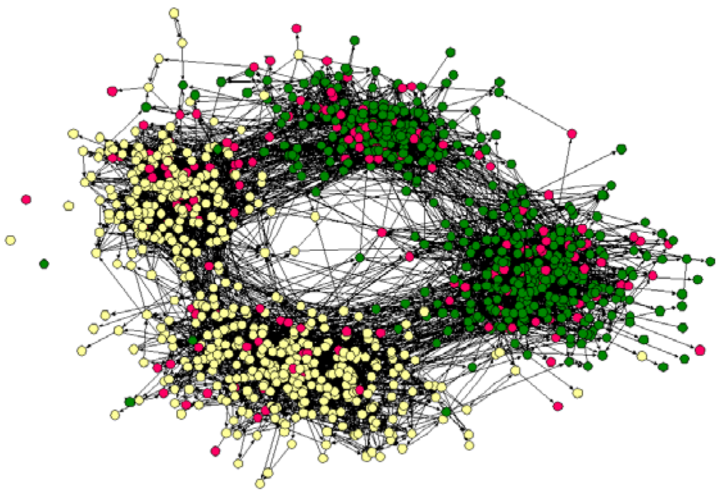
Uncovering the overlapping modular structure of complex networks

Tamás Vicsek

Dept. of Biological Physics, Eötvös University, Hungary

http://angel.elte.hu/~vicsek

http://cfinder.org



Why modules (densely interconnected parts)?

The internal organization of large networks is responsible for their function.

Complex systems/networks are typically *hierarchical.*

The units organize (become more closely connected) into groups which can themselves be regarded as units on a higher level.

We call these densely interconnected groups of nodes as modules/communities/cohesive groups/clusters etc. They are the "building blocks" of the complex networks on many scales.

For example:

Person->group->department->division->company->industrial sector

Letter->word->sentence->paragraph->section->chapter->book

Questions:

How can we recover the hierarchy of overlapping groups/modules/communities in the network if only a (very long) list of links between pairs of units is given?

What are their main characteristics?

Outline

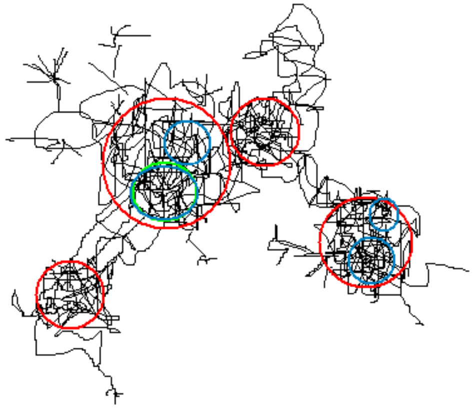
- **Basic facts and principles**
- **Community finding *versus k*-clique percolation**
- **Results for protein interaction, word association, phone calls, school friendship and collaboration networks**

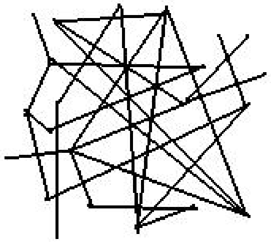
Basic observations:

A large complex network is bounded to be highly structured
(has modules; function follows from structure)

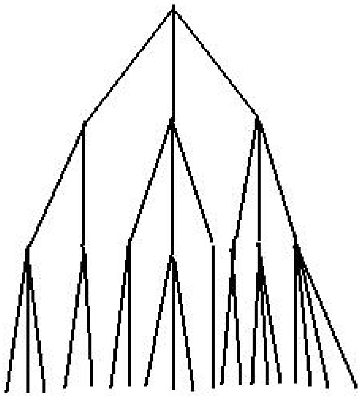
The internal organization is typically hierarchical
(i.e., displays some sort of self-similarity of the structure)

