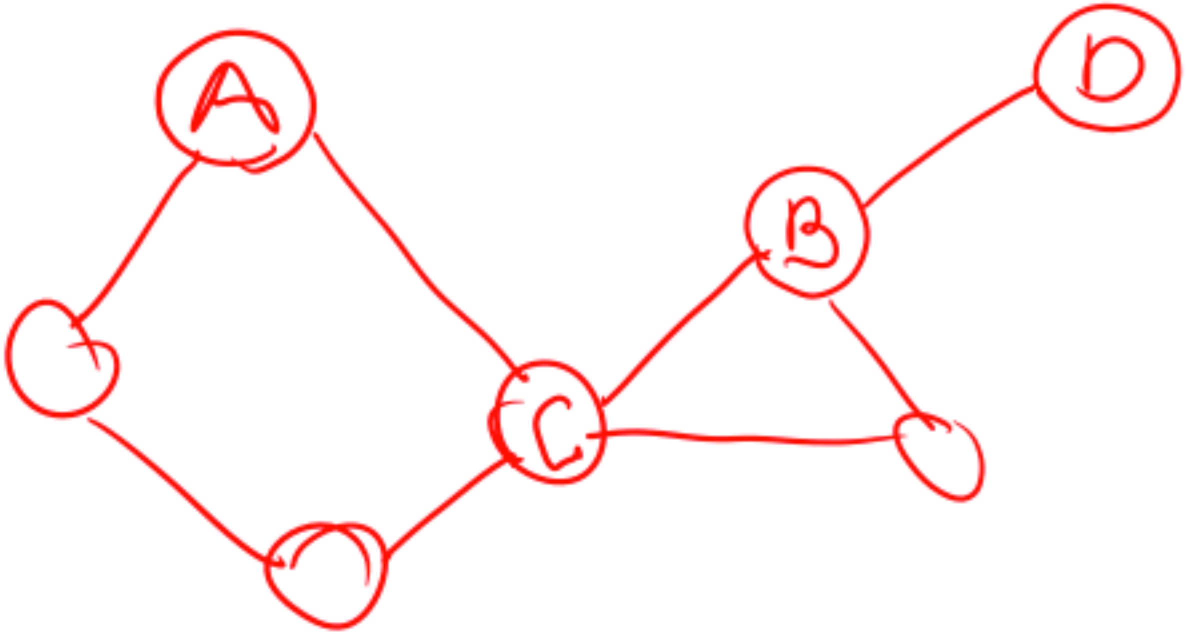


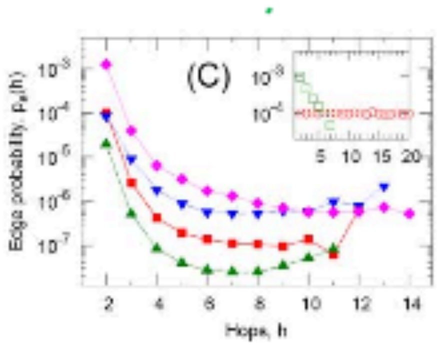
FIG. 2.—Local bridges. *a*, Degree 3; *b*, Degree 13. ——— = strong tie; — — — = weak tie.

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:

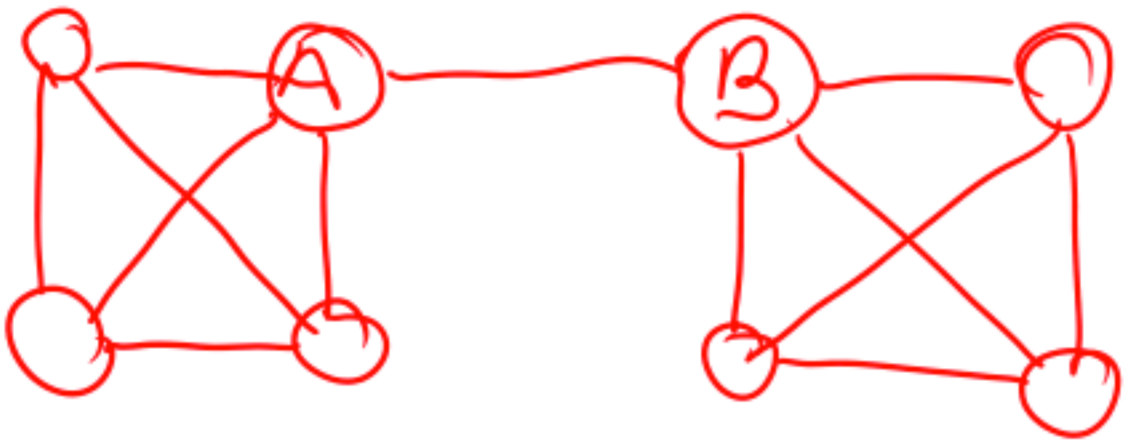
- If B and C have a friend A in common, then:
  - B is more likely to meet C
    - (since they both spend time with A)
  - B and C trust each other
    - (since they have a friend in common)
  - A has incentive to bring B and C together
    - (as it is hard for A to maintain two disjoint relationships)

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)

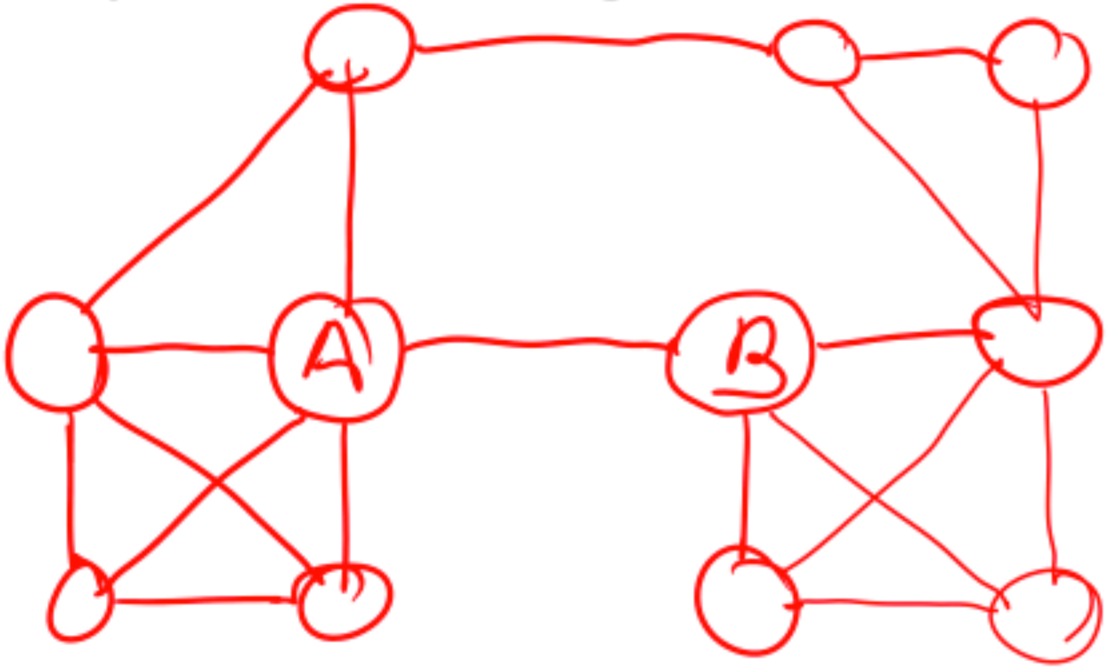
Bridges and Local Bridges

- 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)

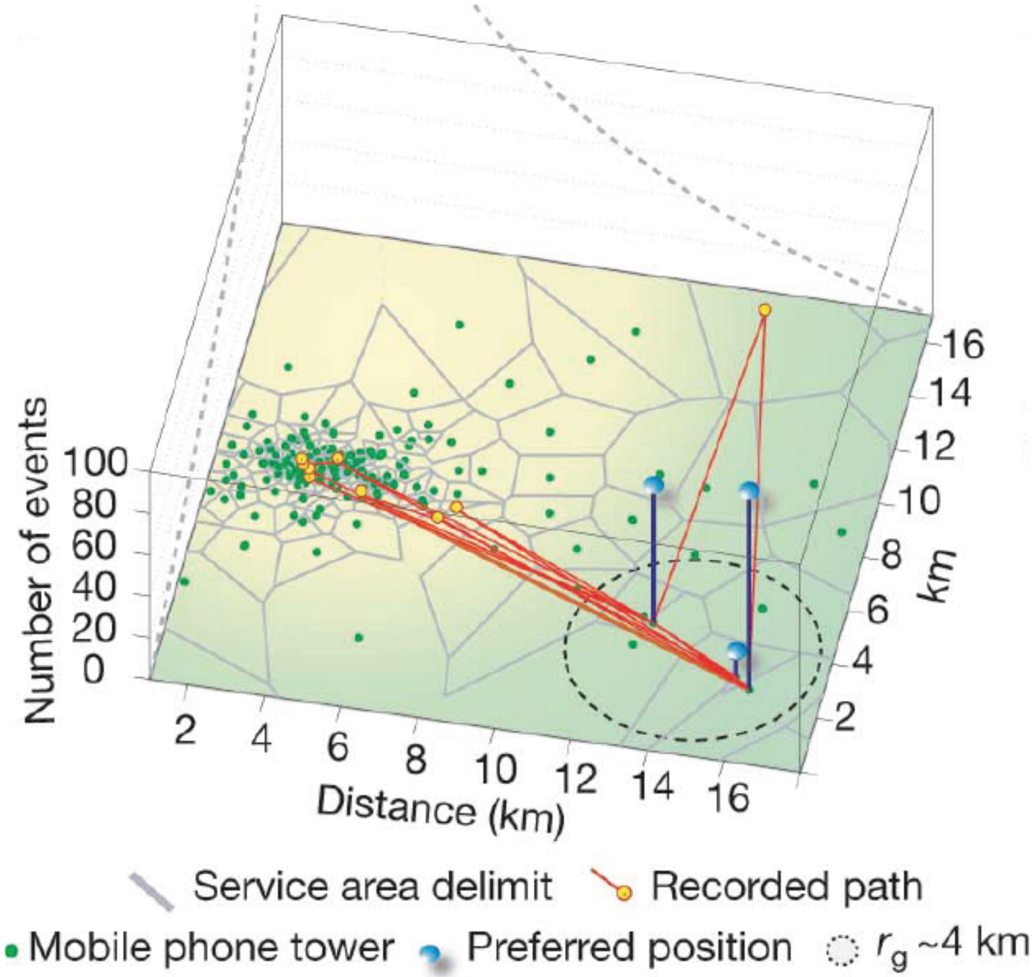
Local Bridges and Weak ties

- <u>Claim:</u> 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).

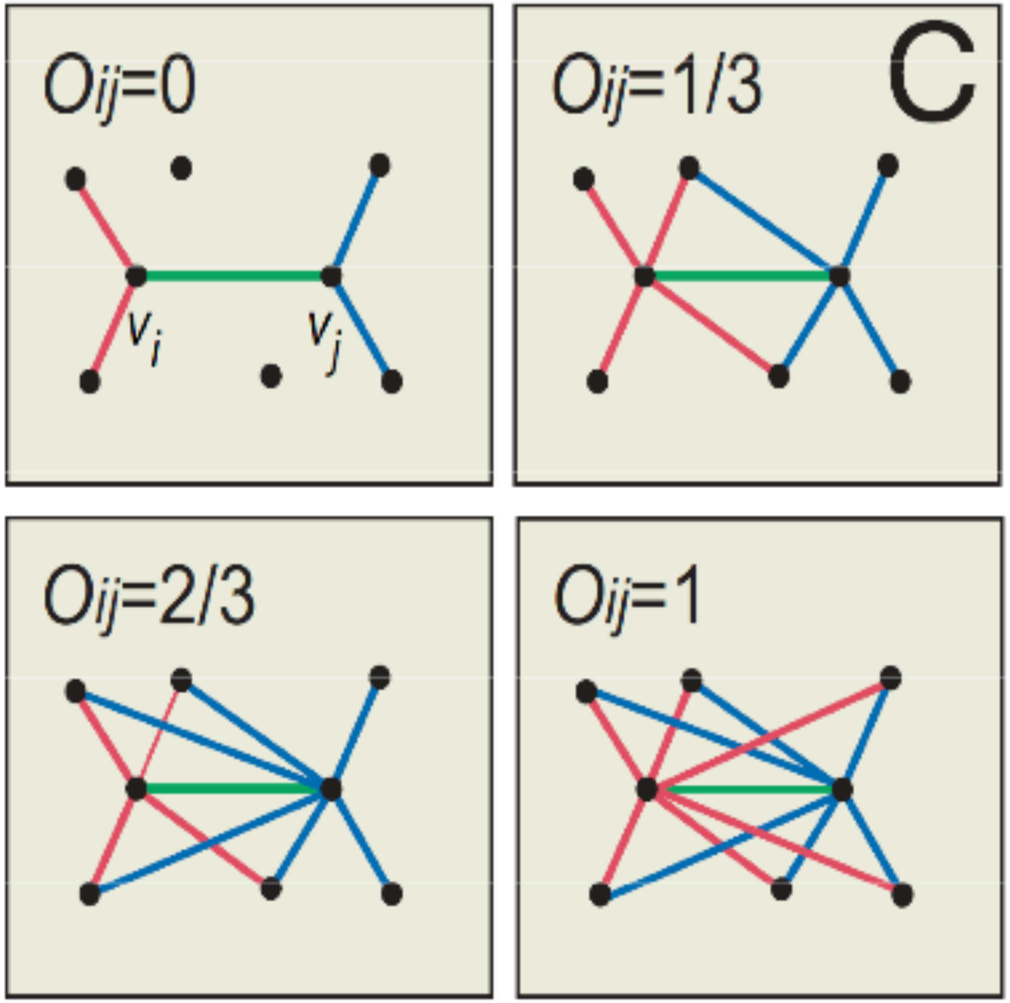
Neighborhood Overlap

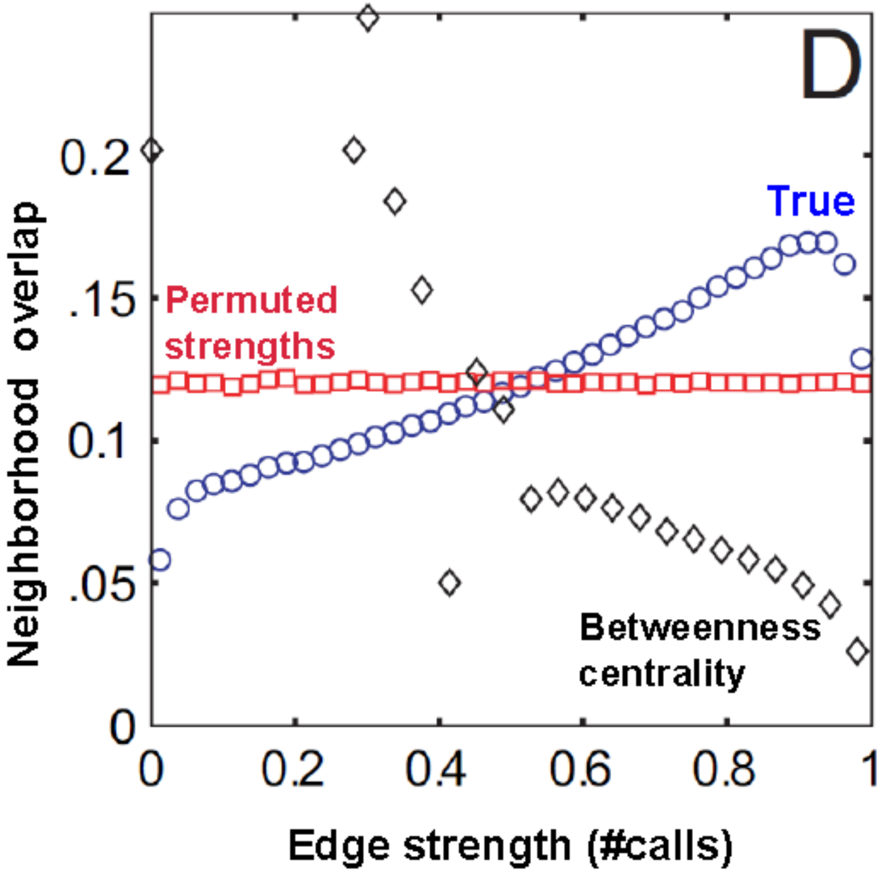
- Overlap:

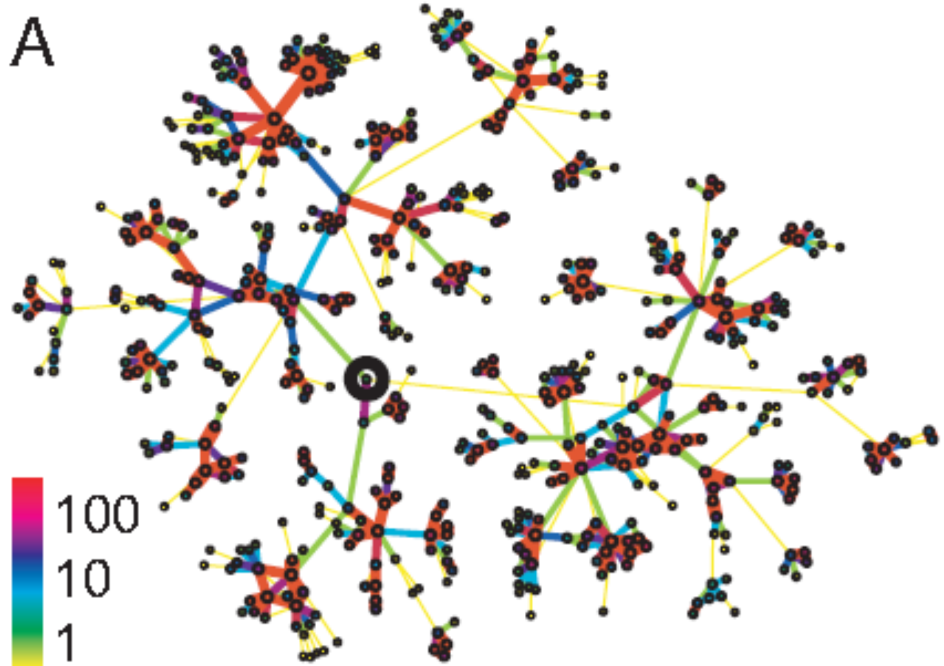
  $$O_{ij} = \frac{n(i) \cap n(j)}{n(i) \cup n(j)}$$

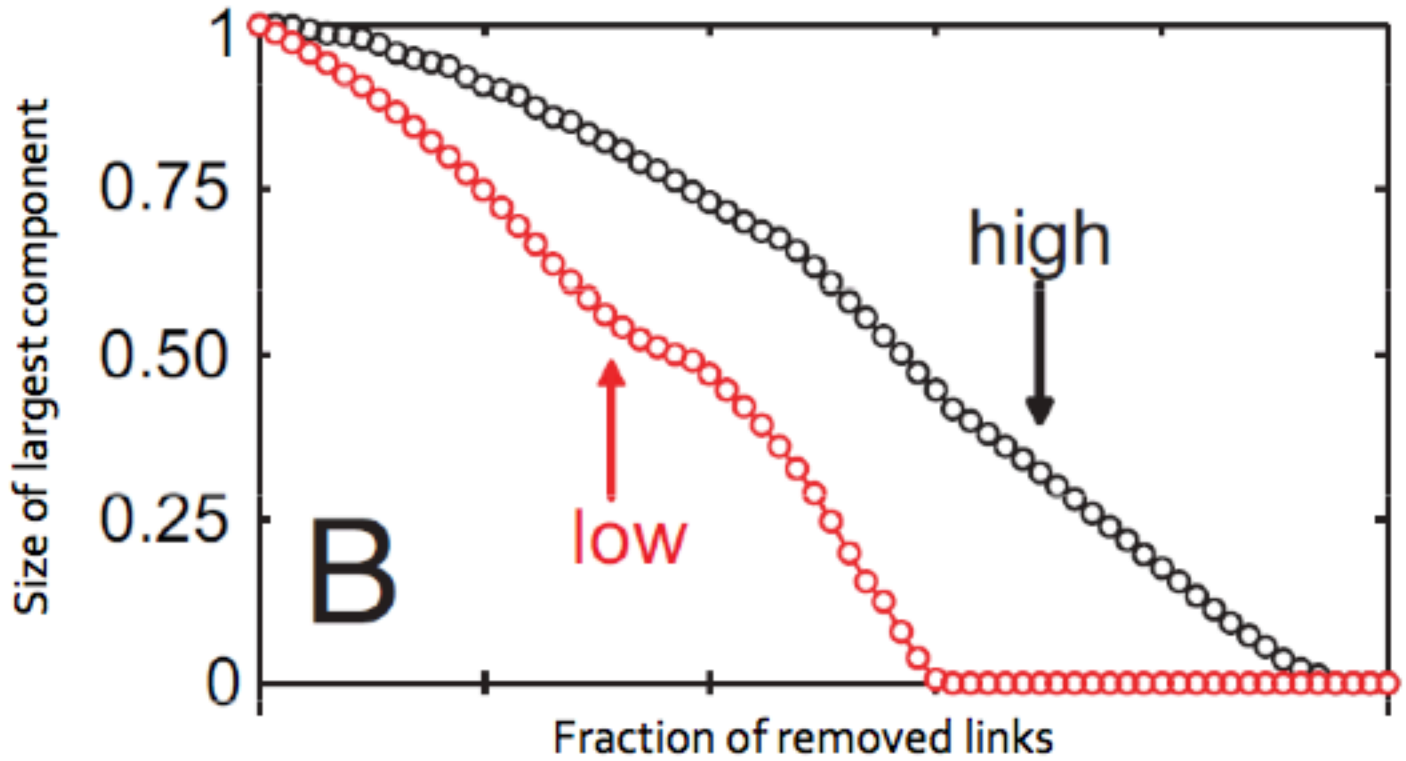
  - n(i) … set of neighbors of A

- 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

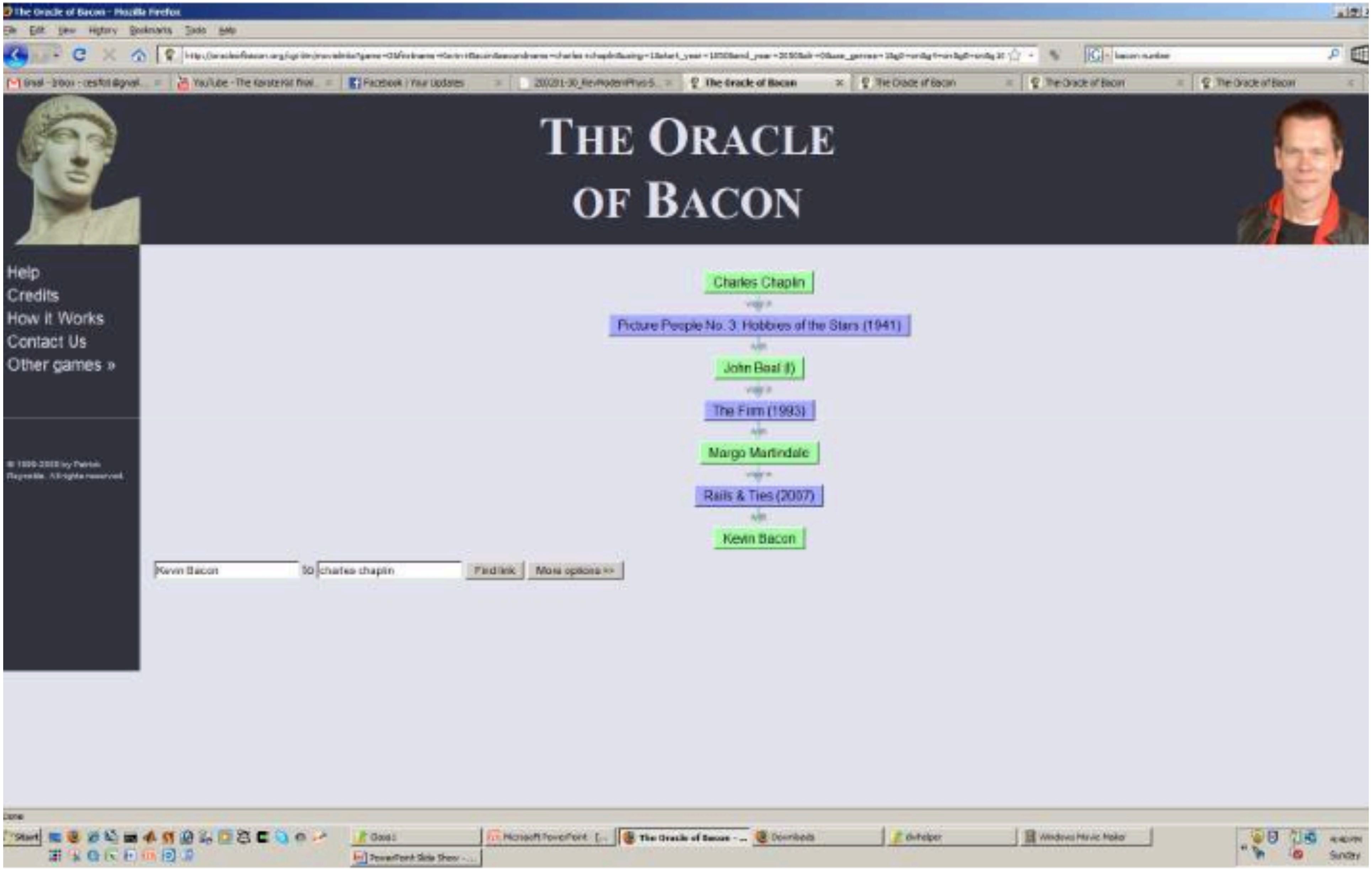

A
100
10
1



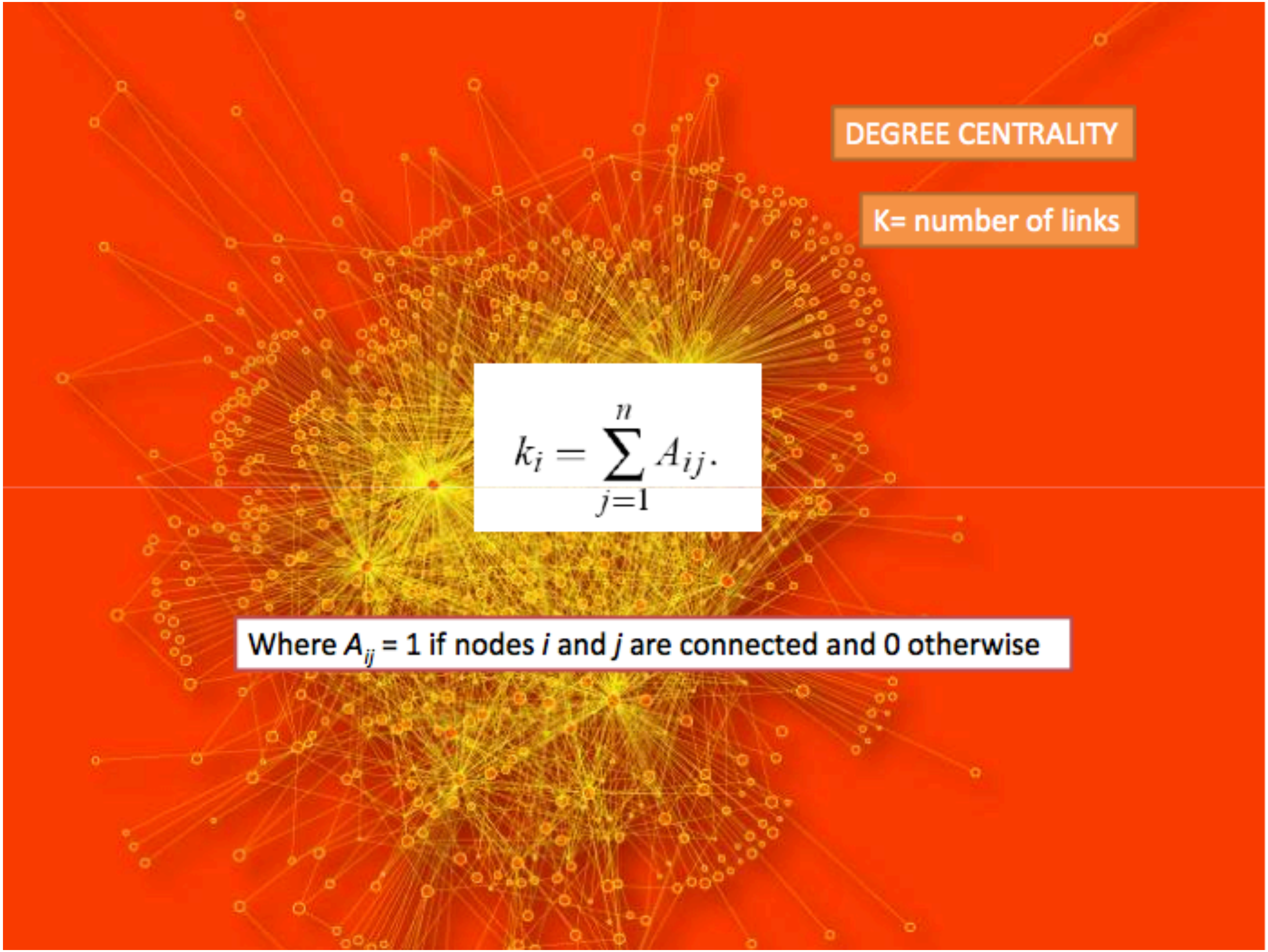
- Removing links based on overlap
  - Low to high
  - High to low

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

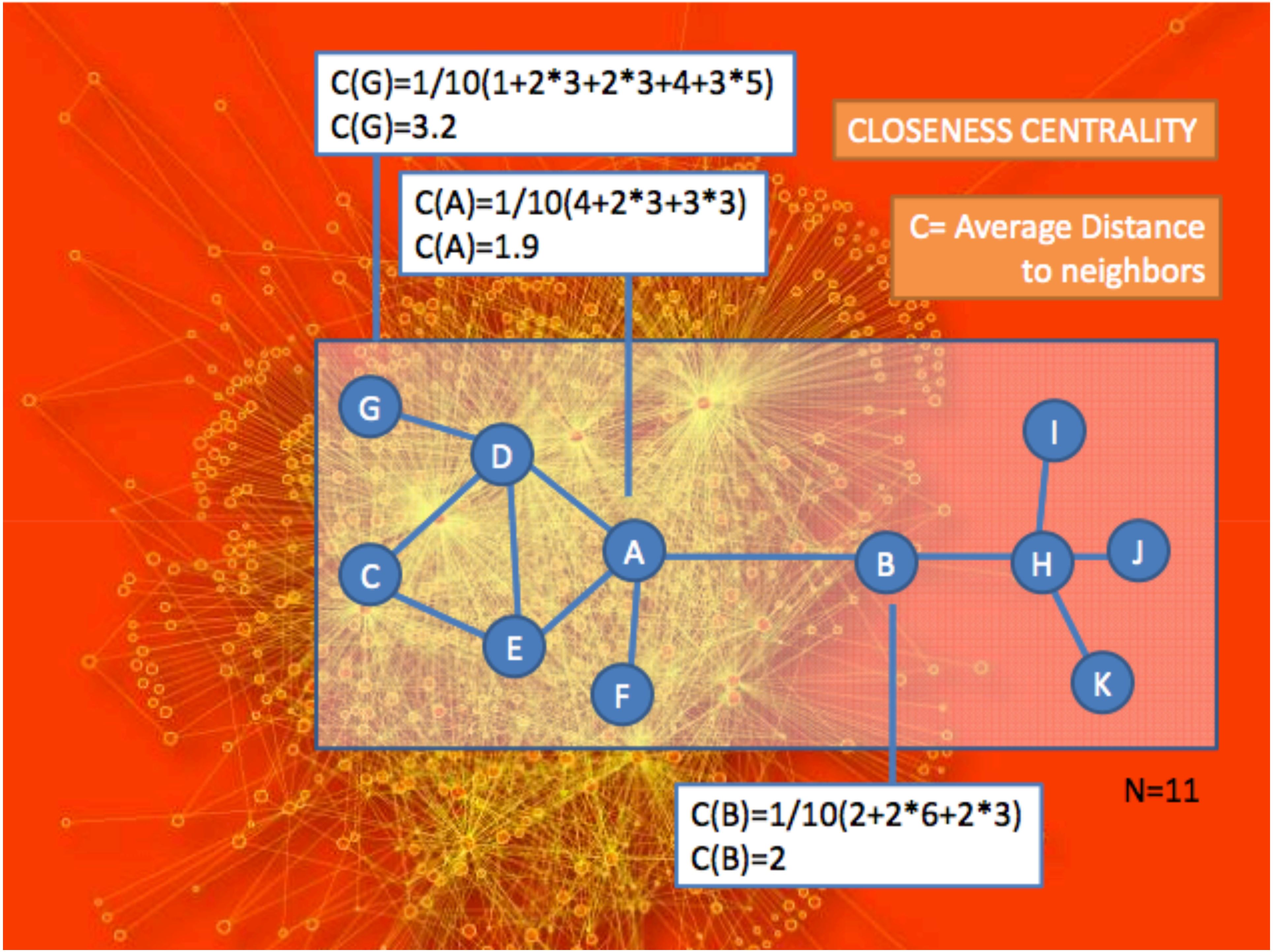
Most Connected Actors in Hollywood

(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

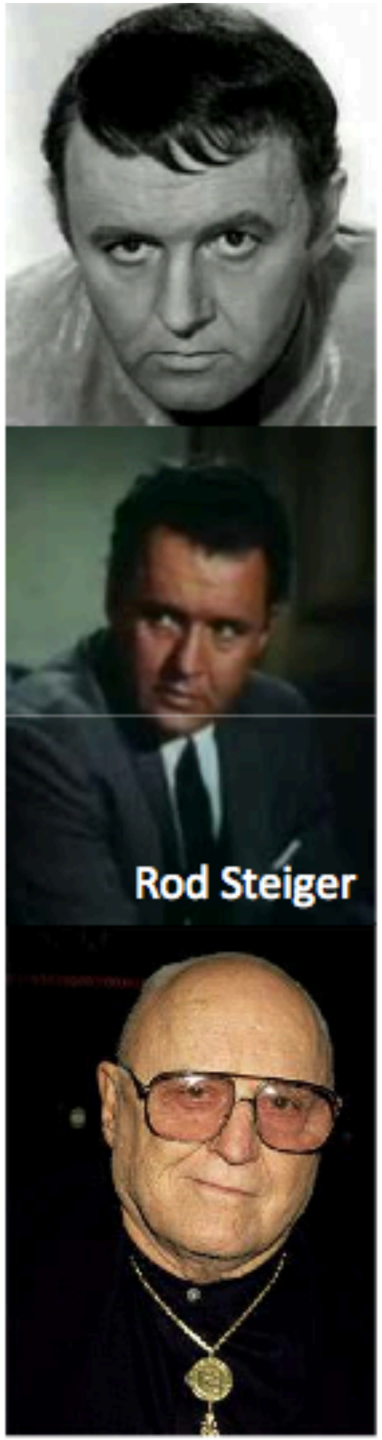
A-L Barabasi, "Linked", 2002

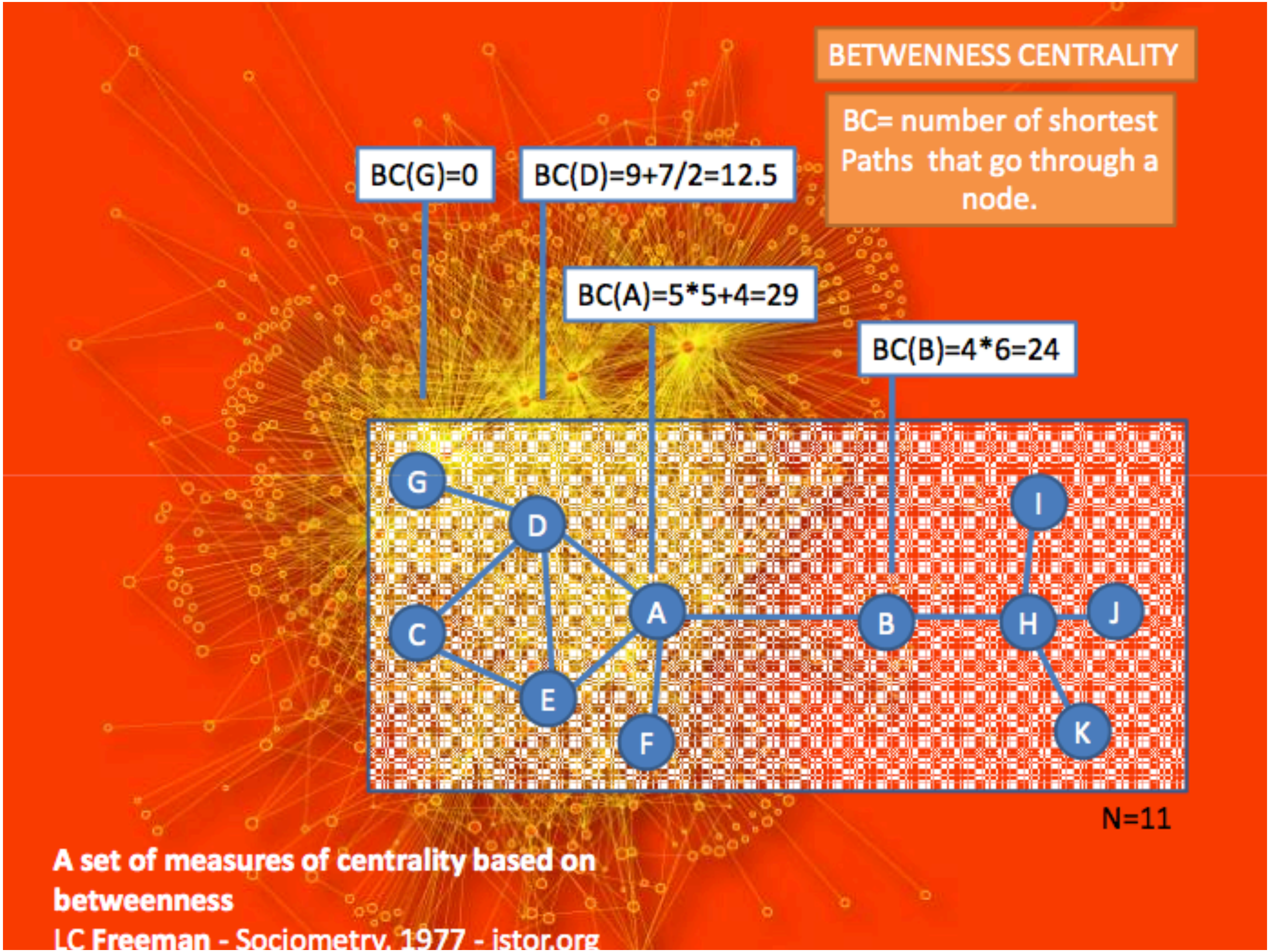


C(G)=1/10(1+2*3+2*3+4+3*5)
C(G)=3.2
CLOSENESS CENTRALITY
C(A)=1/10(4+2*3+3*3)
C(A)=1.9
C= Average Distance to neighbors
G
D
C
A
E
F
B
H
I
J
K
N=11
C(B)=1/10(2+2*6+2*3)
C(B)=2

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)

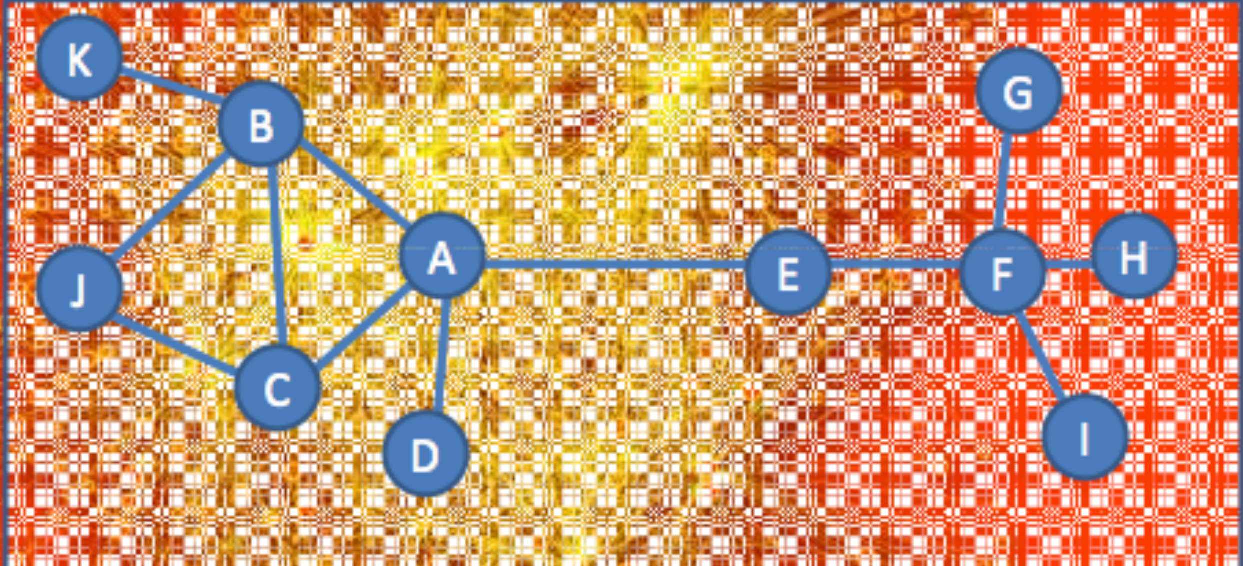



BETWENNESS CENTRALITY
BC= number of shortest Paths that go through a node.
BC(G)=0
BC(D)=9+7/2=12.5
BC(A)=5*5+4=29
BC(B)=4*6=24
G
D
C
E
A
F
B
H
I
J
K
N=11
A set of measures of centrality based on betweenness
LC Freeman - Sociometry, 1977 - istor.org

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

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

PAGE RANK

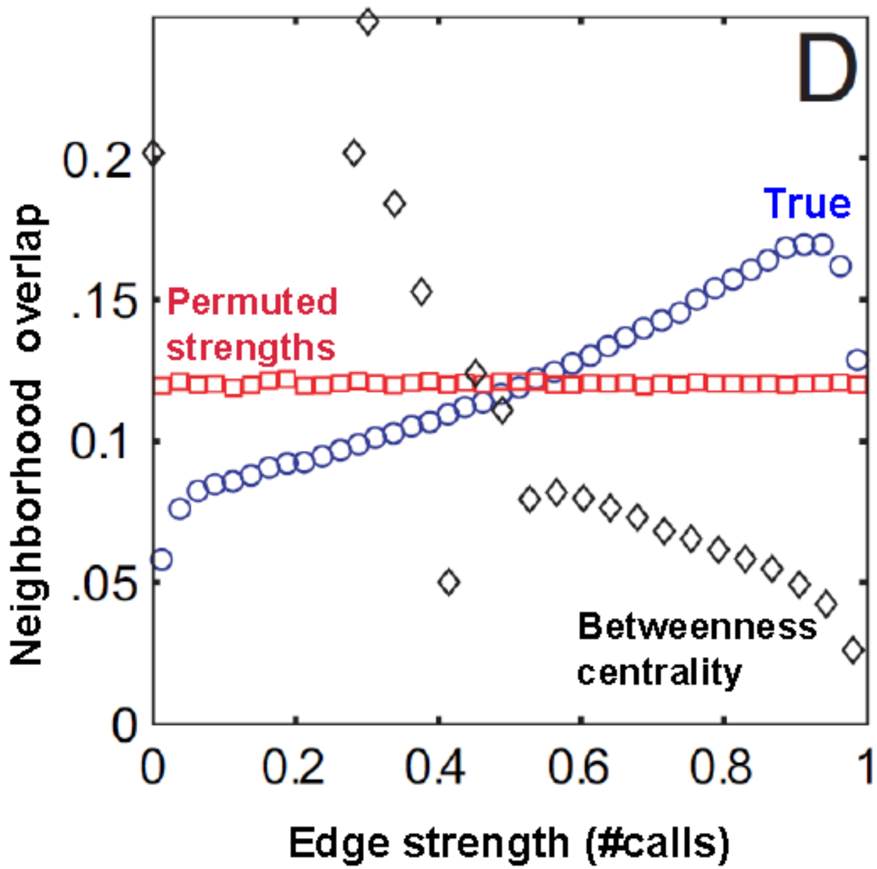
PR=Probability that a random Walker would visit that node.
PR=Each page votes for its neighbors.

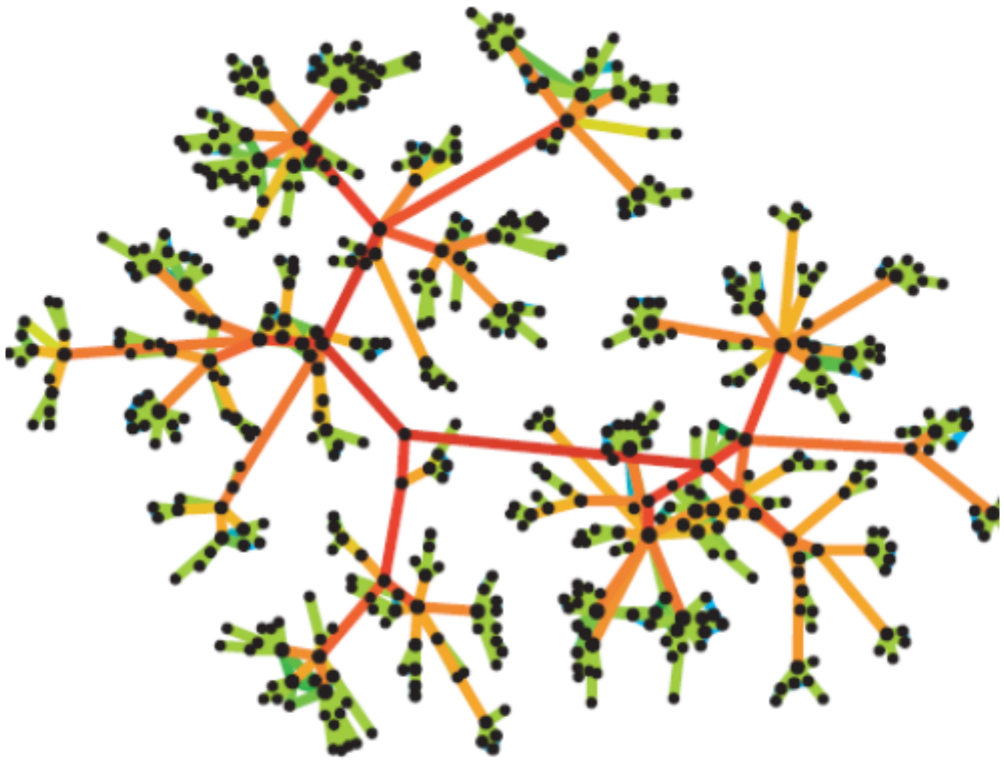
$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

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

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

Back to Granovetter





Human mobility, social ties and link prediction

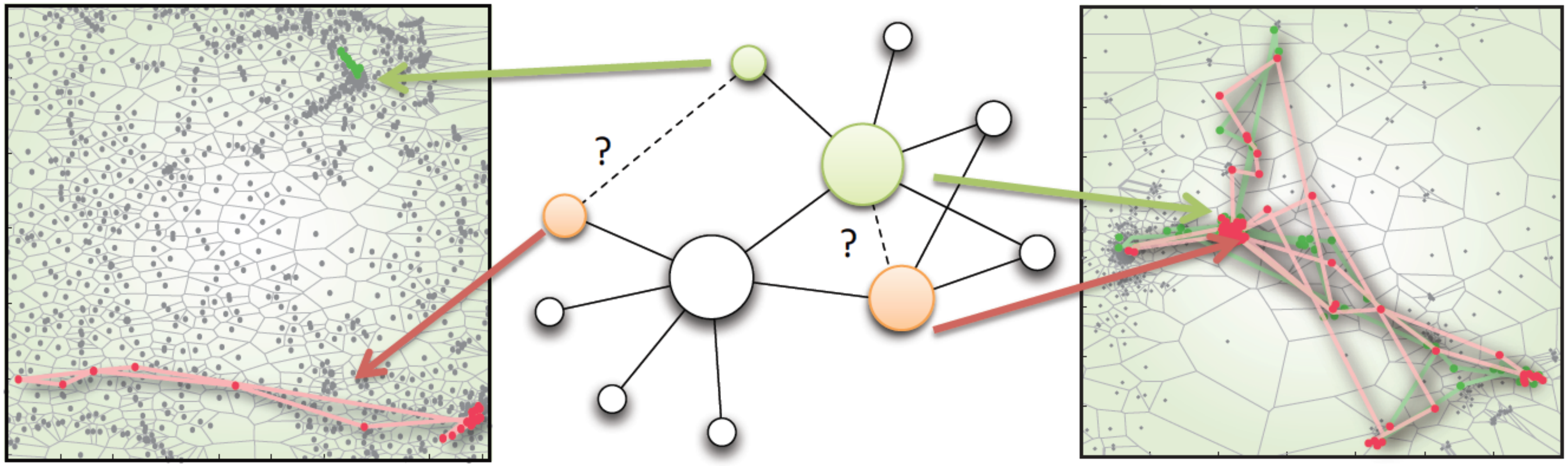
Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, Albert-Lászlo Barabási

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

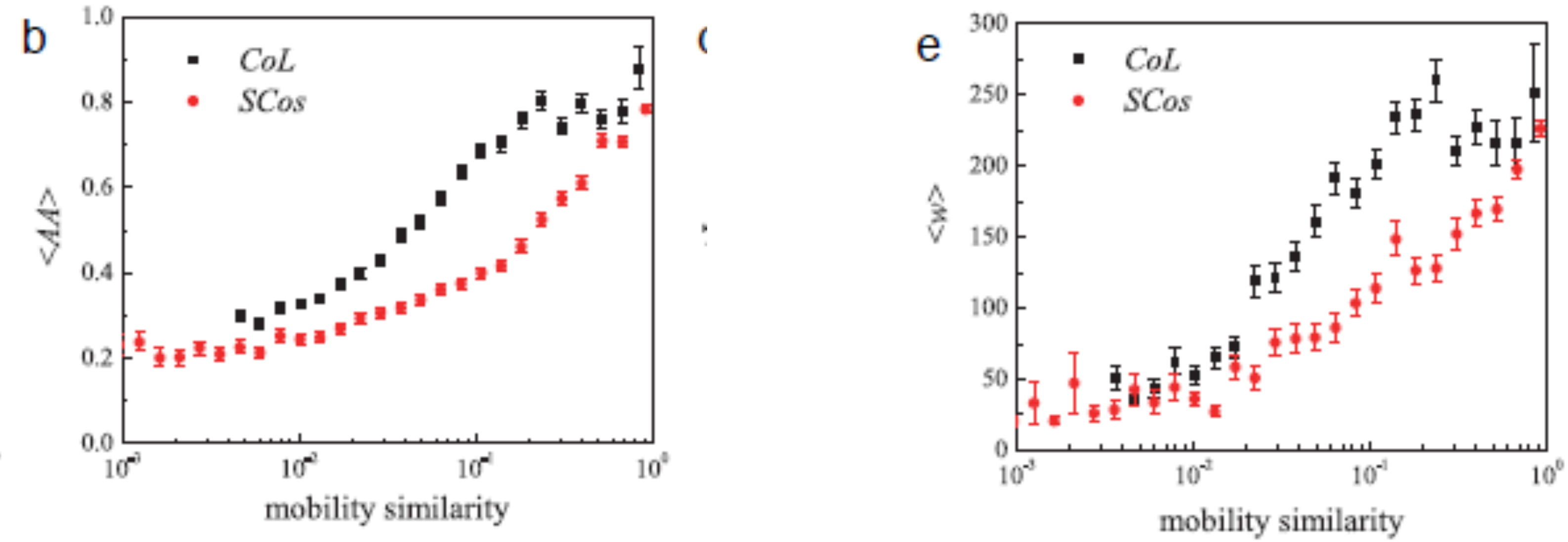
Colocation, social proximity, tie strength

- 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**

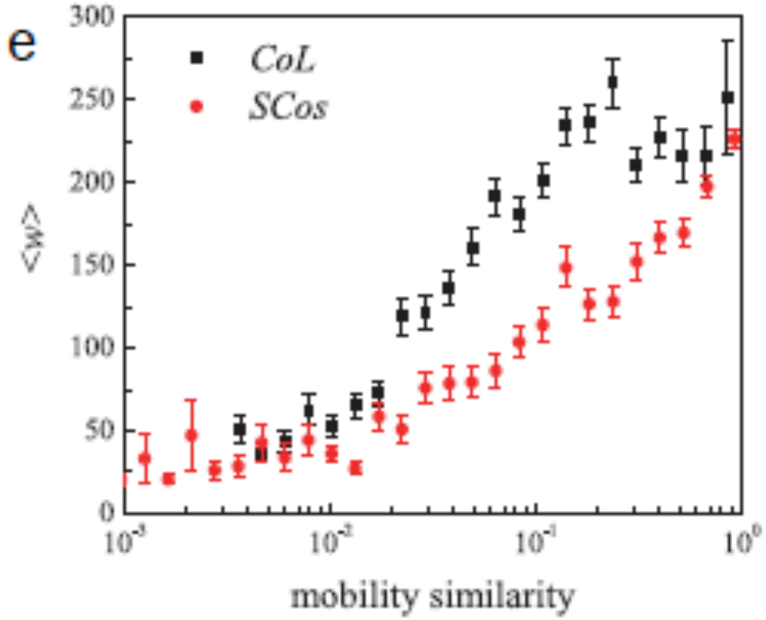
Network proximity vs. mobile homophily

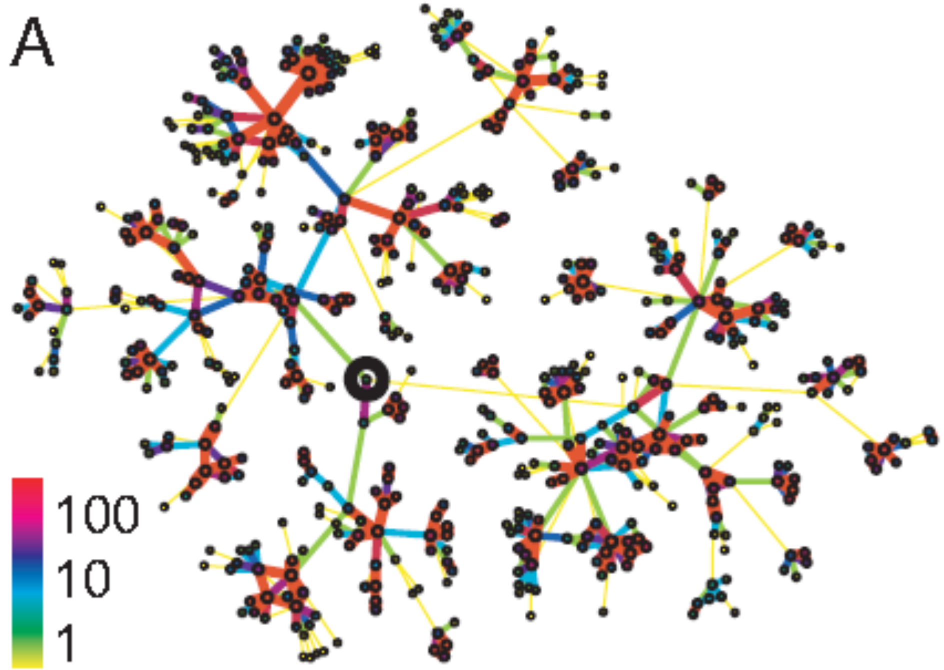


mobility dimension of the “strength of weak ties”



- 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)

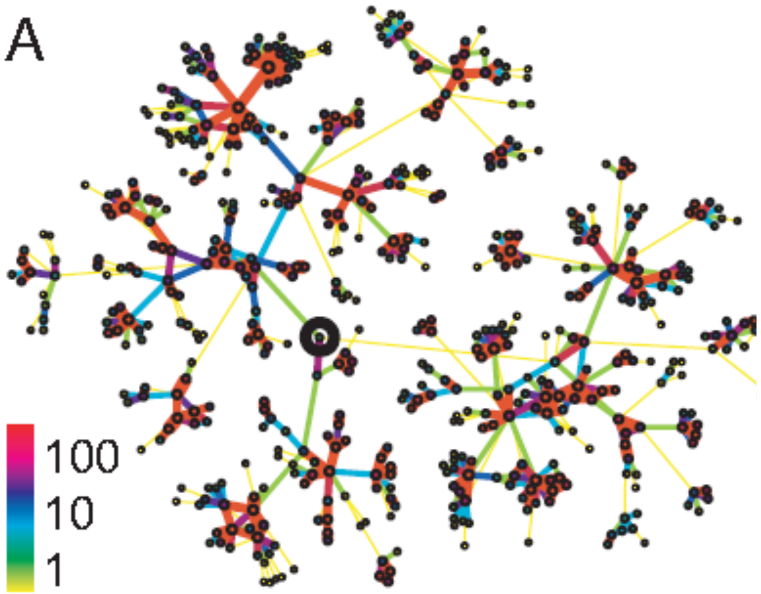

e
CoL
SCos
<w>
mobility similarity
300
250
200
150
100
50
0
10^{-3}
10^{-2}
10^{-1}
10^{0}

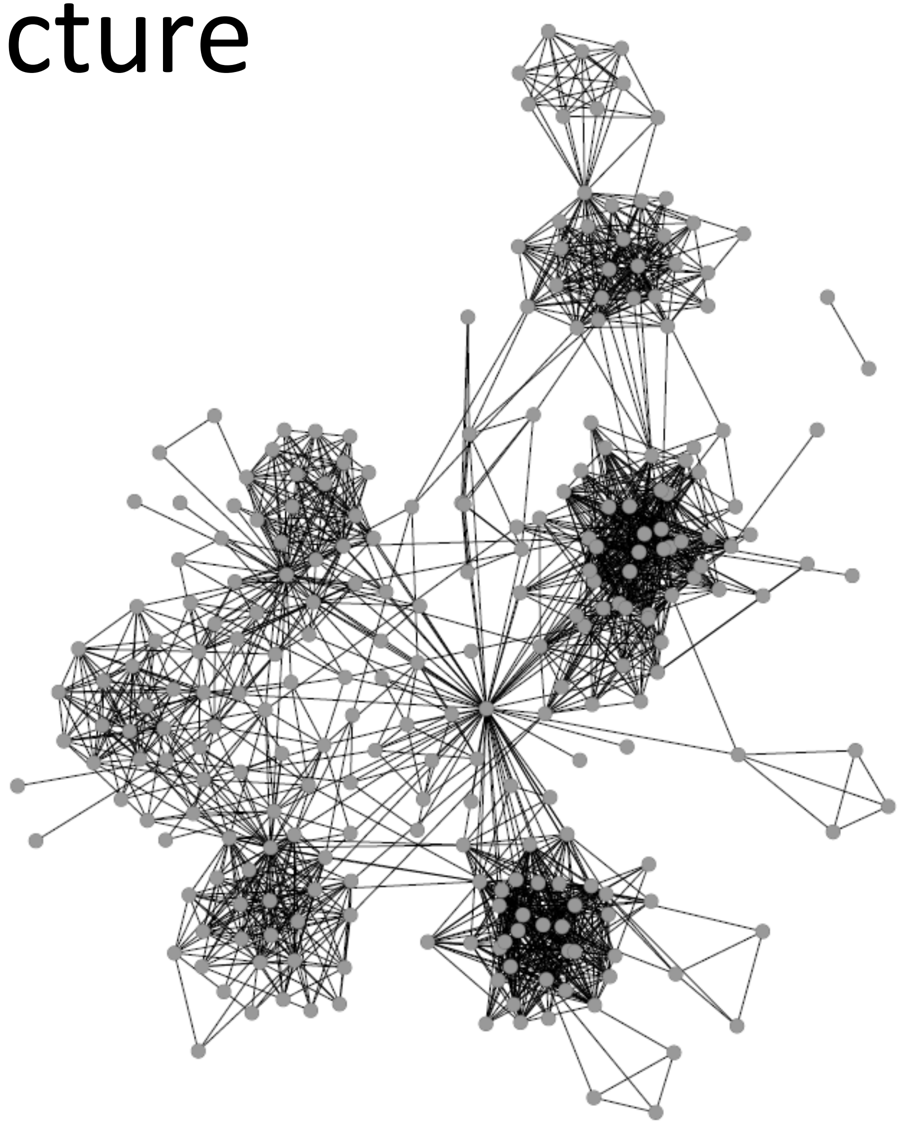

A
100
10
1

Community discovery

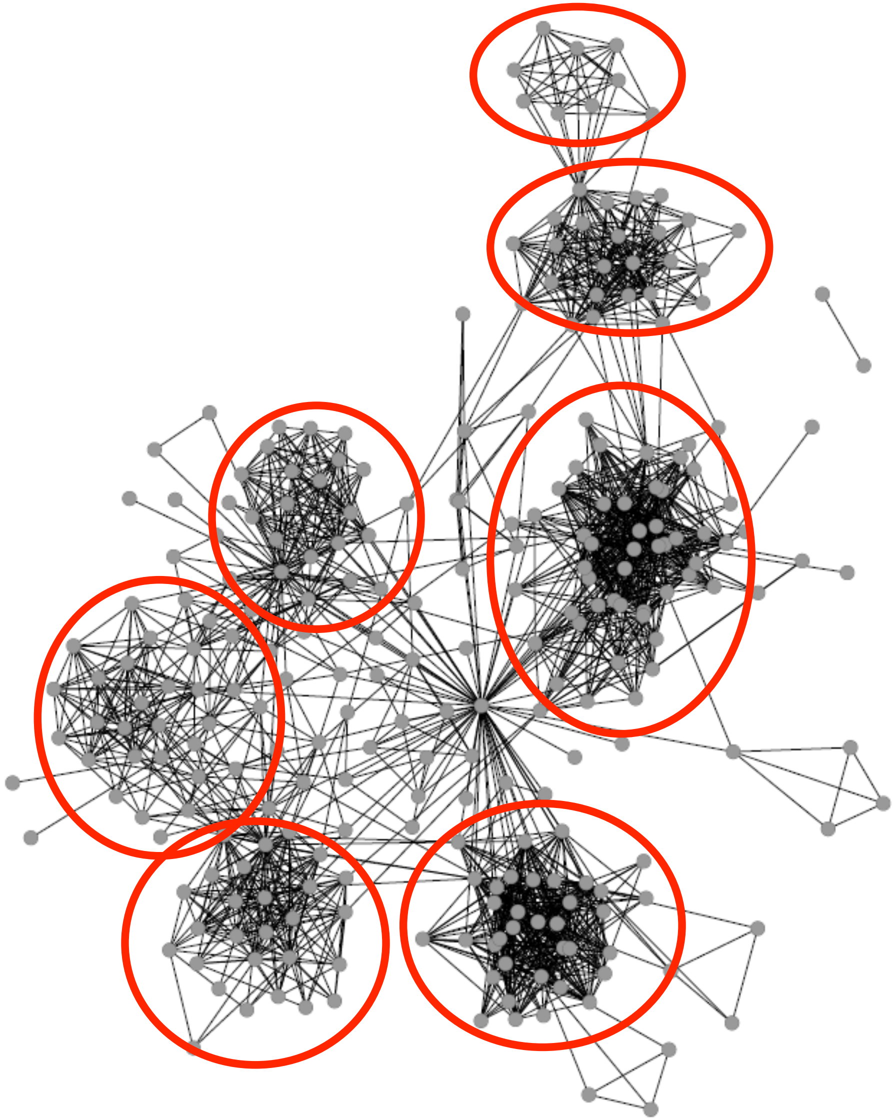
How to highlight the modular structure of a network?

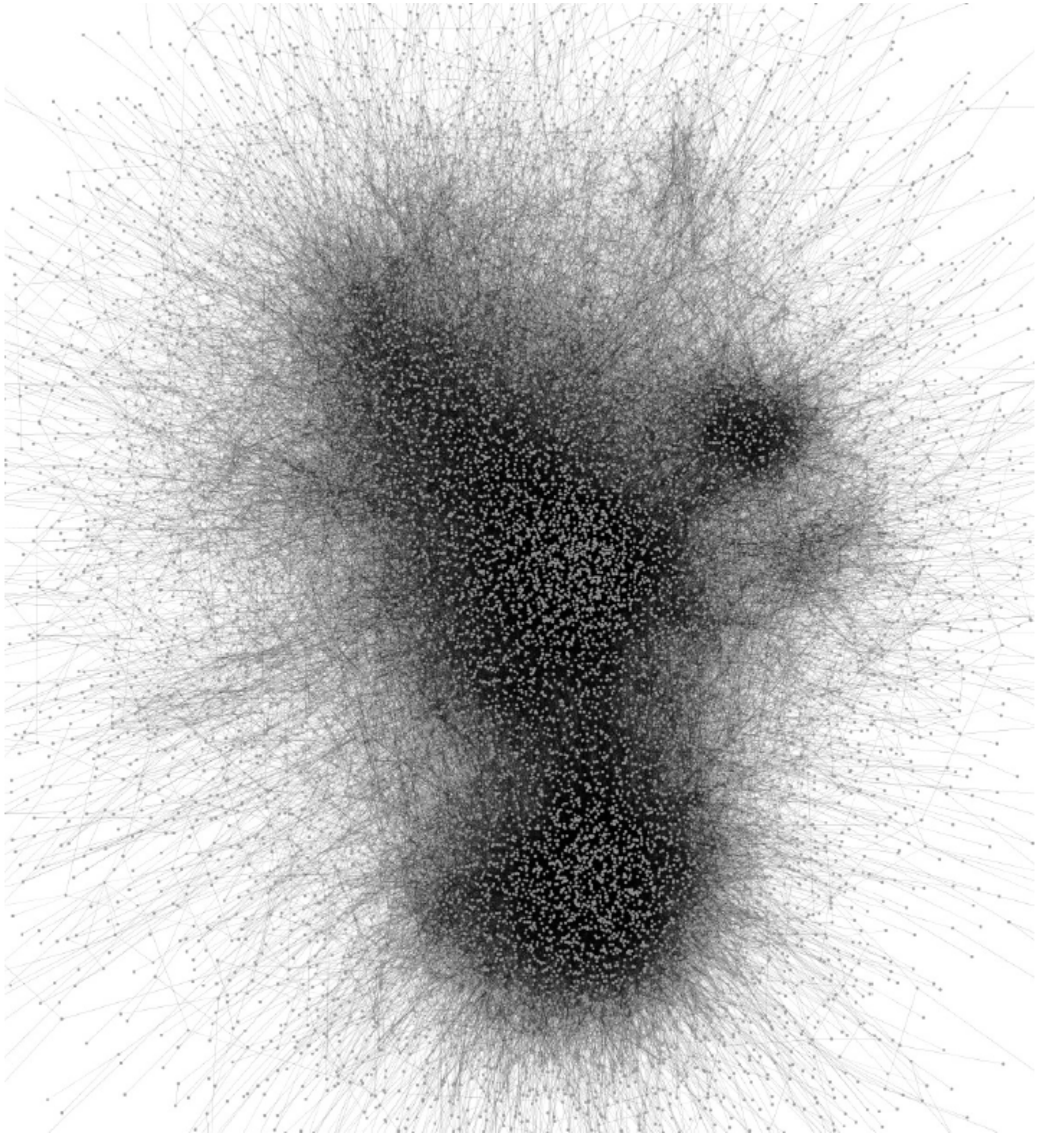
Community structure

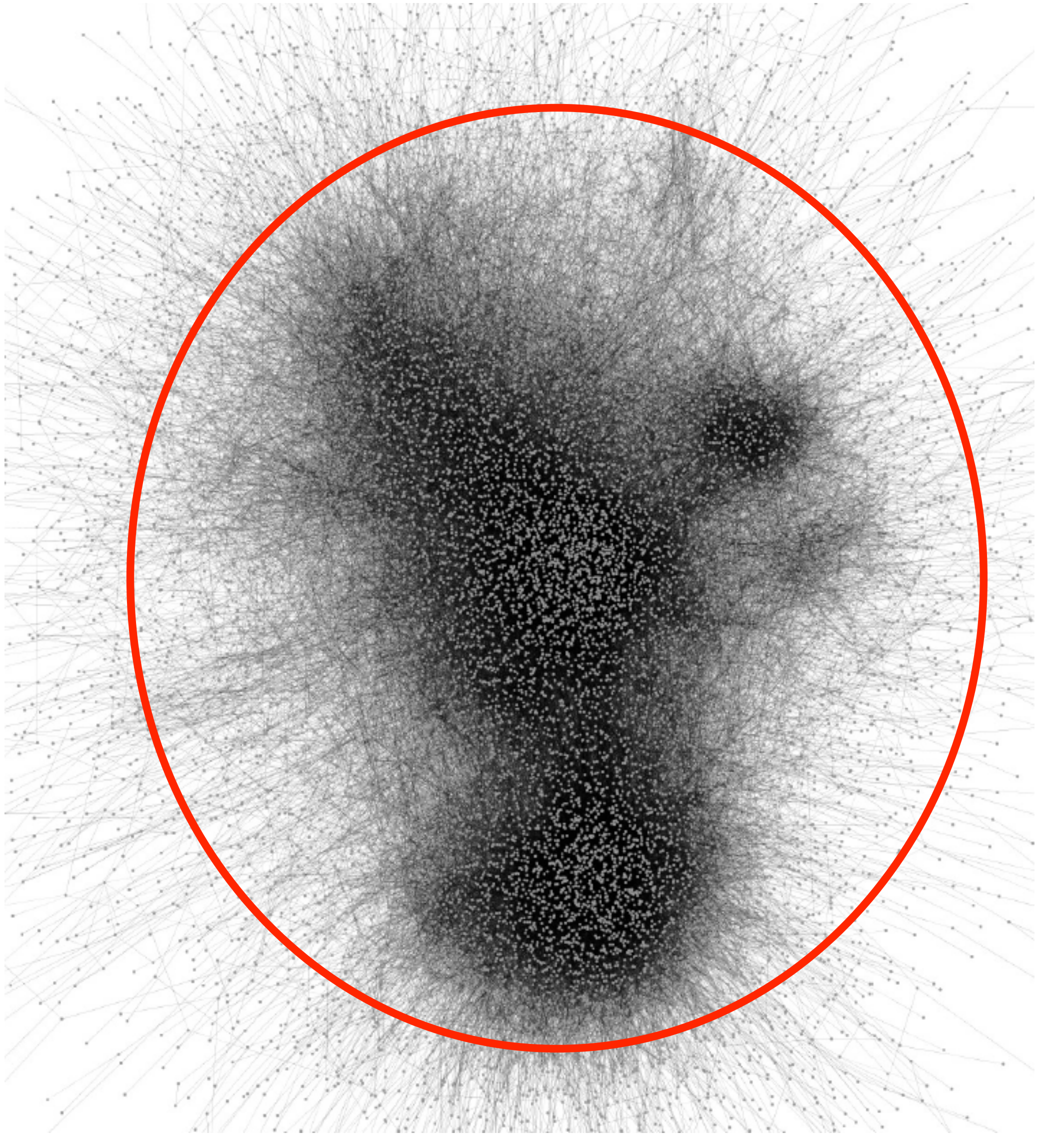




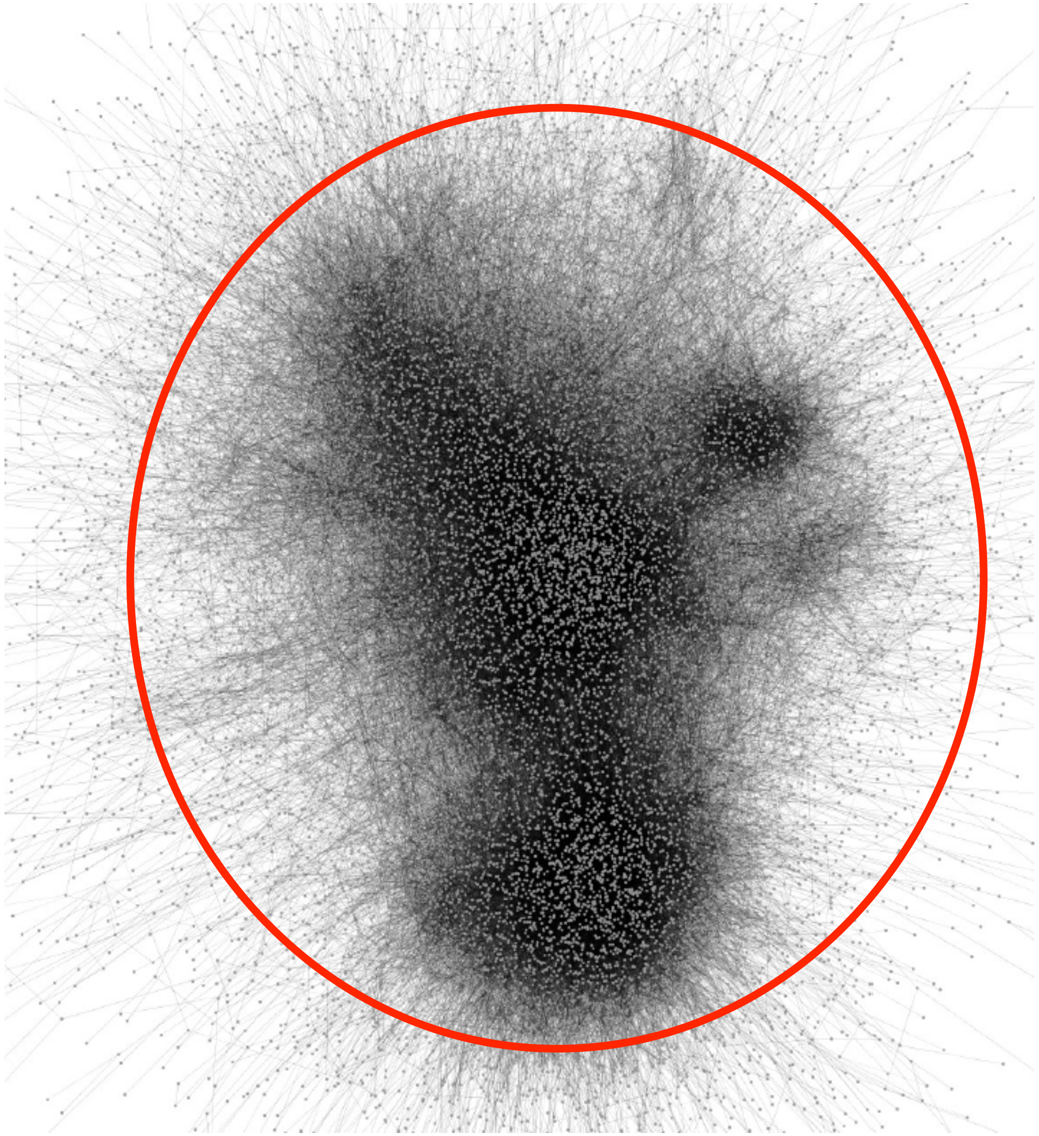
Communities



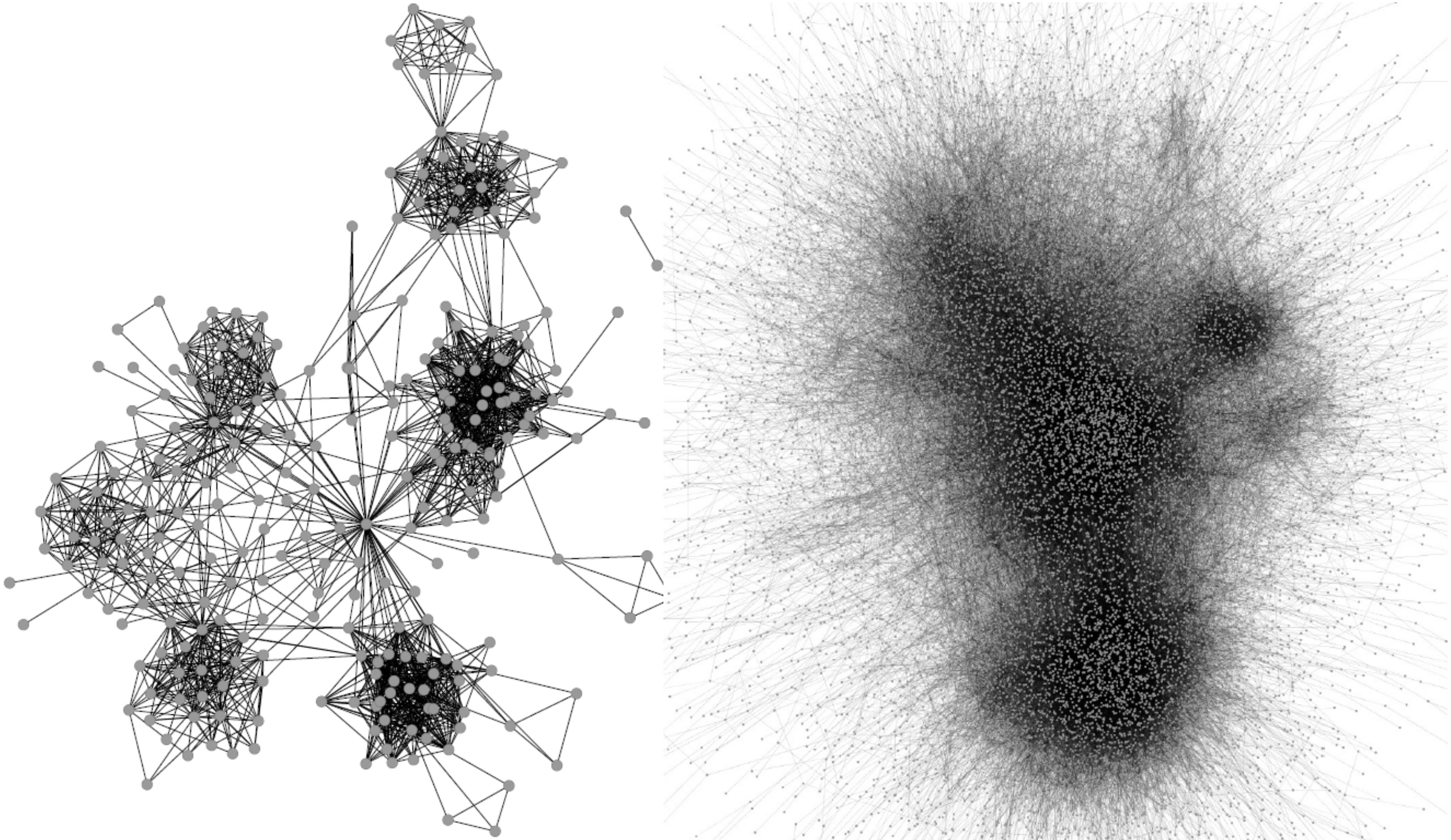




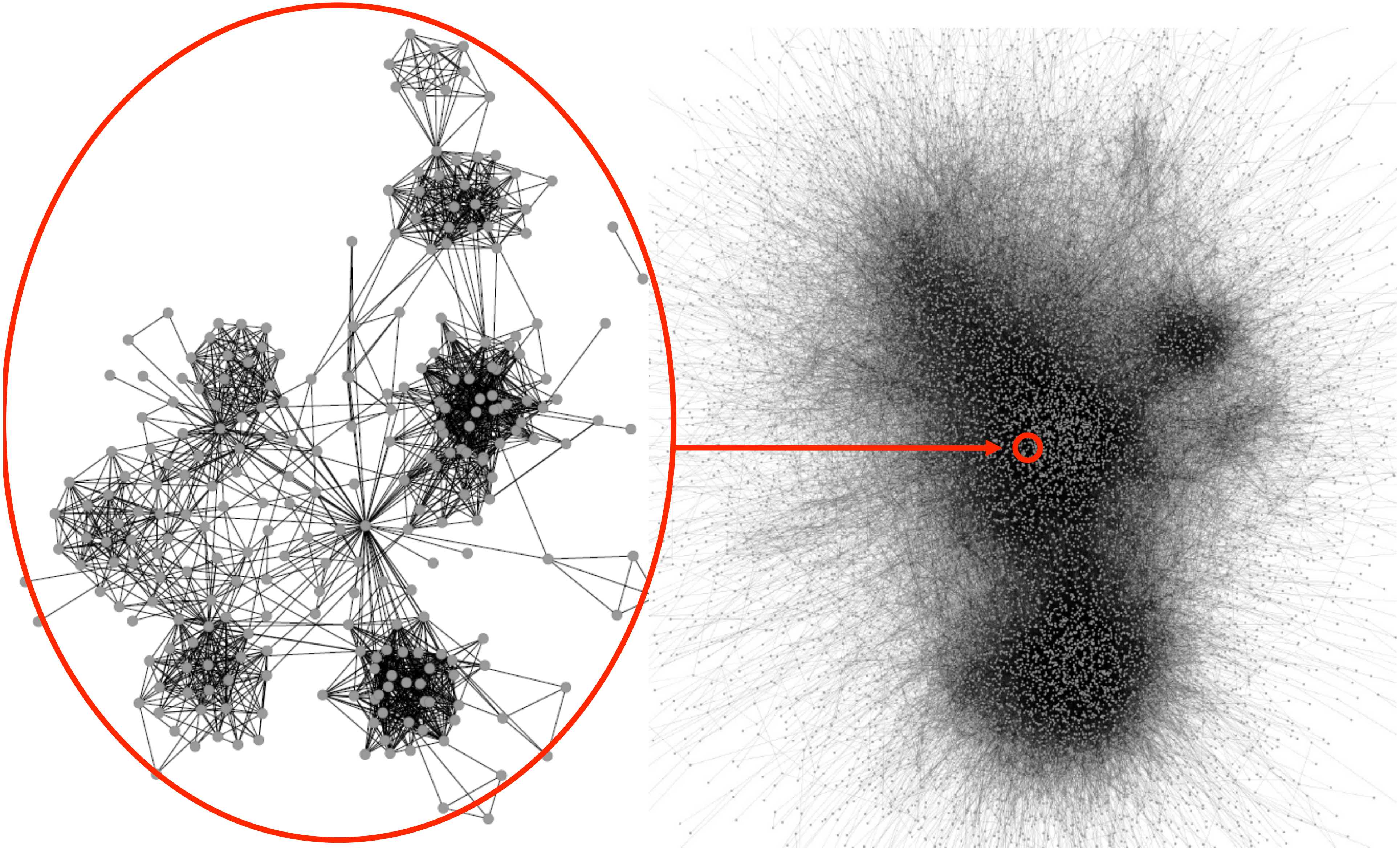
?



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

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 Trans. on Knowledge Discovery from Data TKDD* 9(1): 6 (2014)

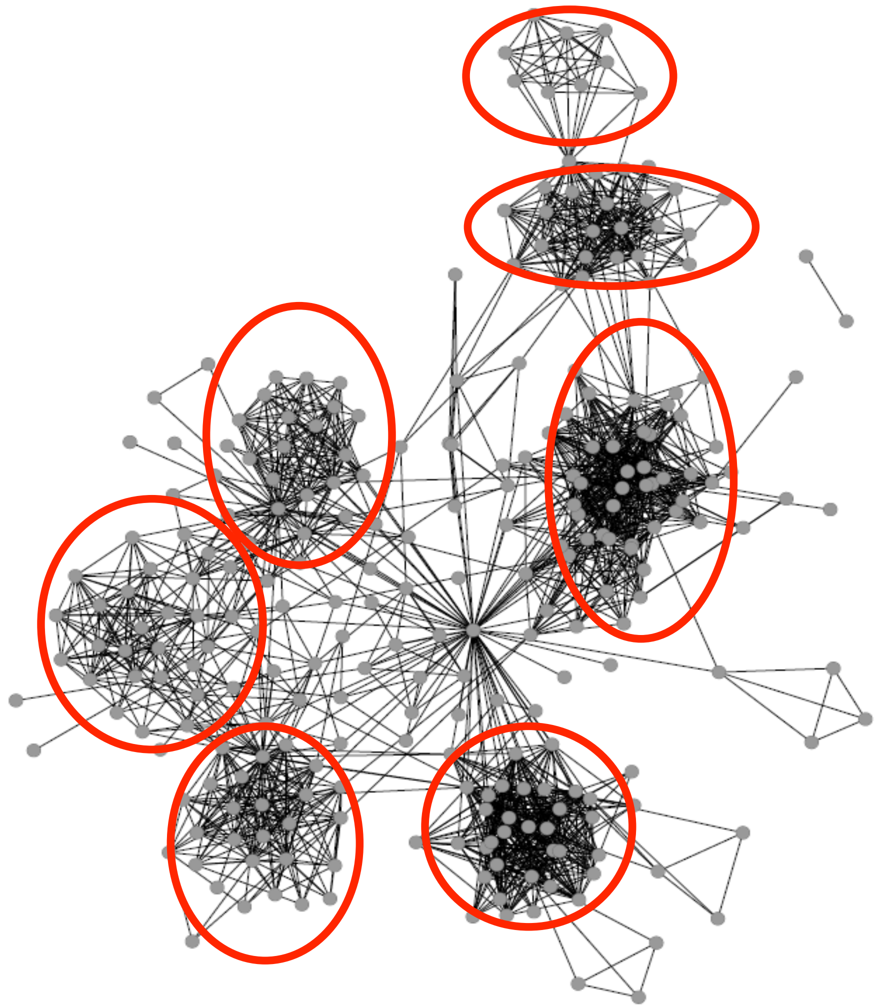
DEMON

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

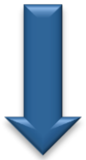
- In simple, small, networks it is easy identify them by looking at the structure.



Reducing the complexity

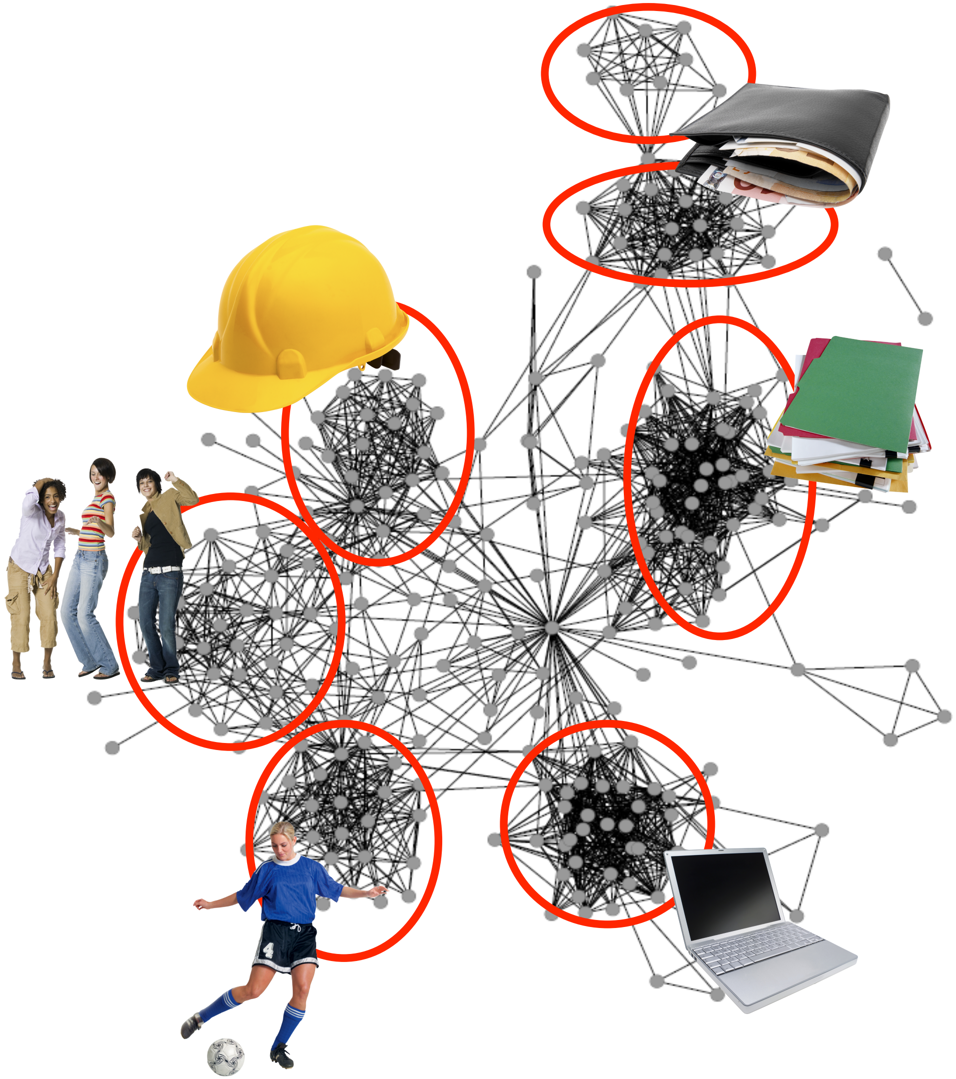
Real Networks are Complex Objects

Can we make them "simpler"?



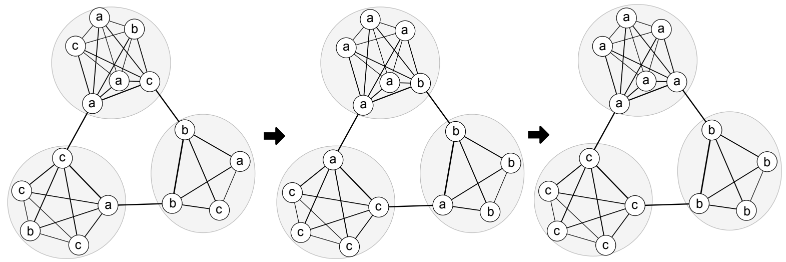
Ego-Networks

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

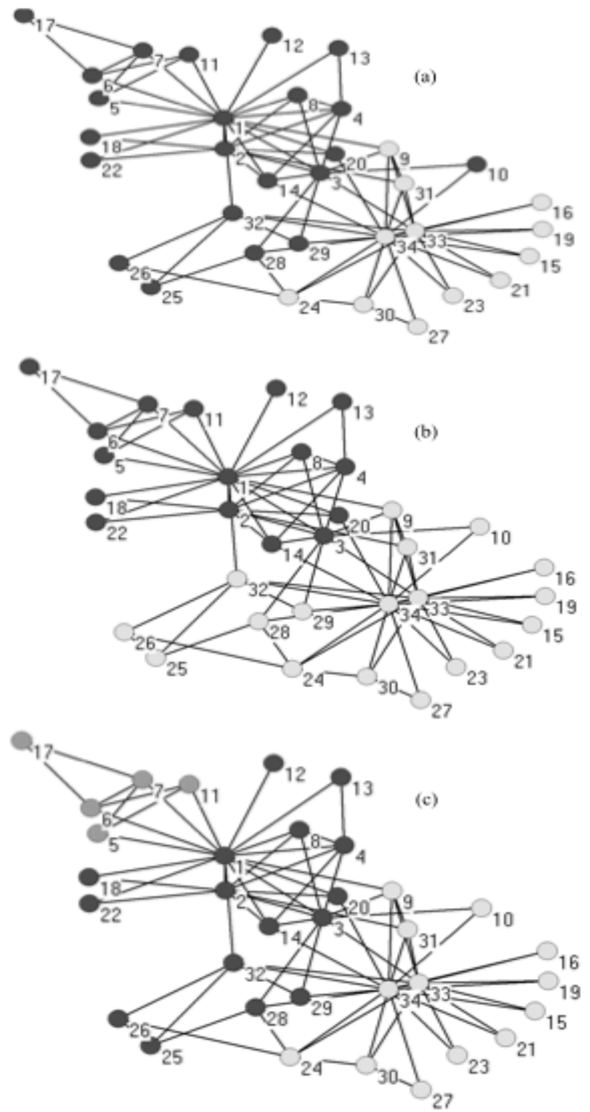


- For each raw community c in C
  1. Merge with “similar” ones in the set (given a threshold)
     (i.e. merge iff at most the ε% of the smaller one is not included in the bigger one)

[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 @ 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

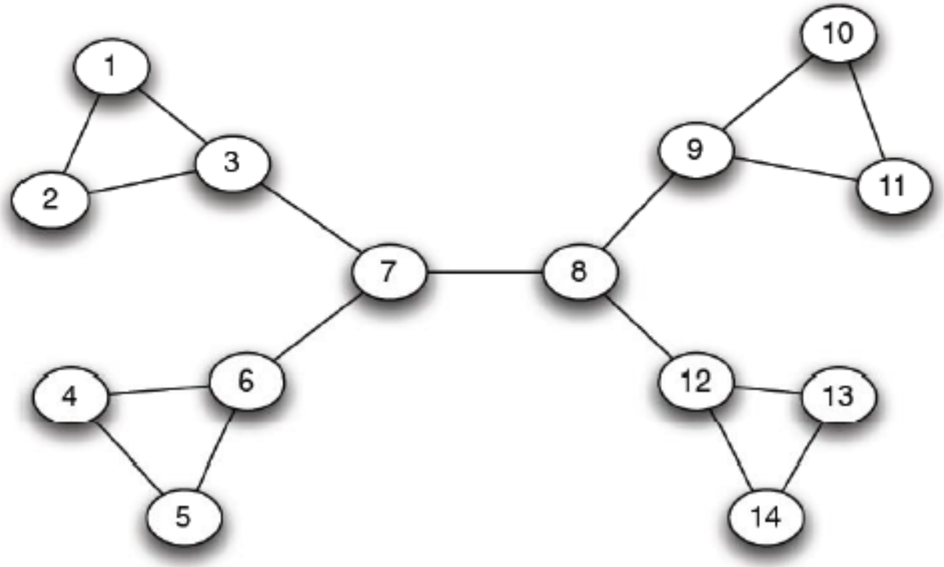
Bottom-up (local) vs top-down (global) community detection

[Girvan-Newman PNAS '02]

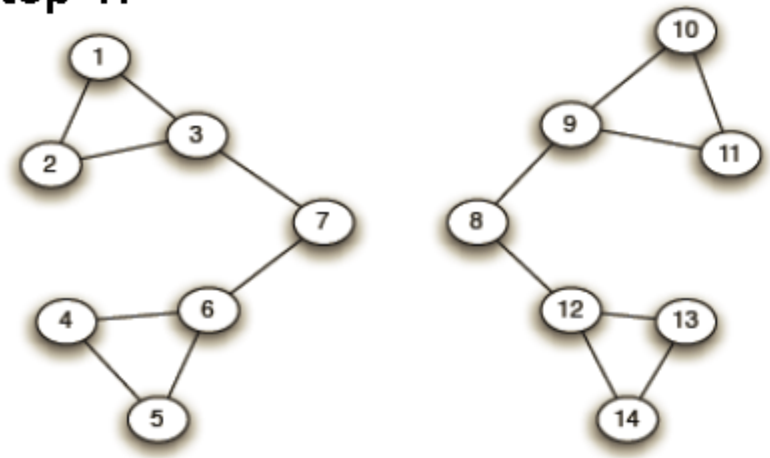
Method 1: Girvan-Newman

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

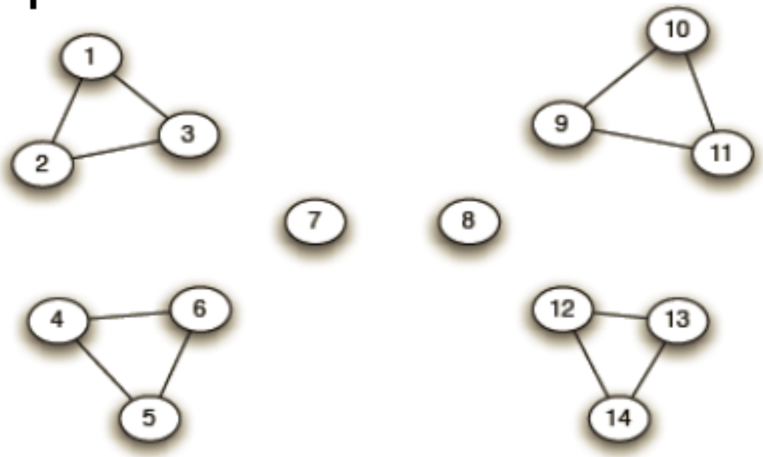
  Number of shortest paths passing through the edge
- Remove edges in decreasing betweenness
- Example:



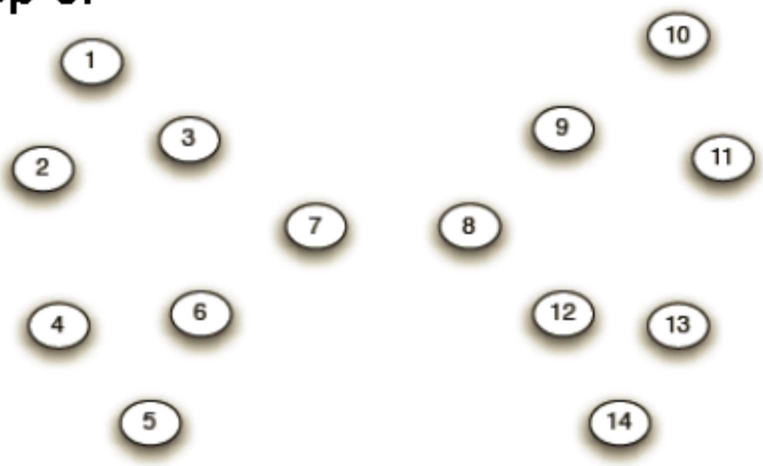
Step 1:



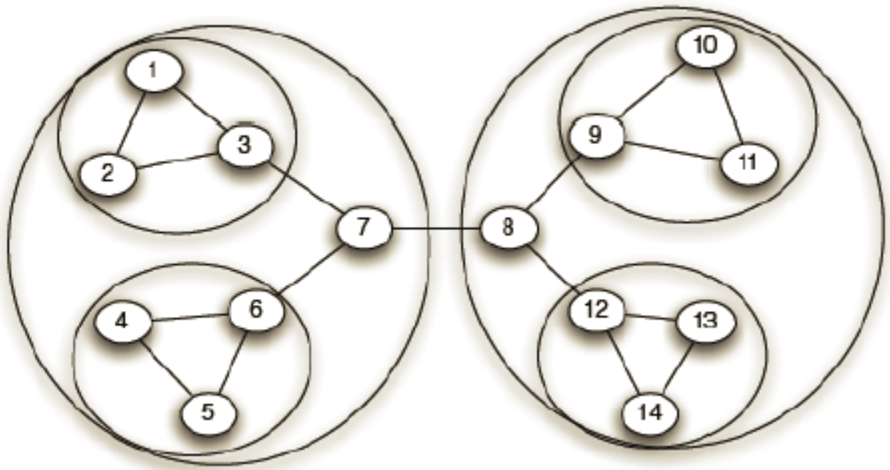
Step 2:

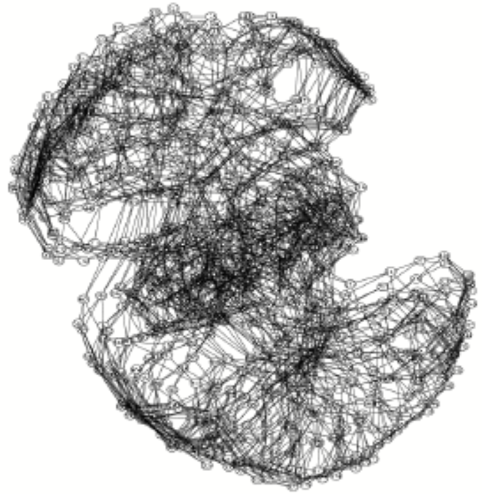


Step 3:

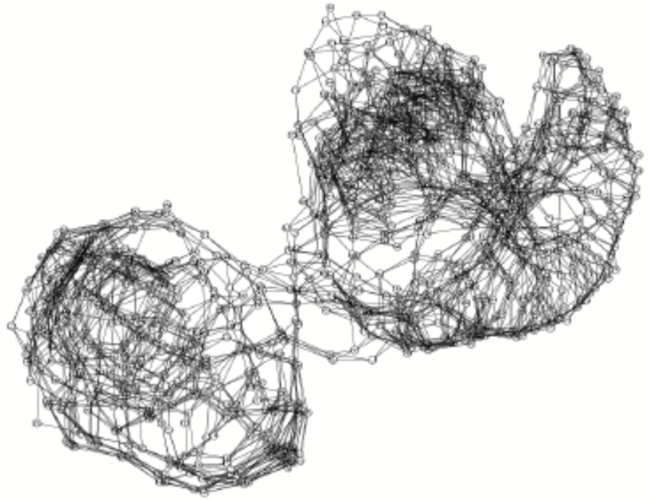


Hierarchical network decomposition:

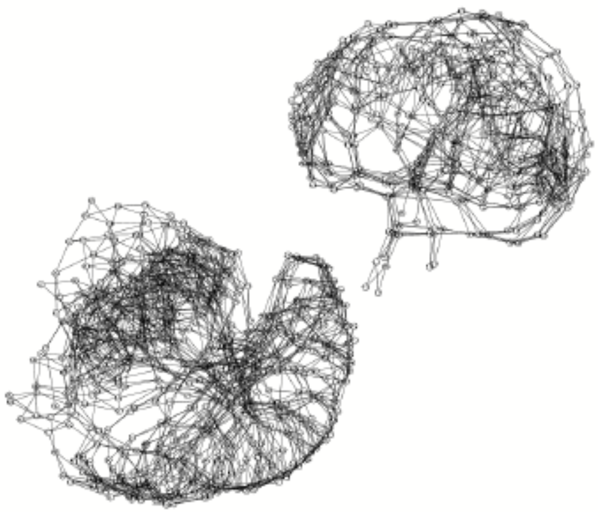




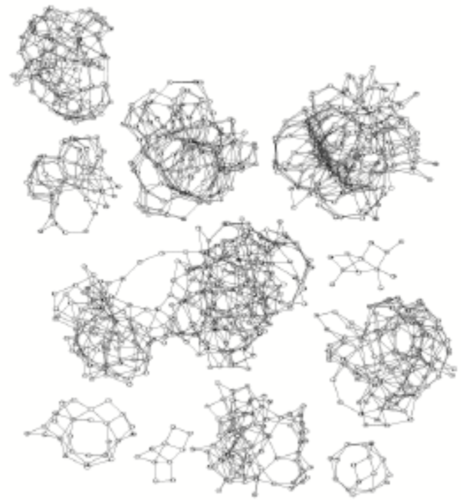
0 cuts



100 cuts



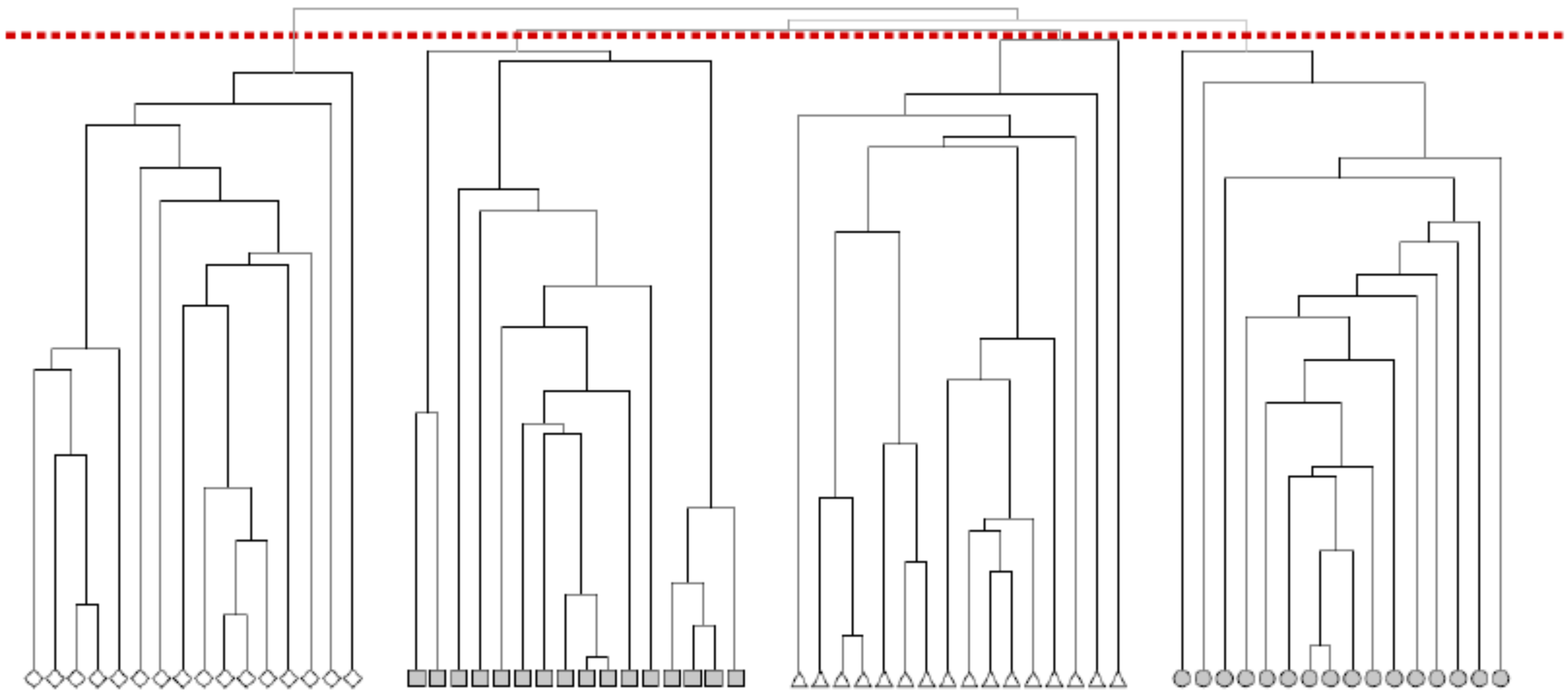
120 cuts



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

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

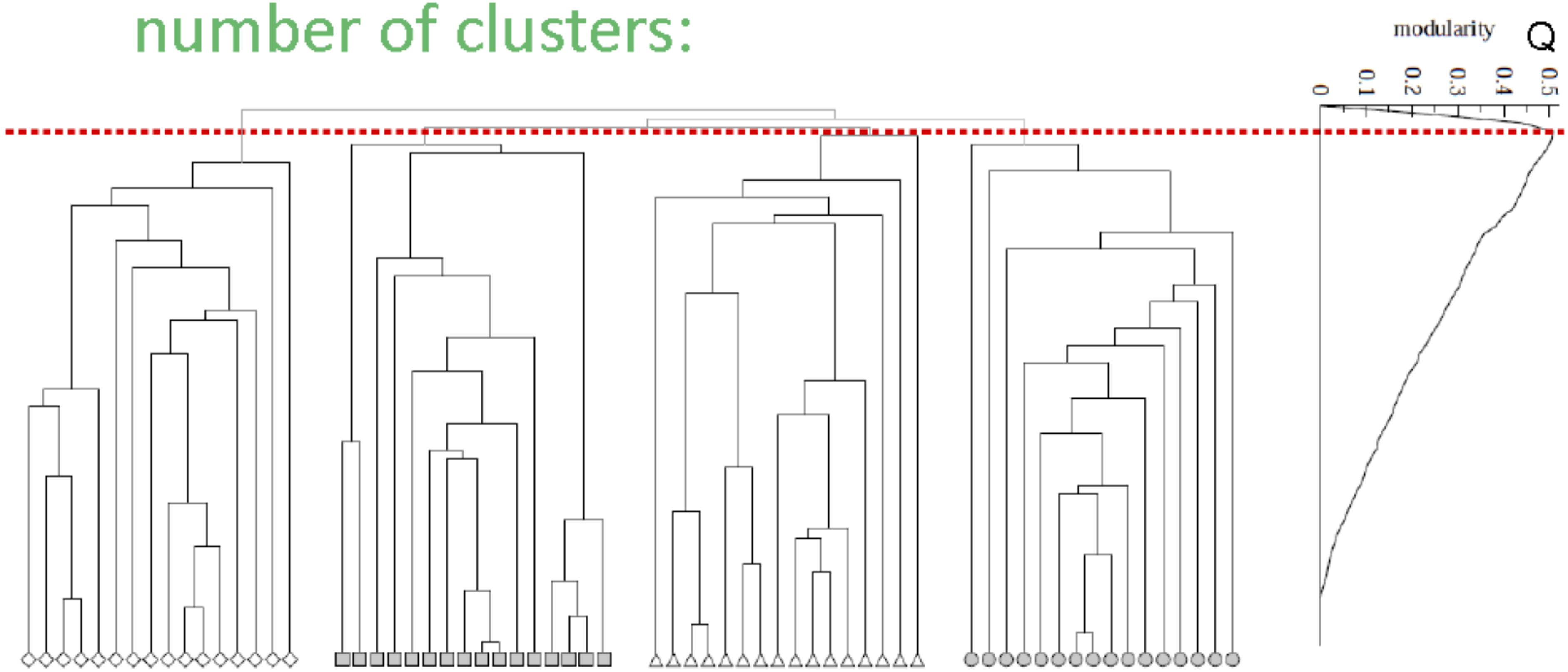
Modularity

- 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:



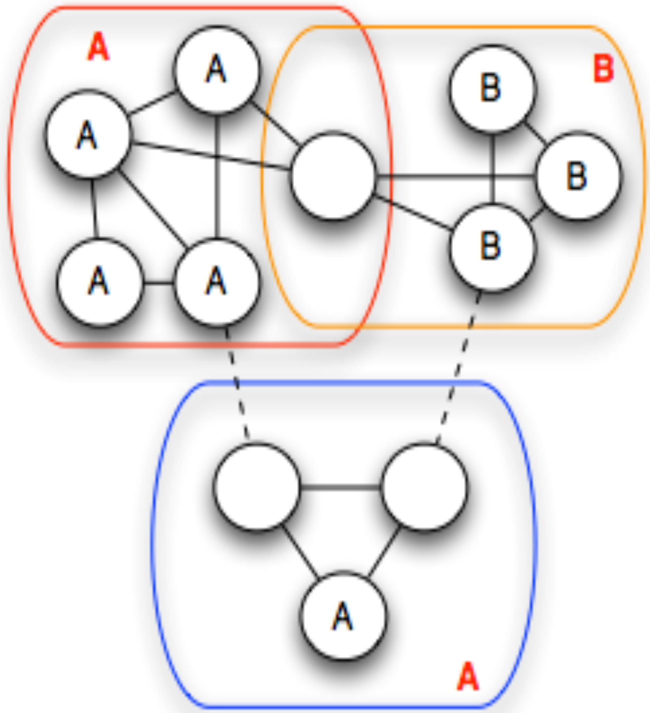
Community discovery

- 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)

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

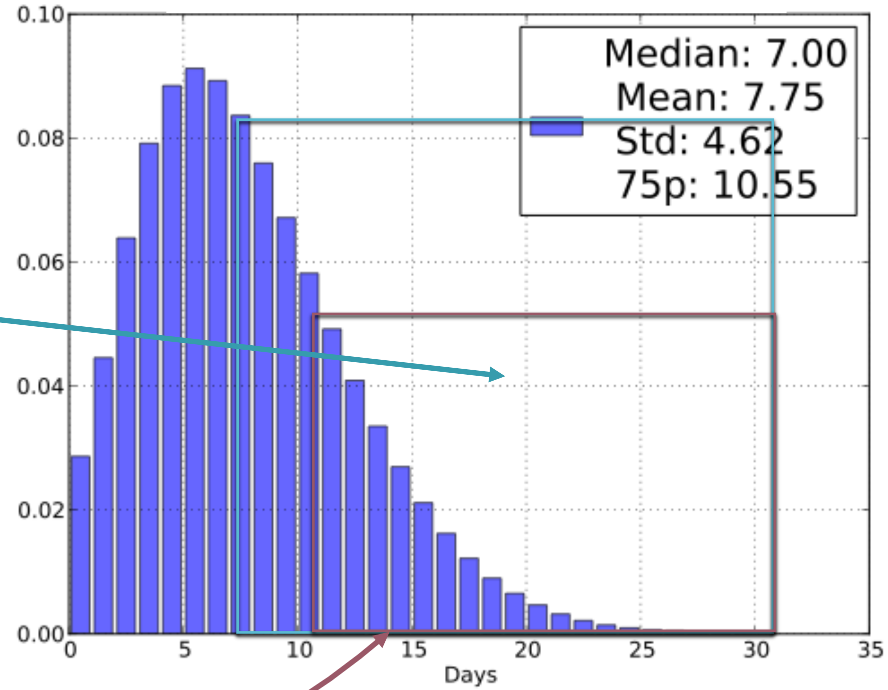
Target Class (for each service)

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

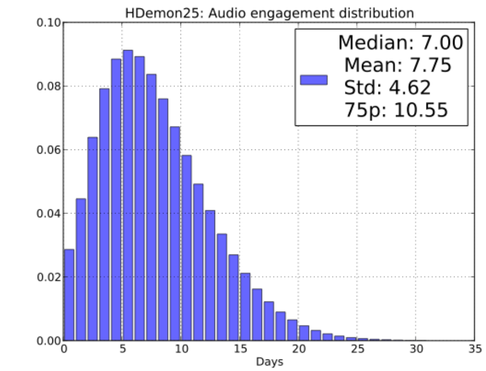


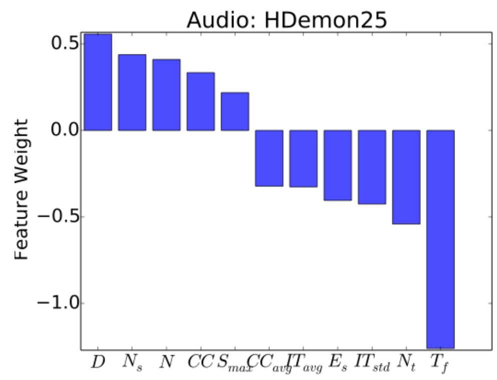
"Social Engagement" : Skype social graph

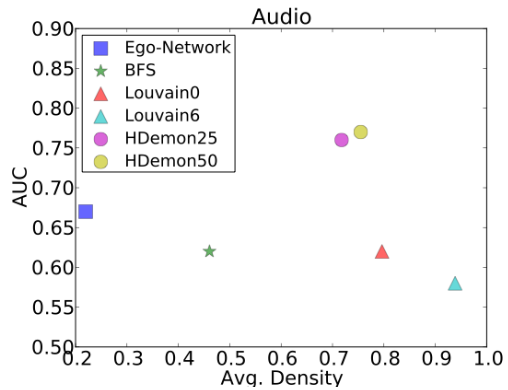
- **Problem:**
  Given the Skype social graph and its user information (i.e., location…) predict average level of community activity for the Audio \Video services.

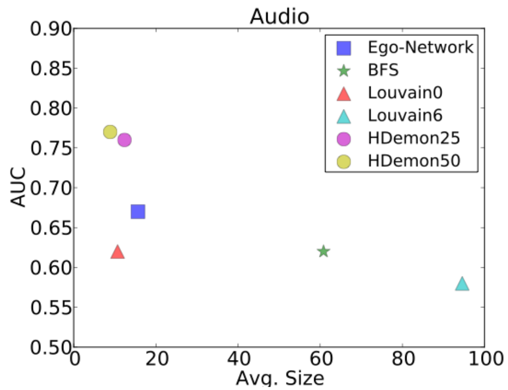
- **Question:**
  The CD method chosen will affect the classification results?

- **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



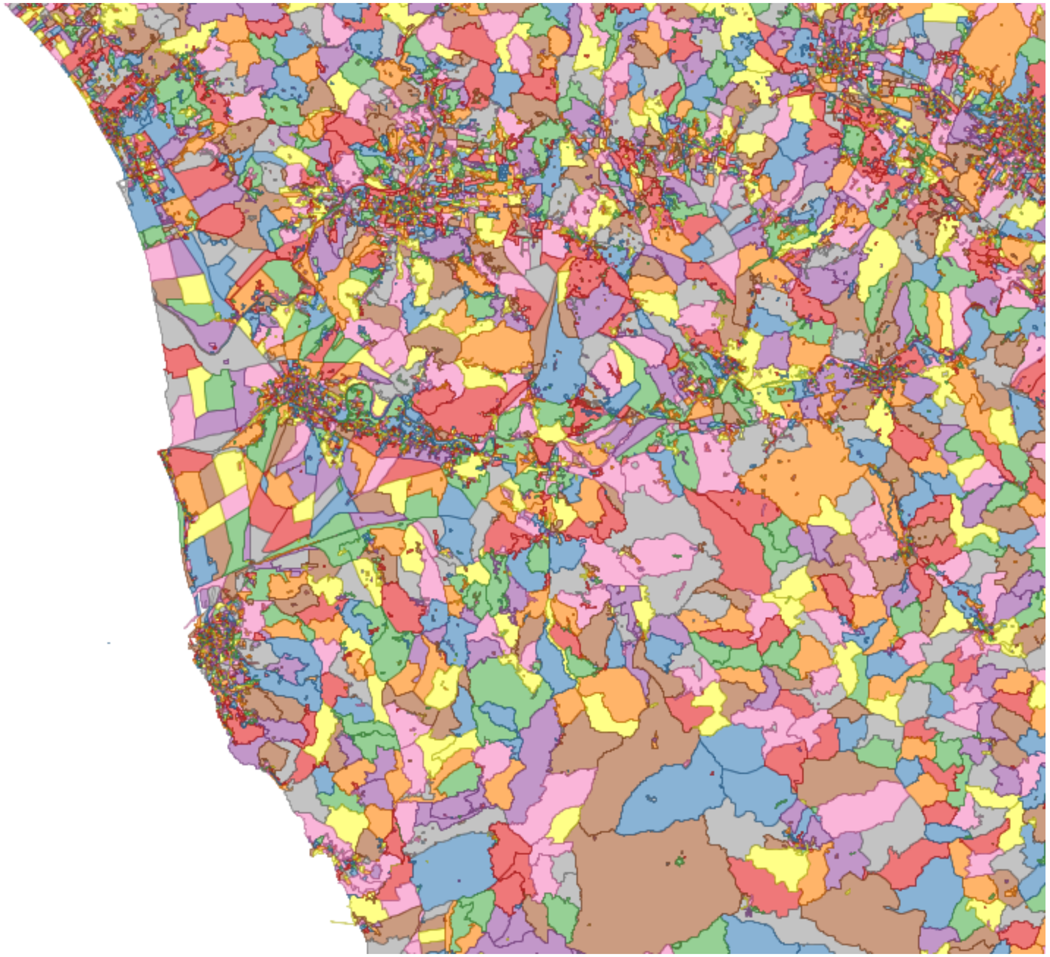




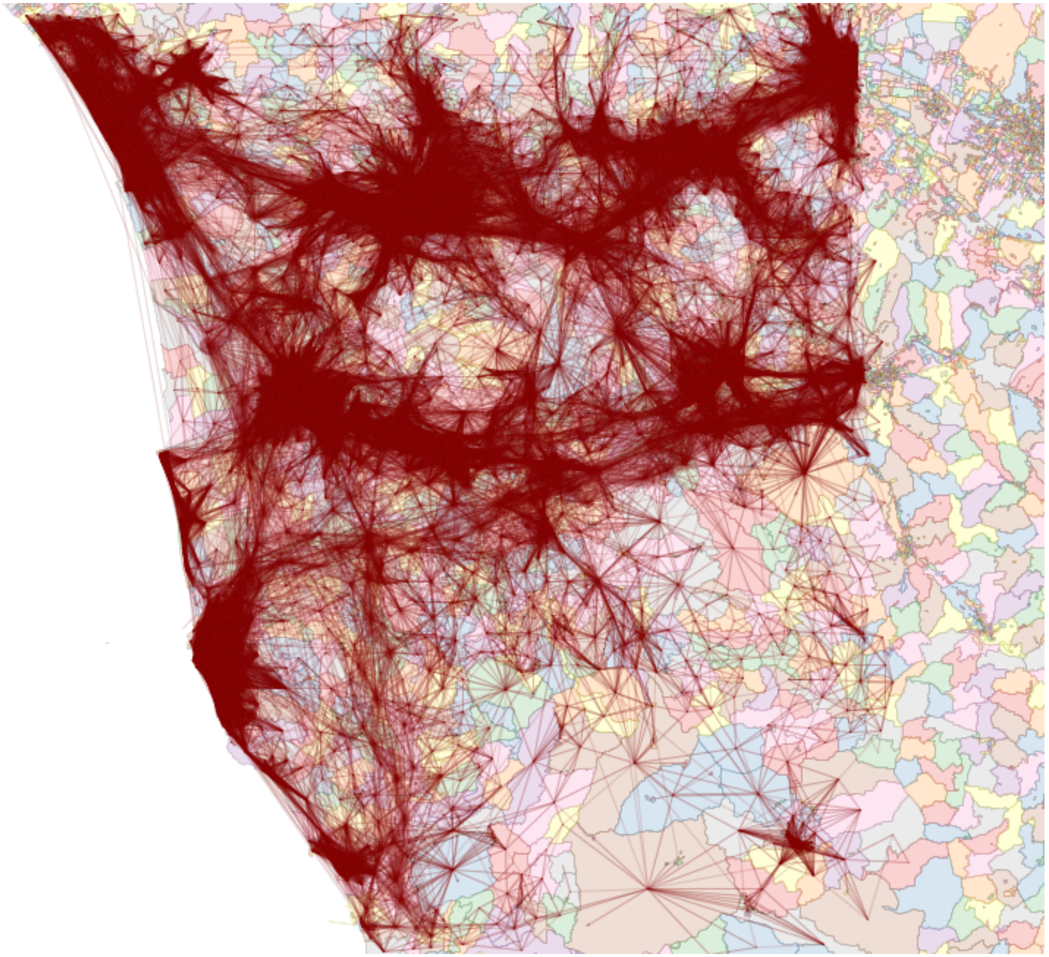


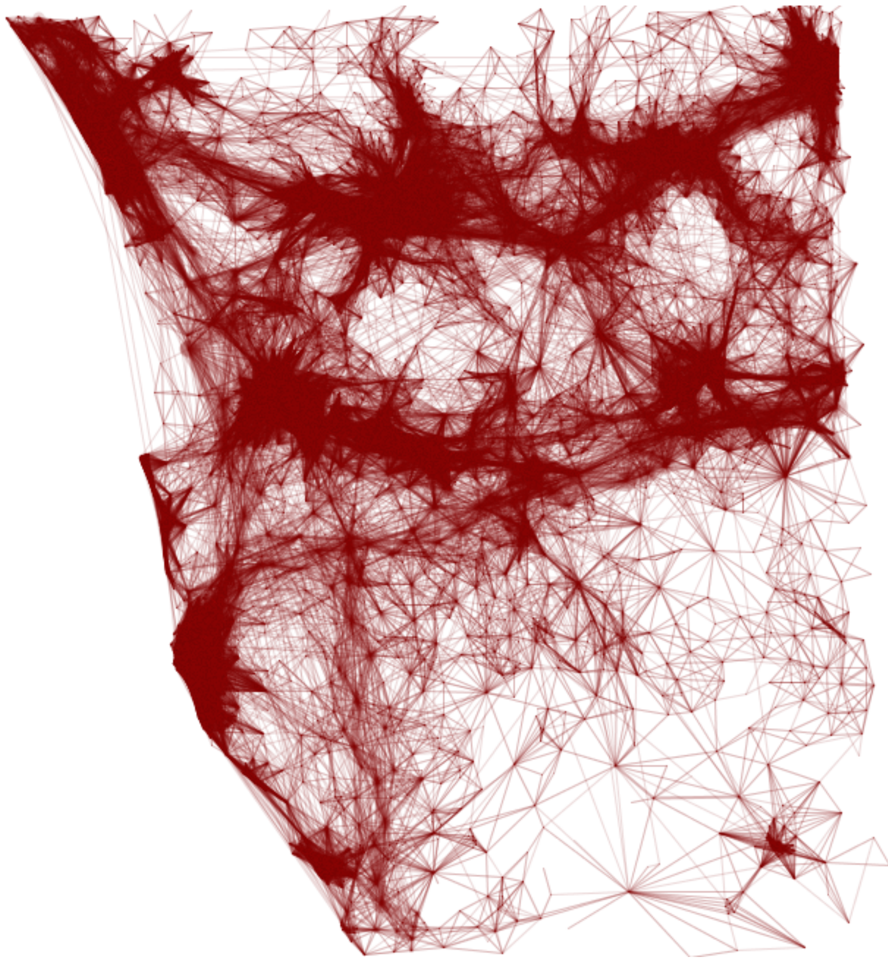
G. Rossetti, L. Pappalardo, R. Kikas, F. Giannotti, D. Pedreschi, M. Dumas *Community-centric analysis of service en- gagement in Skype social networks*. IEEE ASONAM 2015, France (Accepted)

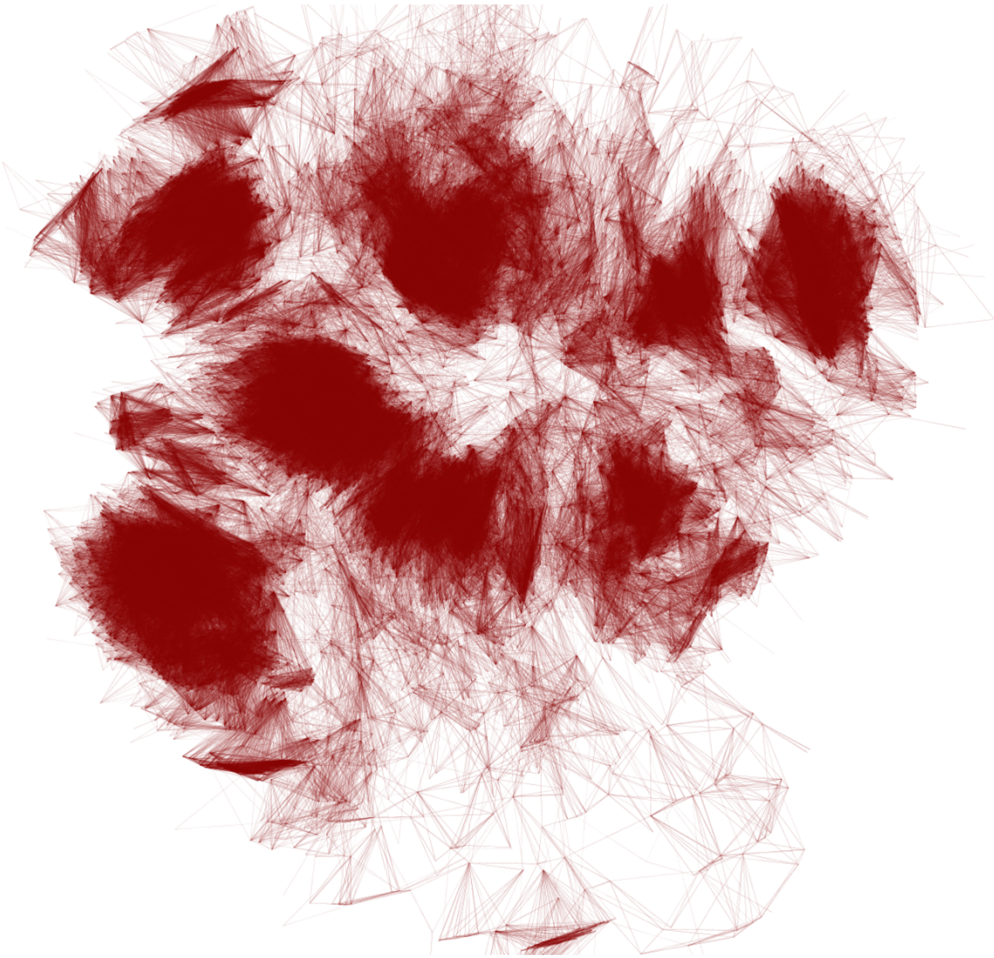
Discover the borders of mobility

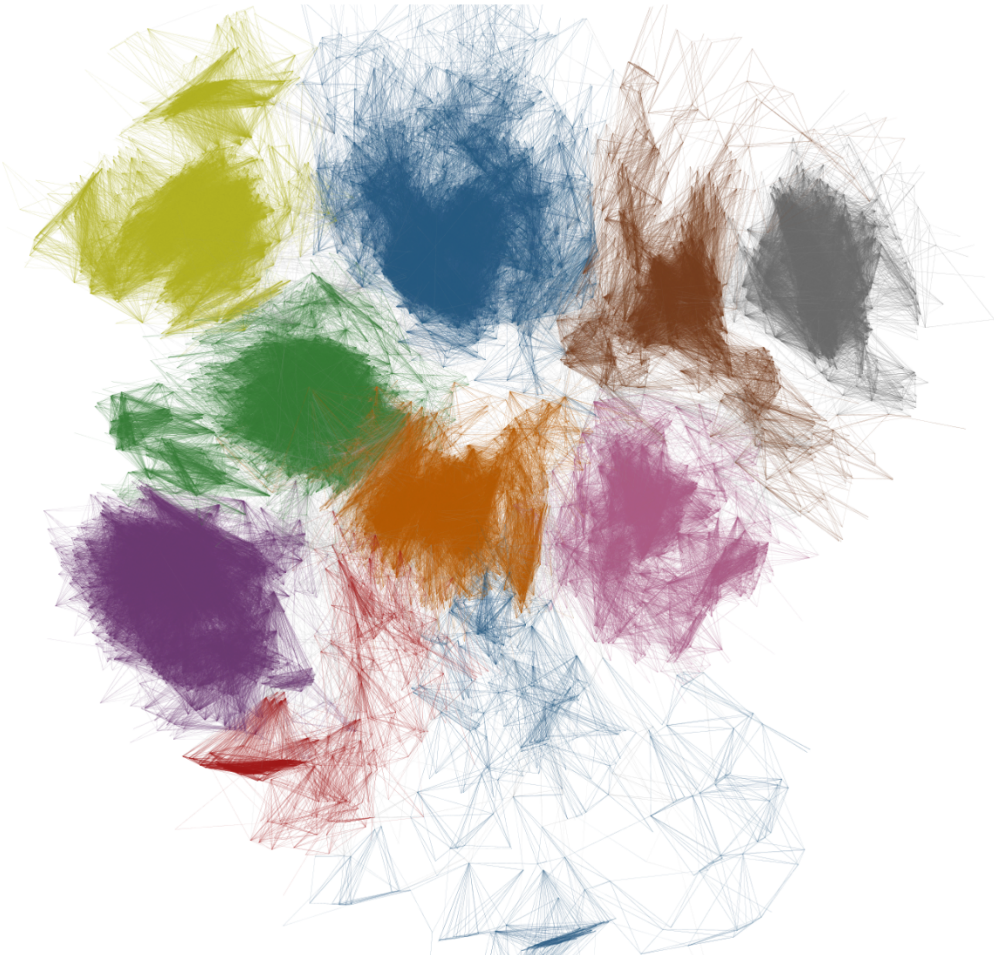


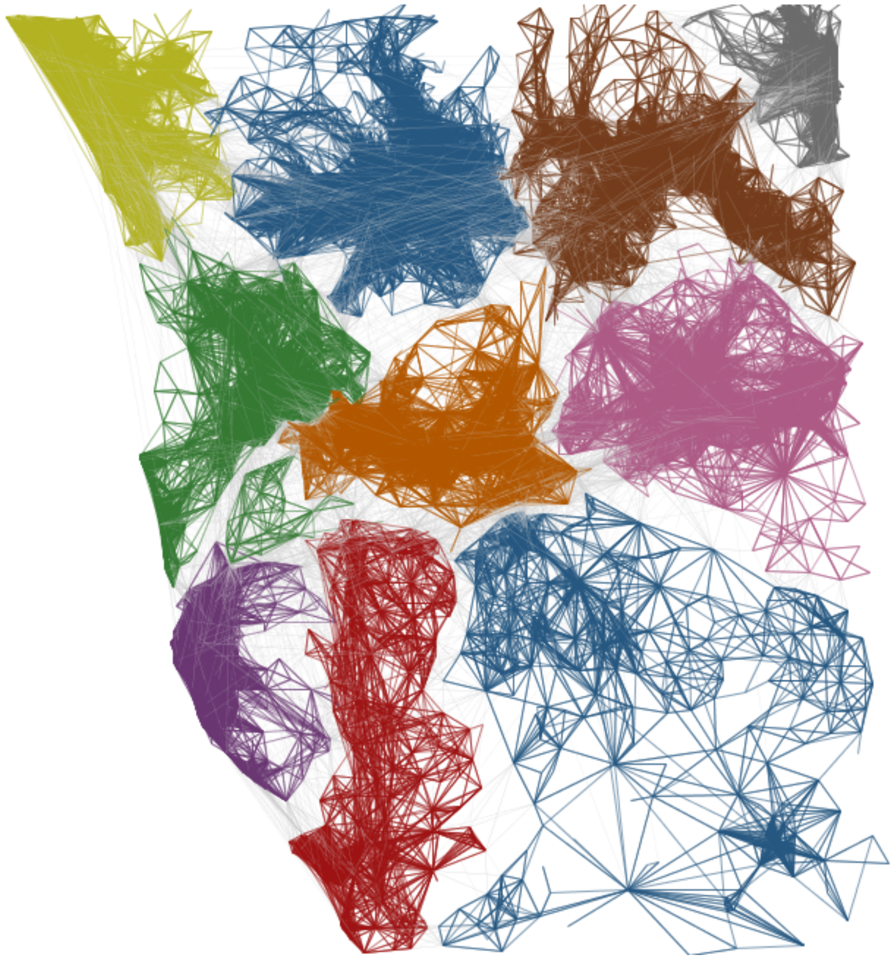
Salvatore Rinzivillo, Mainardi, Pezzoni, Michele Coscia, Dino Pedreschi, Fosca Giannotti: Discovering the Geographical Borders of Human Mobility. KI 26(3): 253-260 (2012)