An important new aspect: Overlaps of modules are essential

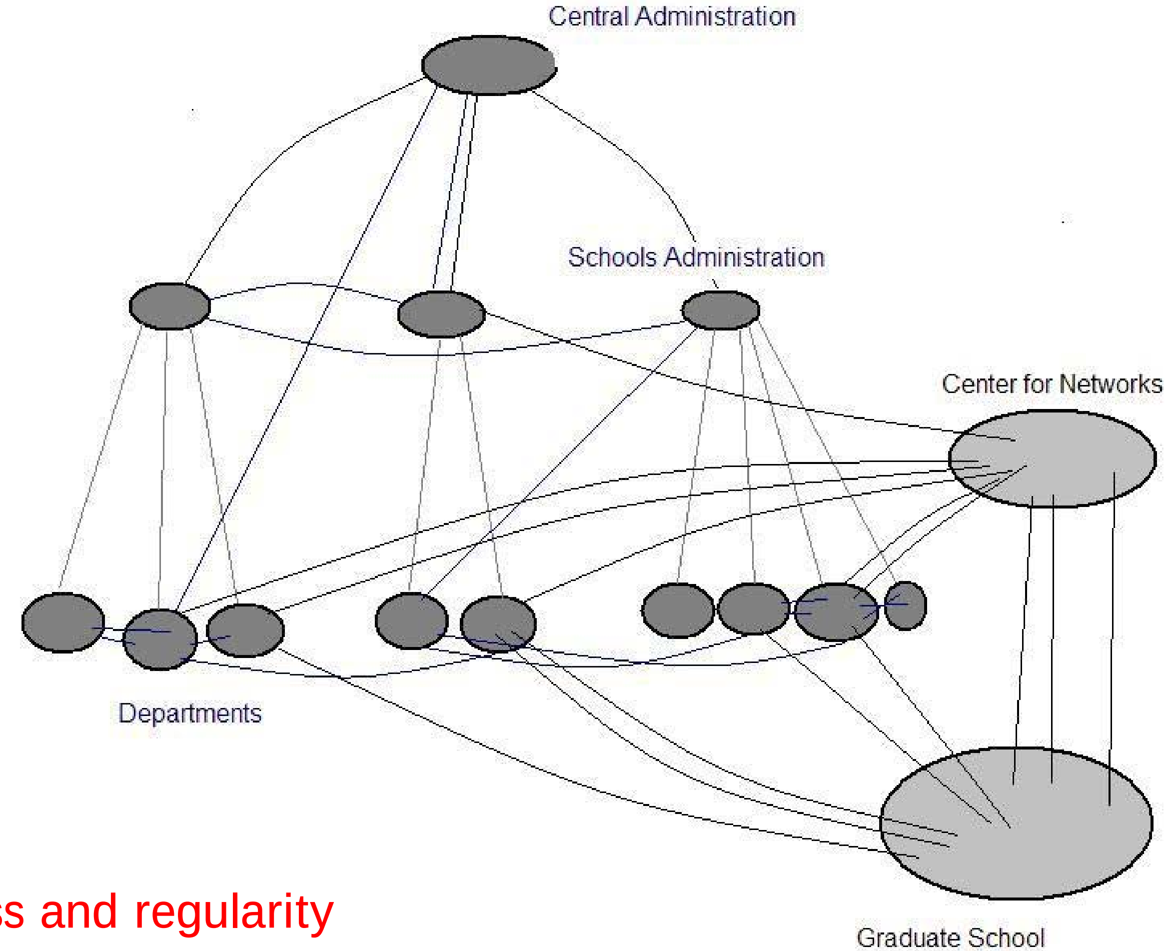




"mess", no function

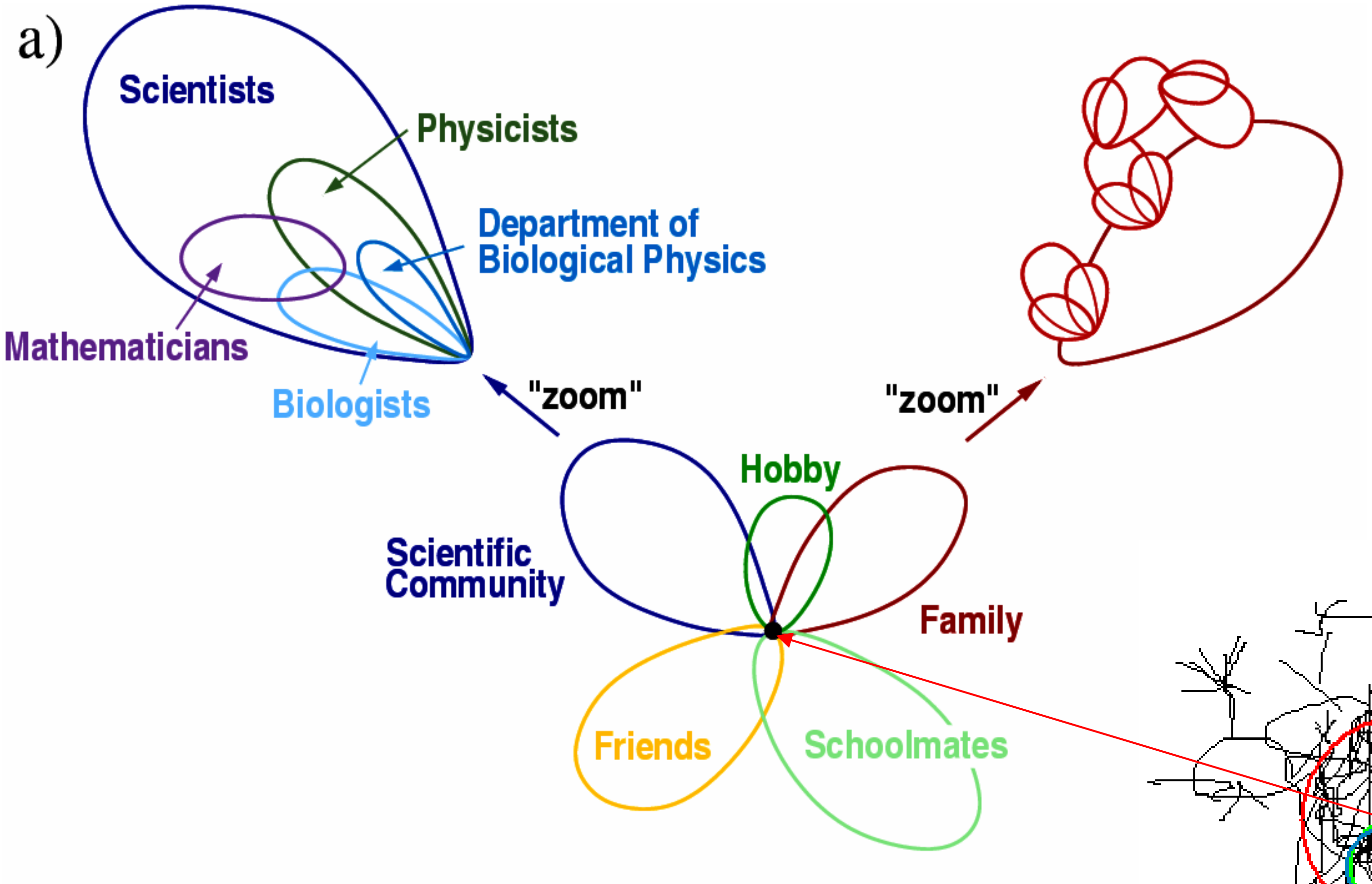


Too constrained, limited function

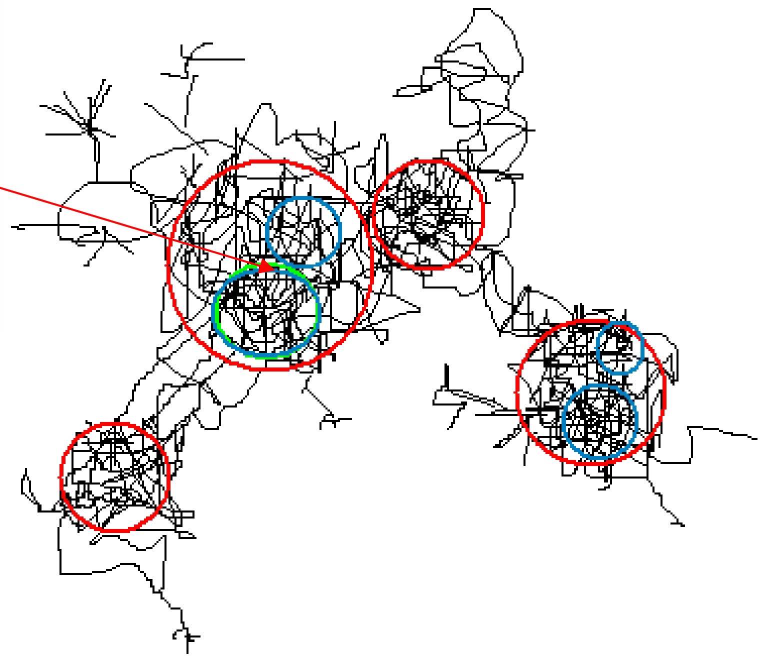


Complexity is between randomness and regularity

Role of overlaps

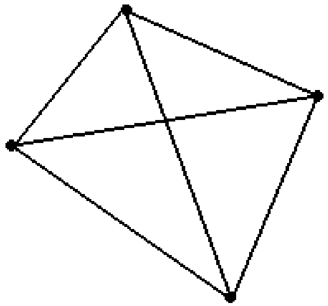


Is this like a tree? (hierarchical methods)