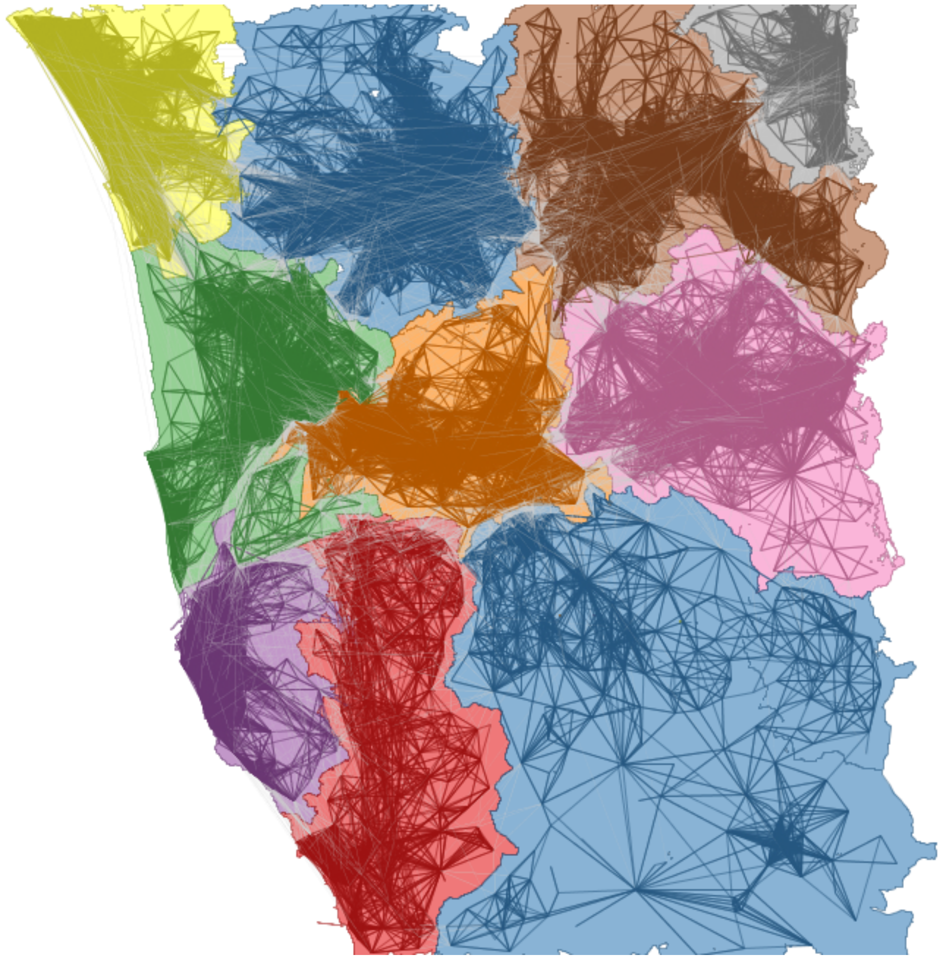


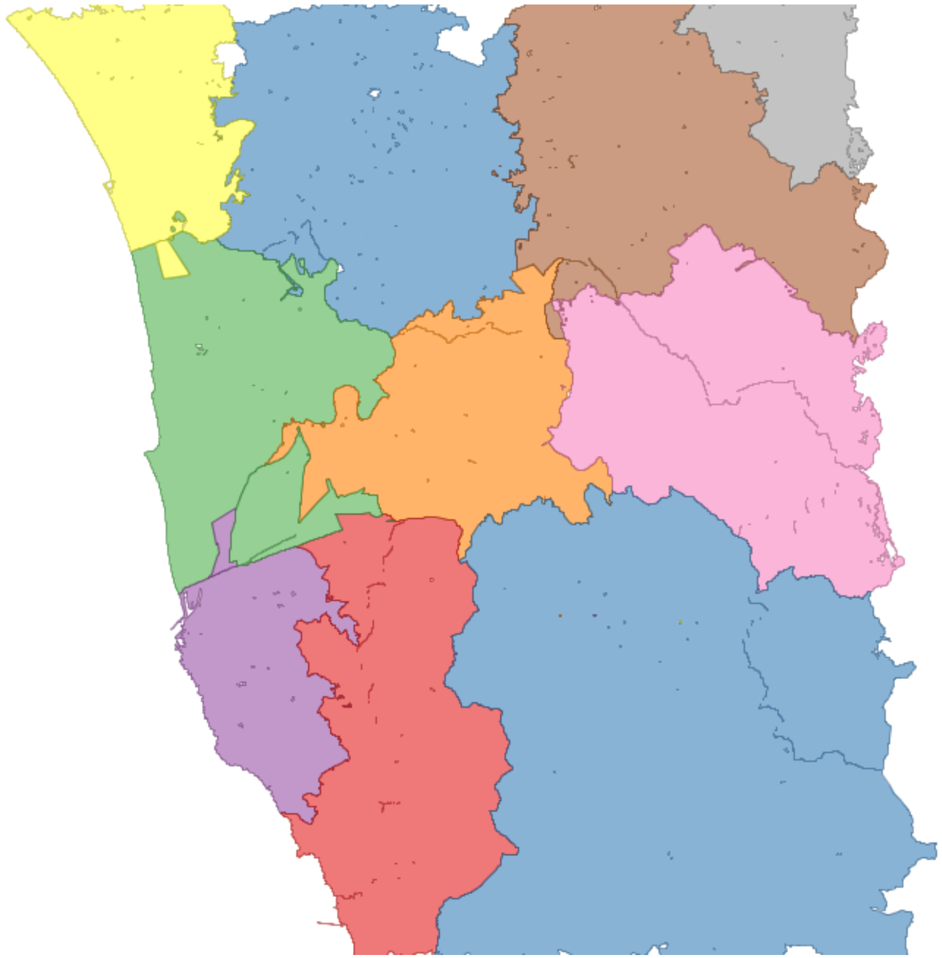


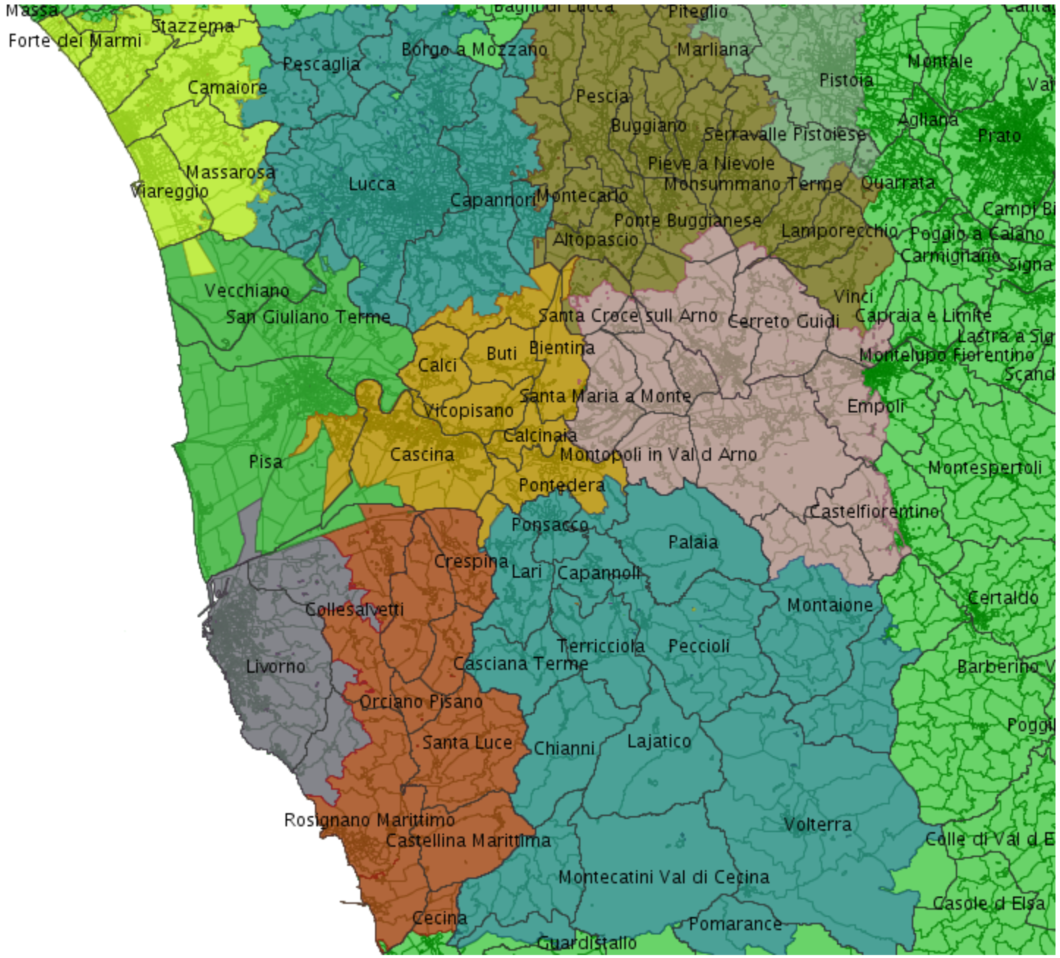




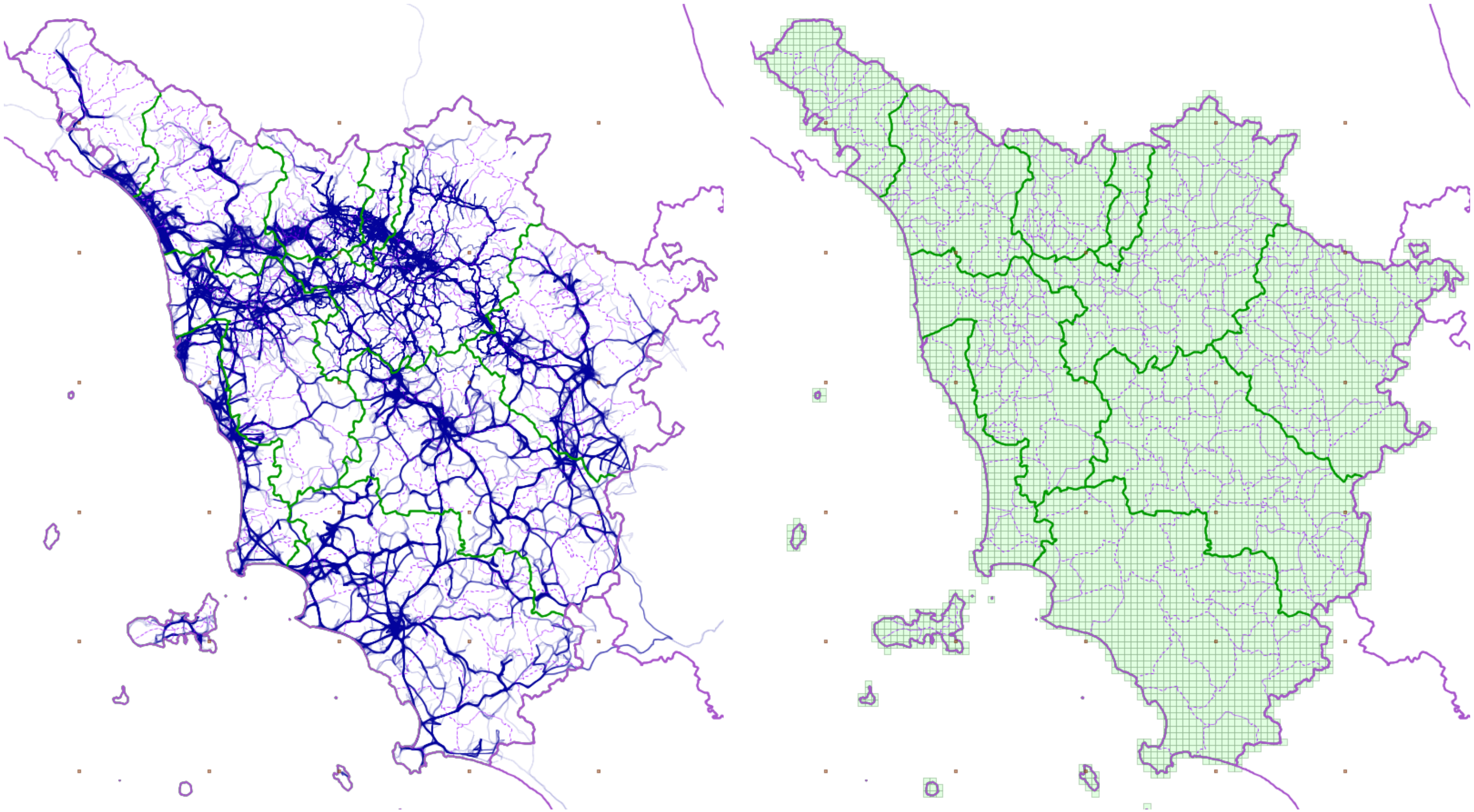


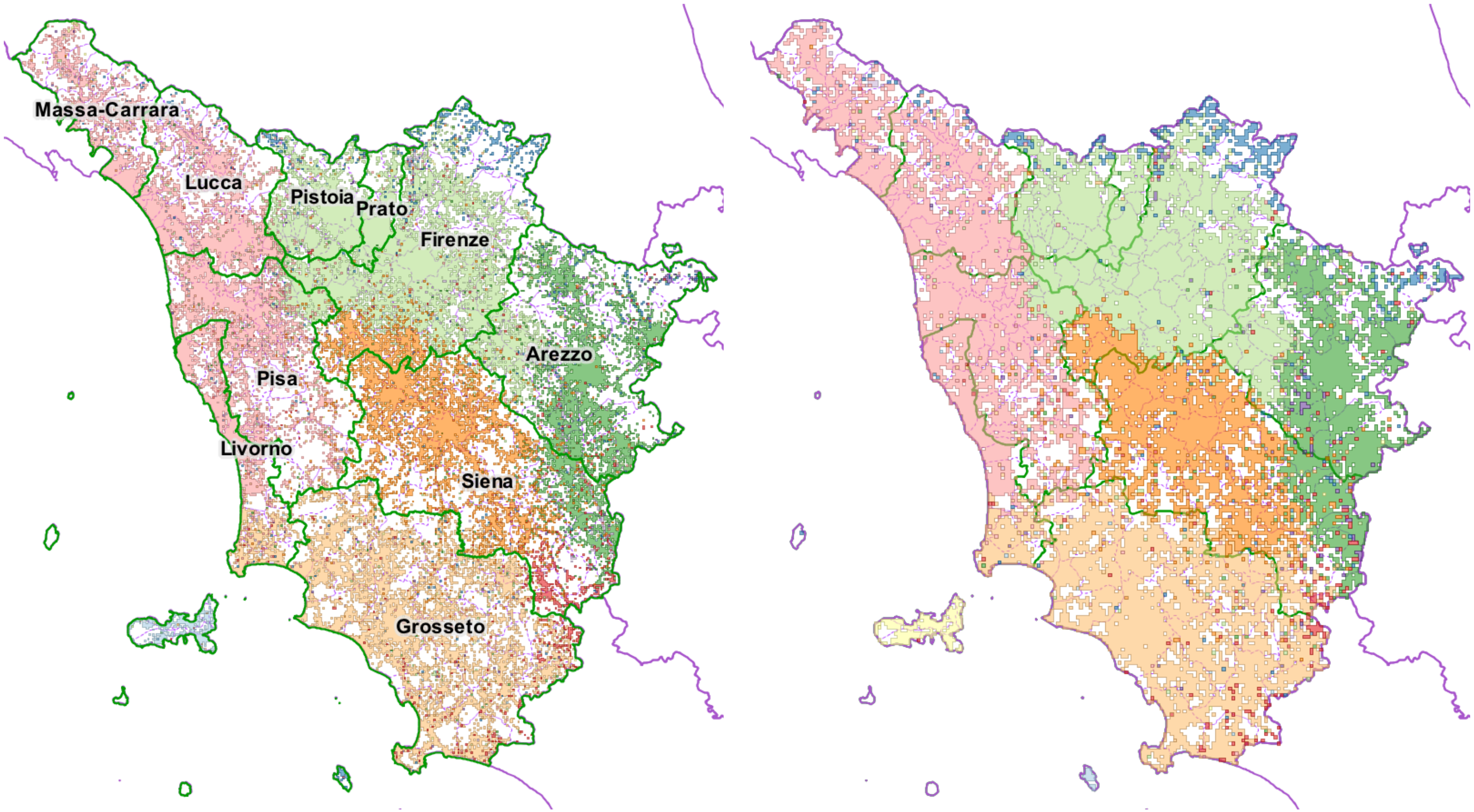


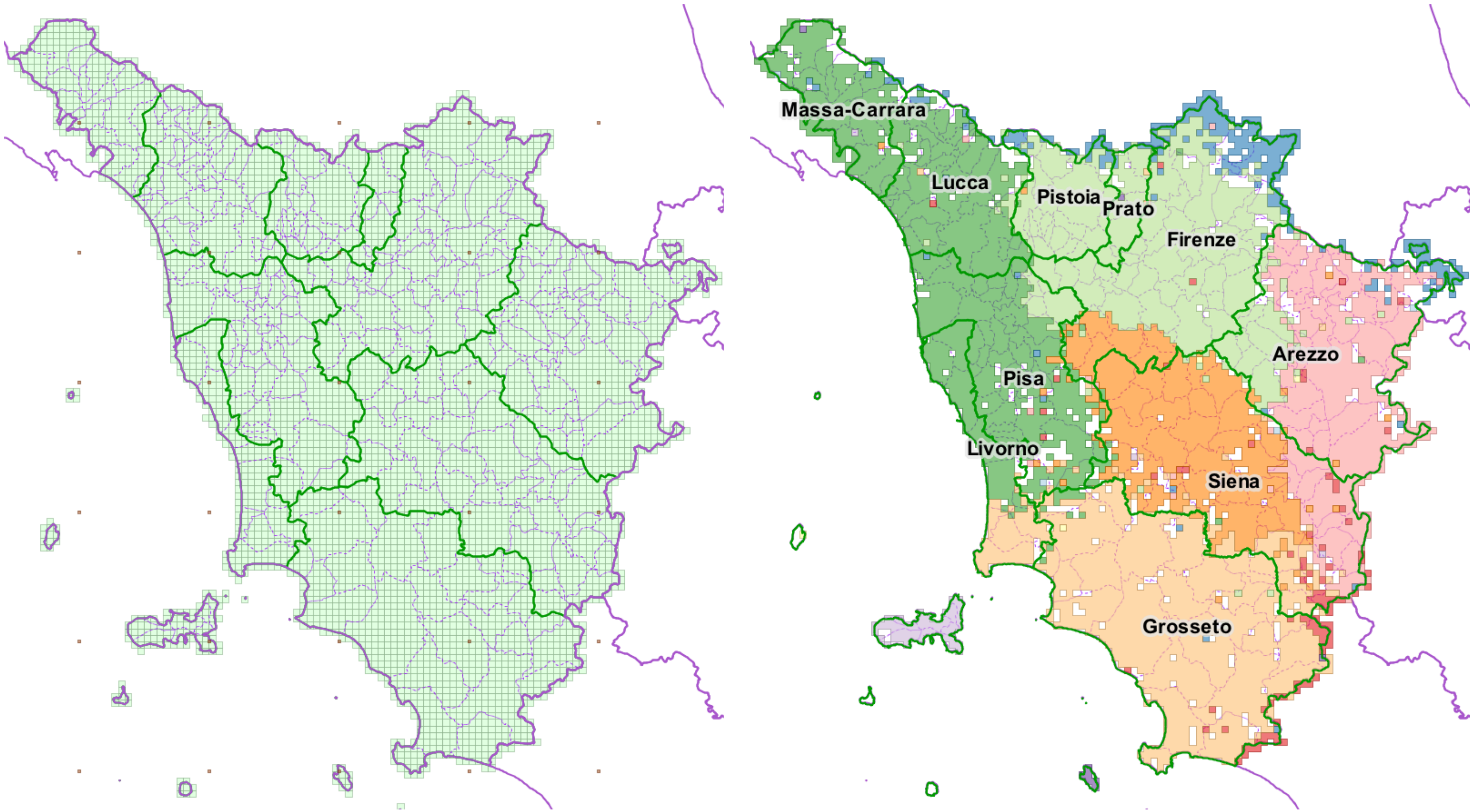


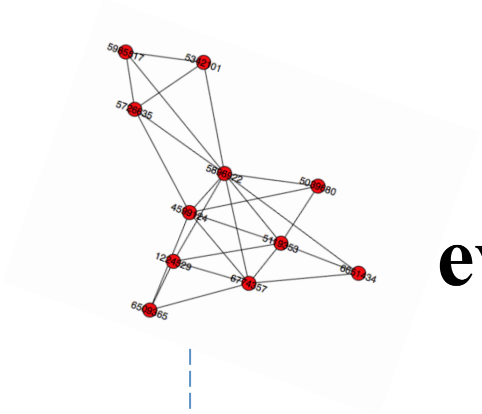


Massa
Stazzema
Forte dei Marmi
Pieve a Nievole
Pescaglia
Borgo a Mozzano
Marliana
Pistoia
Montale
Camaiore
Pescia
Buggiano
Serravalle Pistoiese
Agliana
Prato
Massarosa
Viareggio
Lucca
Capannori
Montecarlo
Monsummano Terme
Quarrata
Ponte Buggianese
Altopascio
Lamporecchio
Poggio a Caiano
Carmignano
Signa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



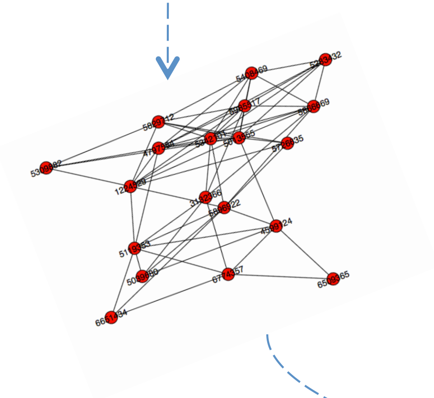

Massa-Carrara
Lucca
Pistoia
Prato
Firenze
Arezzo
Pisa
Livorno
Siena
Grosseto

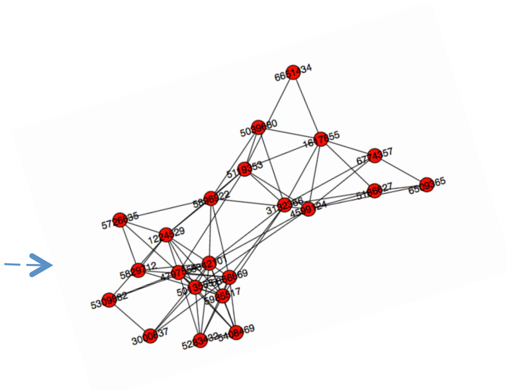

Massa-Carrara
Lucca
Pistoia
Prato
Firenze
Arezzo
Pisa
Livorno
Siena
Grosseto



The frontier: evolutionary community discovery

G Rossetti, L Pappalardo, D Pedreschi, F Giannotti
Tiles: an online algorithm for community discovery in dynamic social networks
Machine Learning, 1-29, 2016

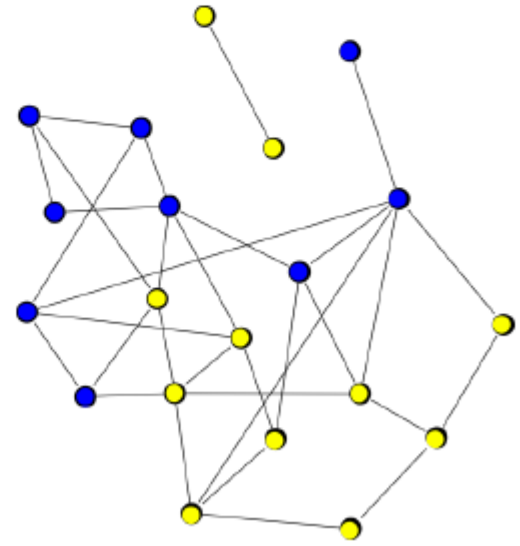


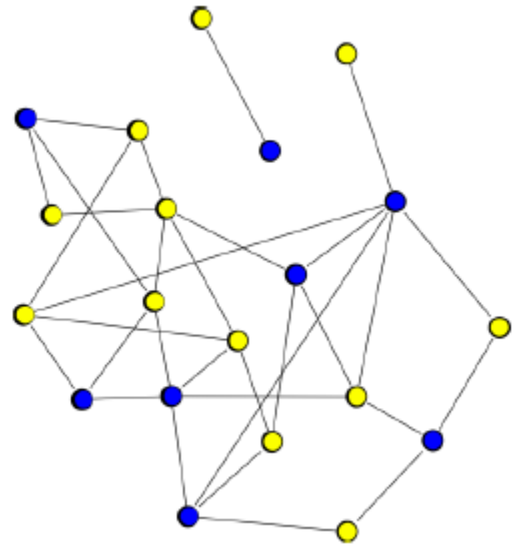


Group formation dynamics

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?

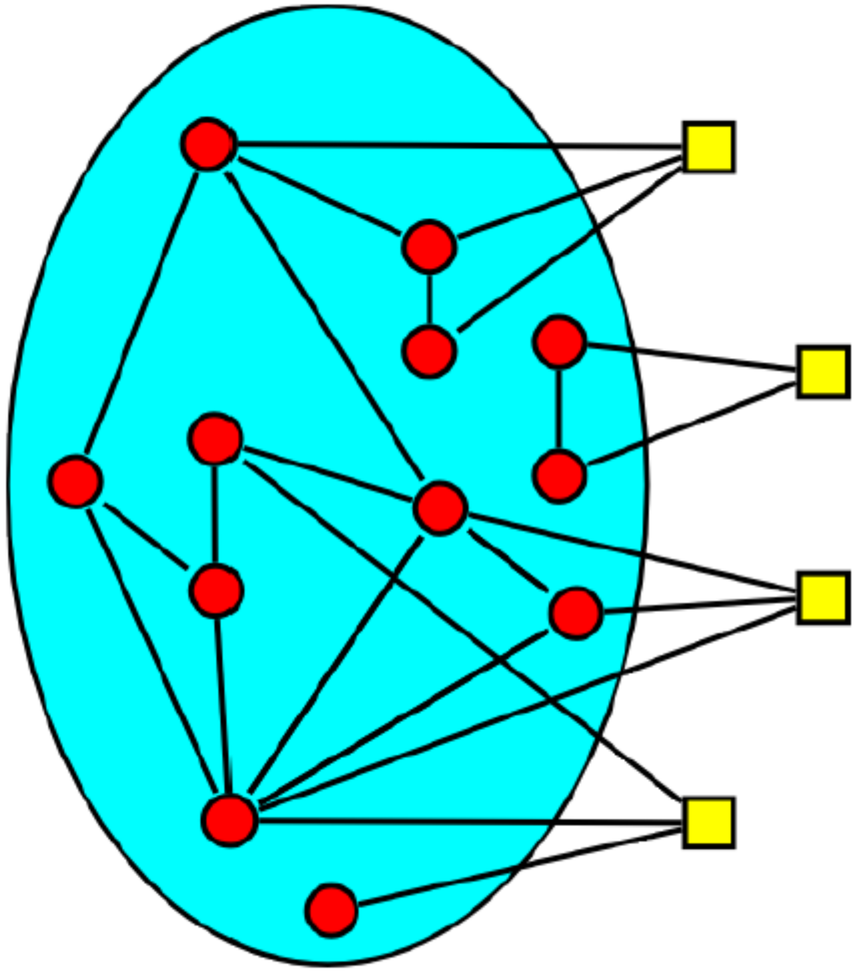


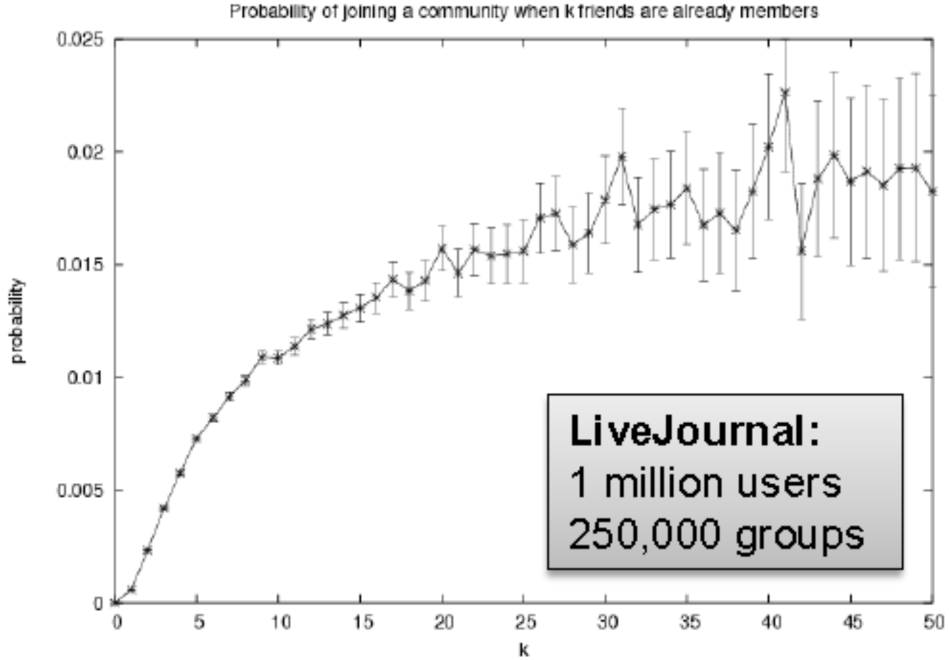


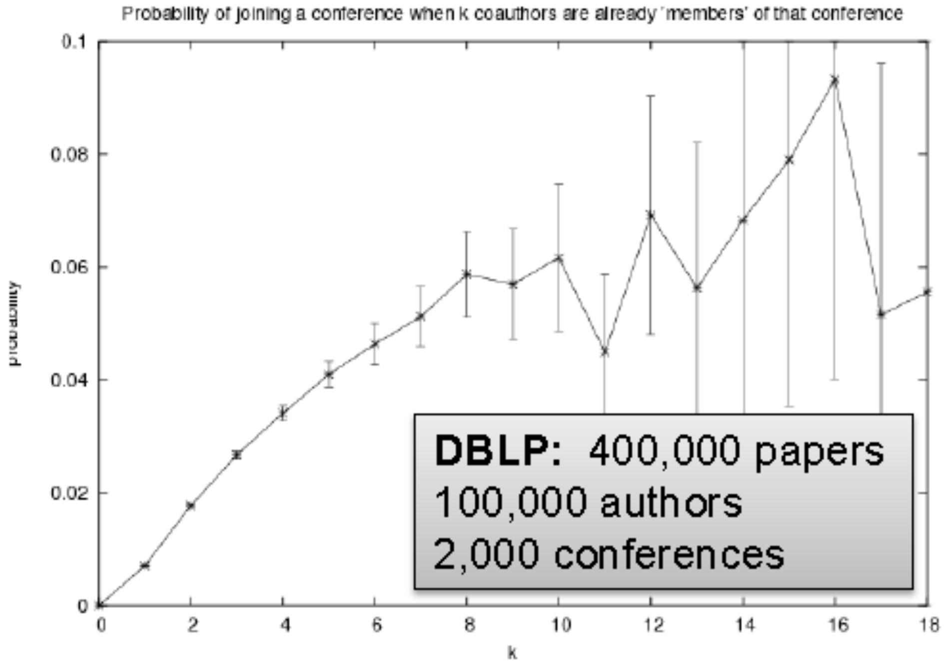
Group memberships spread over the network:

- 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?





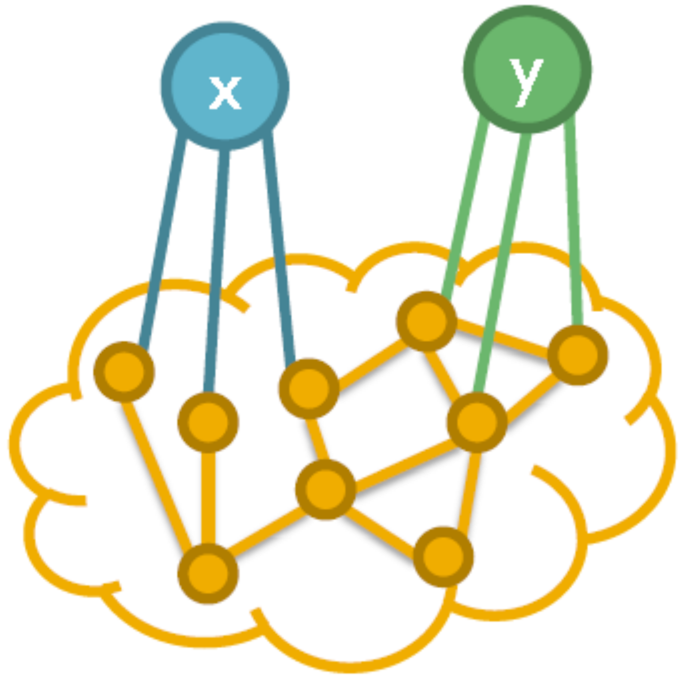


- **Diminishing returns:**
  - Probability of joining increases with the number of friends in the group
  - But increases get smaller and smaller

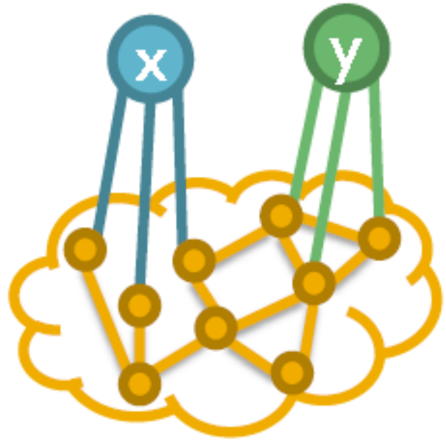
Connectedness of friends and group membership

- *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?



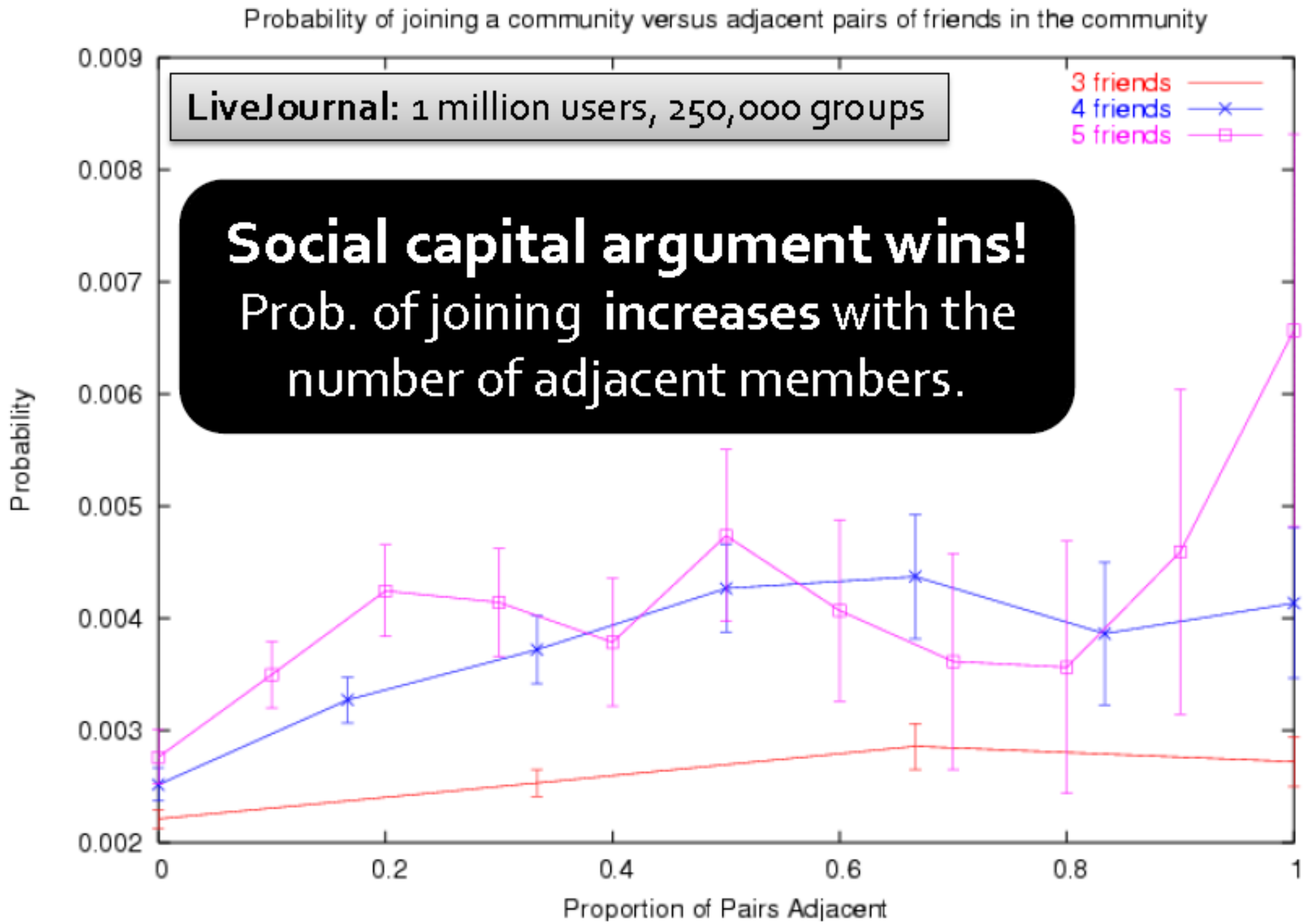
- **Competing sociological theories:**
  - Information argument [Granovetter '73]
  - Social capital argument [Coleman '88]



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

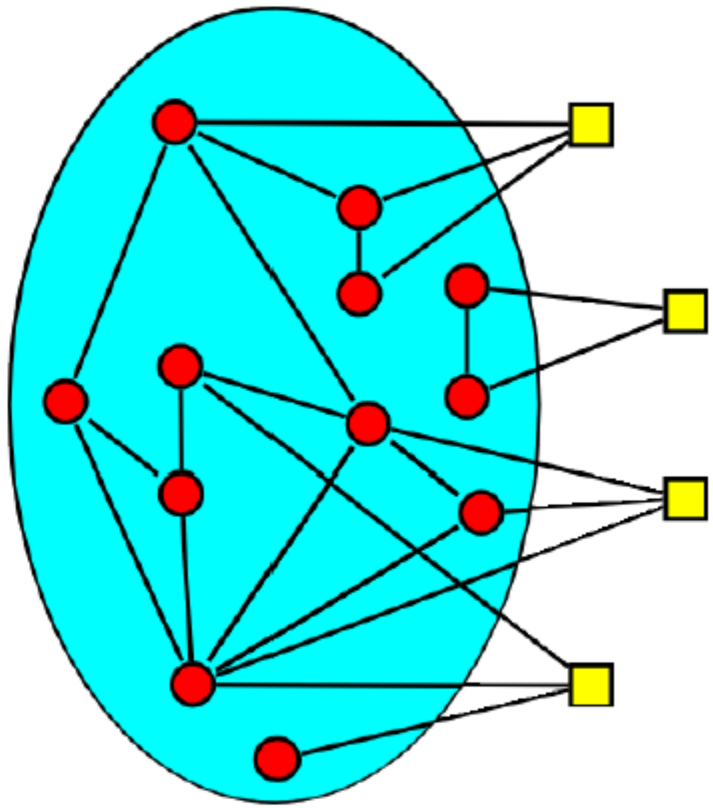
… and the winner is …

[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**

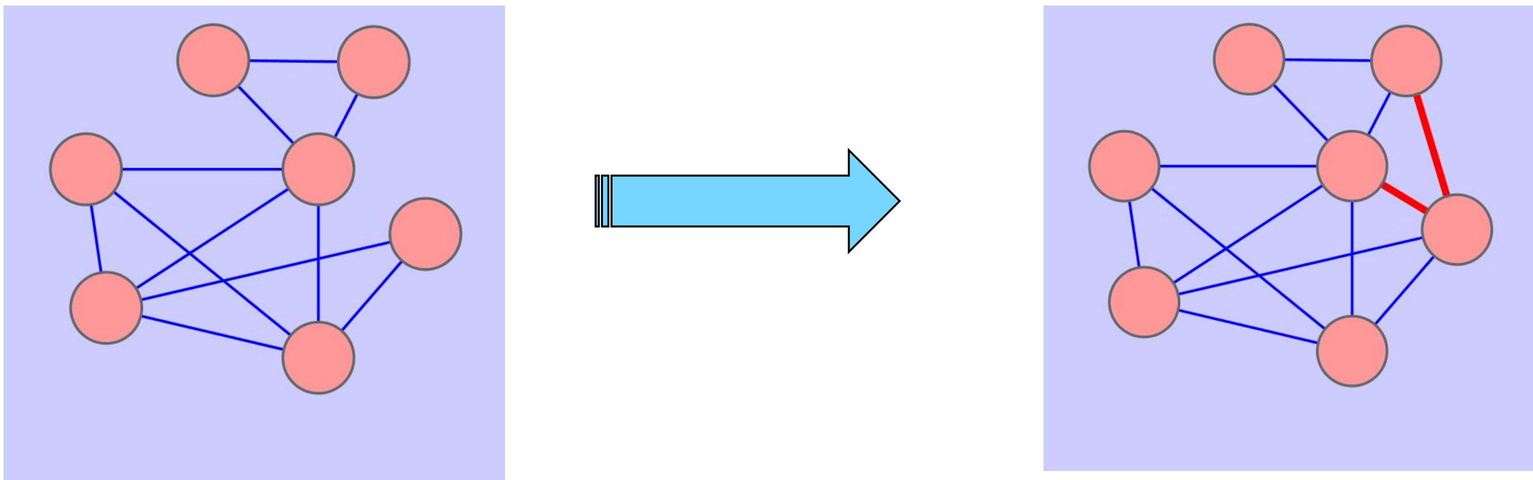


Link prediction

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

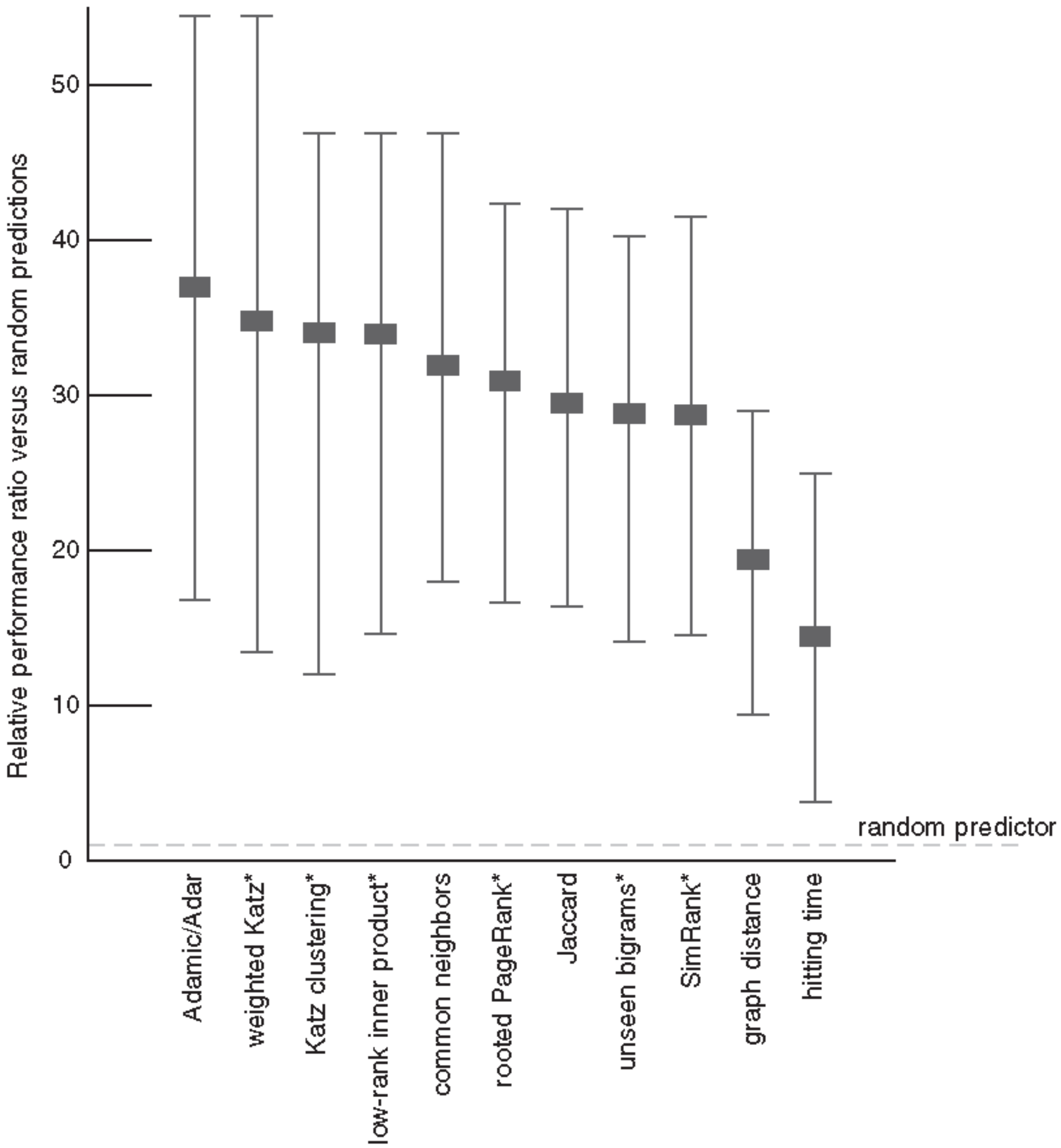
- **Unsupervised** link prediction, based on scores of topology measures such as common neighbors, Jaccard coefficient, Adamic/Adar measure, Katz
  - D. Liben-Nowell, J. Kleinberg. The link prediction problem for social networks. *J. of Am. Soc. for Information Science and Technology*, 58(7):1019-1031, 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?

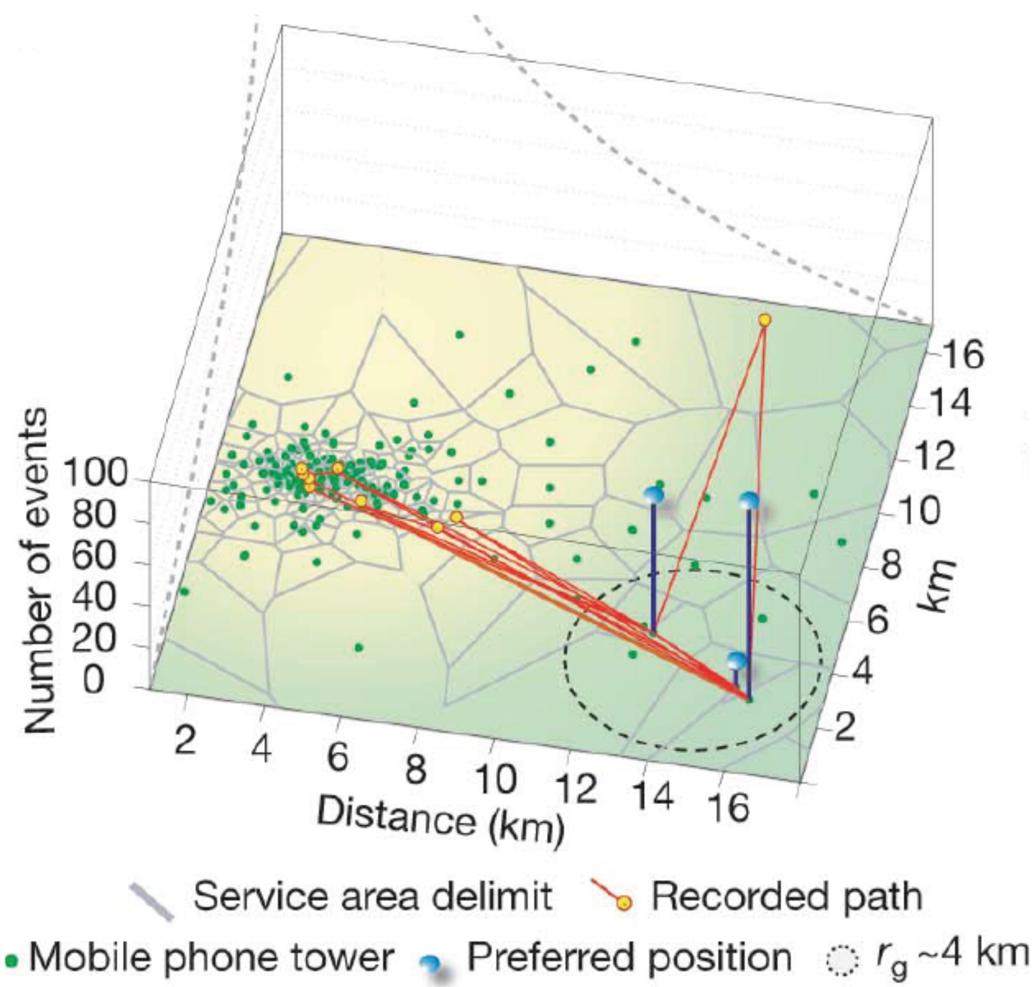
- [Liben-Nowell and Kleinberg 2006]

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$_\beta$	$\sum_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\mathsf{paths}_{x,y}^{\langle\ell\rangle}\rvert$
	where $\mathsf{paths}_{x,y}^{\langle\ell\rangle} := \{$paths of length exactly ℓ from x to $y\}$

Performance of predictors (wrt random)



Country-wide tele-communication data



when you call

where you call

who you call

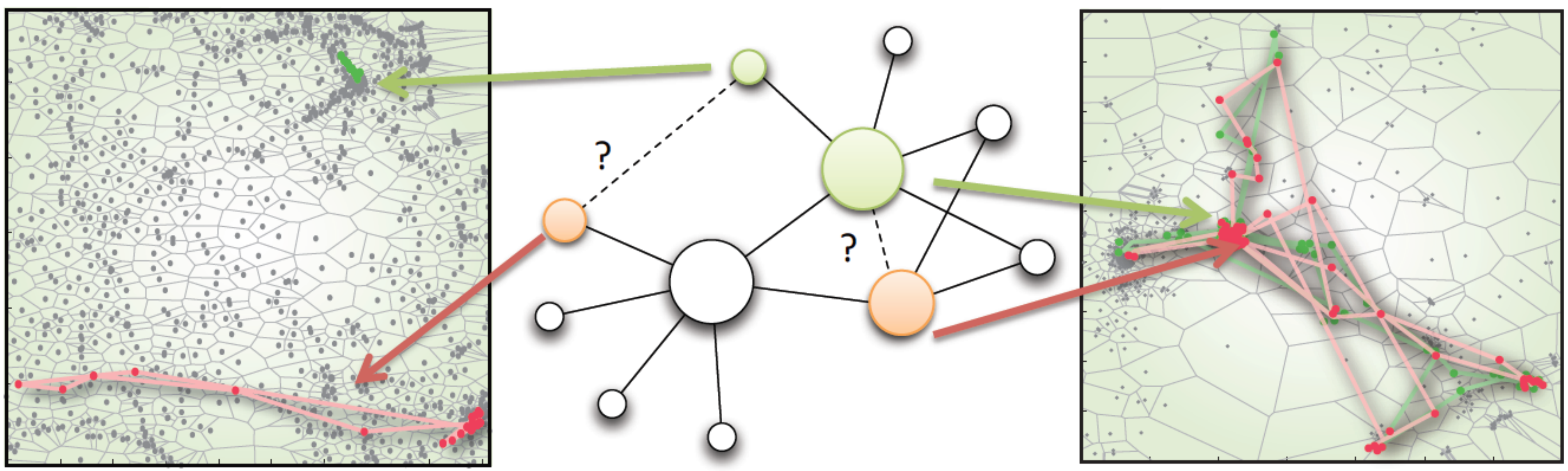
Link prediction in **mobile** social networks

- 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

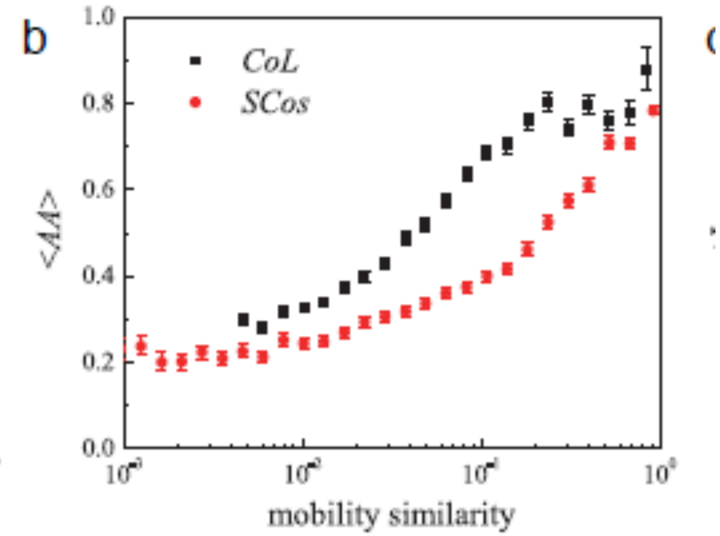


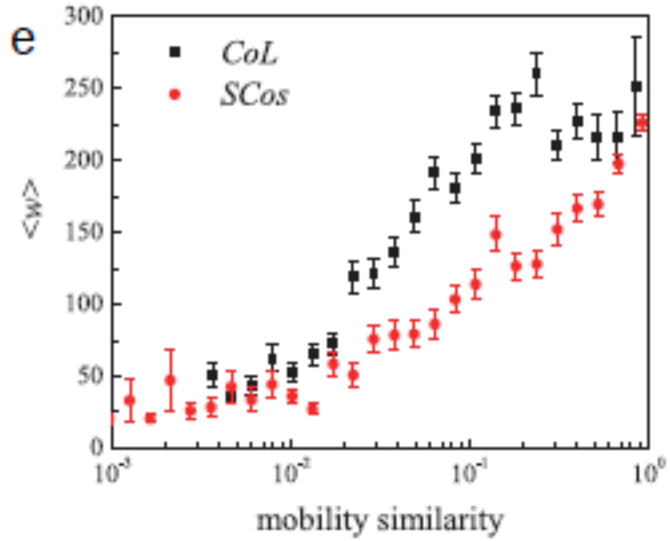
Correlation: Colocation, social proximity, tie strength

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

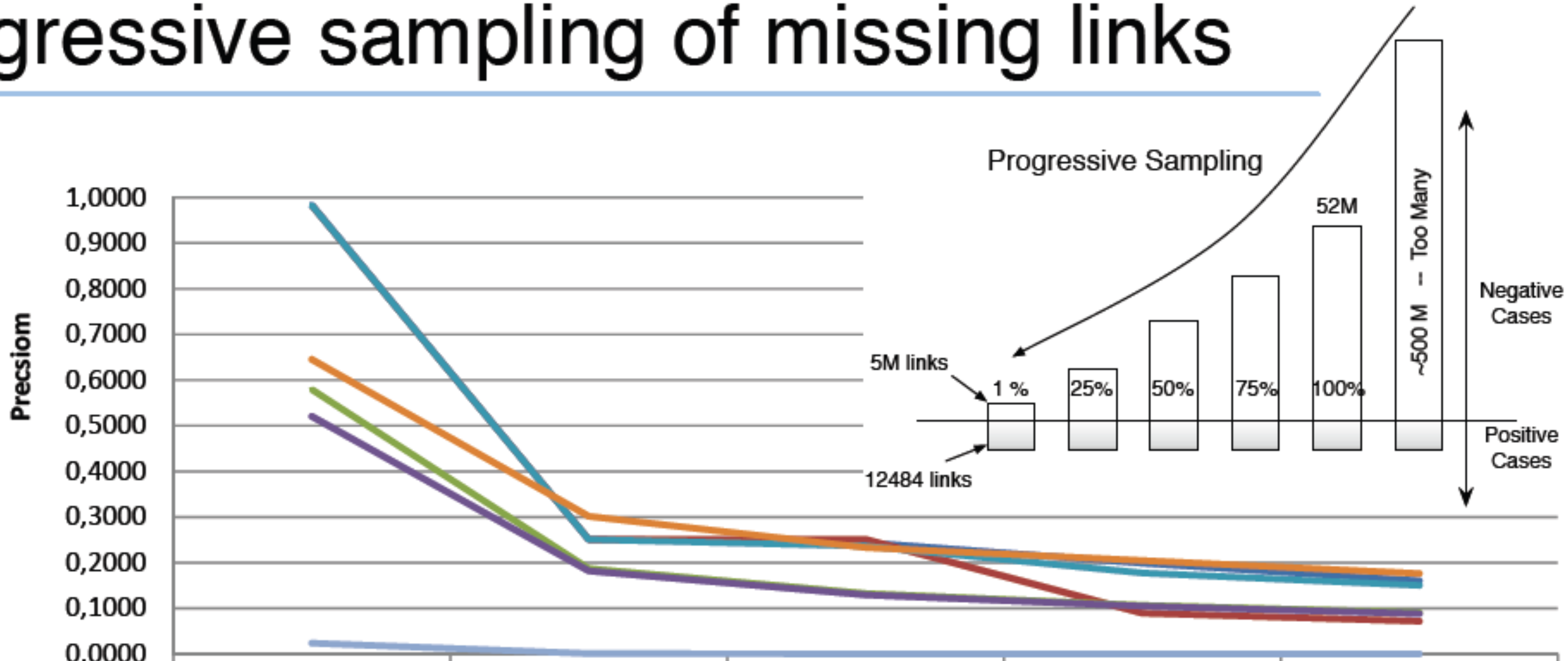




- co-location, network proximity and tie strength strongly correlate with each other
- measured on 3 months of calls, 6 Million users, nation-wide (large European country)
- **mobility dimension of the "strength of weak ties"**

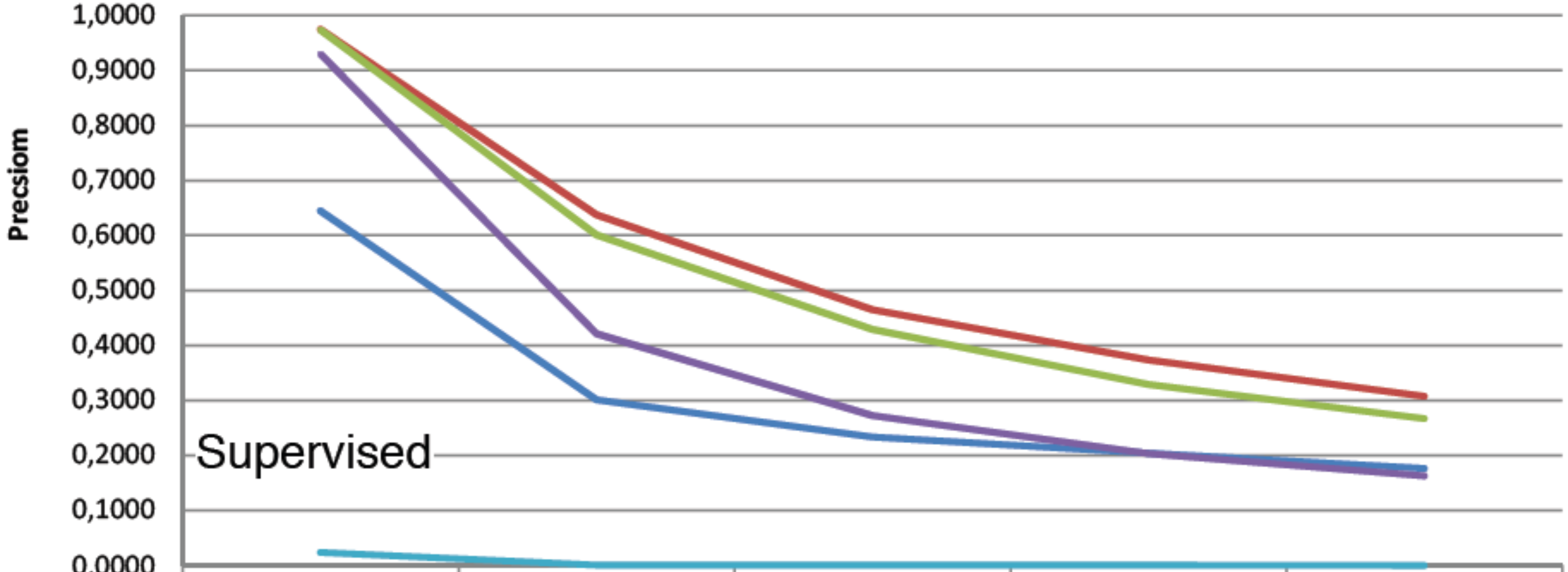
Unsupervised link prediction

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:

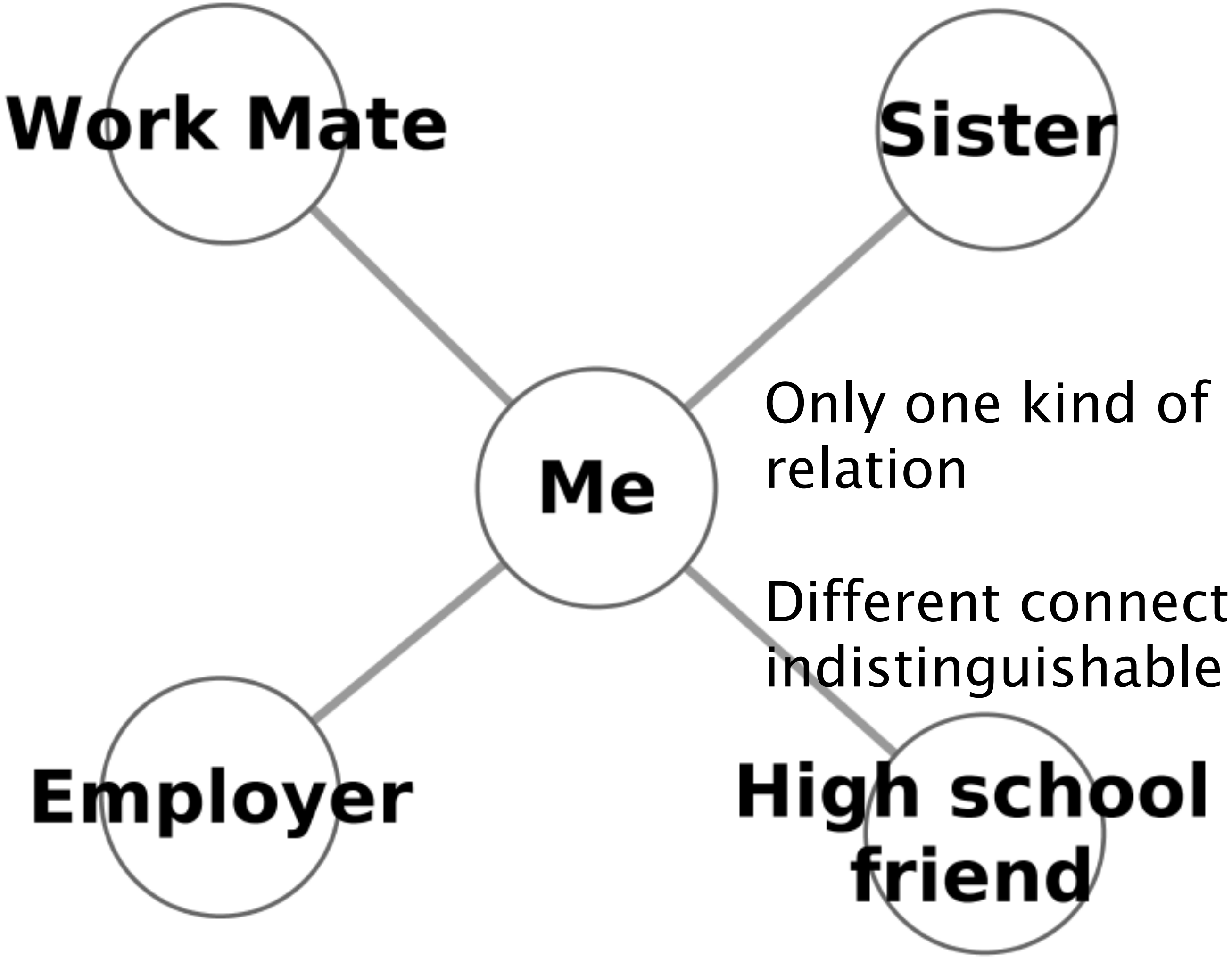
  75%

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)

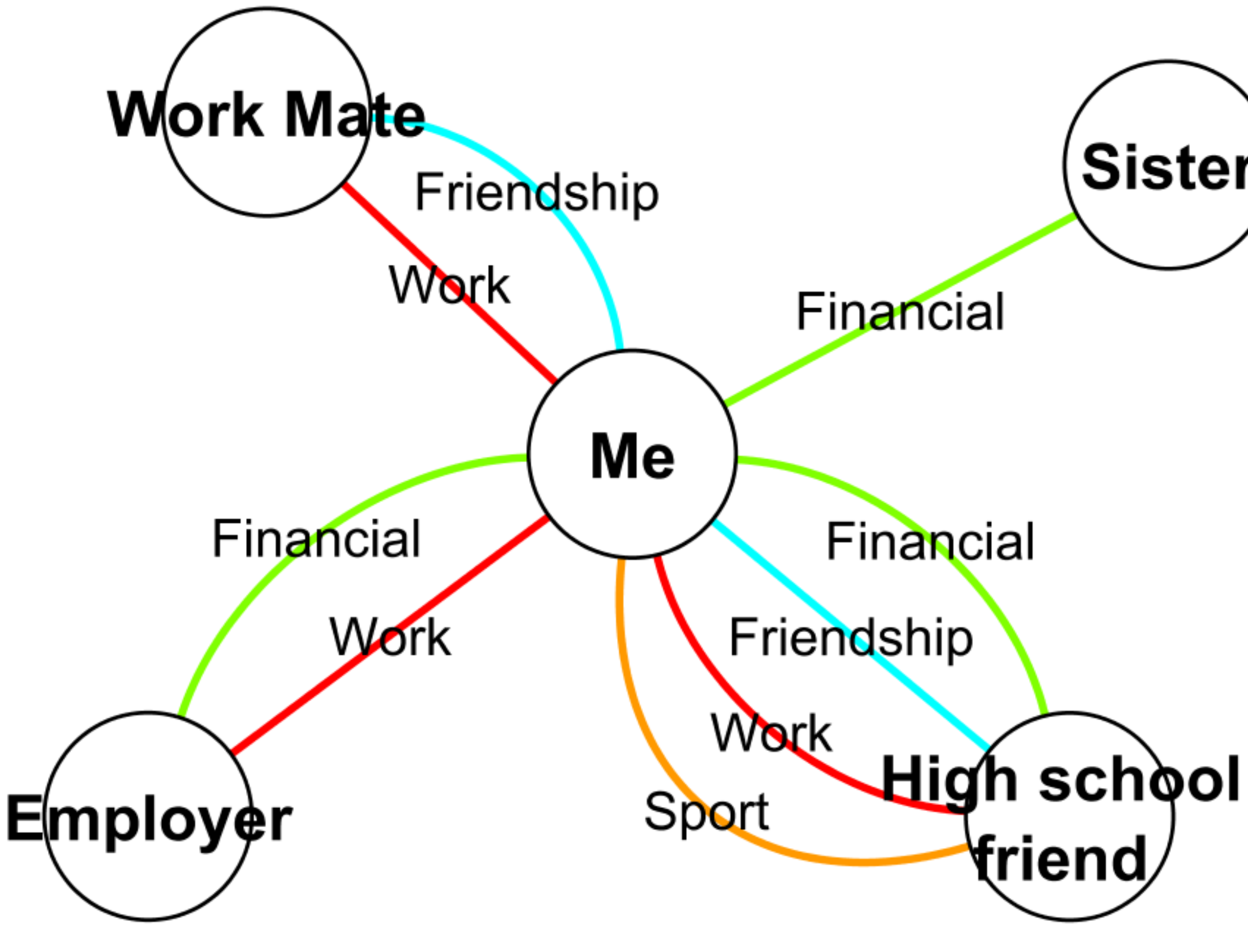
Classical Network Representation



Only one kind of relation

Different connections indistinguishable

Multigraphs as multidimensional networks



Network robustness

A SIMPLE STORY (3):

ROBUSTNESS IN COMPLEX SYSTEMS

Complex systems maintain their basic functions even under errors and failures

cell → mutations

There are uncountable number of mutations and other errors in our cells, yet, we do not notice their consequences.

Internet → router breakdowns

At any moment hundreds of routers on the internet are broken, yet, the internet as a whole does not loose its functionality.

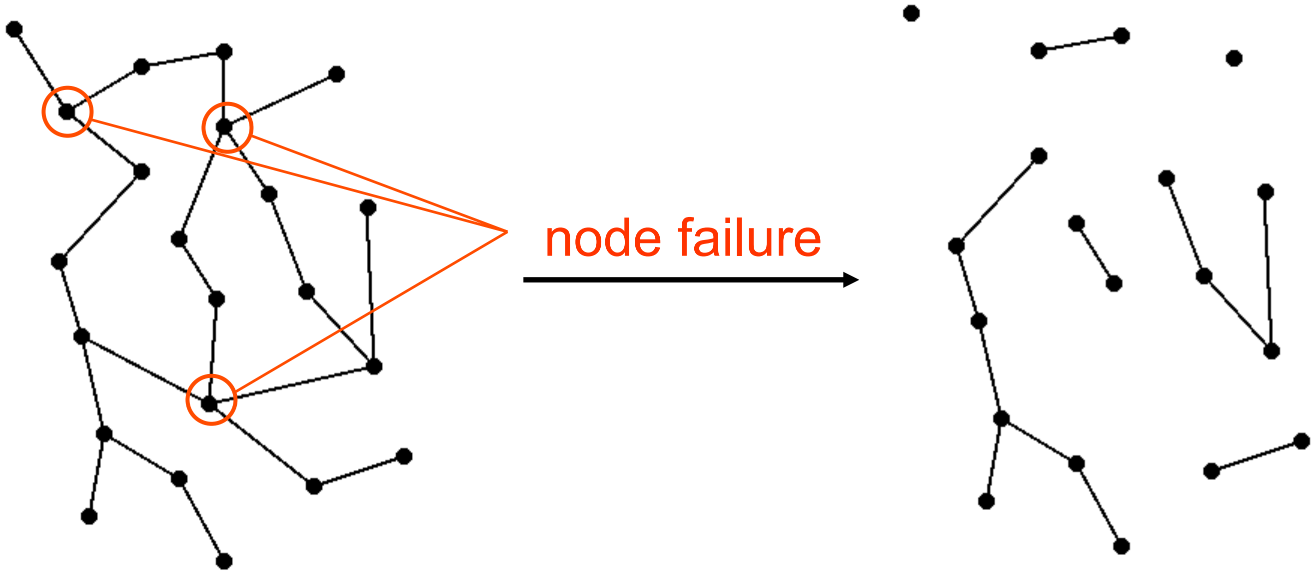
Where does robustness come from?

There are feedback loops in most complex systems that keep tab on the component's and the system's 'health'.

Could the network structure affect a system's robustness?

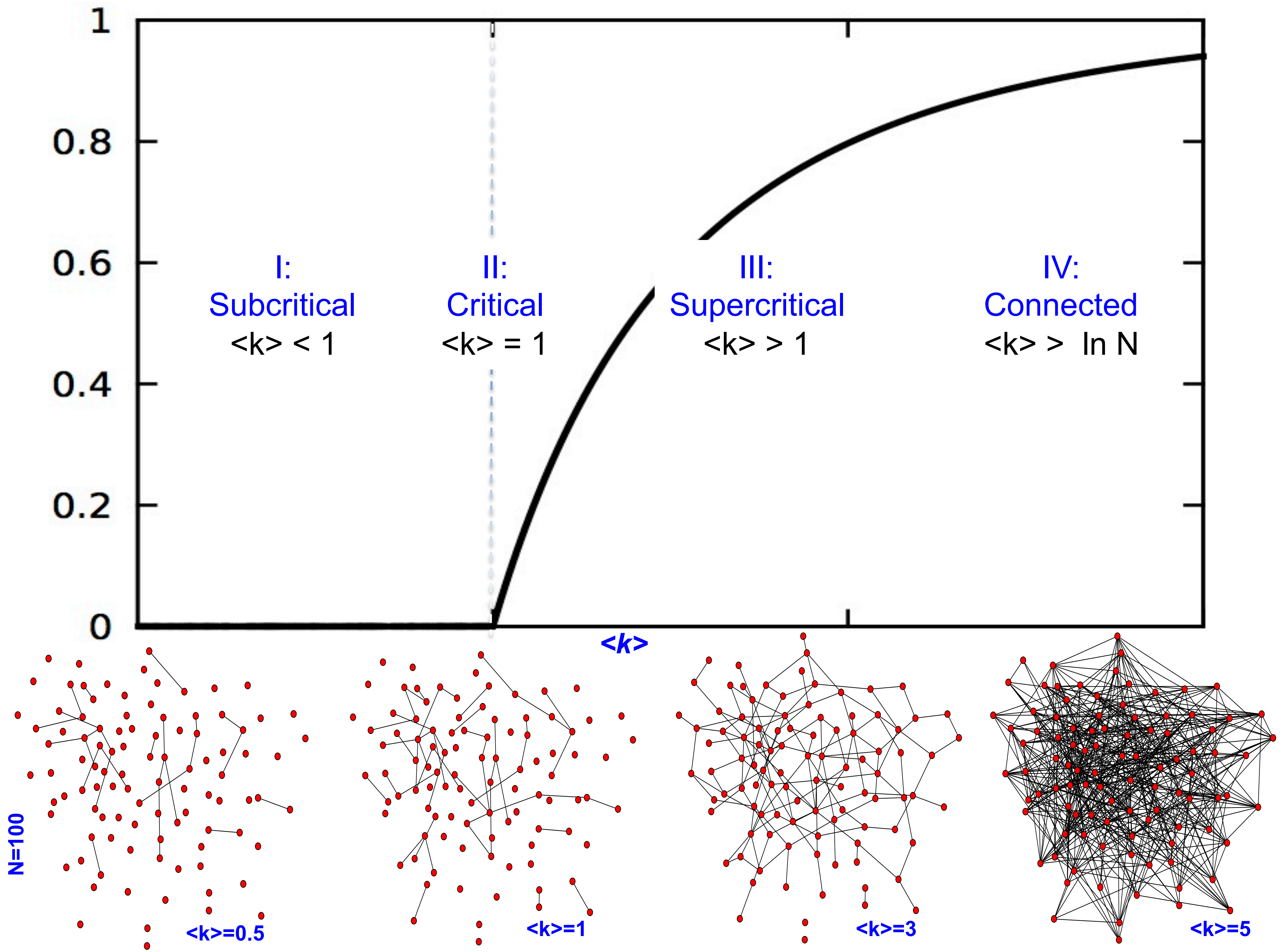
ROBUSTNESS

Could the network structure affect a system's robustness?



How do we describe in quantitave terms the breakdown of a network under node or link removal?

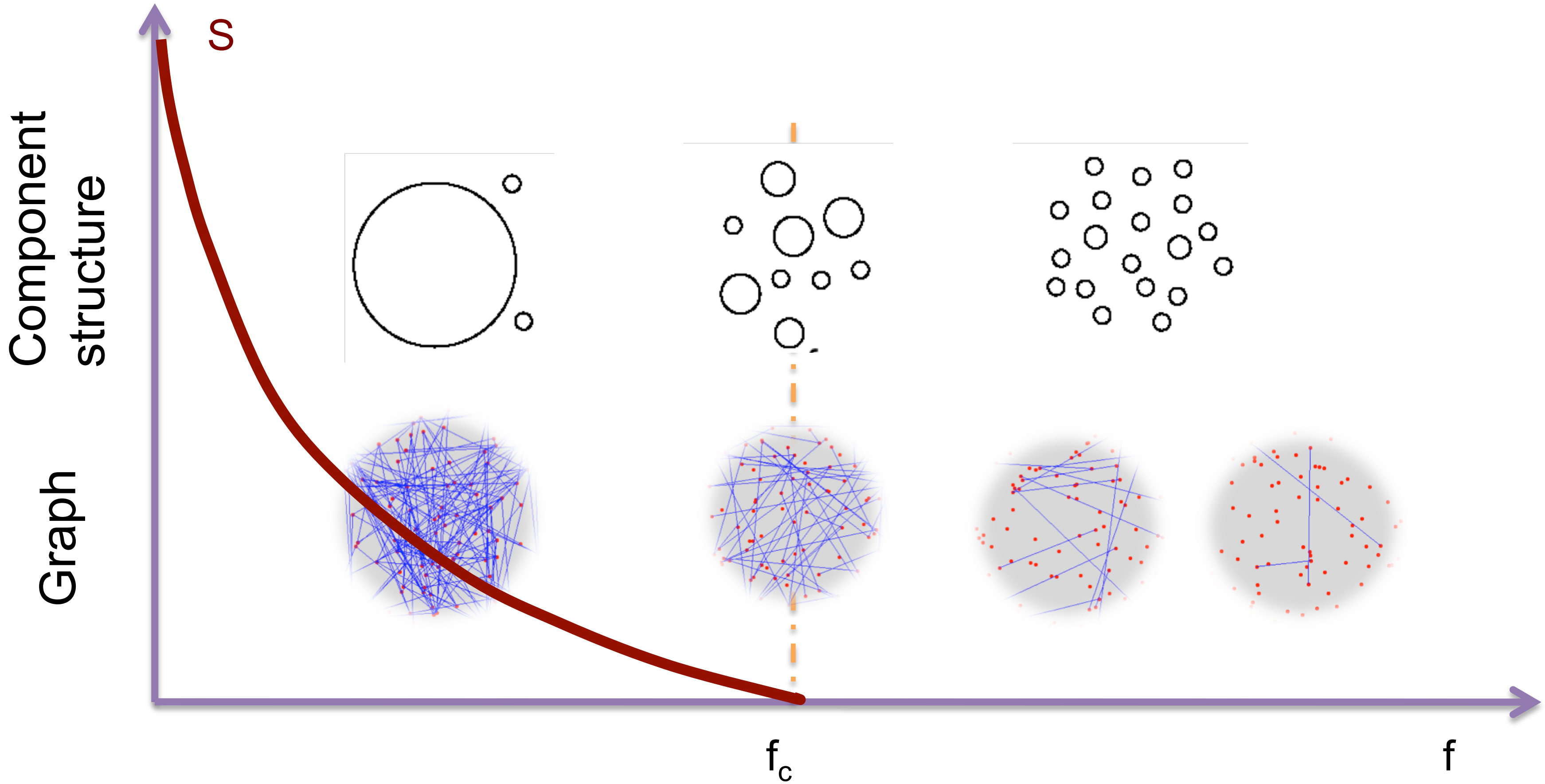
~percolation theory~



1
0.8
0.6
0.4
0.2
0
I:
Subcritical
<k> < 1
II:
Critical
<k> = 1
III:
Supercritical
<k> > 1
IV:
Connected
<k> > ln N
<k>
N=100
<k>=0.5
<k>=1
<k>=3
<k>=5

Damage is modeled as an inverse percolation process

f= fraction of removed nodes



(Inverse Percolation phase transition)

ROBUSTNESS: OF SCALE-FREE NETWORKS

The interest in the robustness problem has three origins:

→Robustness of complex systems is an important problem in many areas

→Many real networks are not regular, but have a scale-free topology

→*In scale-free networks the scenario described above is not valid*

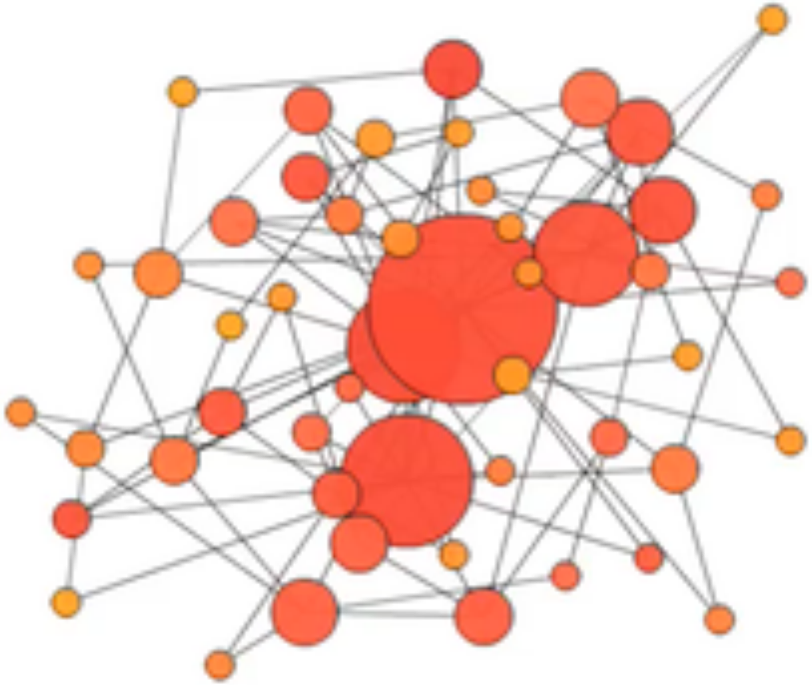
Albert, Jeong, Barabási, *Nature* **406** 378 (2000)

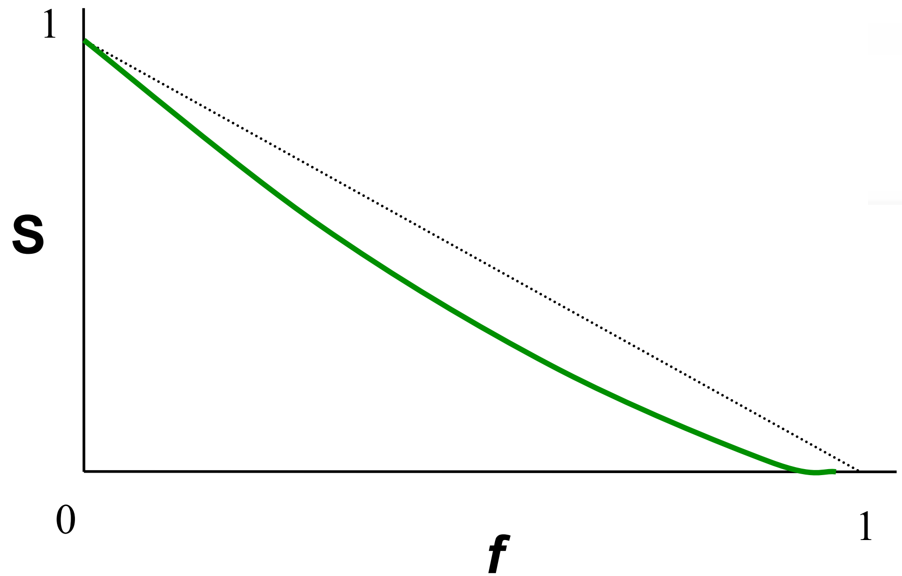
ROBUSTNESS OF SCALE-FREE NETWORKS

Scale-free networks do not appear to break apart under random failures.

Reason: the hubs.

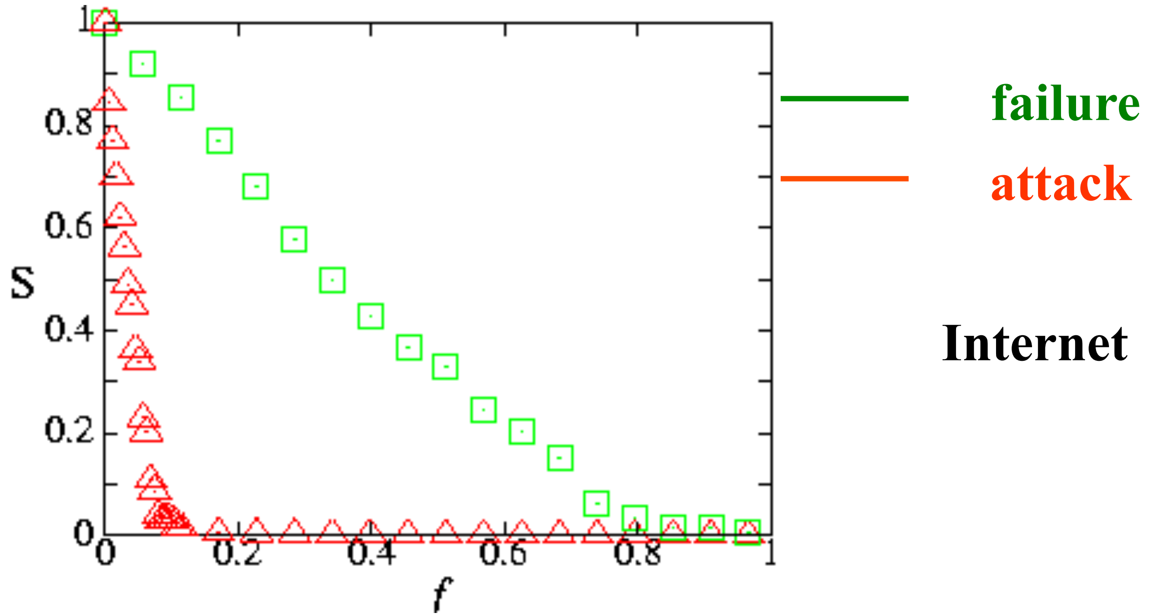
The likelihood of removing a hub is small.





Albert, Jeong, Barabási, *Nature* **406** 378 (2000)

INTERNET'S ROBUSTNESS TO RANDOM FAILURES



$$f_c = 1 - \frac{1}{\kappa - 1}$$

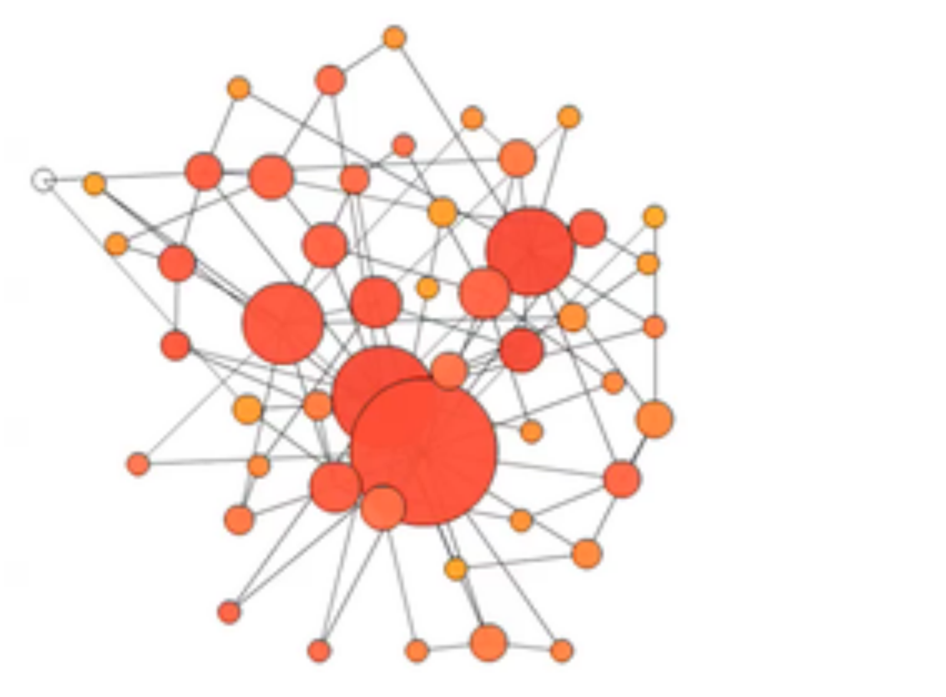
R. Albert, H. Jeong, A.L. Barabasi, *Nature* **406** 378 (2000)

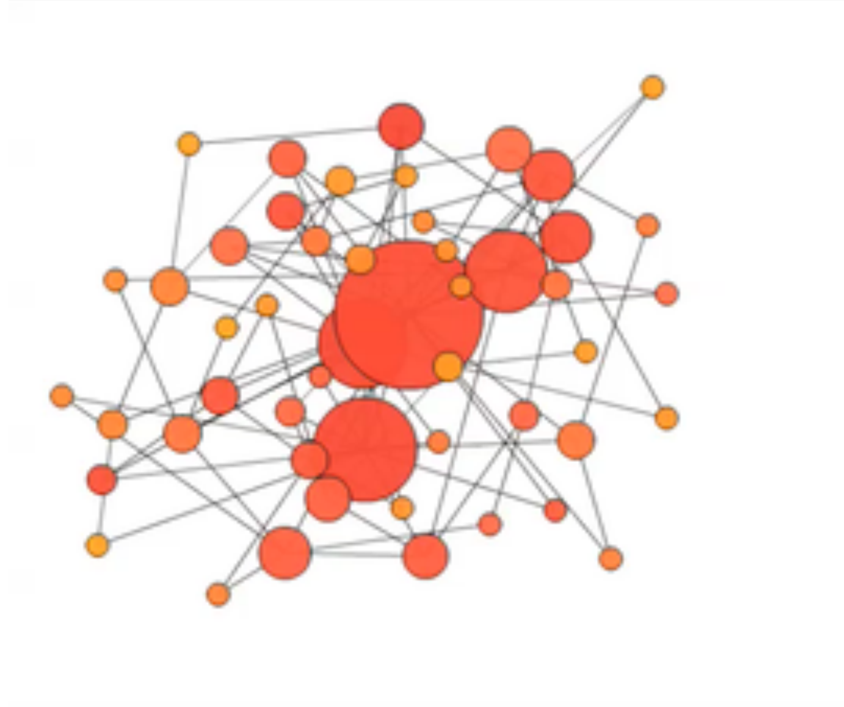
Internet: Router level map, N=228,263; γ=2.1±0.1; *κ=28* → ***f_c=0.962***

AS level map, N= 11,164; γ=2.1±0.1; *κ=264* → ***f_c=0.996***

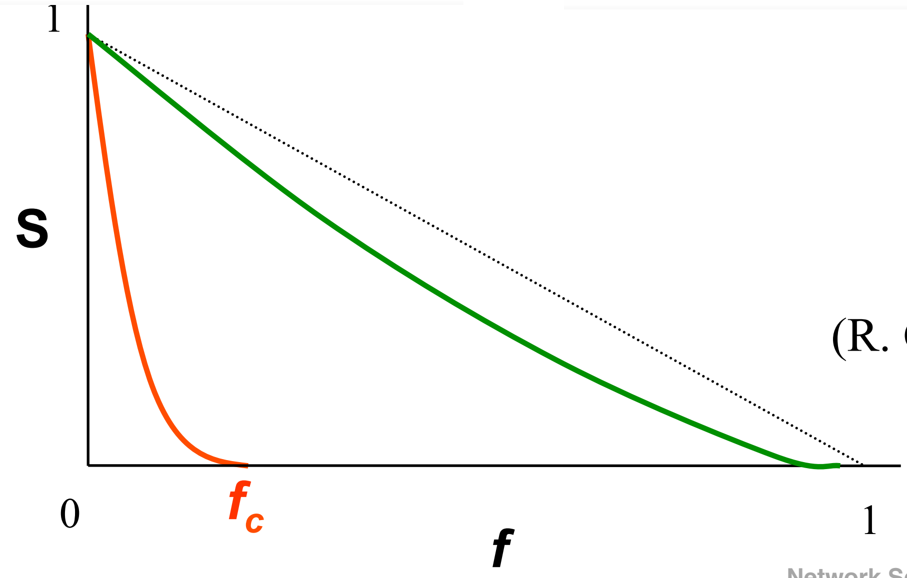
Internet parameters: Pastor-Satorras & Vespignani, *Evolution and Structure of the Internet*: Table 4.1 & 4.4

Achilles' Heel of scale-free networks





Attacks



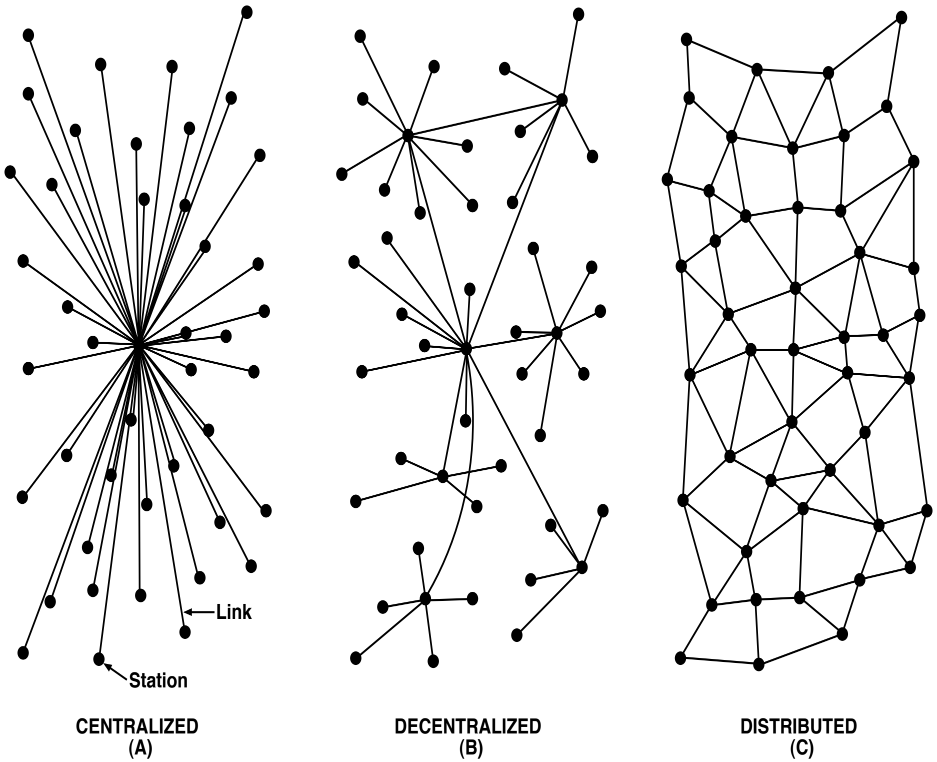
Failures

$\gamma \leq 3 : f_c = 1$

(R. Cohen et al PRL, 2000)

Albert, Jeong, Barabási, *Nature* **406** 378 (2000)

Historical Detour: Paul Baran and Internet



1958

Cascades

Cascades

Potentially large events triggered by small initial shocks

- **Information cascades**

  social and economic systems

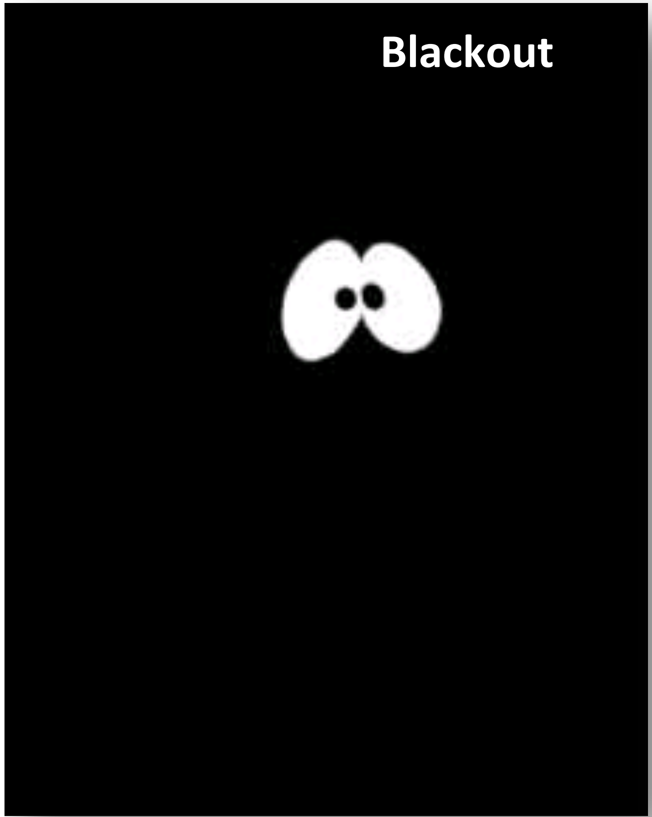
  diffusion of innovations

- **Cascading failures**

  infrastructural networks

  complex organizations

Cascading Failures in Nature and Technology



Flows of physical quantities

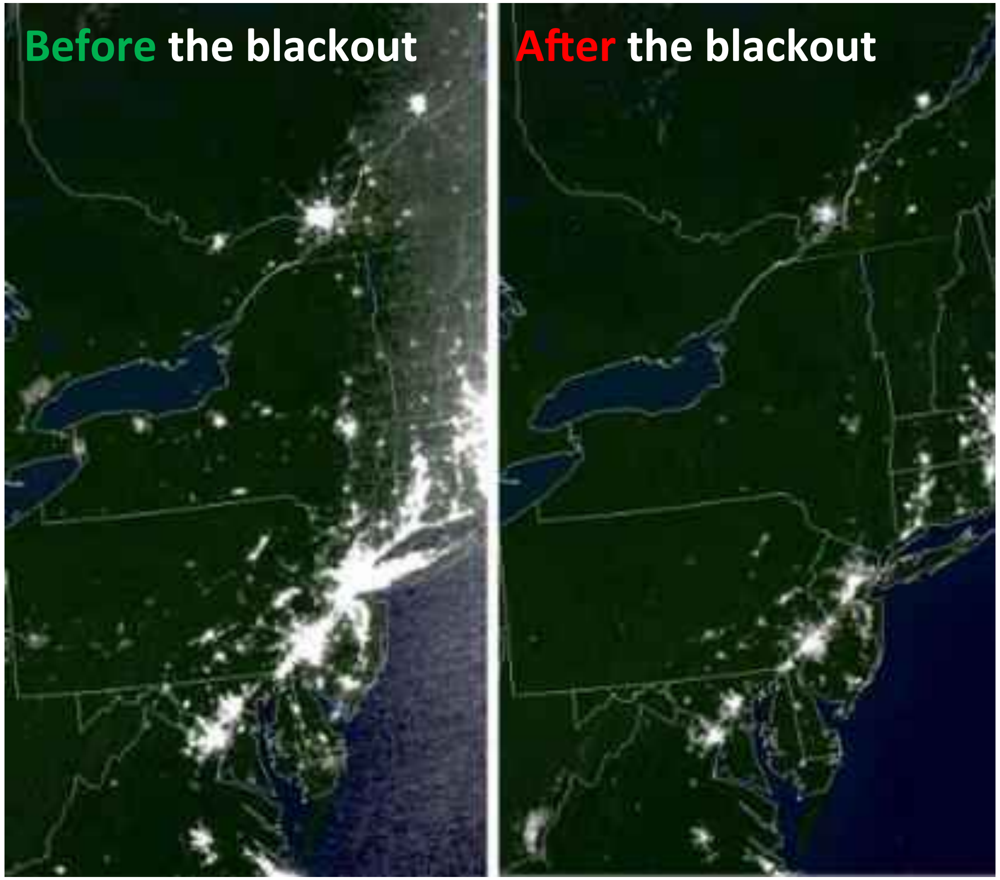
- congestions
- instabilities
- Overloads

Cascades depend on

- Structure of the network
- Properties of the flow
- Properties of the net elements
- Breakdown mechanism

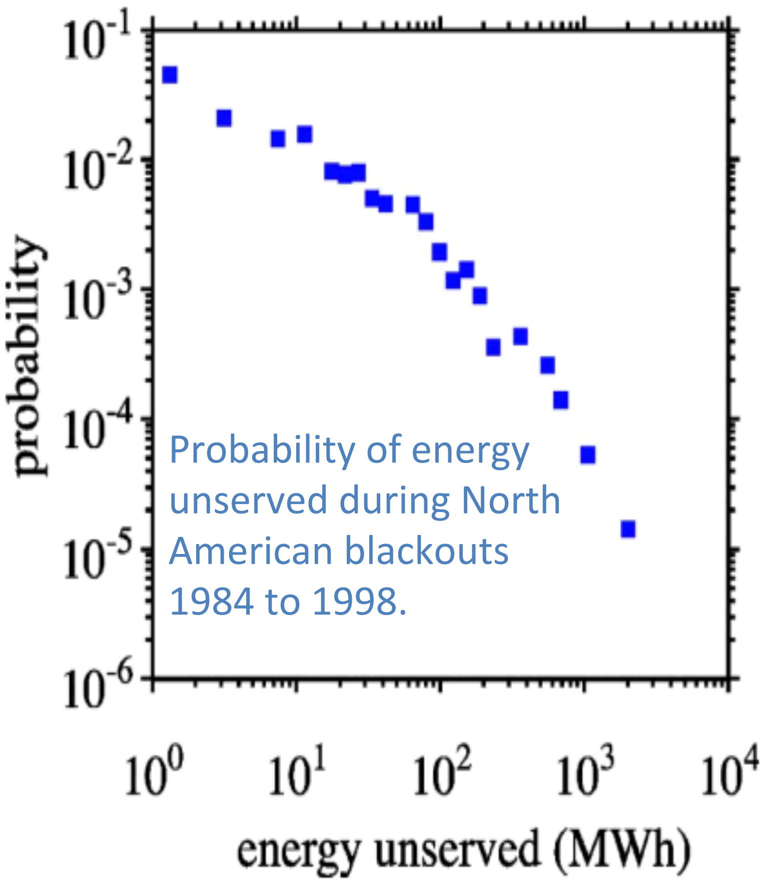
Northeast Blackout of 2003

Origin
A 3,500 MW power surge (towards Ontario) affected the transmission grid at 4:10:39 p.m. EDT. (Aug-14-2003)



Consequences
More than 508 generating units at 265 power plants shut down during the outage. In the minutes before the event, the NYISO-managed power system was carrying 28,700 MW of load. At the height of the outage, the load had dropped to 5,716 MW, a loss of 80%.

Cascades Size Distribution of Blackouts



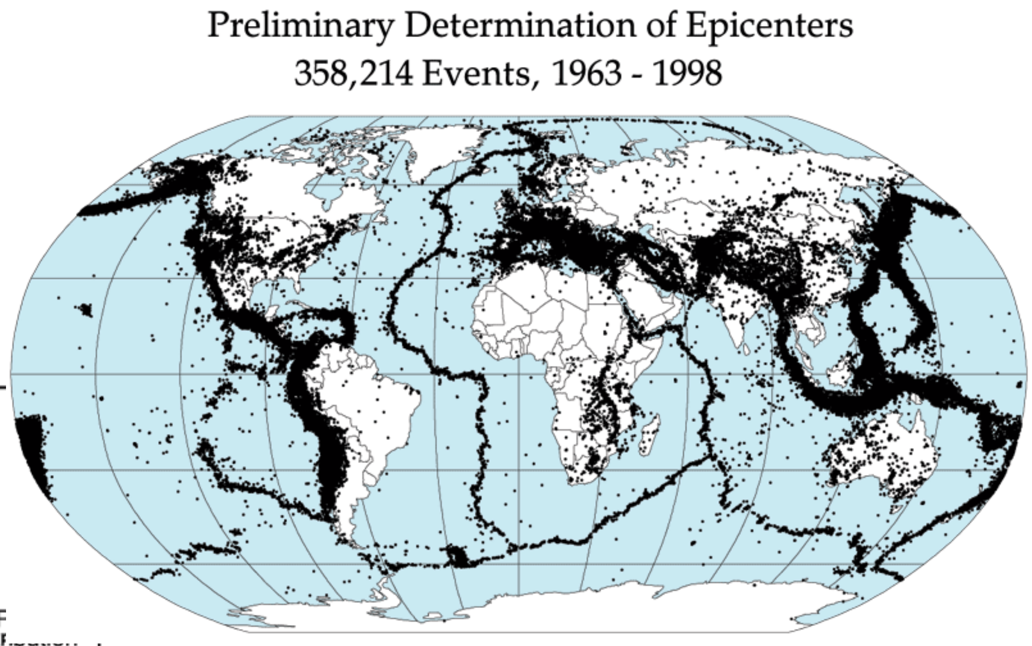
Unserved energy/power magnitude (*S*) distribution

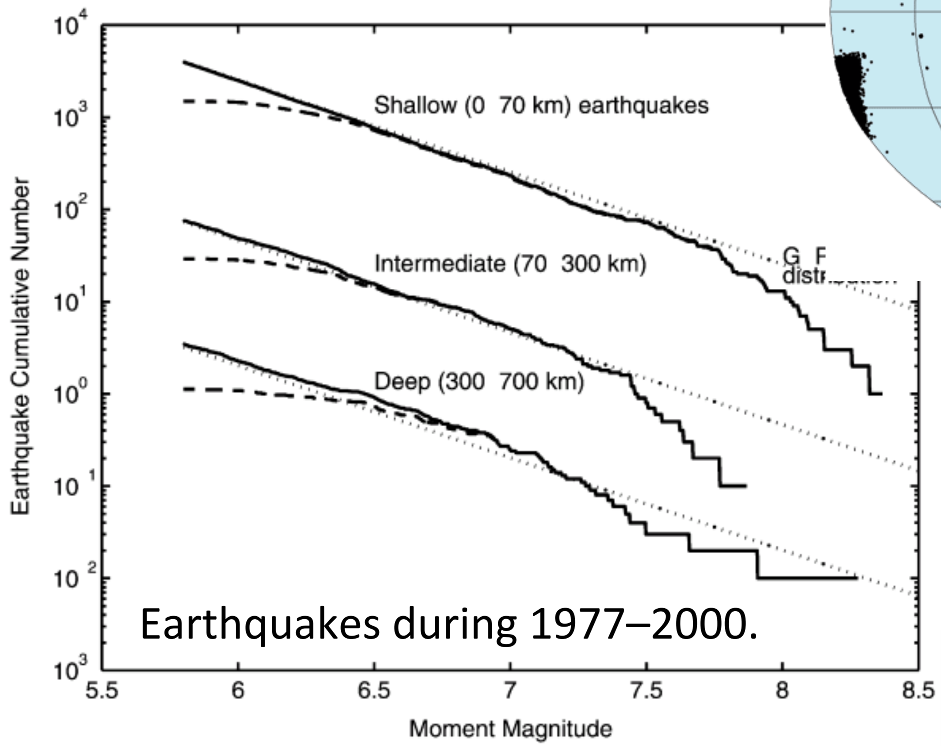
$$P(S) \sim S^{-\alpha},\ 1 < \alpha < 2$$

Source	Exponent	Quantity
North America	2.0	Power
Sweden	1.6	Energy
Norway	1.7	Power
New Zealand	1.6	Energy
China	1.8	Energy

I. Dobson, B. A. Carreras, V. E. Lynch, D. E. Newman, *CHAOS* 17, 026103 (2007)

Cascades Size Distribution of Earthquakes





Earthquakes during 1977–2000.

Earthquake size *S* distribution

$$P(S) \sim S^{-\alpha}, \alpha \approx 1.67$$

Y. Y. Kagan, *Phys. Earth Planet. Inter.* 135 (2–3), 173–209 (2003)

Short Summary of Models: Universality

Models	Networks	Exponents
Failure Prorogation Model	ER	1.5
Overload Model	Complete Graph	1.5
BTW Sandpile Model	ER/SF	1.5 (ER) $\gamma/(\gamma - 1)$(SF)
Branching Process Model	ER/SF	1.5 (ER) $\gamma/(\gamma - 1)$(SF)

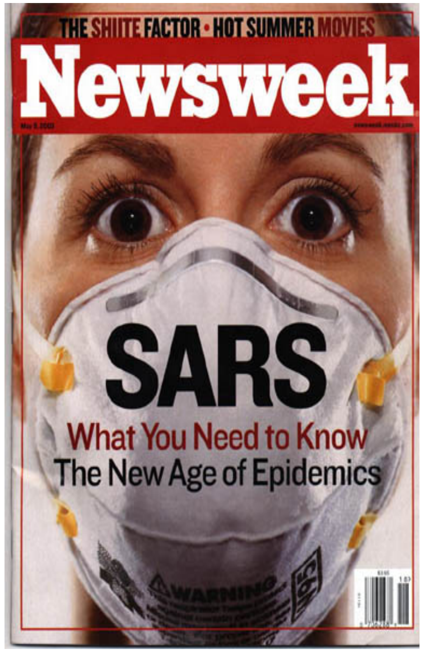
Universal for homogenous networks

$$P(S) \sim S^{-3/2}$$

Same exponent for percolation too
(random failure, attacking, etc.)

Epidemics and spreading

Why is the spreading process important?



"Epidemic"

Epi + demos

upon *people*

Biological:

Airborne diseases (flu, SARS, …)

- Venereal diseases (HIV, …)
- Other infectious diseases including some cancers (HPV, …)
- Parasites (bedbugs, malaria, …)

Digital:

- Computer viruses, worms
- Mobile phone viruses

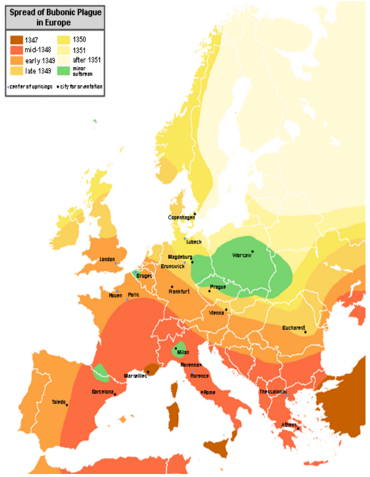
Conceptual/Intellectual:

- Diffusion of innovations
- Rumors
- Memes
- Business practices

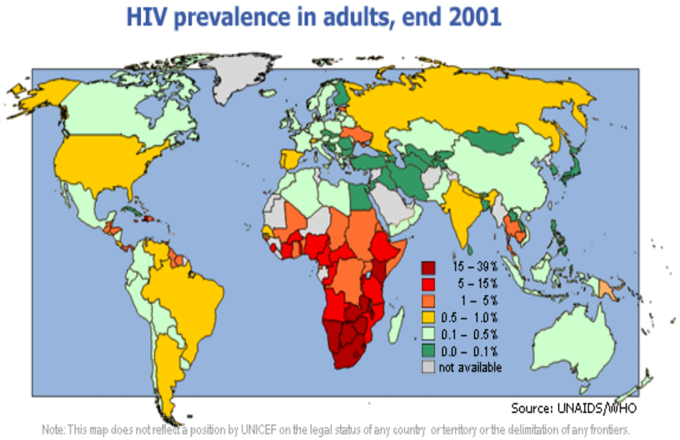
http://en.wikipedia.org/wiki/Epidemic

Biological: Notable Epidemic Outbreaks

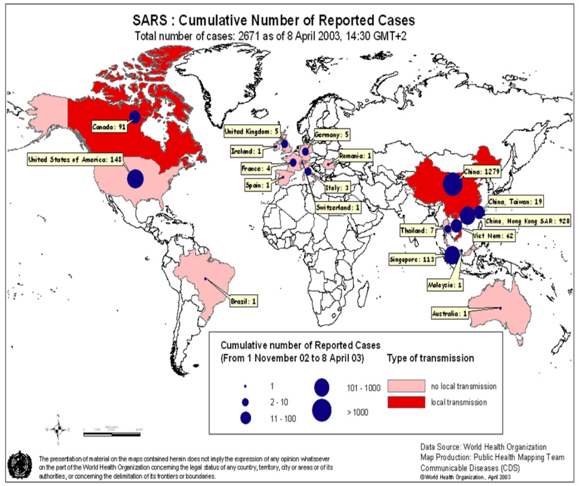
The Great Plague



HIV



SARS



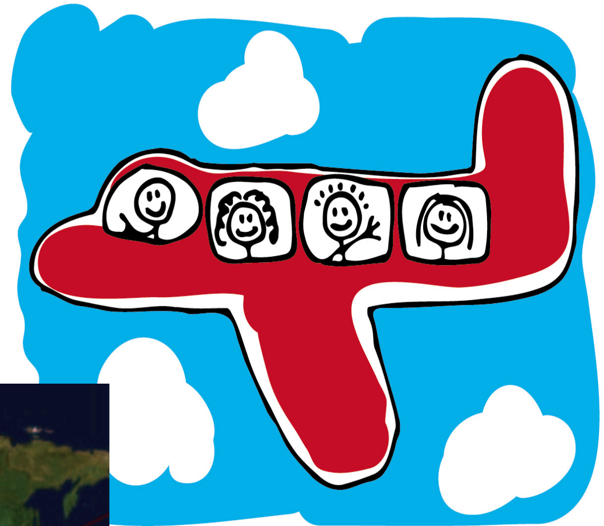
1918 Spanish flu

H1N1 flu

Epidemic spreading – Why does it matter now?

High population density

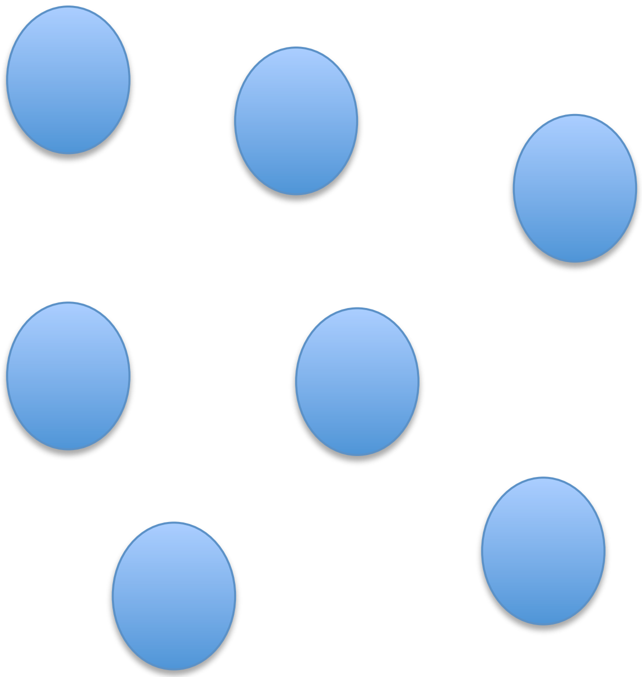
High mobility

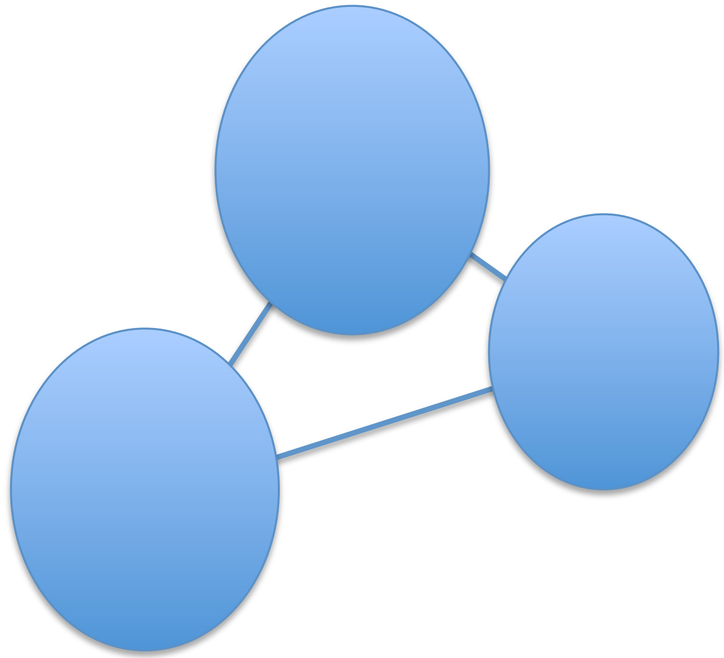


→ perfect conditions for epidemic spreading.

Airline figure: L. Hufnagel et al. *PNAS* **101,** 15124 (2004)

Large population can provide the "fuel"



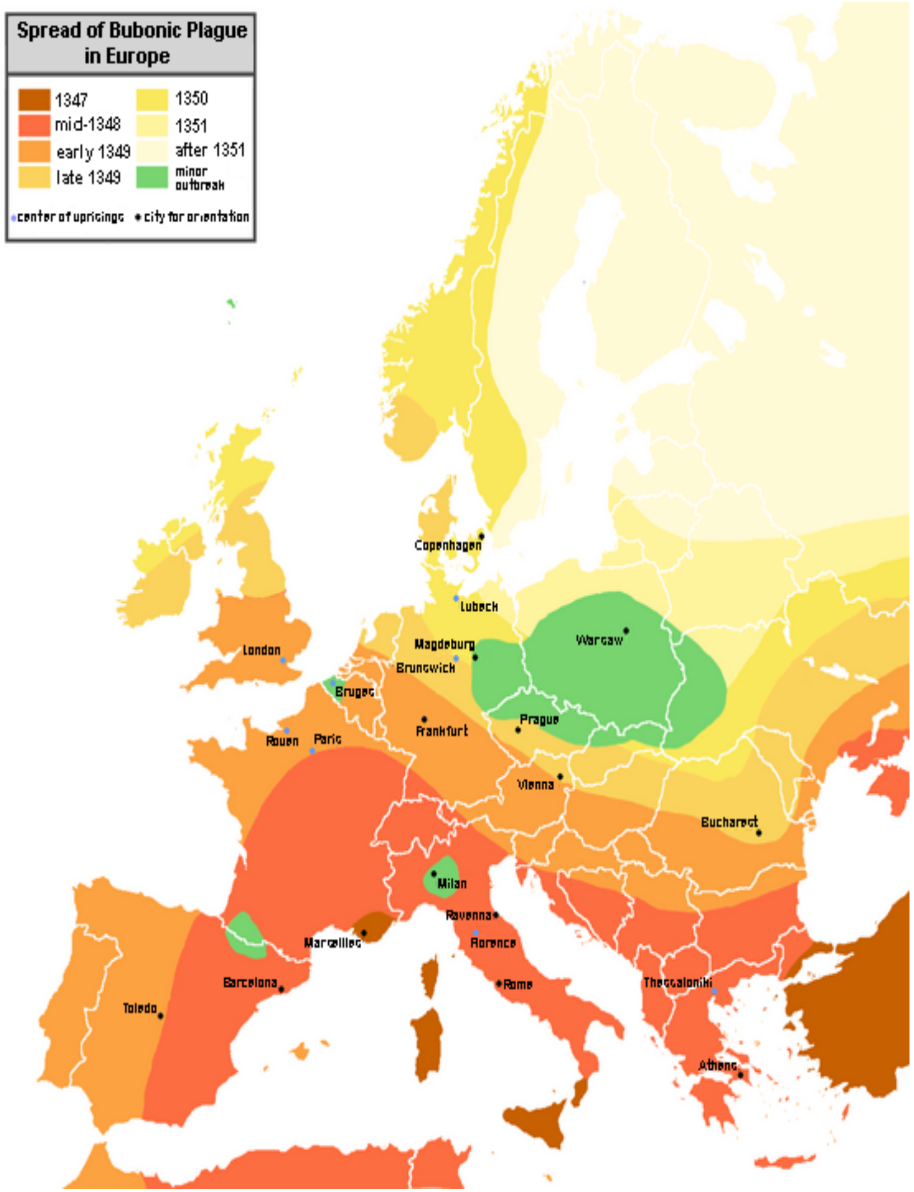


Separate, small population
(hunter-gatherer society, wild animals)

Connected, highly populated areas
(cities)

Human societies have "**crowd diseases**", which are the consequences of large, interconnected populations (Measles, tuberculosis, smallpox, influenza, common cold, …)

14th Century – The Great Plague

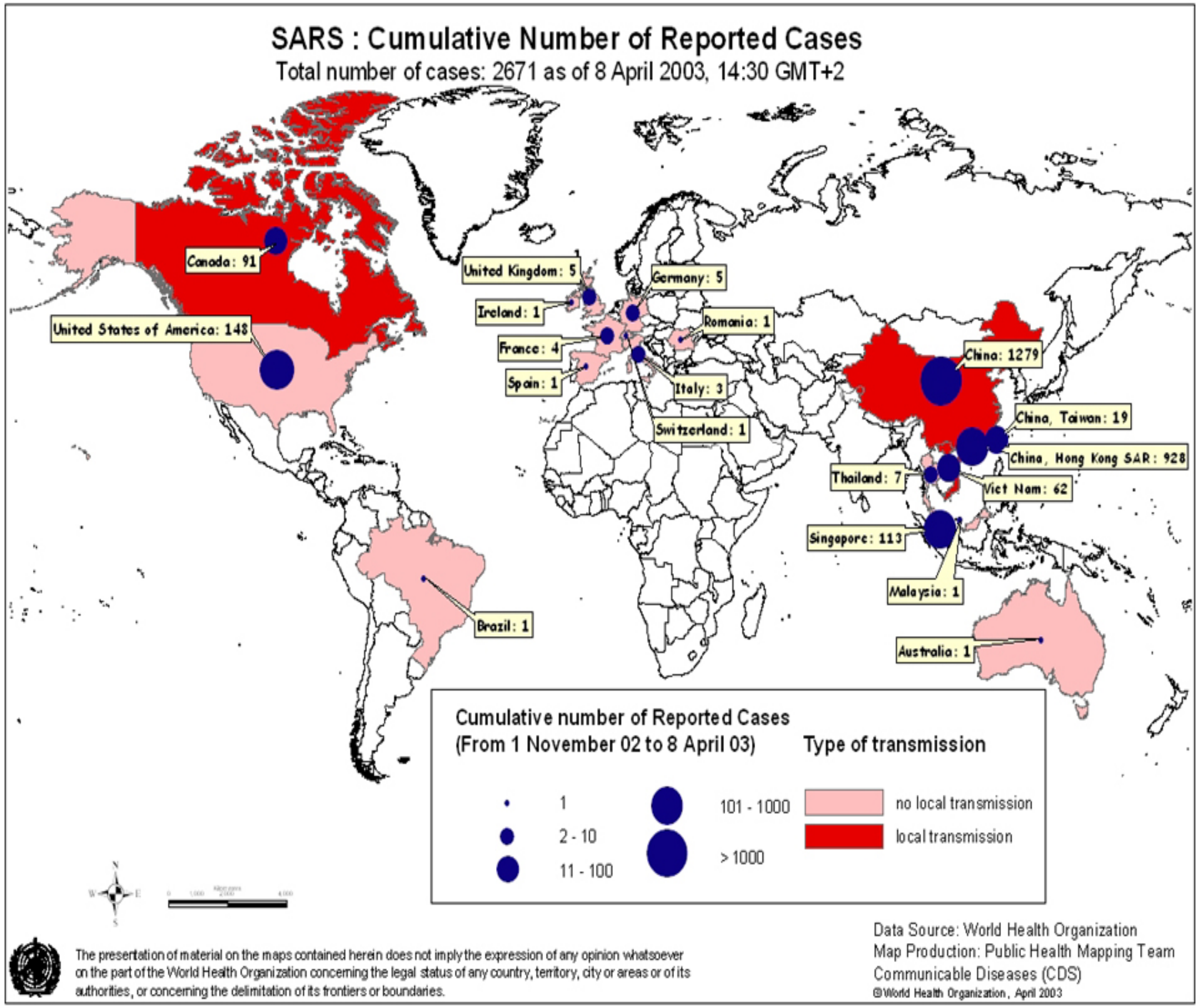


4 years from France to Sweden

Limited by the speed of human travel

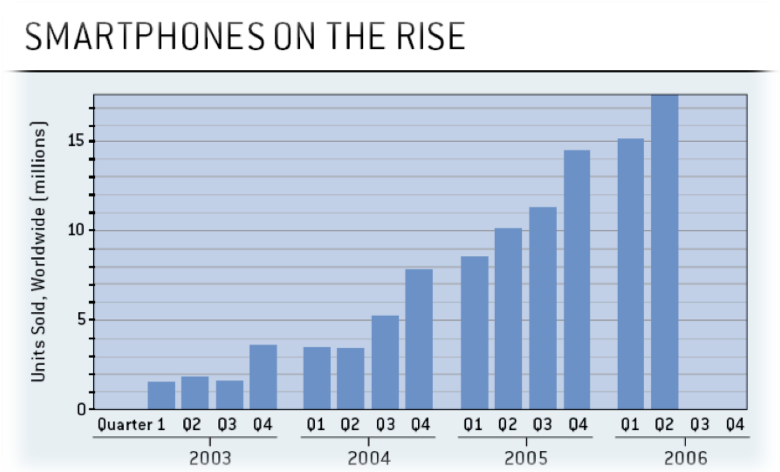
http://en.wikipedia.org/wiki/Black_Death
http://de.wikipedia.org/wiki/Schwarzer_Tod

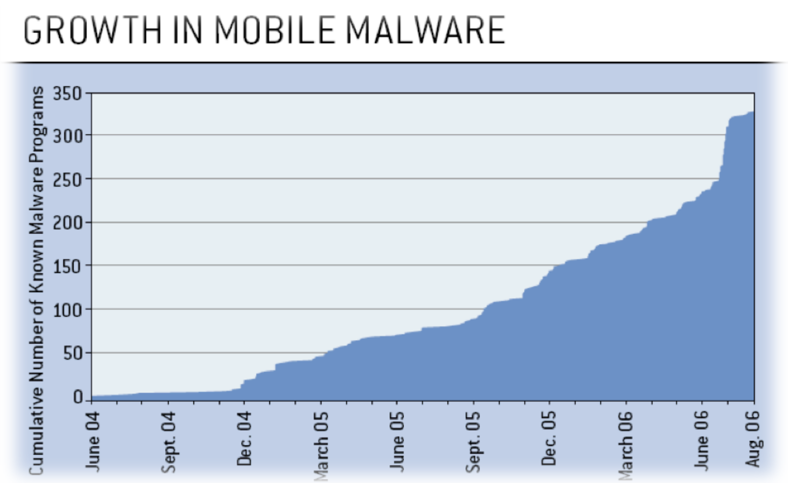
21st Century – SARS



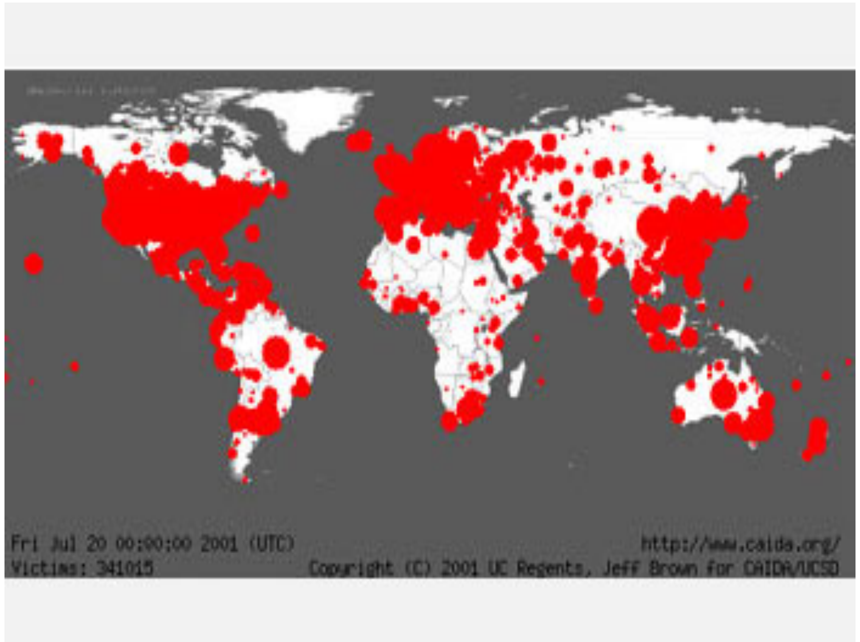
Source: World Health Organization

Computer Viruses, Worms, Mobile Phone Viruses





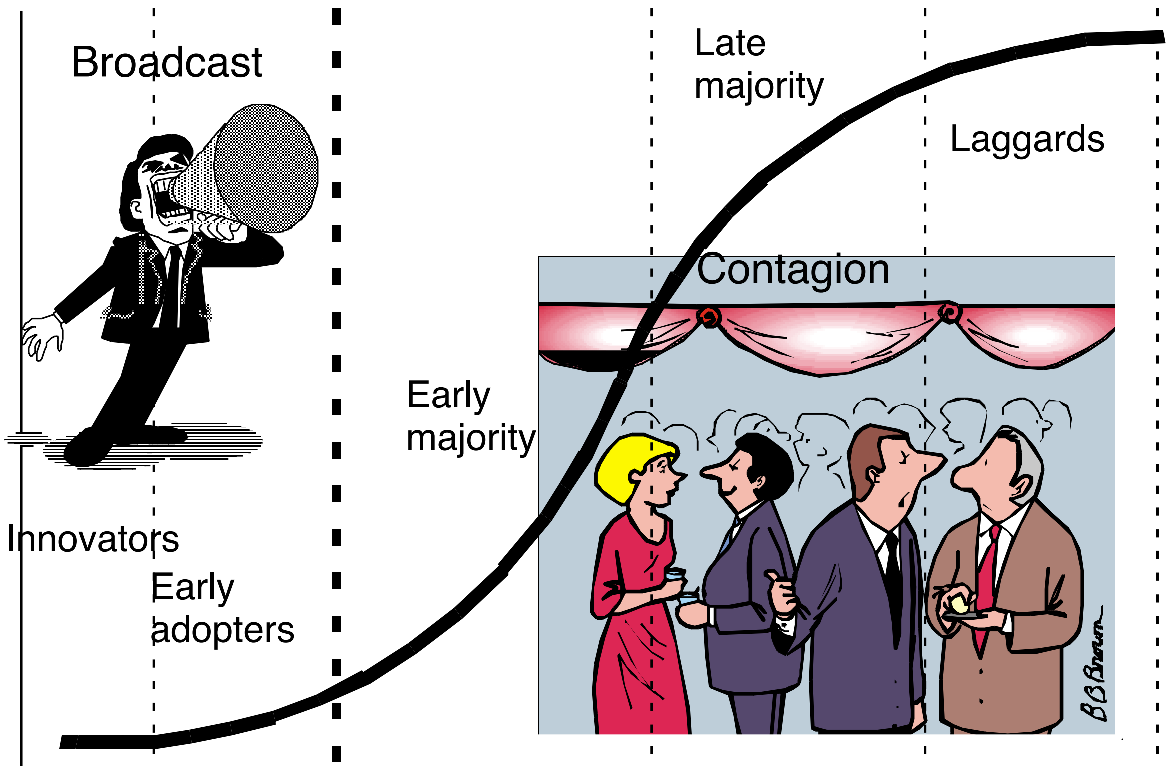
Code Red Worm paralyzed many countries' Internet



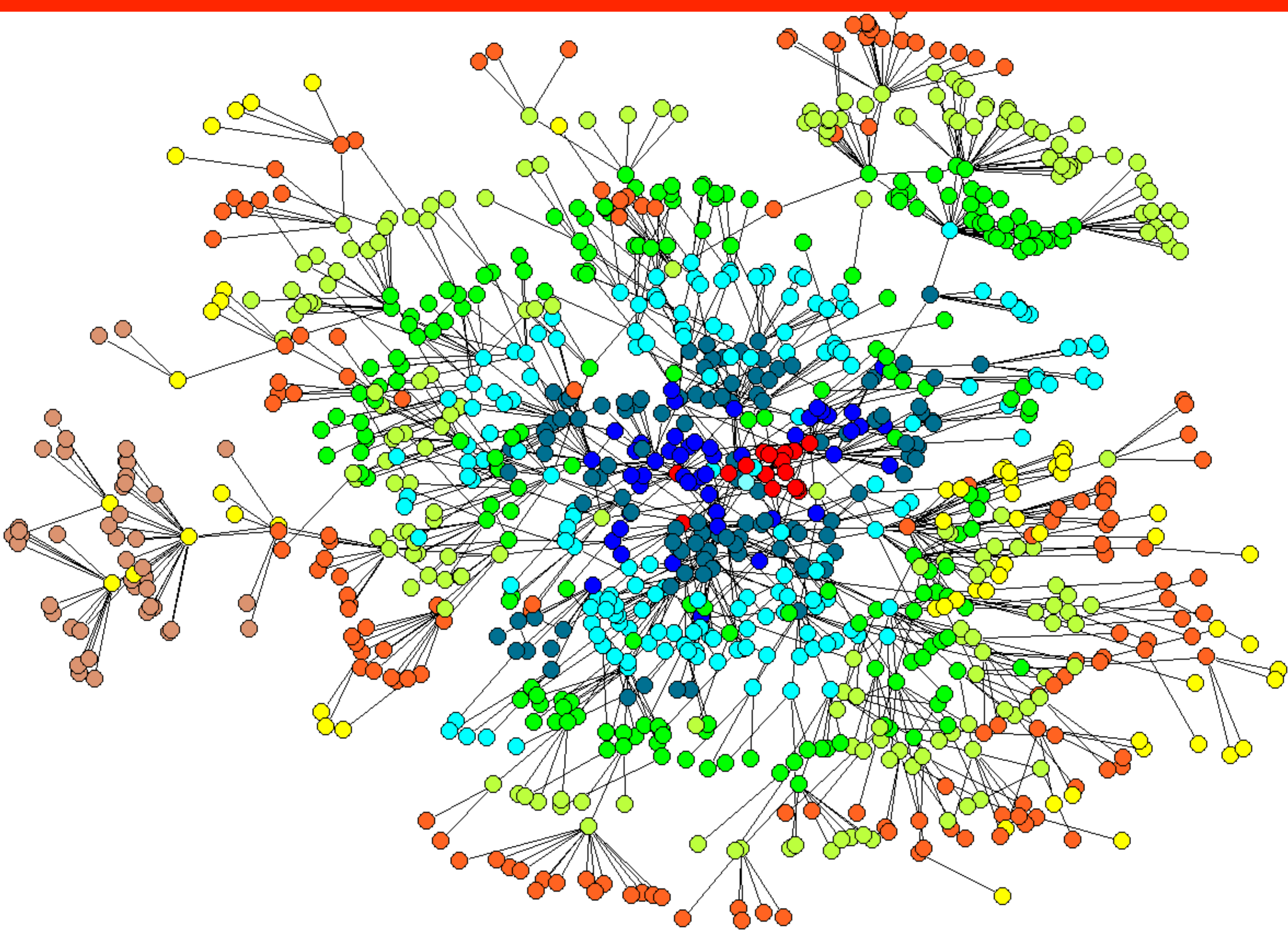
http://www.caida.org/publications/visualizations/

Hypponen M. *Scientific American* Nov. 70-77 (2006).

Diffusion of Innovation – The Adoption Curve

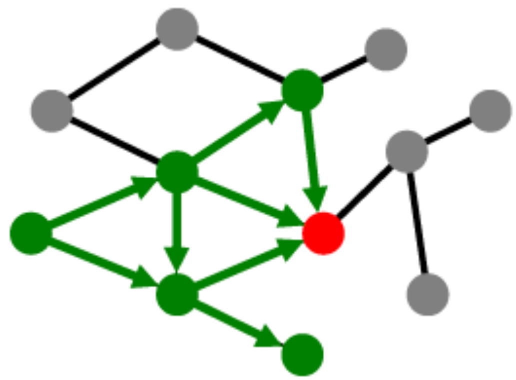


Information Spreading



How to model diffusion?

- **Probabilistic models:**
  - Models of influence or disease spreading
    - An infected node tries to "push" the contagion to an uninfected node
  - **Example:**
    - You "catch" a disease with some prob. from each active neighbor in the network

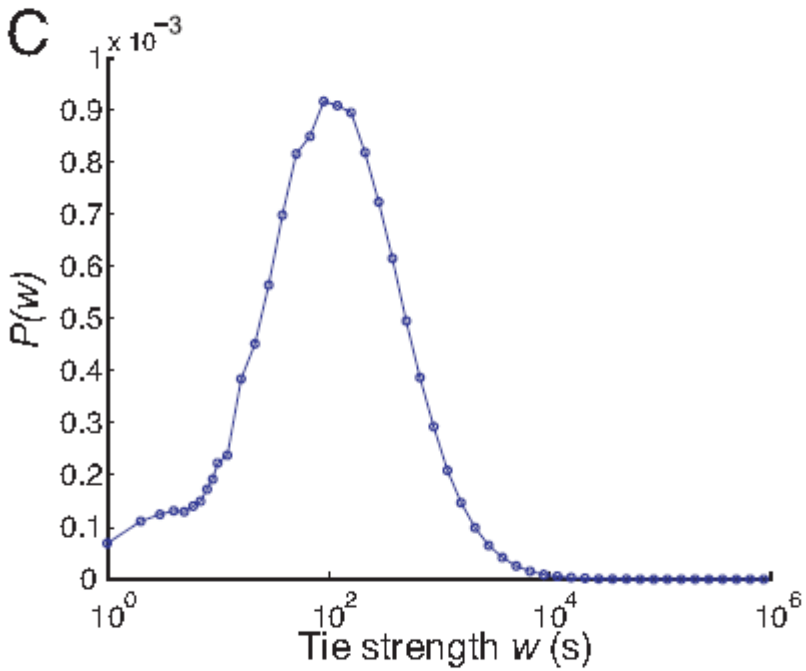


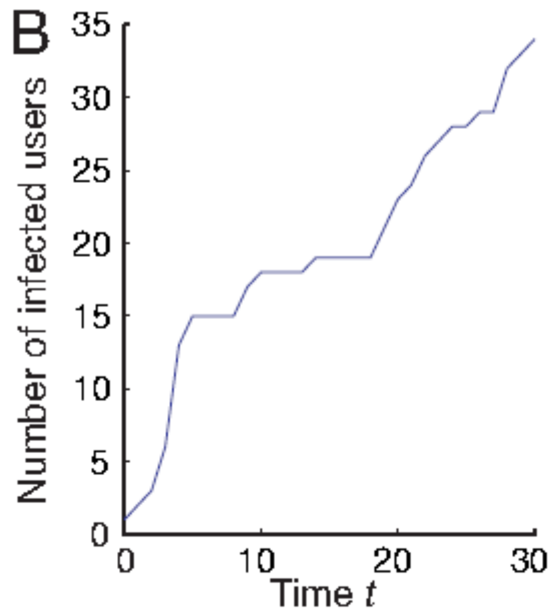
- **Decision based models:**
  - Models of product adoption, decision making
    - A node observes decisions of its neighbors and makes its own decision
  - **Example:**
    - You join demonstrations if *k* of your friends do so too

Empirical studies of cascading behavior

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**

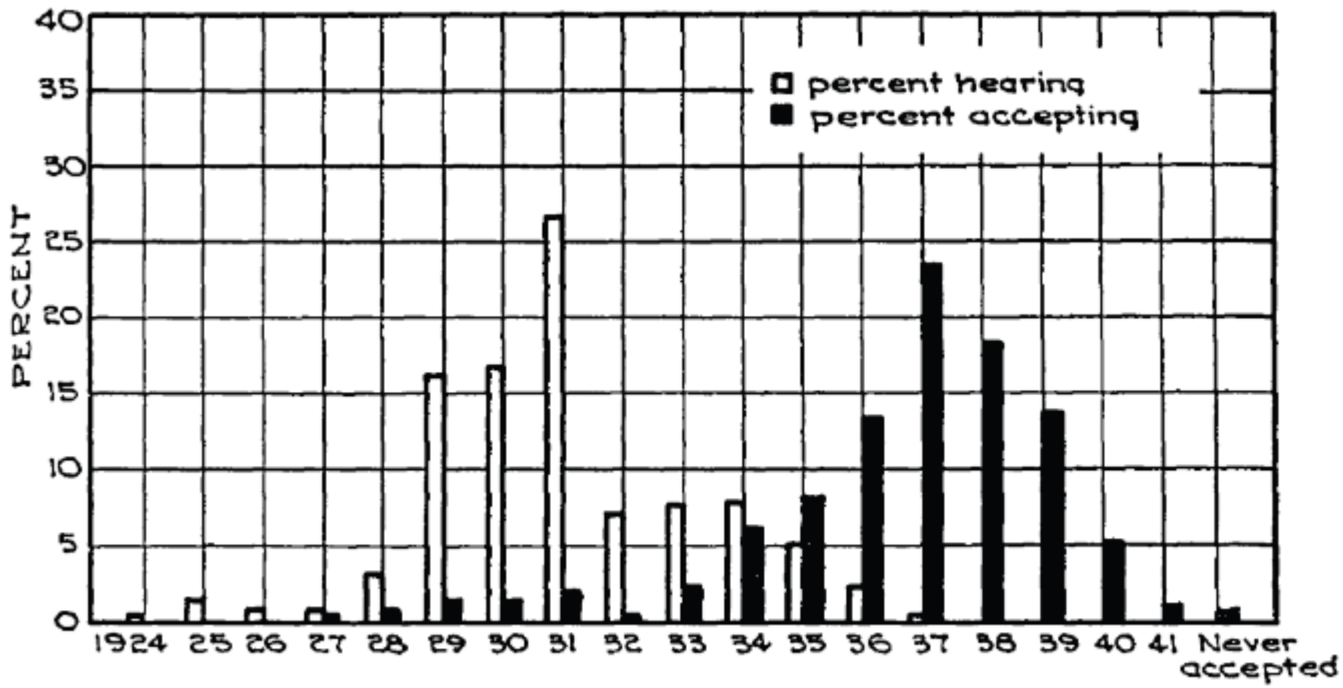
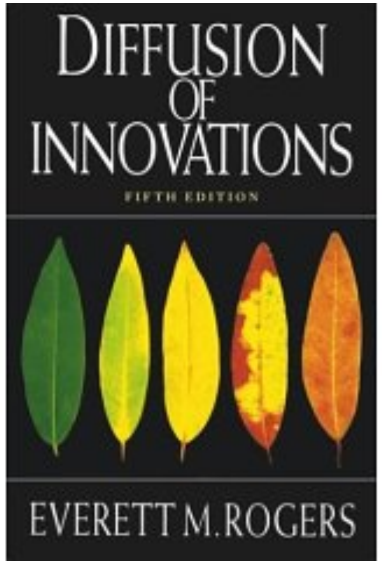
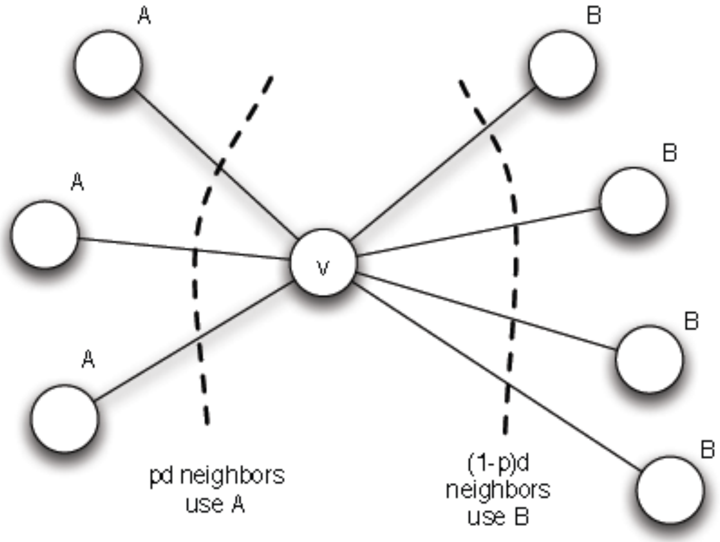
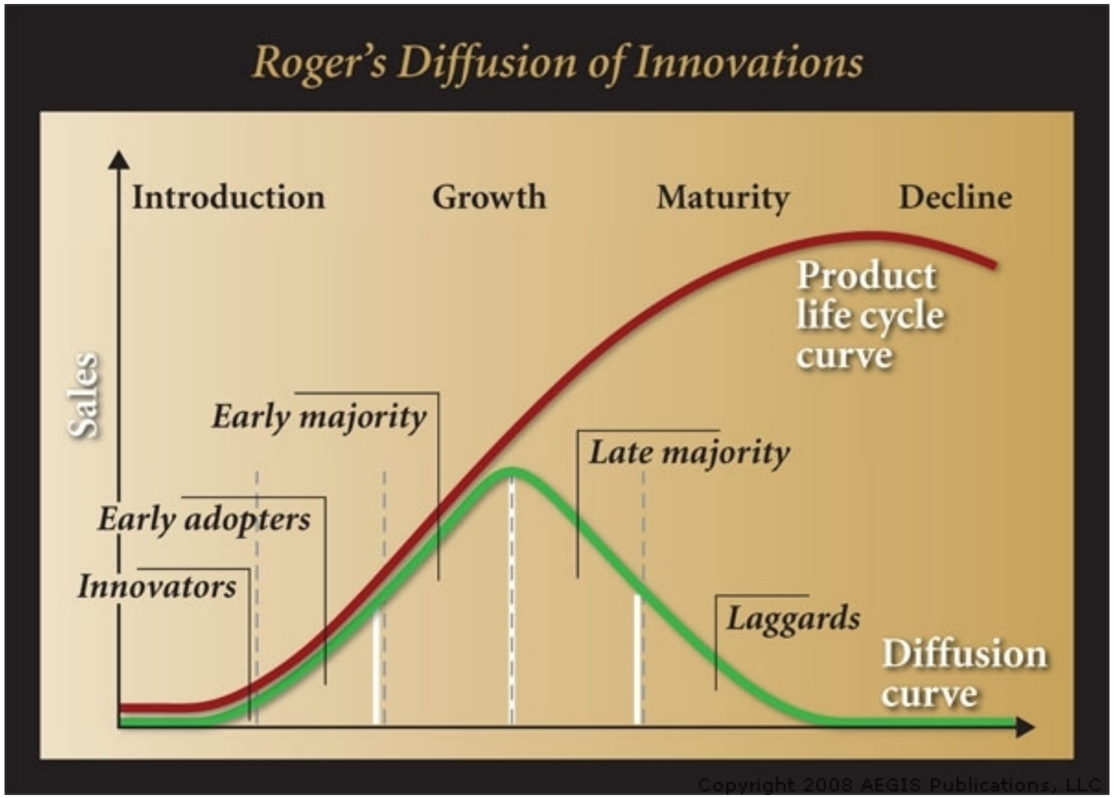


Figure 19.10: The years of first awareness and first adoption for hybrid seed corn in the Ryan-Gross study. (Image from [358].)

The strength of the strong ties for the



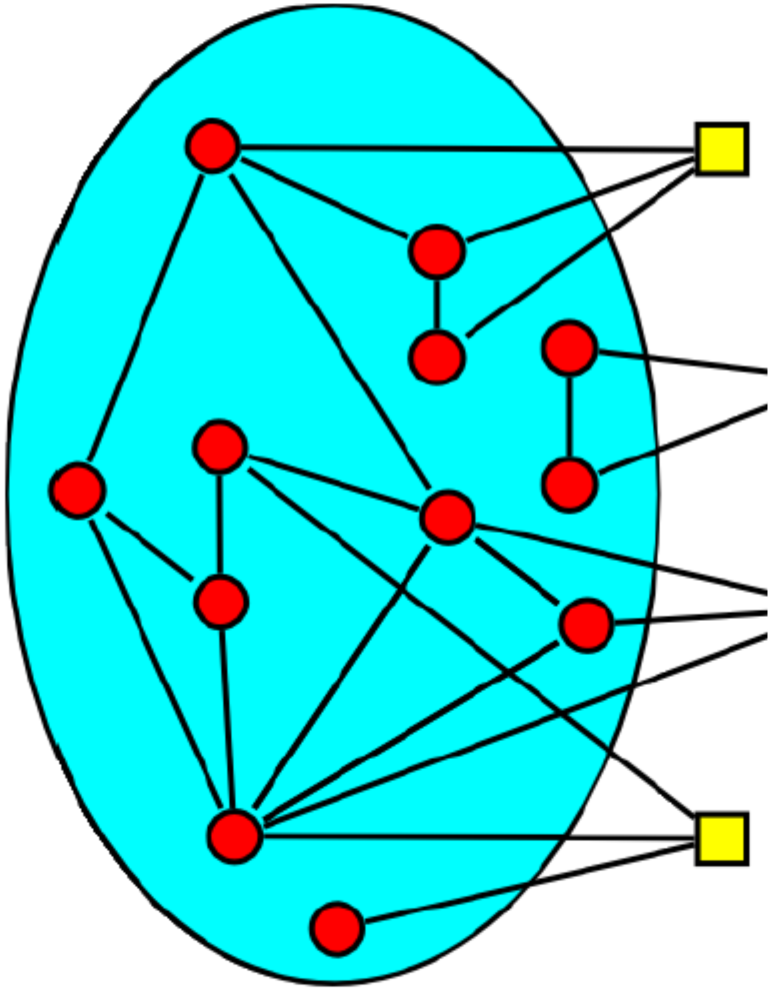




[Backstrom et al. KD

Adoption Curve: LiveJournal

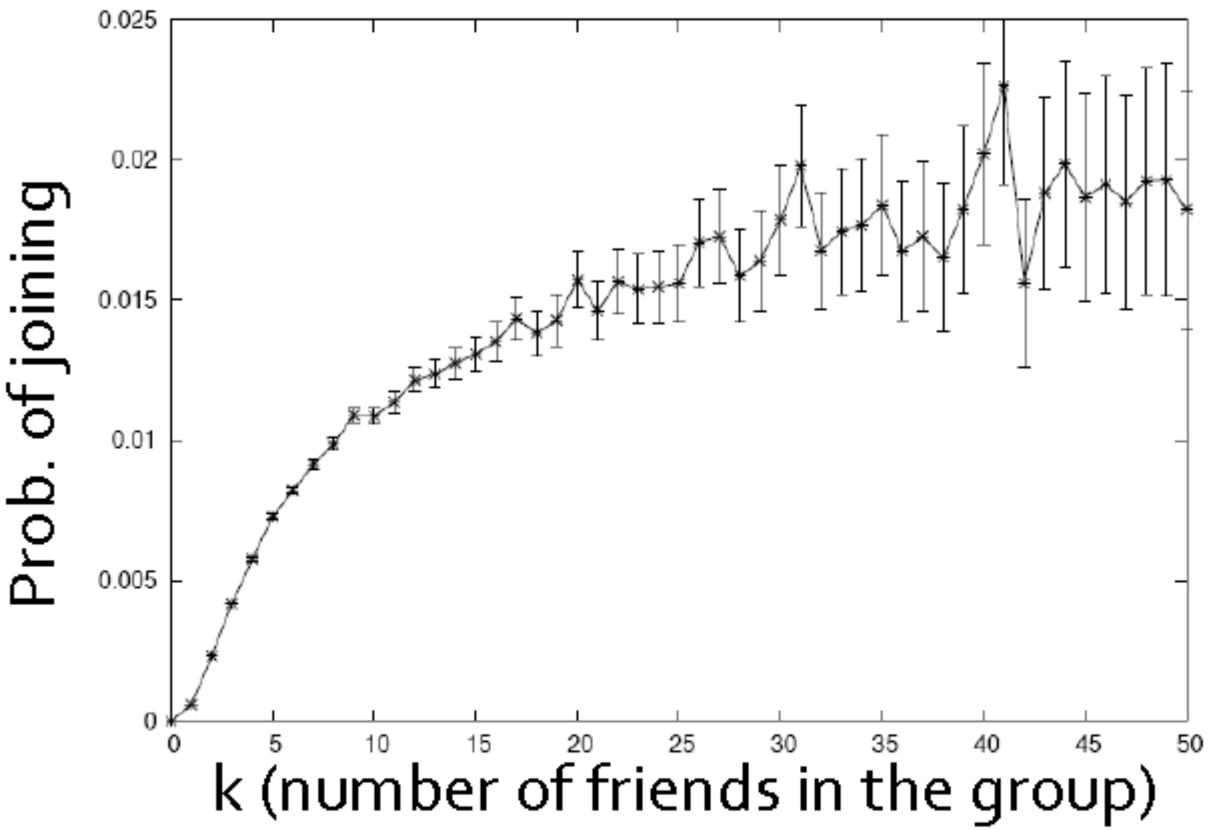
- **Group memberships spread over the network:**
  - Red circles represent existing group members
  - Yellow squares may join
- **Question:**
  - How does prob. of joining a group depend on the number of friends already in the group?



[Backstrom et al., KDD 'c

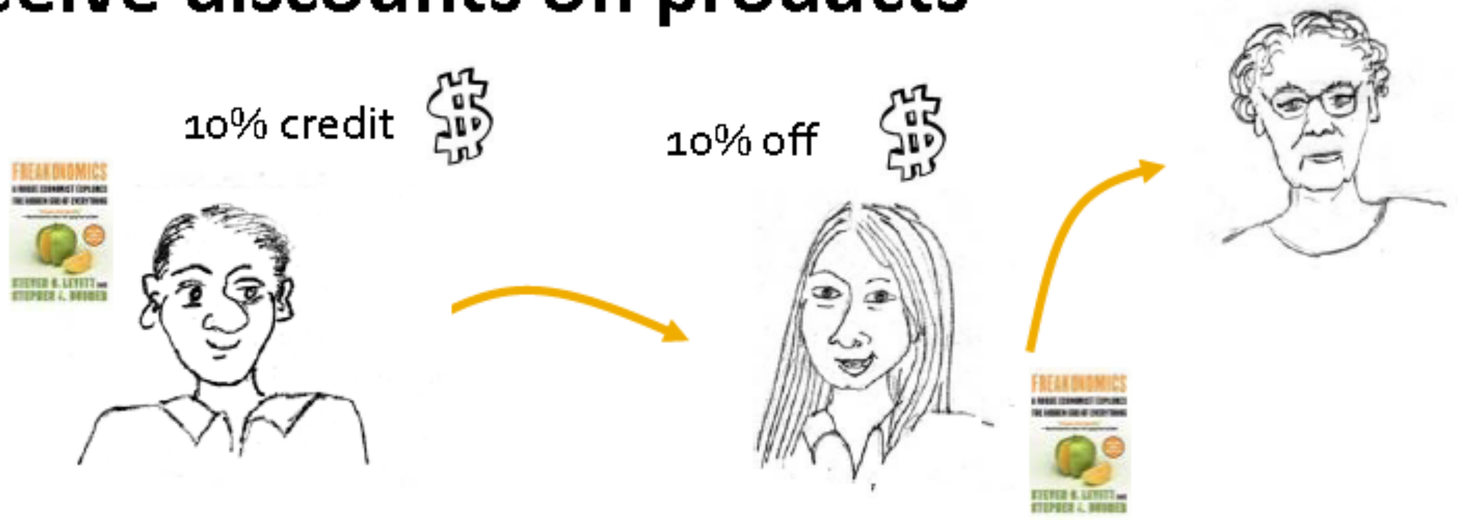
Adoption Curve: LiveJournal

- LiveJournal group membership



Diffusion in Viral Marketing

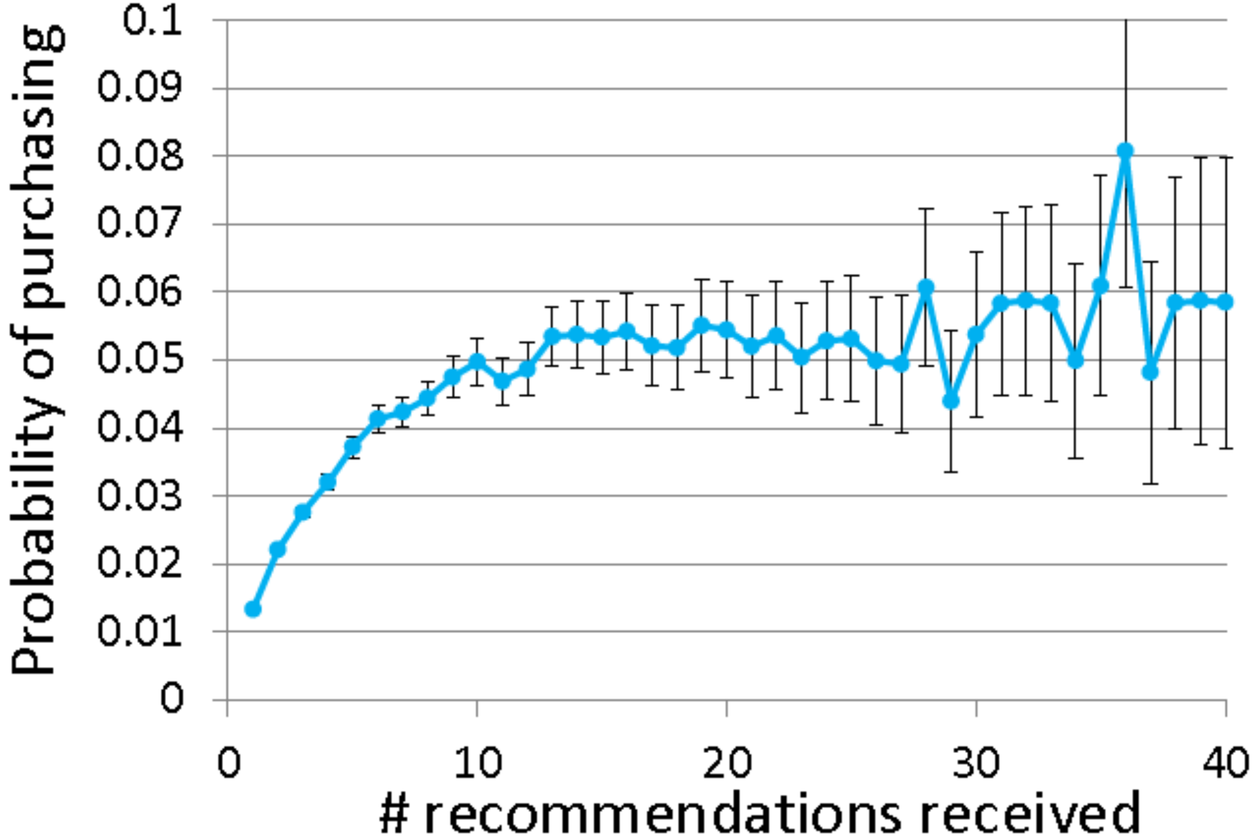
- **Senders and followers of recommendations receive discounts on products**

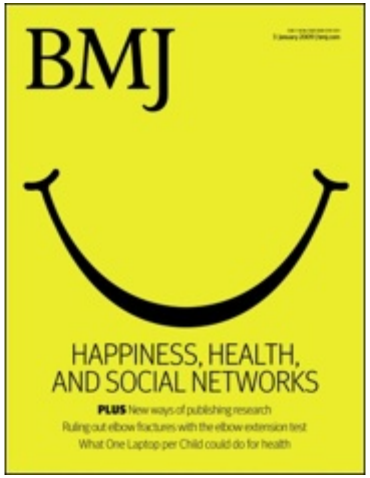


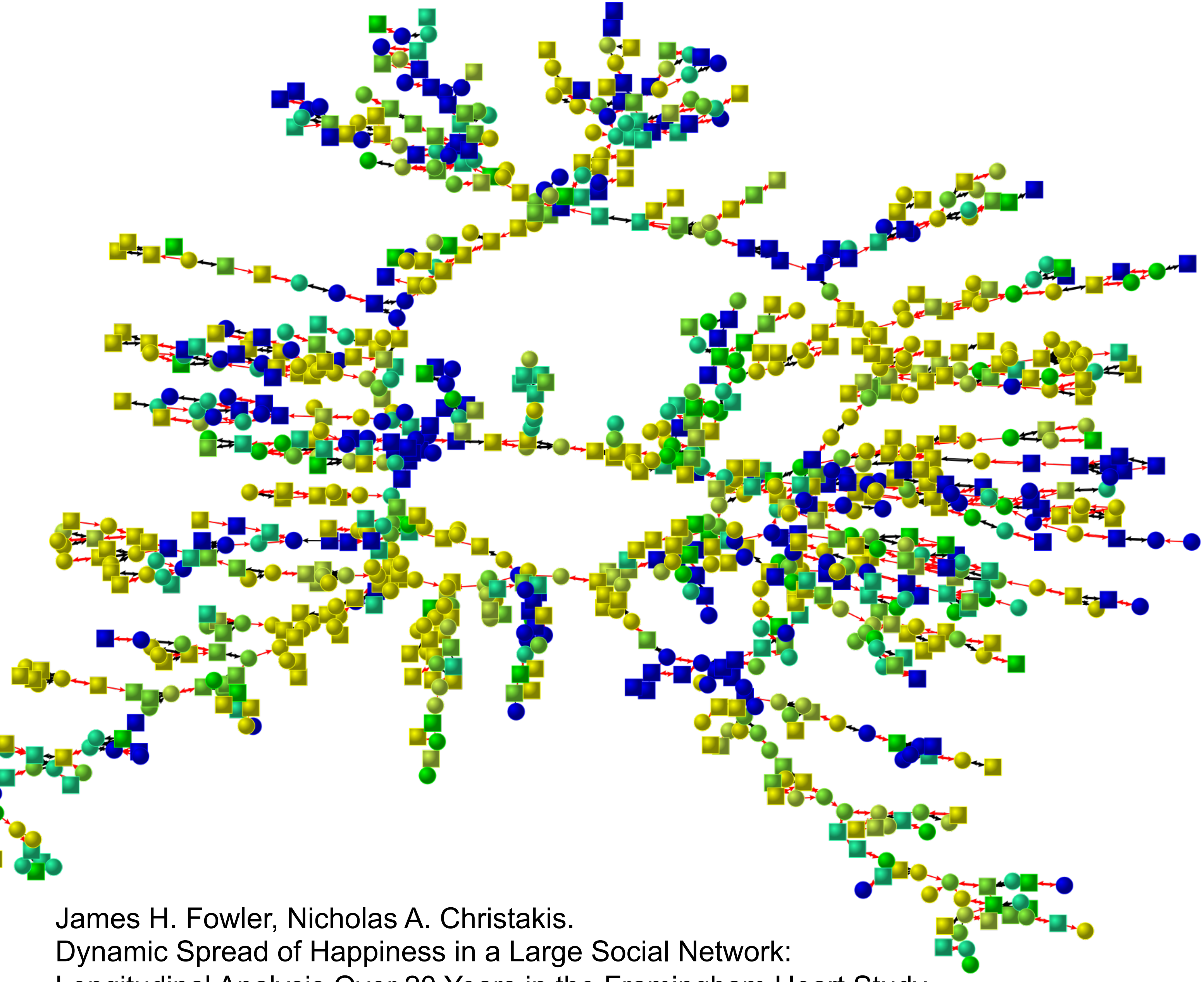
- **Data: Incentivized Viral Marketing program**
  - 16 million recommendations
  - 4 million people, 500k products

[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)

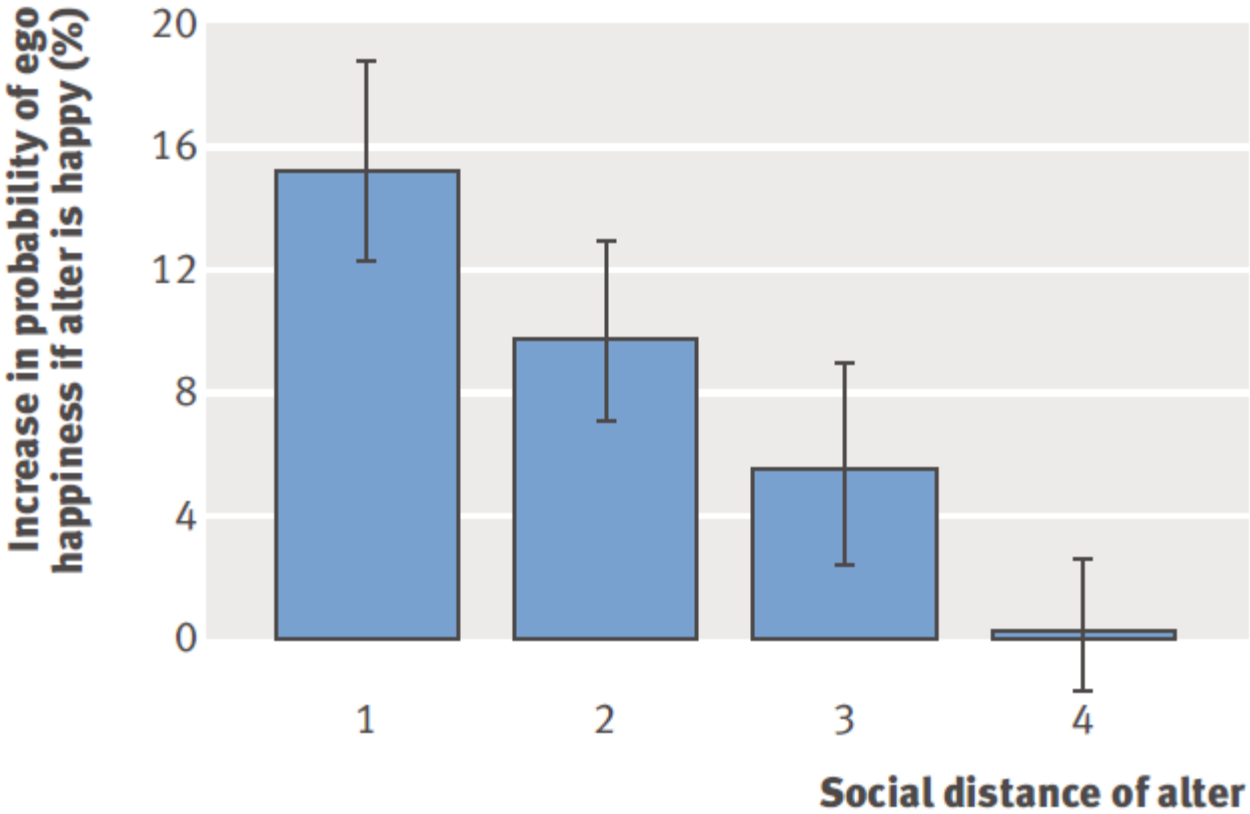
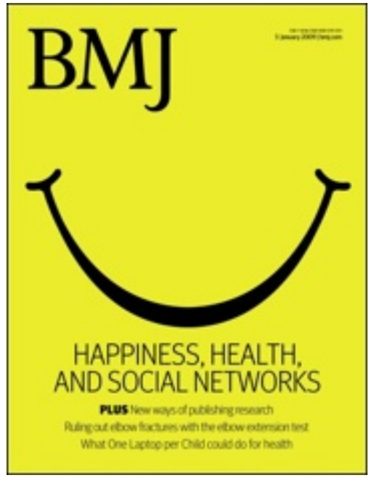


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >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



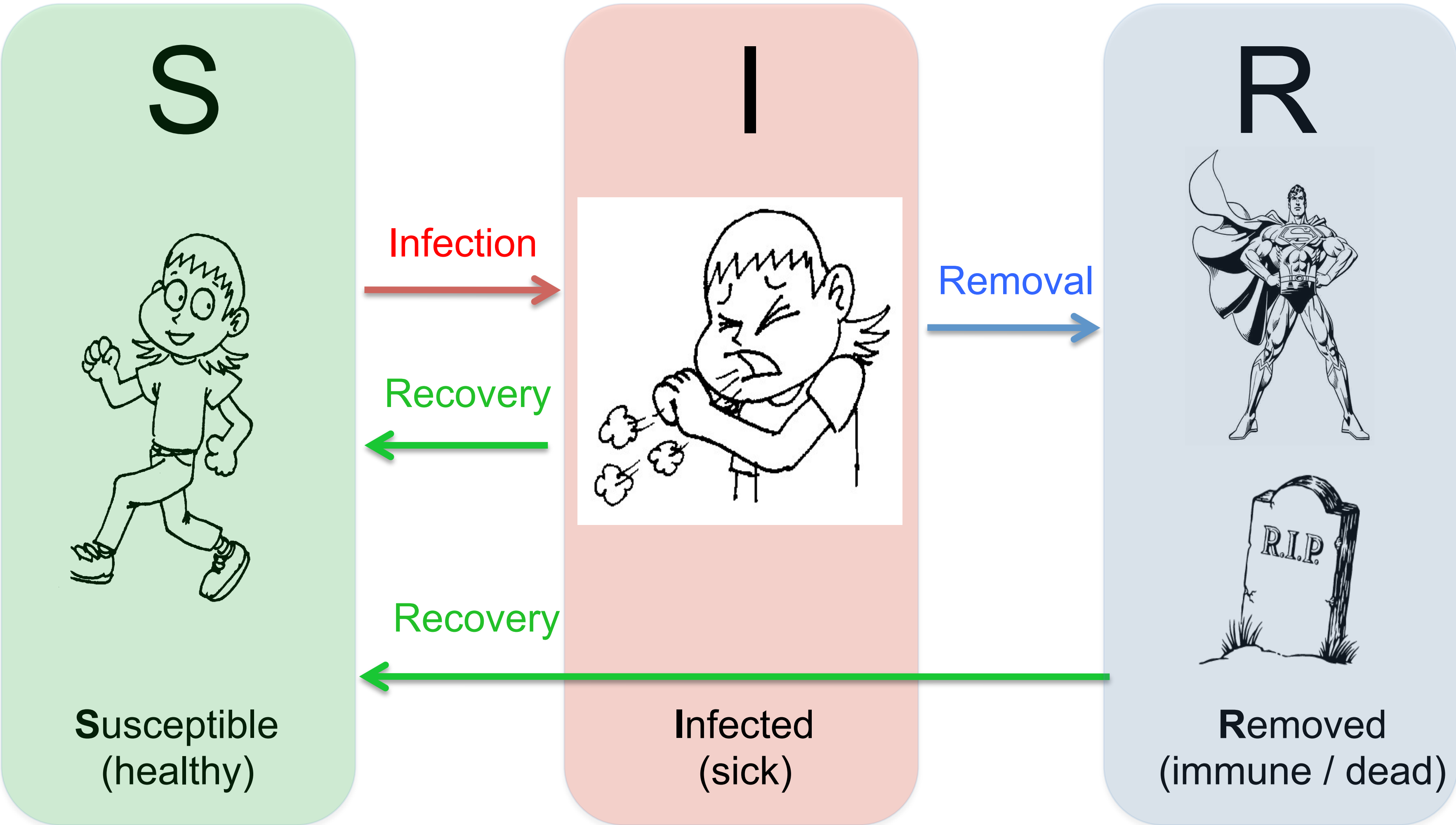
Social influence
or
homophily?

Probabilistic models of diffusion

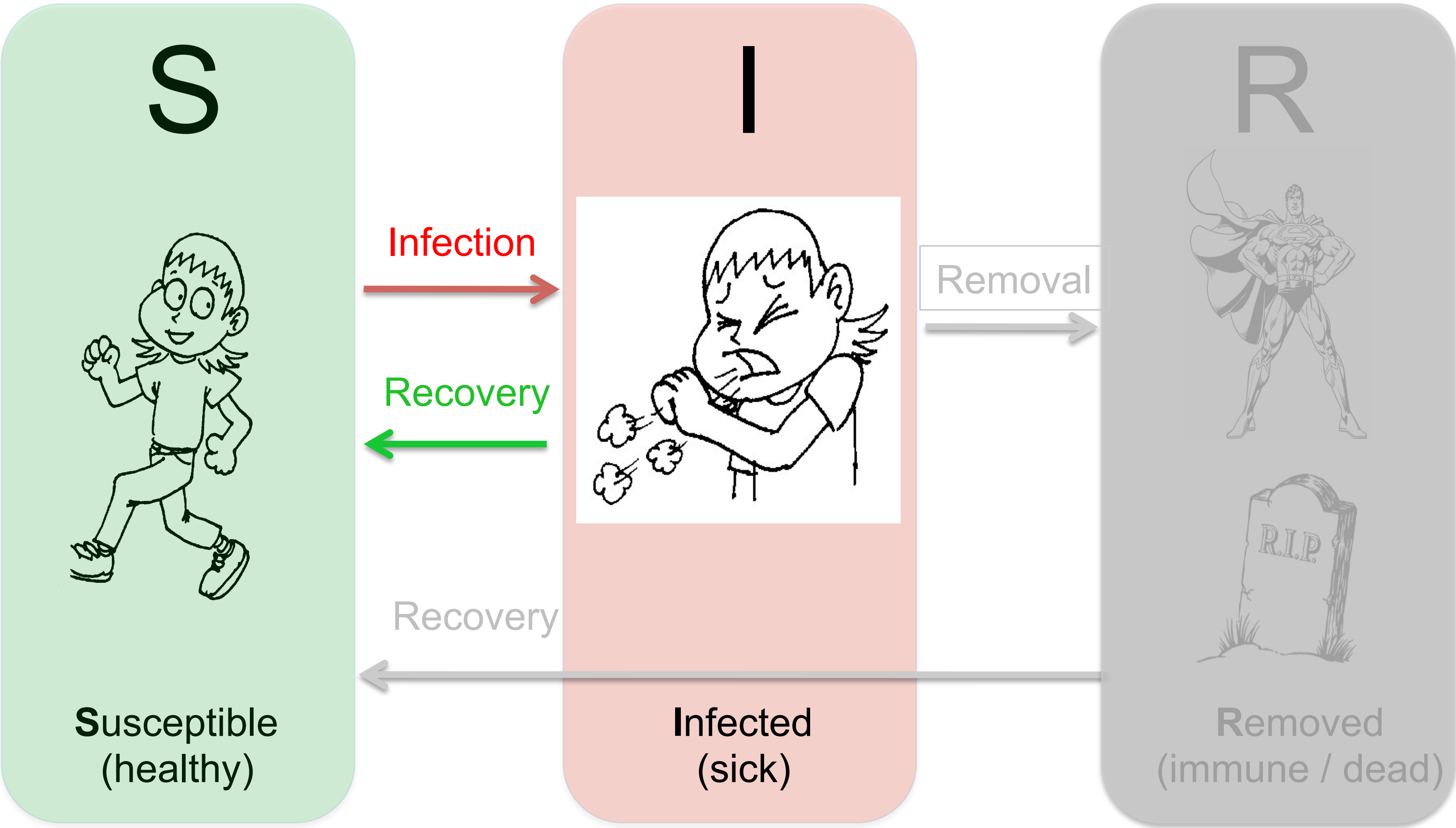
Epidemic modeling

Epidemic Modeling (classical models)

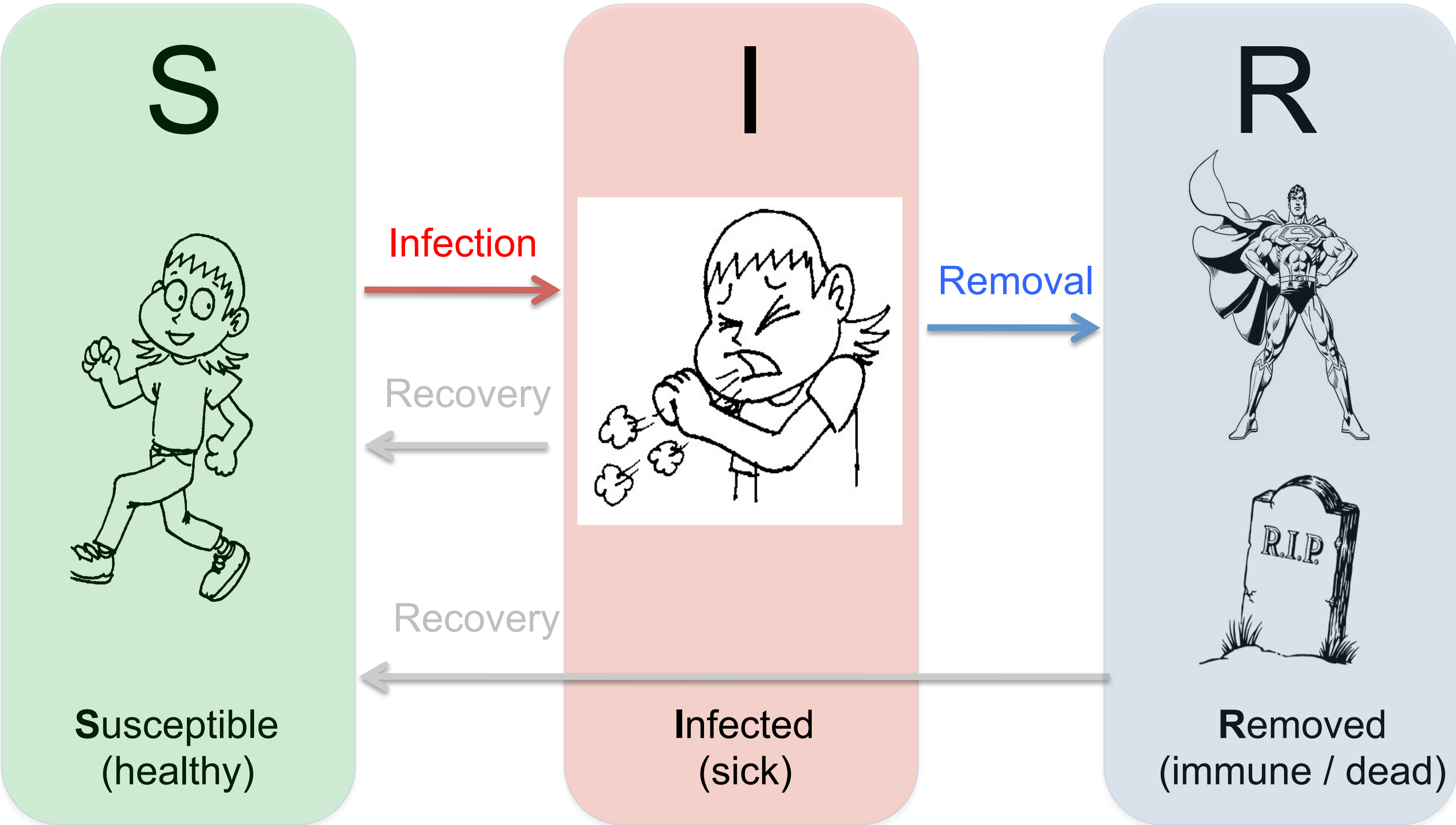
Classical Epidemic Models – Basic States



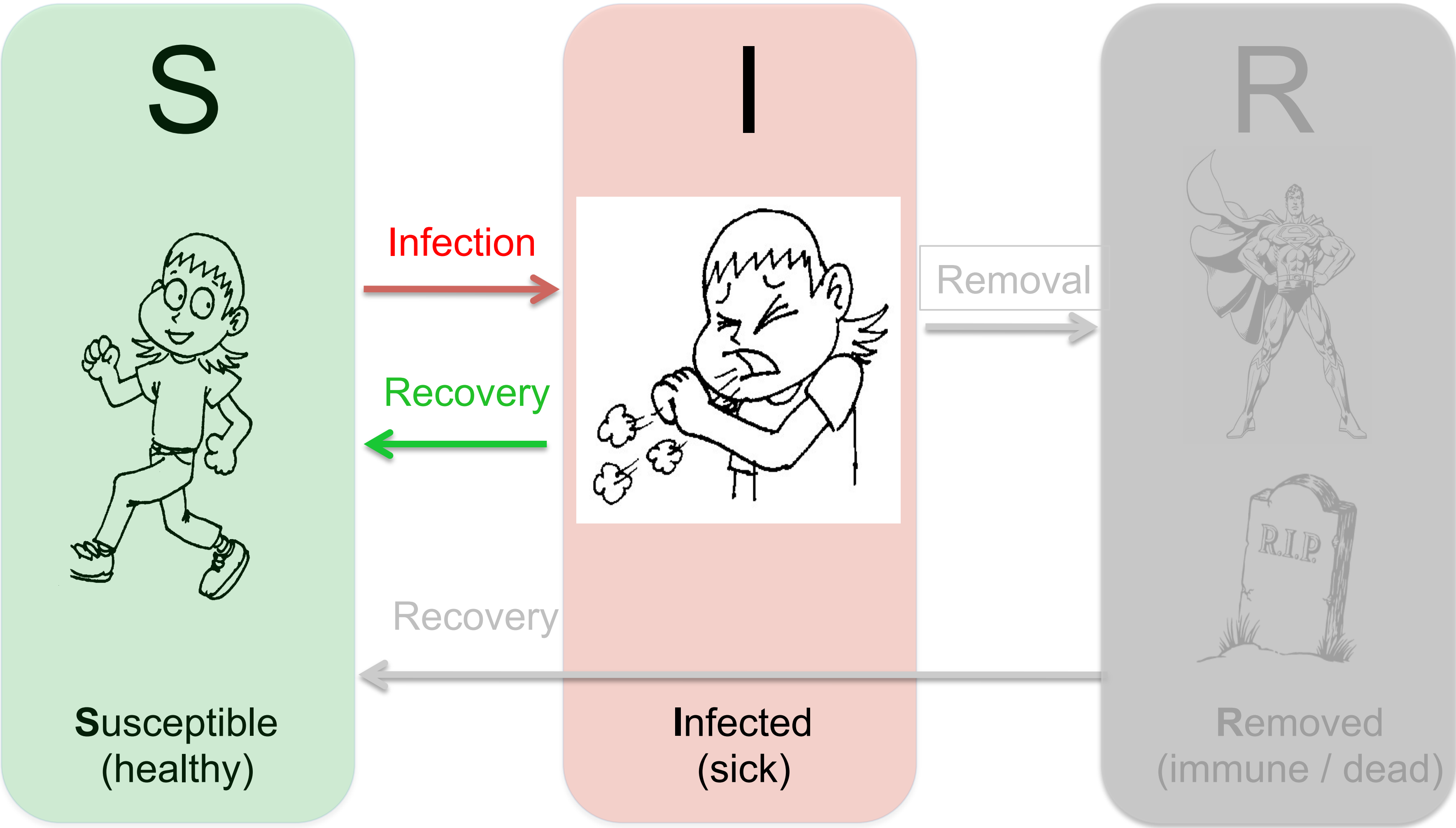
SIS Model: Common Cold



Example 2: Flu, SARS, Plague, …



SIS Model: Common Cold



SIS Model Dynamics

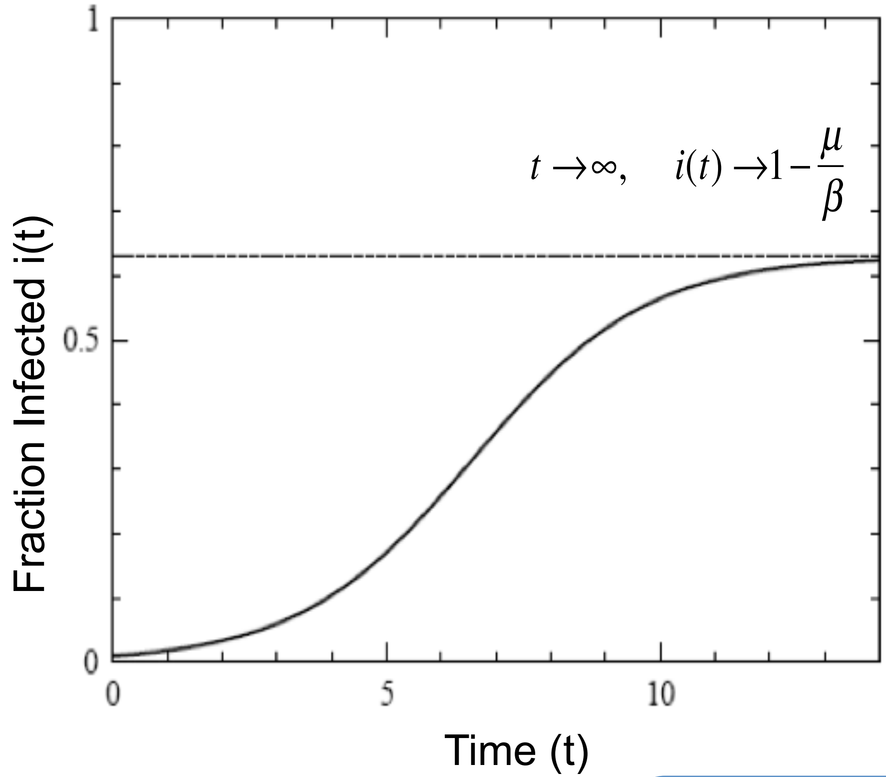
$$\frac{di}{dt} = \underbrace{\beta i}_{\text{I}}\underbrace{(1-i)}_{\text{S}} - \underbrace{\mu i}_{\text{I}\to\text{S}} = i(\beta - \mu - \beta i)$$

$$\frac{di}{i} + \frac{di}{1-\mu/\beta - i} = (\beta - \mu)dt$$

$$\ln(i) - \ln(1-\mu/\beta - i) = (\beta - \mu)t + c$$

$$\frac{i}{1-\mu/\beta - i} = Ce^{(\beta-\mu)t} \qquad C = e^{c}$$

$$\therefore i(t) = \left(1 - \frac{\mu}{\beta}\right)\frac{Ce^{(\beta-\mu)t}}{1+Ce^{(\beta-\mu)t}}$$



Stationary state:

$$\frac{di}{dt} = \beta i(1-i) - \mu i = 0$$

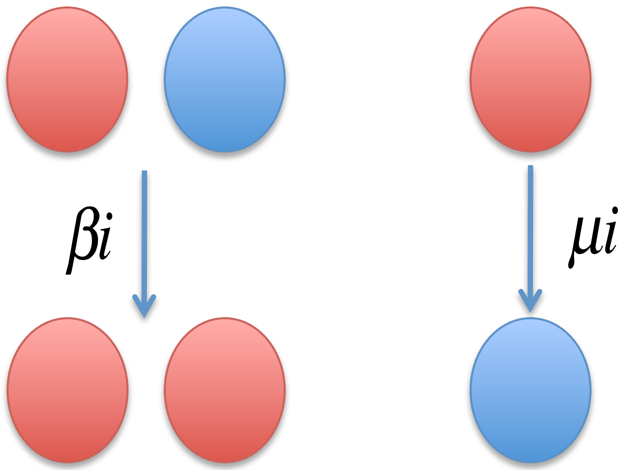
SIS model: fraction infected individuals saturates below 1.

SIS Model: Epidemic Threshold and Basic Reproductive Number

$$\frac{di}{dt} = \beta \underbrace{i}_{\text{I}} \underbrace{(1-i)}_{\text{S}} - \underbrace{\mu i}_{\text{I}\rightarrow\text{S}}$$

If $\mu \approx \beta, \quad i \to 0$

"Epidemic threshold"



$I = I + 1$ $\qquad$ $I = I - 1$

$$\lambda \equiv \frac{\beta}{\mu}$$

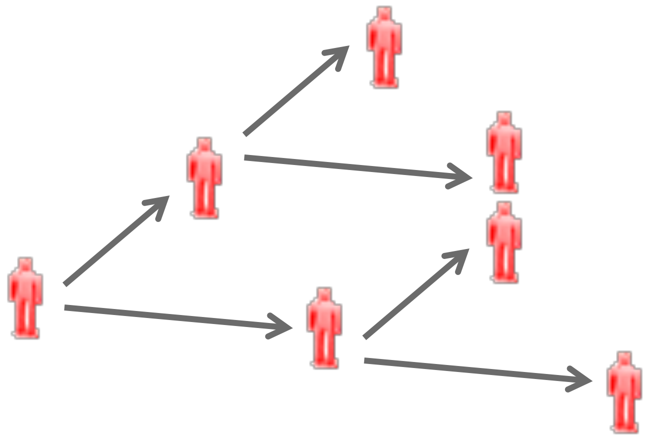
"Basic reproductive number"

On average, how many infected individuals will be infected by one infected individual?

$\lambda > 1$: Outbreak, $\lambda < 1$: Die out

reproductive number λ: average # of infectious individuals generated by one infected in a fully susceptible population.

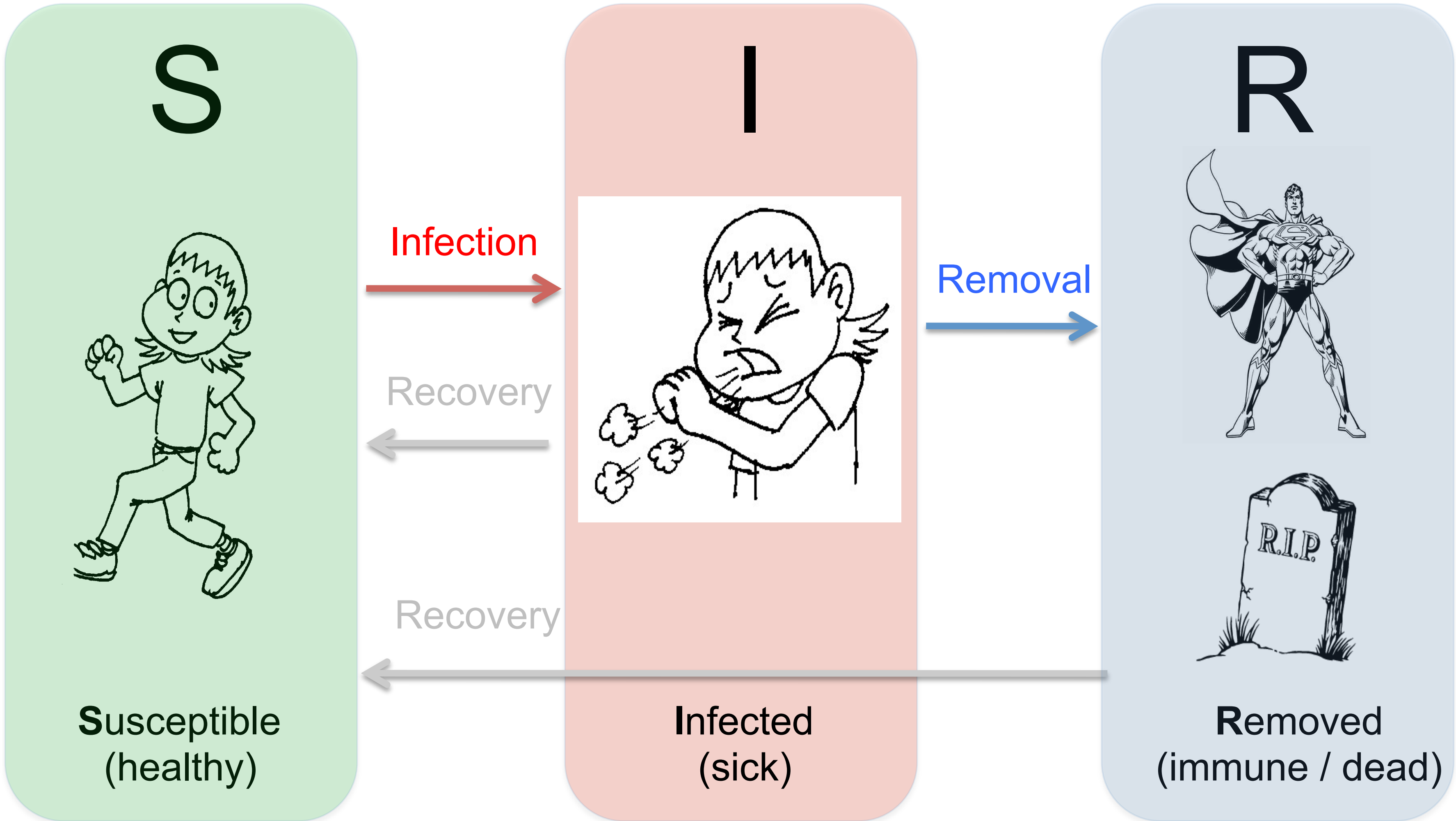
e.g. λ =2



Choose transmission scenario 2	mild	medium	high	very high
	λ =1.5	λ =1.9	λ =2.3	λ =2.7

Vespignani

Example 2: Flu, SARS, Plague, …

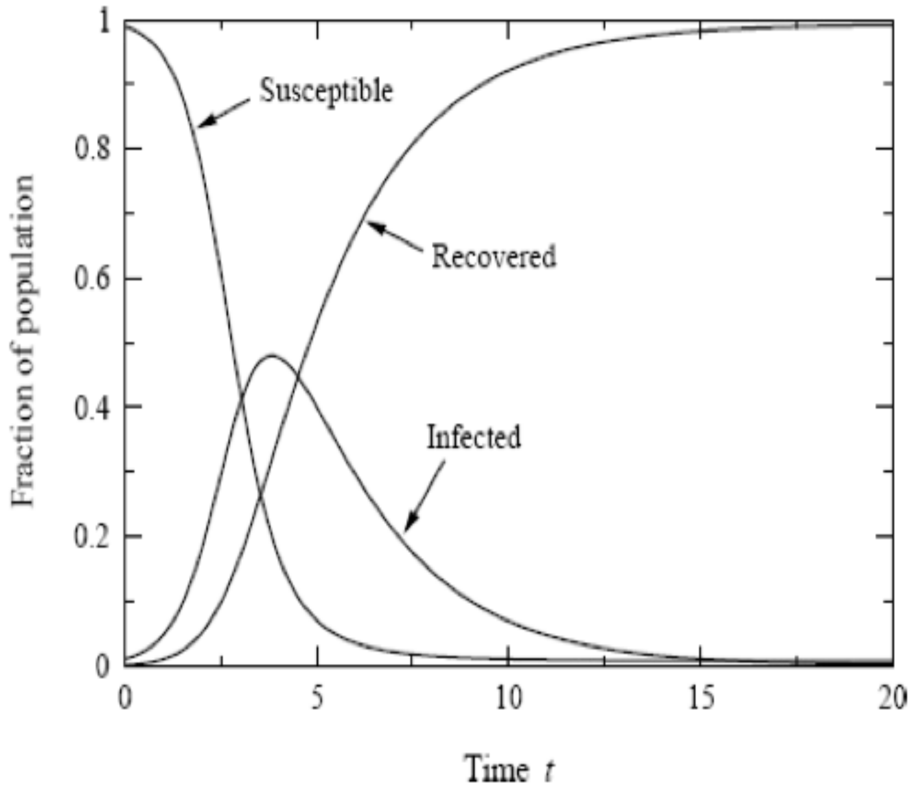


SIR Model

$$\frac{\mathrm{d}s(t)}{\mathrm{d}t} = \beta\langle k\rangle i(t)\,[1 - r(t) - i(t)]$$

$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} = -\mu i(t) + \beta\langle k\rangle i(t)\,[1 - r(t) - i(t)]$$

$$\frac{\mathrm{d}r(t)}{\mathrm{d}t} = \mu i(t).$$

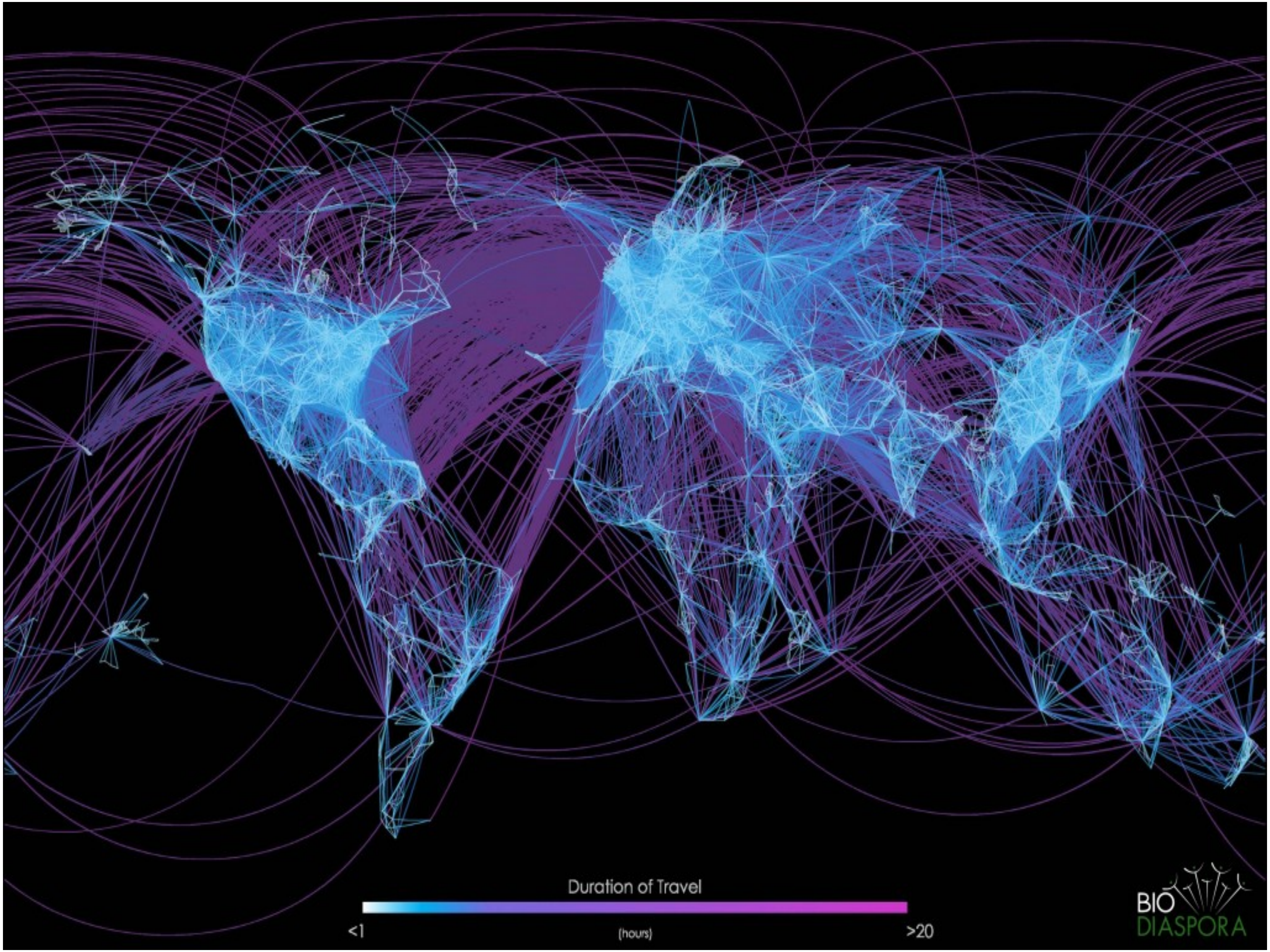


- SIR model: the fraction infected peaks and the fraction recovered saturates.

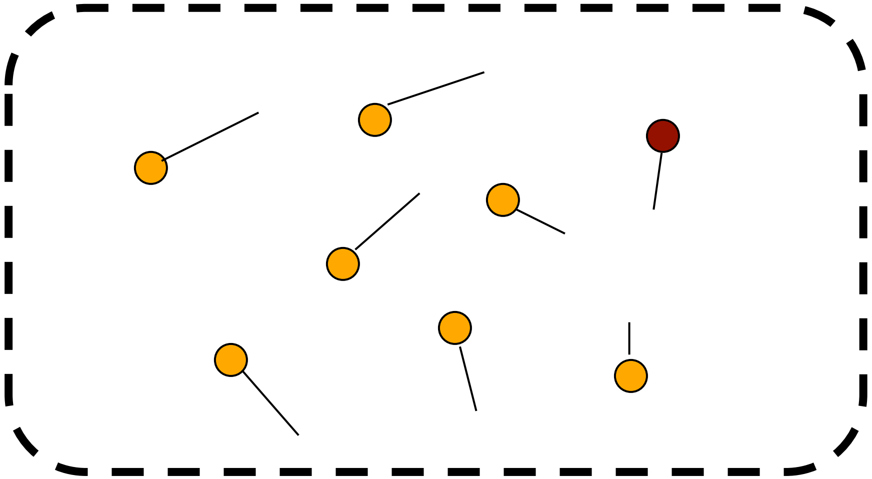
Epidemic modeling on networks

[Vespignani et al., since 2002]

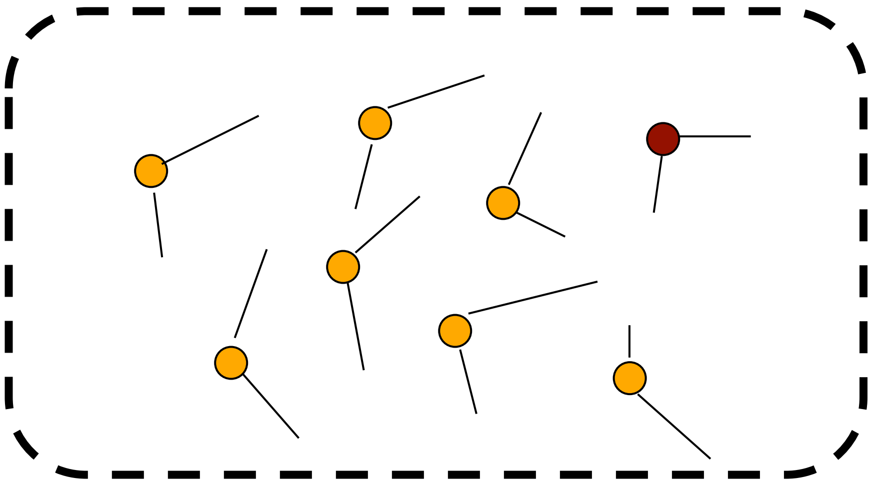
Gleamviz


Duration of Travel
<1
(hours)
>20
BIO
DIASPORA

SIS model on a network: Degree based representation



Class of nodes with degree *k=1*



Class of nodes with degree k=2

Split nodes by their degrees

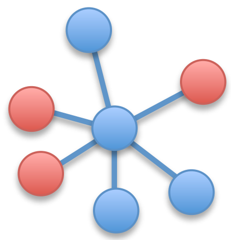
$$i_k = \frac{I_k}{N_k}, \quad i = \sum_k P(k) i_k$$

SIS model:

$$\frac{di_k(t)}{dt} = \beta(1 - i_k(t))k\Theta_k(t) - \mu i_k(t)$$

Proportional to *k*

Density of infected neighbors of nodes with degree *k*



I am susceptible with *k* neighbors, and $\Theta_k(t)$ of my neighbors are infected.

(Vespignani)

Early time behavior – SI model – the characteristic time vanishes!

$$\tau = \frac{\langle k \rangle}{\beta\left(\langle k^2 \rangle - \langle k \rangle\right)}$$

The timescale it takes for an epidemics to grow. The smaller is τ, the faster it grows.

ER network:
$\langle k^2 \rangle = \langle k \rangle (\langle k \rangle - 1)$

$$\tau_{ER} = \frac{1}{\beta\langle k \rangle}$$

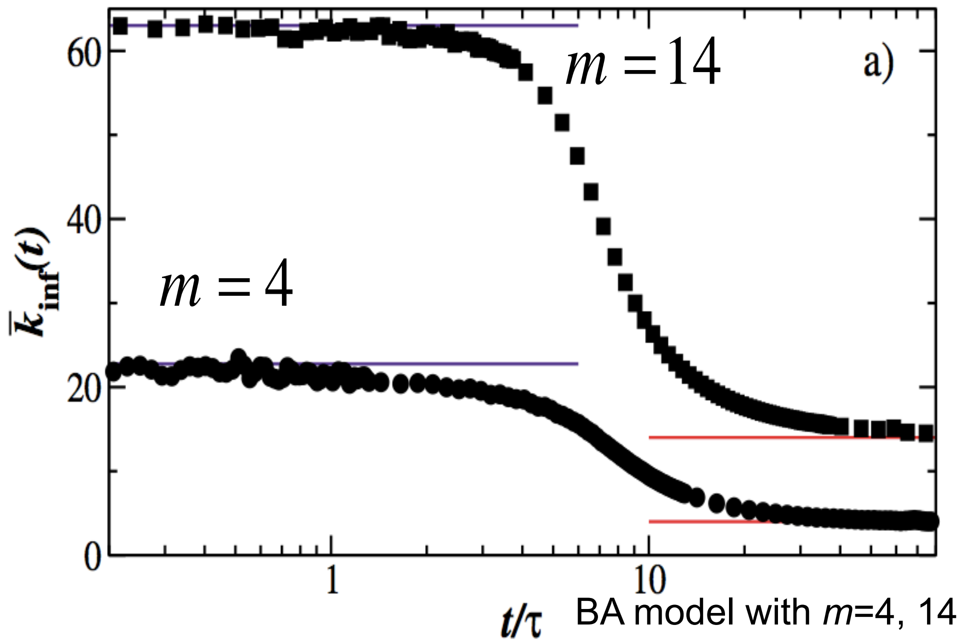
→ The more connected the network is, the faster does the epidemic spread.

SF network (γ<3):
$\langle k^2 \rangle \to \infty$ for N→ ∞ ➔ τ→ 0

For scale-free networks, the characteristic time vanishes, which means that the epidemic becomes instantaneous. The reason: the hubs get infected first, which then rapidly reach most nodes.

Numerical Test:
The average degree of newly infected nodes at time *t*:

$$\bar{k}_{\inf}(t) = \frac{\sum_k k\left(I_k(t) - I_k(t-1)\right)}{I(t) - I(t-1)}$$

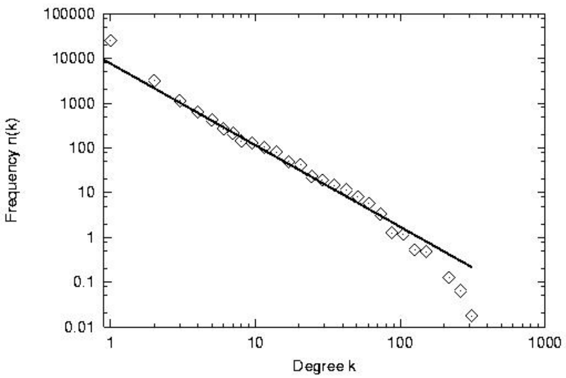


BA model with *m*=4, 14

M. Barthélemy et al., *PRL* **92**, 178701 (2004)

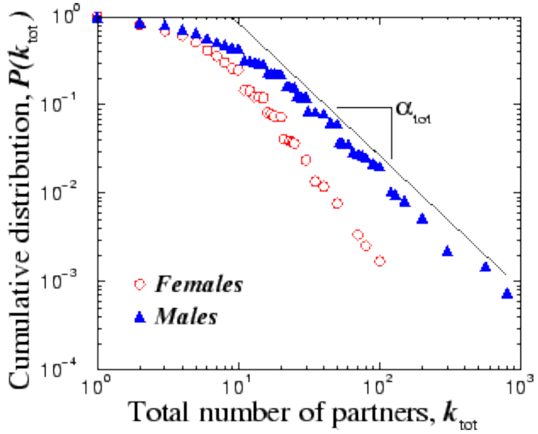
SIS Model – Absence of Epidemic Threshold

Email network



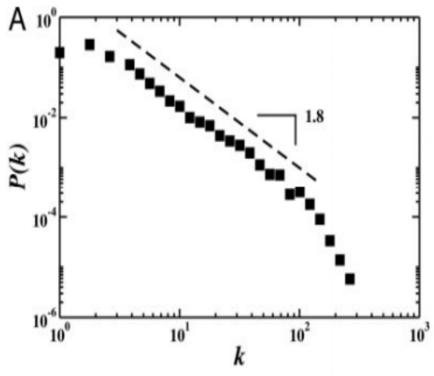
Ebel et al. (2002)

Human sexual network



Lilijeros et al., Nature (2001), Schneeberger et al. STD (2004)

Air transportation network



Colizza et al., PNAS 2006

Many networks will have small or vanishing epidemic threshold!

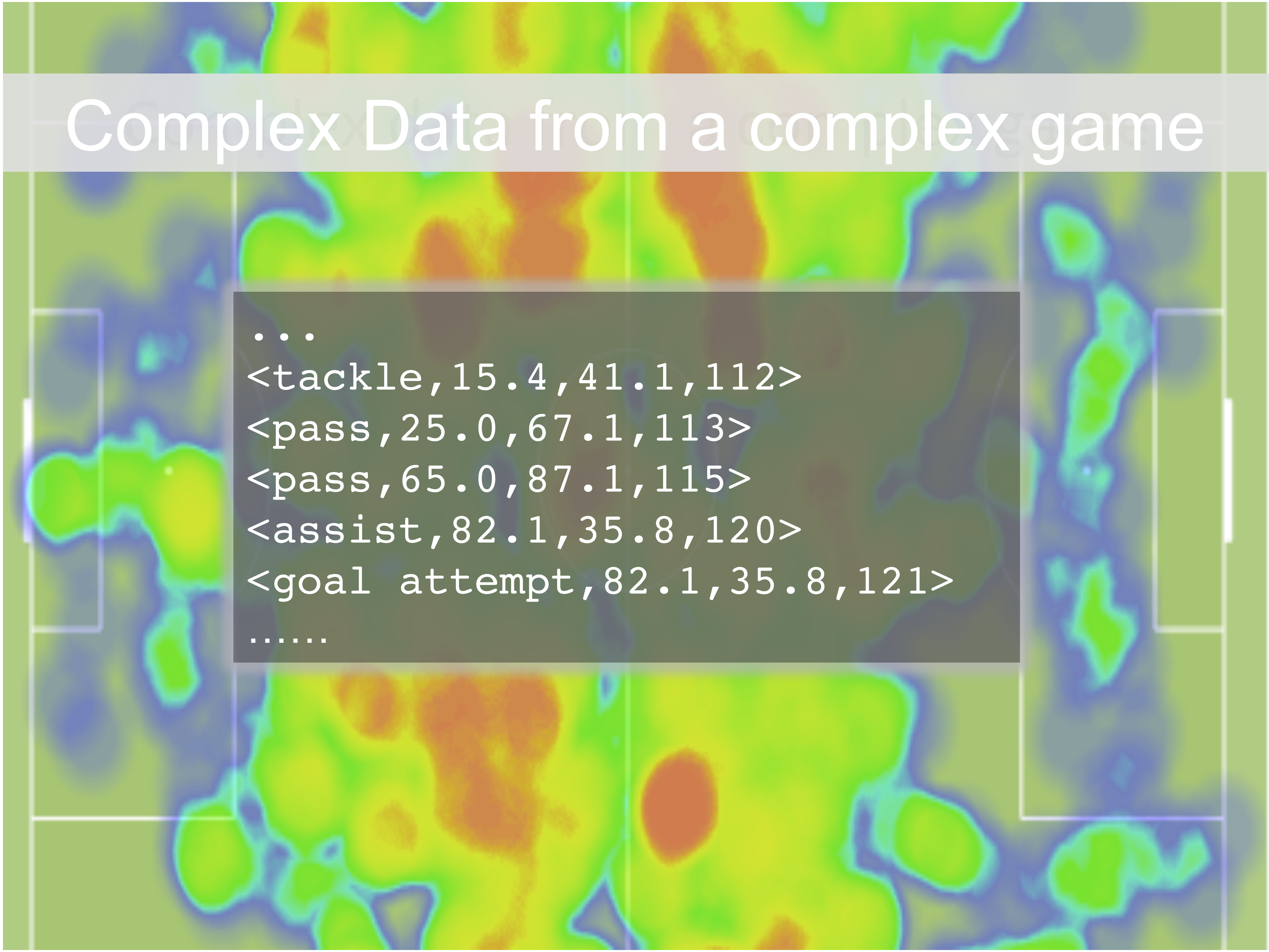
Sport data analytics

[Pappalardo, Cintia et al. @KDD Lab, since 2013]


CNR
La dura legge dei DATI!
IN SUPREMÆ DIGNITATIS
1343
Consiglio Nazionale delle Ricerche
SoBigData
Euro Lab on Big Data Analytics & Social Mining
Paolo Cintia
Marco Malvaldi
Luca Pappalardo
con la partecipazione di
Dino Pedreschi
Fosca Giannotti
...CONTINUA?

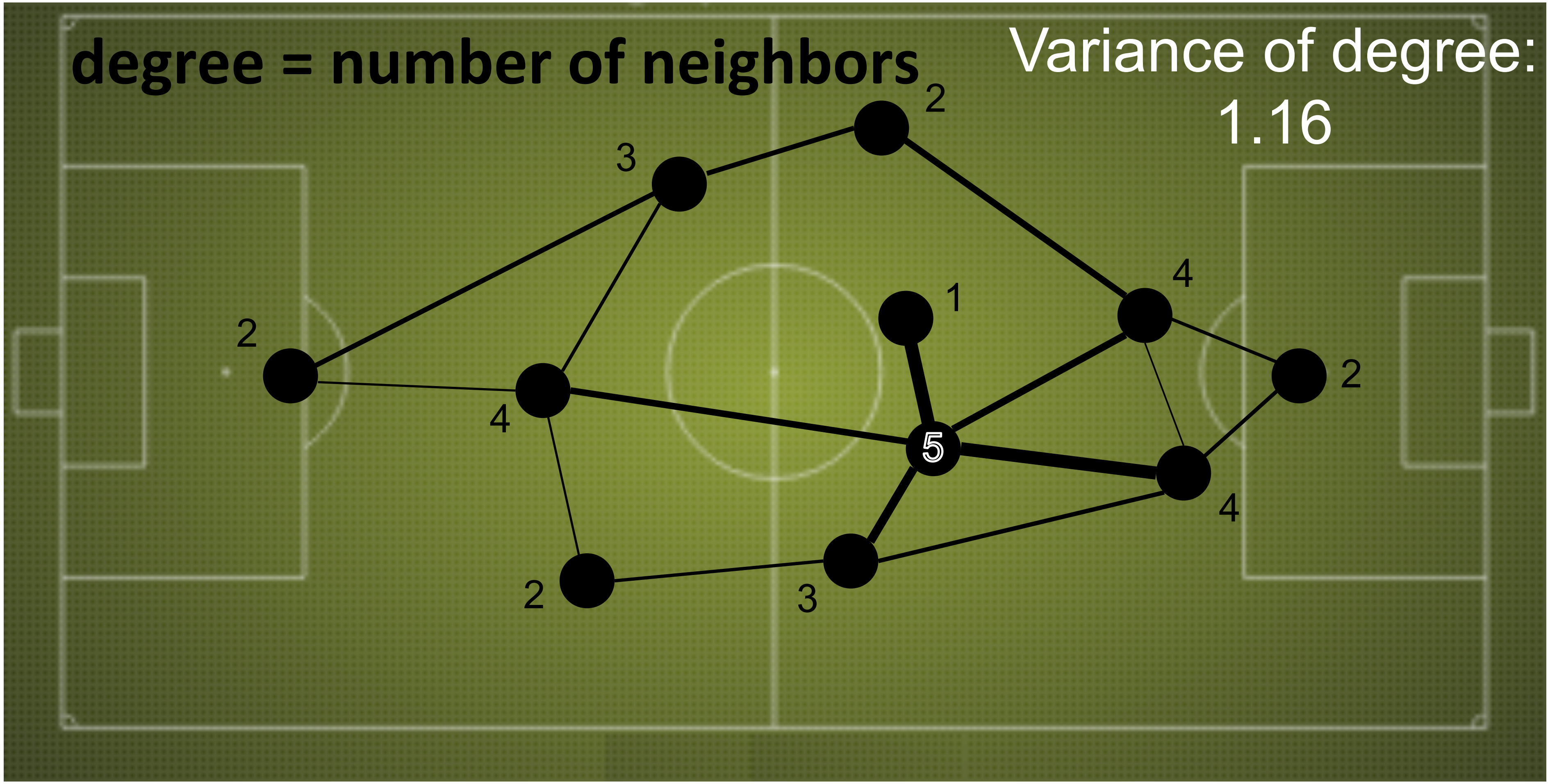

We are the champions
...I need a
Data Scientist...
FECHA
LUGAR
J. M.
Football and Big Data

Complex Data from a complex game

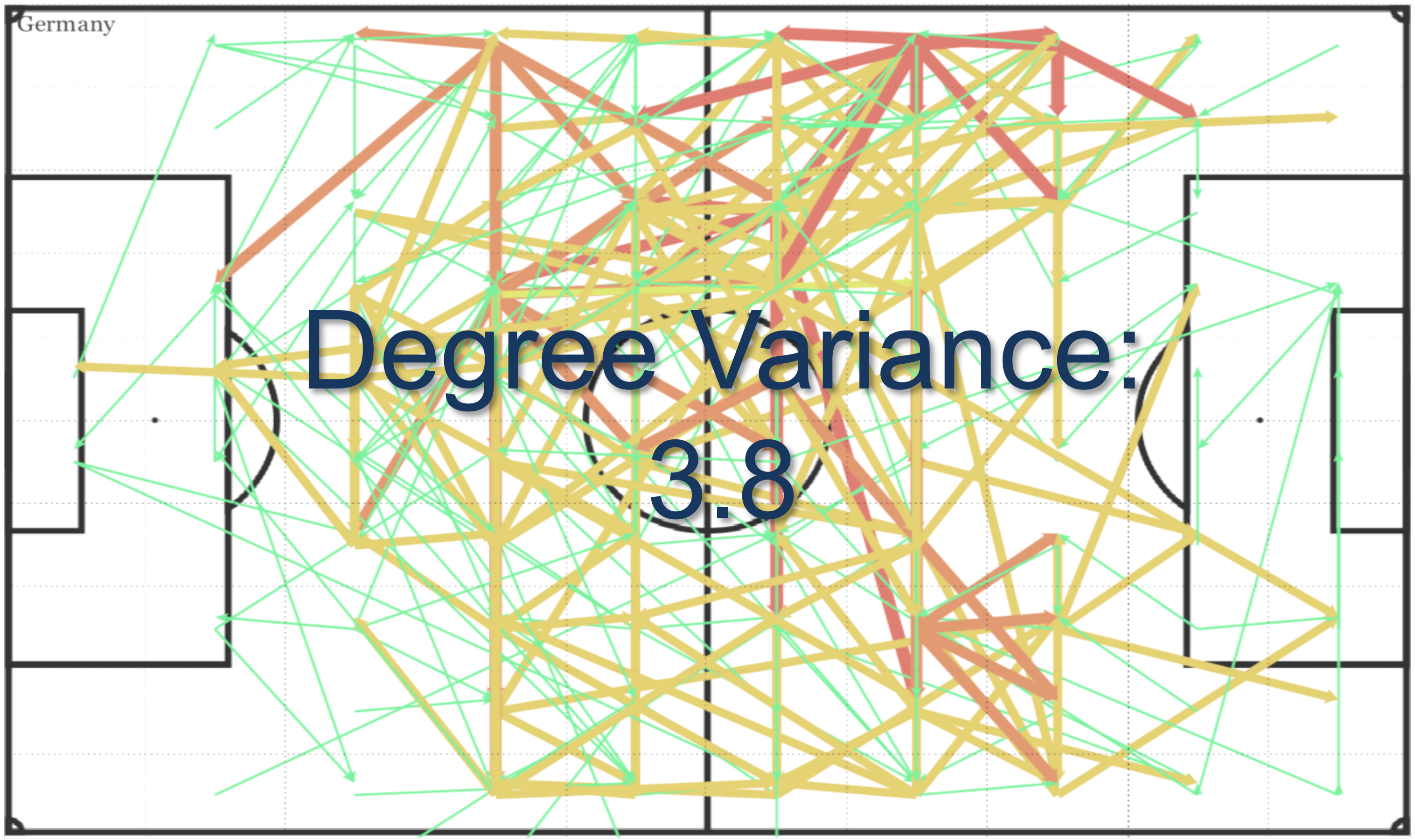


```
...
<tackle,15.4,41.1,112>
<pass,25.0,67.1,113>
<pass,65.0,87.1,115>
<assist,82.1,35.8,120>
<goal attempt,82.1,35.8,121>
......
```

The passes network among players

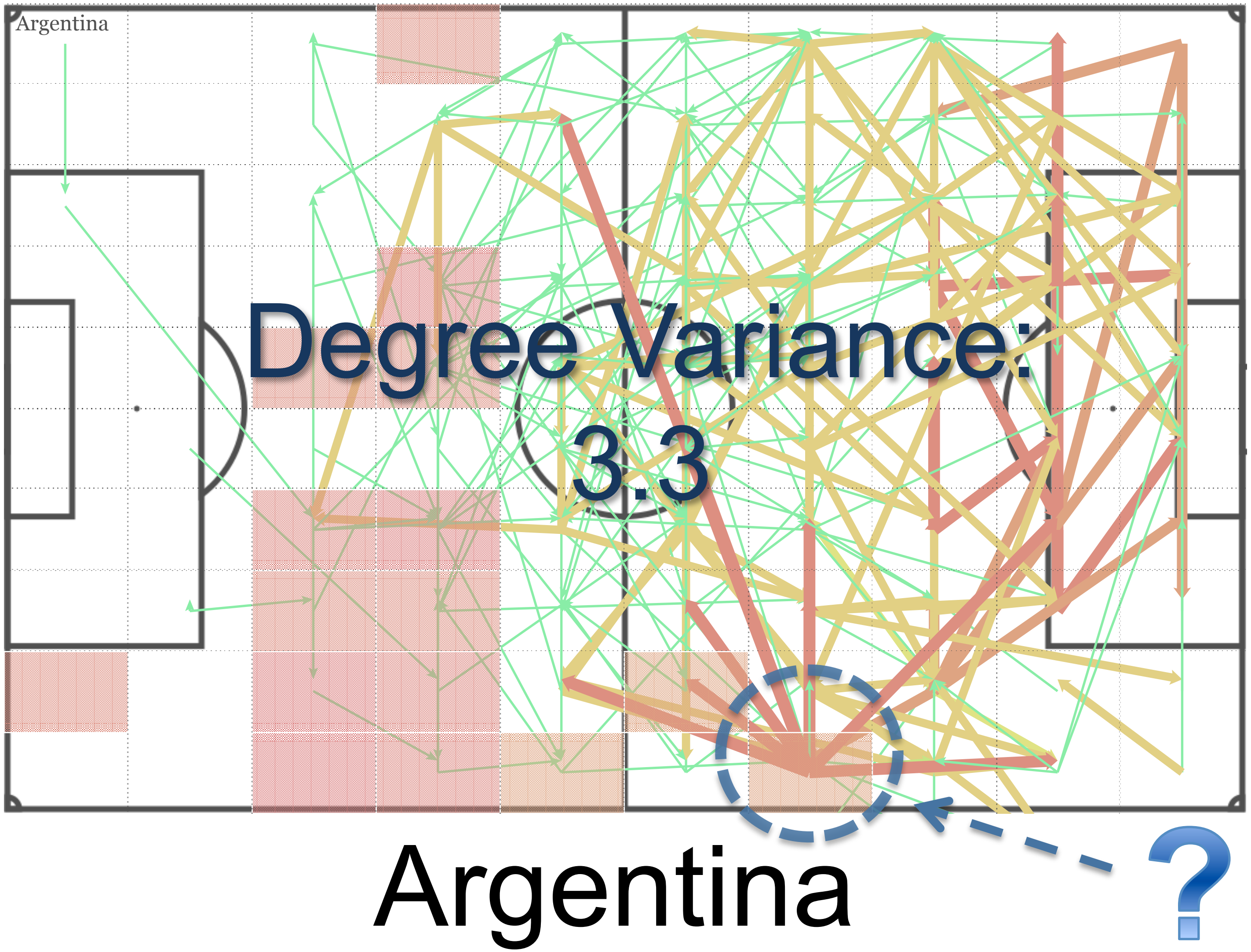


The passes network among zones

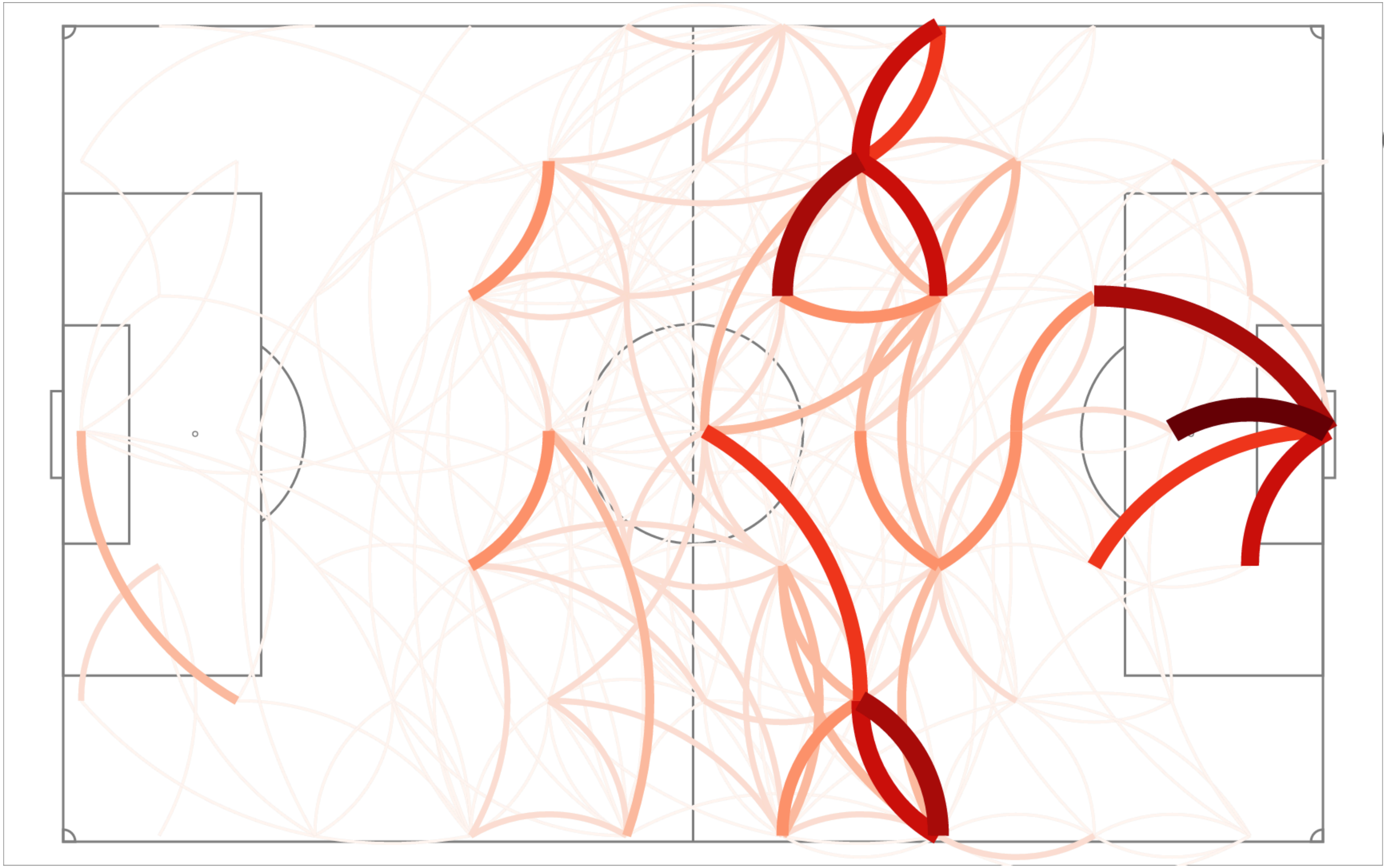


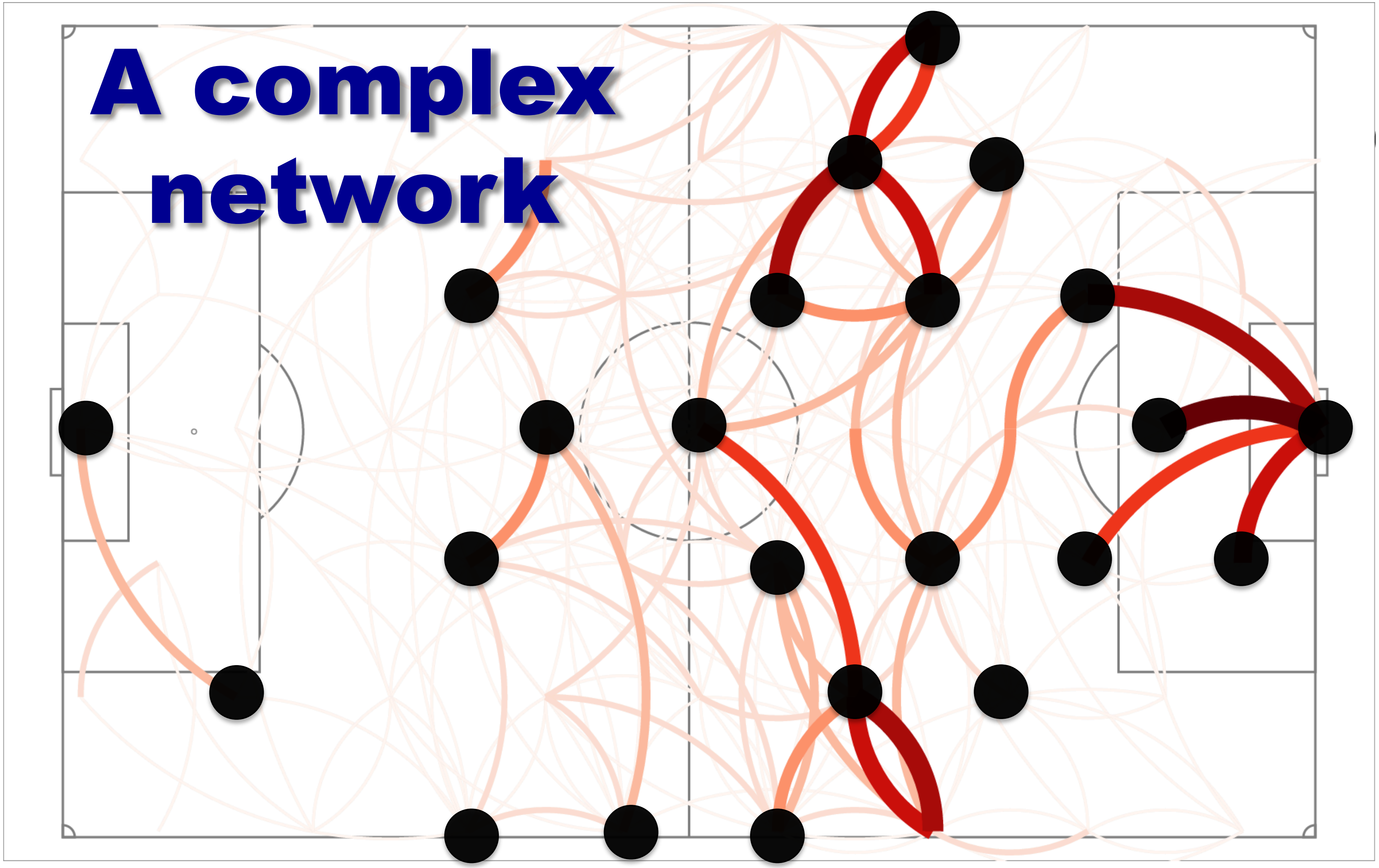
Germany

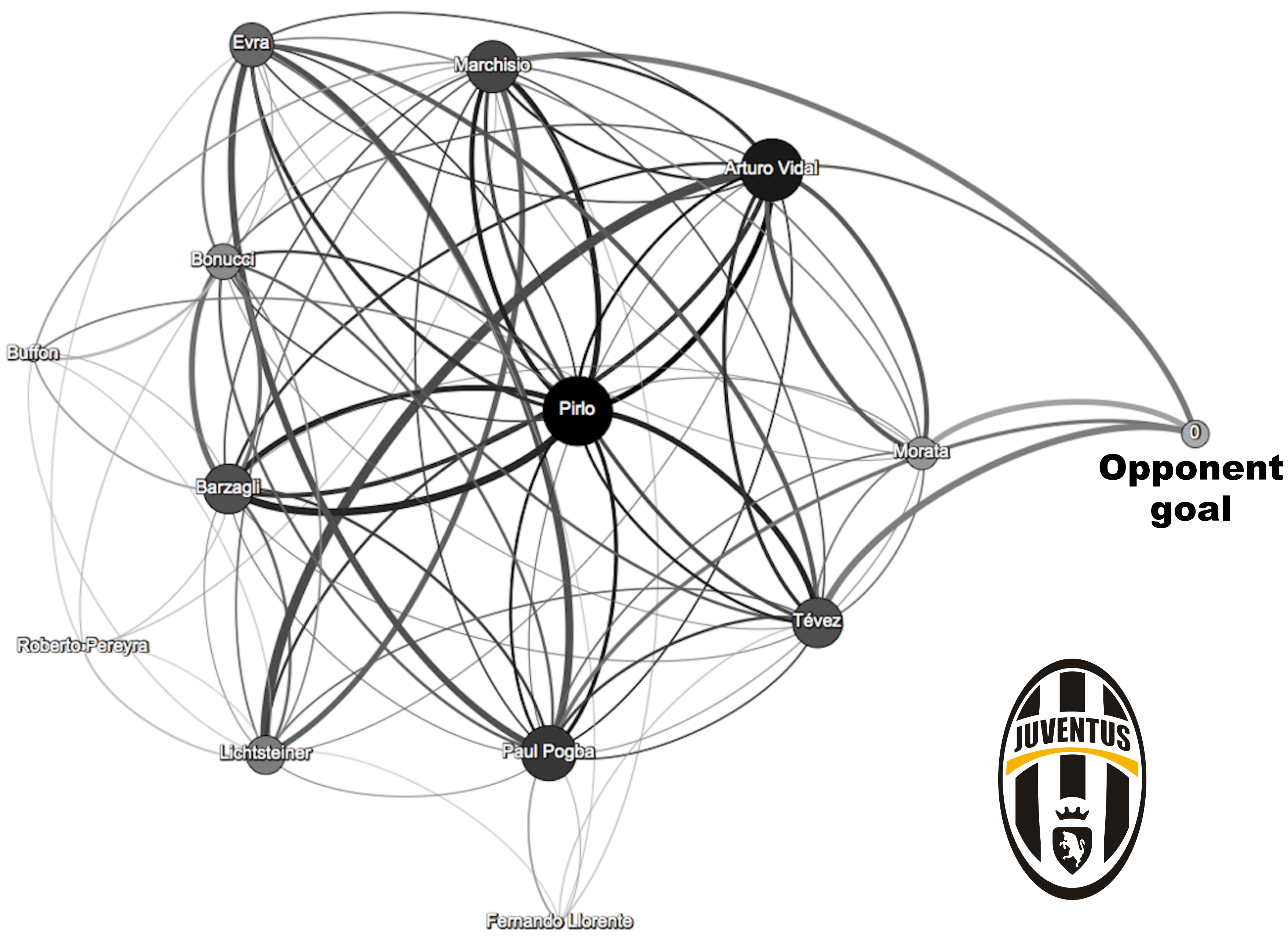
The passes network among zones

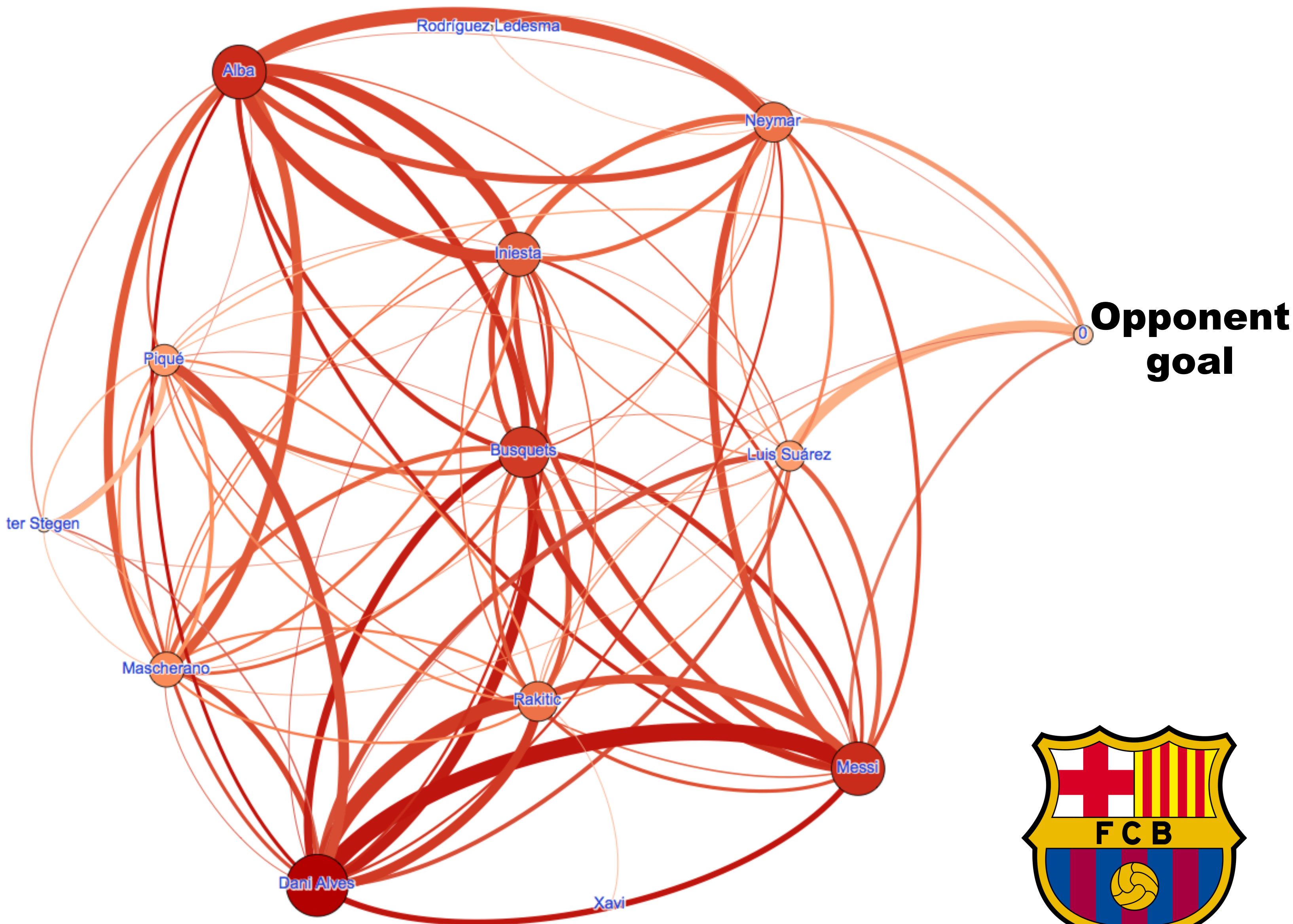


Argentina




A complex
network


Evra
Marchisio
Arturo Vidal
Bonucci
Buffon
Pirlo
0
Opponent goal
Morata
Barzagli
Tévez
Roberto Pereyra
Lichtsteiner
Paul Pogba
Fernando Llorente
JUVENTUS


Rodríguez Ledesma
Alba
Neymar
Iniesta
Opponent goal
0
Piqué
Busquets
Luis Suárez
ter Stegen
Mascherano
Rakitic
Messi
Dani Alves
Xavi

F C B

We computed the variance for each team during the World Cup 2014

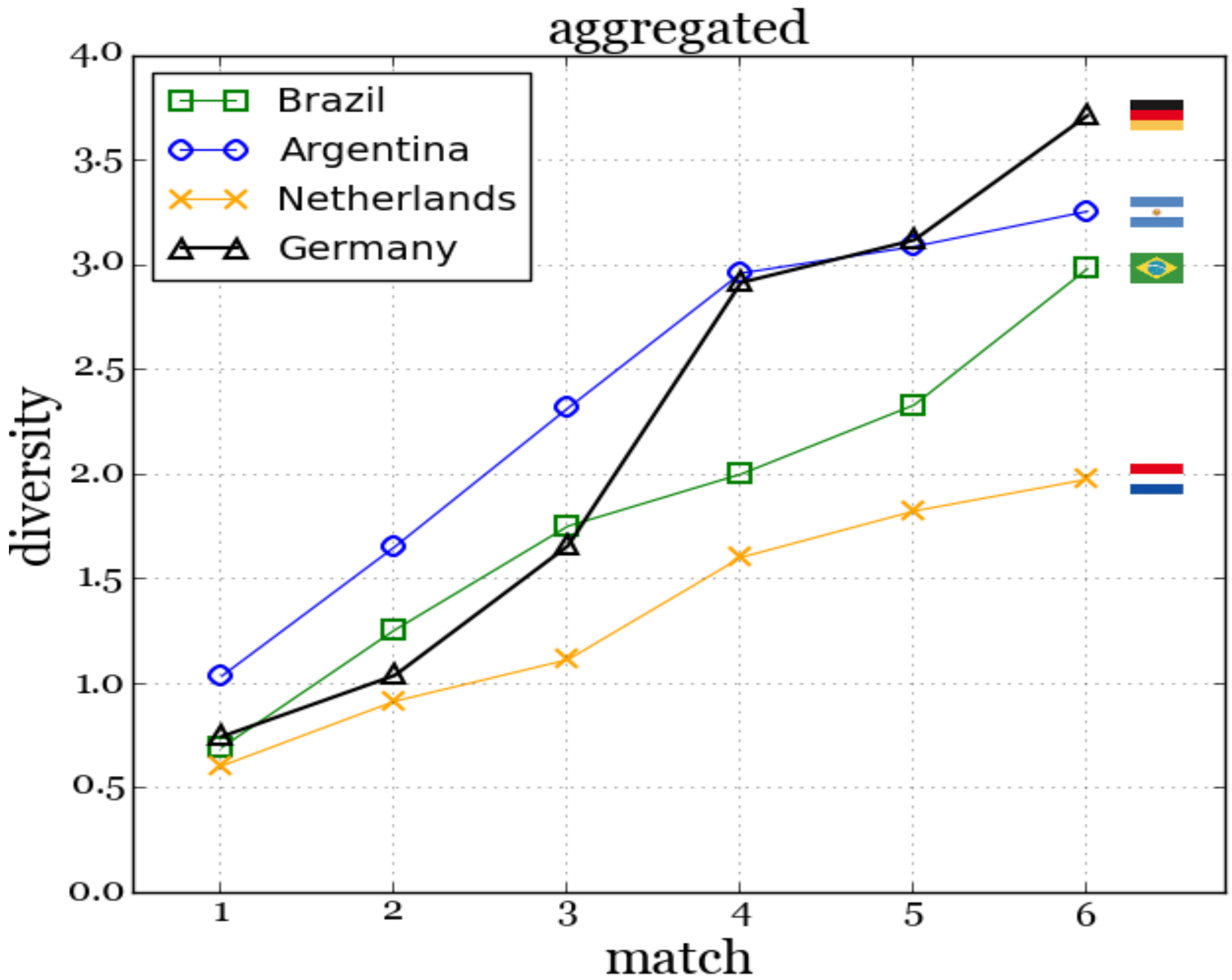
World Cup 2014

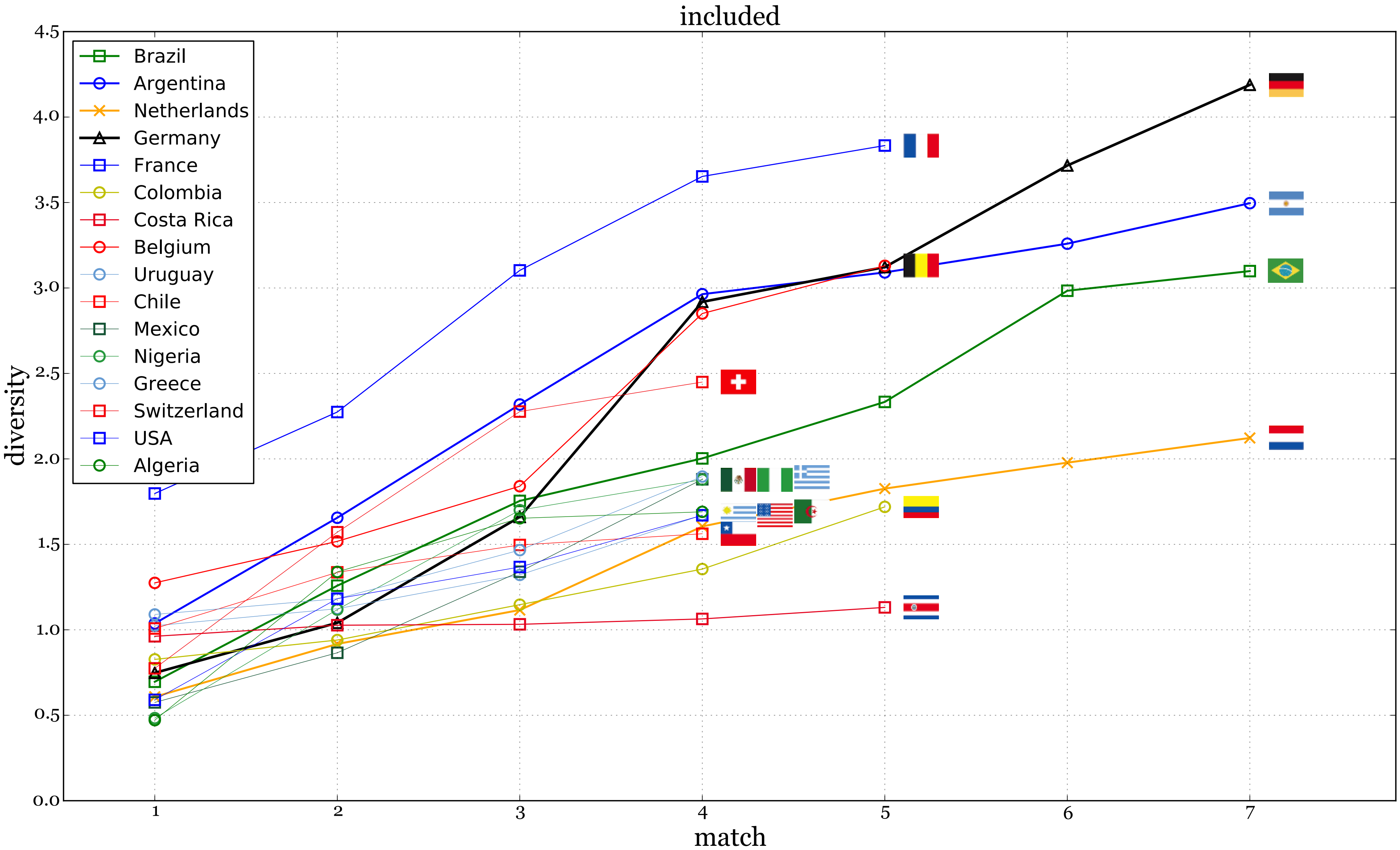
Big Data Tales
@bigdatatales

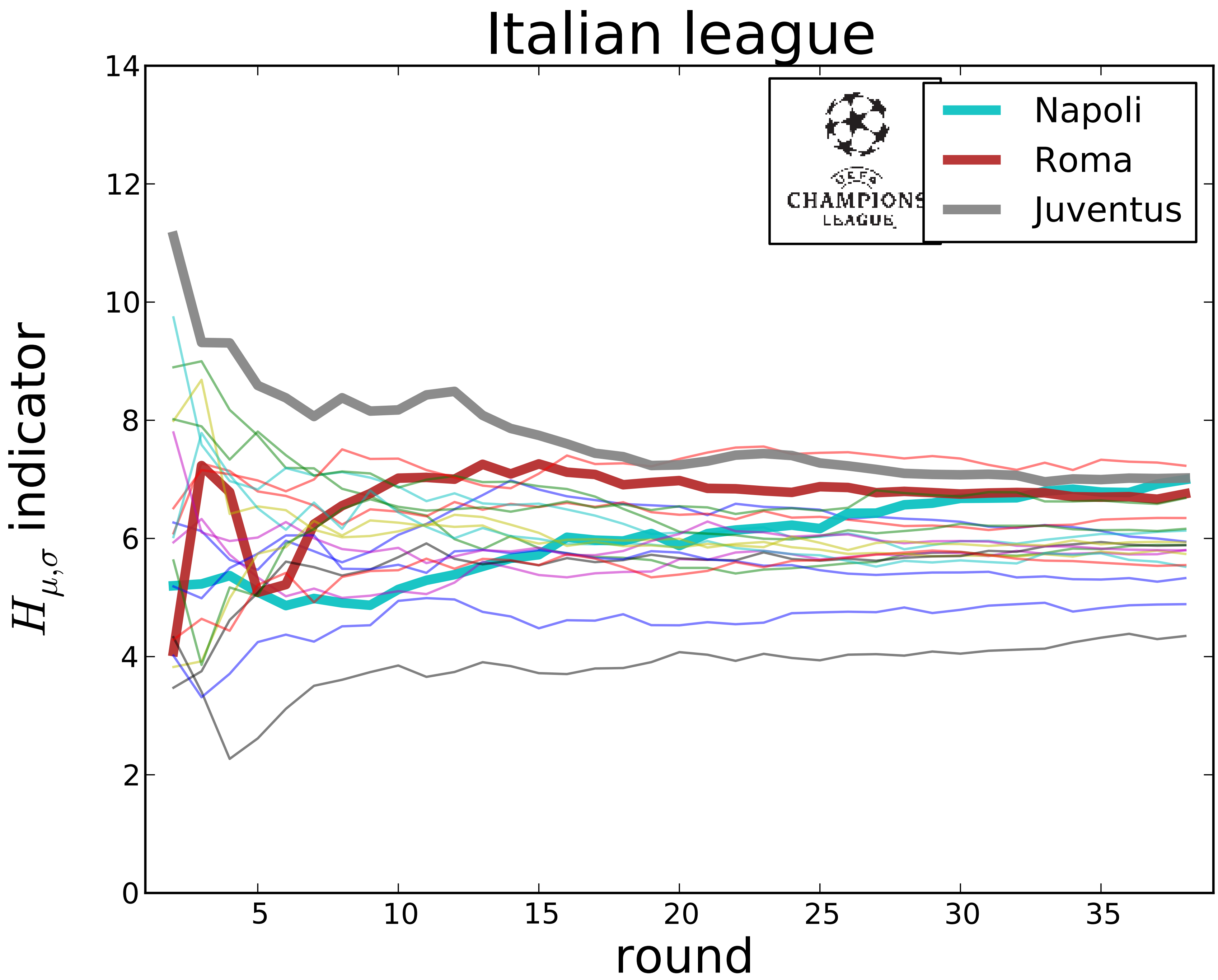
Segui

According to our models the final will be Germany-Argentina. Are our data-driven models correct ? Let's see what happens!!! #WorldCup2014

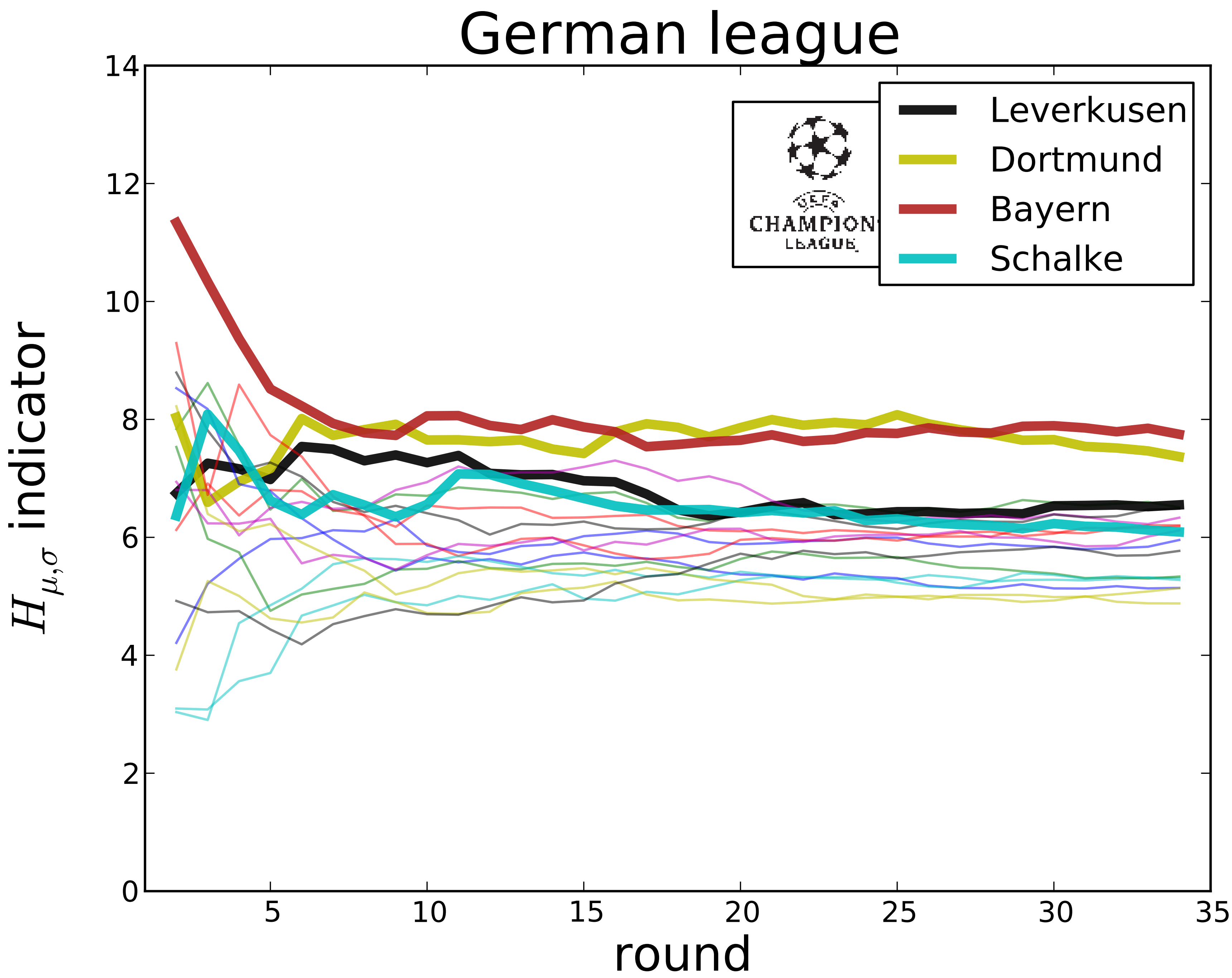
21:00 - 8 Lug 2014 Pisa, Italia


aggregated
Brazil
Argentina
Netherlands
Germany
diversity
match
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6

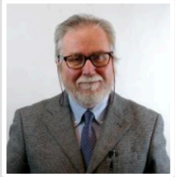

included
diversity
match
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
Brazil
Argentina
Netherlands
Germany
France
Colombia
Costa Rica
Belgium
Uruguay
Chile
Mexico
Nigeria
Greece
Switzerland
USA
Algeria

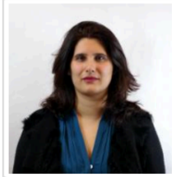


Italian league
14
12
10
8
6
4
2
0
5
10
15
20
25
30
35
round
$H_{\mu,\sigma}$ indicator
CHAMPIONS
LEAGUE
Napoli
Roma
Juventus


German league
Leverkusen
Dortmund
Bayern
Schalke
UEFA
CHAMPIONS
LEAGUE
$H_{\mu,\sigma}$ indicator
round
0
2
4
6
8
10
12
14
5
10
15
20
25
30
35

simulated ranking		real ranking	
Bayern	91	**Bayern**	90
Leverkusen	72	Dortmund	71
Dortmund	68	Schalke	64
Wolfsburg	59	Leverkusen	61
Augsburg	58	Wolfsburg	60
Hoffenheim	49	Mönchengladbach	55
Hertha	49	Mainz	53
Mainz	48	Augsburg	52
Schalke	47	Hoffenheim	44
Frankfurt	46	Hannover	42
Mönchengladbach	42	Hertha	41
Hannover	41	Werder	39
Hamburg	38	Freiburg	36
Stuttgart	35	Frankfurt	36
Freiburg	31	Stuttgart	32
Werder	24	Hamburg	27
Braunschweig	22	Nürnberg	26
Nürnberg	17	Braunschweig	25


Dino Pedreschi
Full Professor
pedre@di.unipi.it

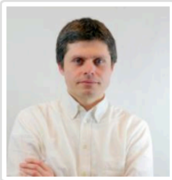
Franco Turini
Full Professor
turini@di.unipi.it

Fosca Giannotti
Director of Research
fosca.giannotti@isti.cnr.it

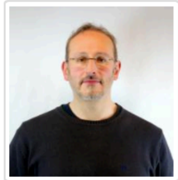
Salvatore Ruggieri
Full Professor
ruggieri@di.unipi.it

Anna Monreale
Assistant Professor
annam@di.unipi.it

Alina Sîrbu
Assistant Professor
alina.sirbu@unipi.it





Barbara Furletti
Researcher
barbara.furletti@isti.cnr.it

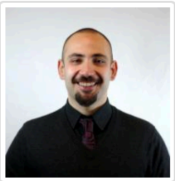
Mirco Nanni
Researcher
mirco.nanni@isti.cnr.it

Salvatore Rinzivillo
Researcher
rinzivillo@isti.cnr.it

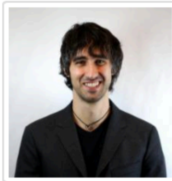
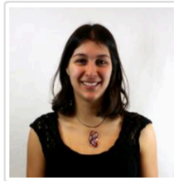
Roberto Trasarti
Researcher
roberto.trasarti@isti.cnr.it

Paolo Cintia
Post Doc
paolo.cintia@isti.cnr.it

Valerio Grossi
Post Doc
vgrossi@di.unipi.it





Luca Pappalardo
Post Doc
luca.pappalardo@isti.cnr.it

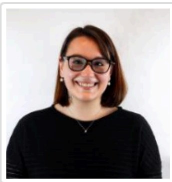
Giulio Rossetti
Post Doc
giulio.rossetti@isti.cnr.it

Alessandro Baroni
Ph.D. Student
baroni@di.unipi.it

Lorenzo Gabrielli
Ph.D. Student
lorenzo.gabrielli@isti.cnr.it

Riccardo Guidotti
Ph.D. Student
riccardo.guidotti@isti.cnr.it

Ioanna Miliou
Ph.D. Student
ioanna.miliou@for.unipi.it



Letizia Milli
Ph.D. Student
letizia.milli@isti.cnr.it

Laura Pollacci
Ph.D. Student
laura.pollacci@di.unipi.it

Francesca Pratesi
Ph.D. Student
francesca.pratesi@isti.cnr.it

Farzad Vaziri
Ph.D. Student
vaziri@di.unipi.it

Viola Bachini
Collaborator
viola.bachini@isti.cnr.it

Daniele Fadda
Collaborator
daniele.fadda@isti.cnr.it


http://kdd.isti.cnr.it/people

Knowledge Discovery & Data Mining Lab
http://kdd.isti.cnr.it