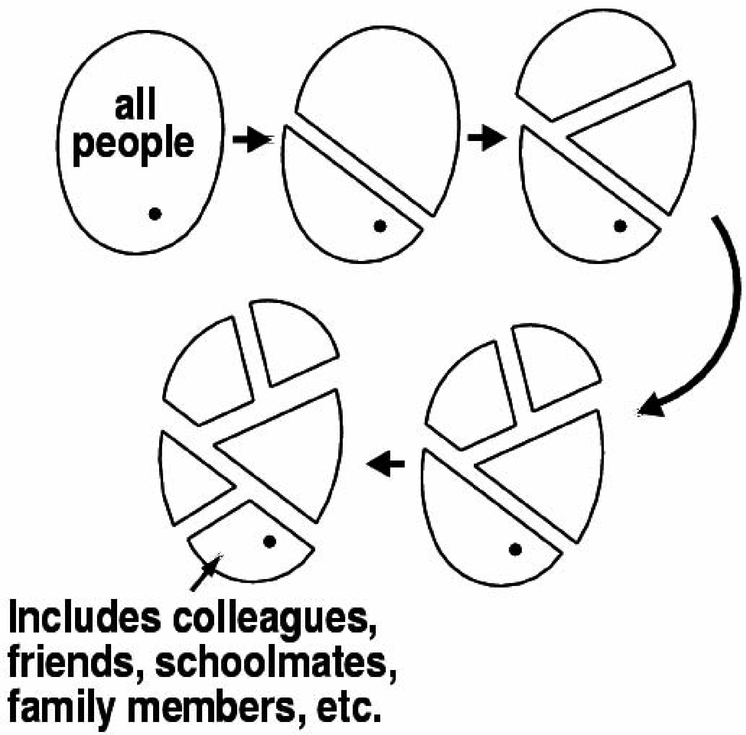


Finding communities

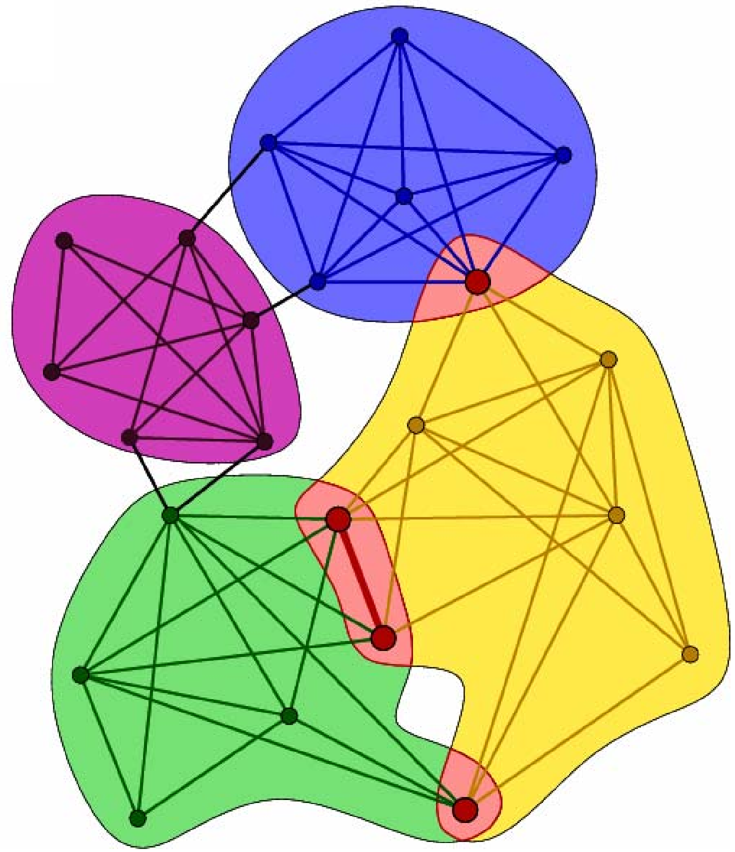
a 4-clique



Hierarchical methods



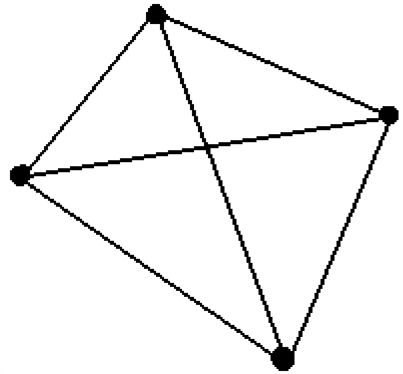
***k*-clique template rolling**



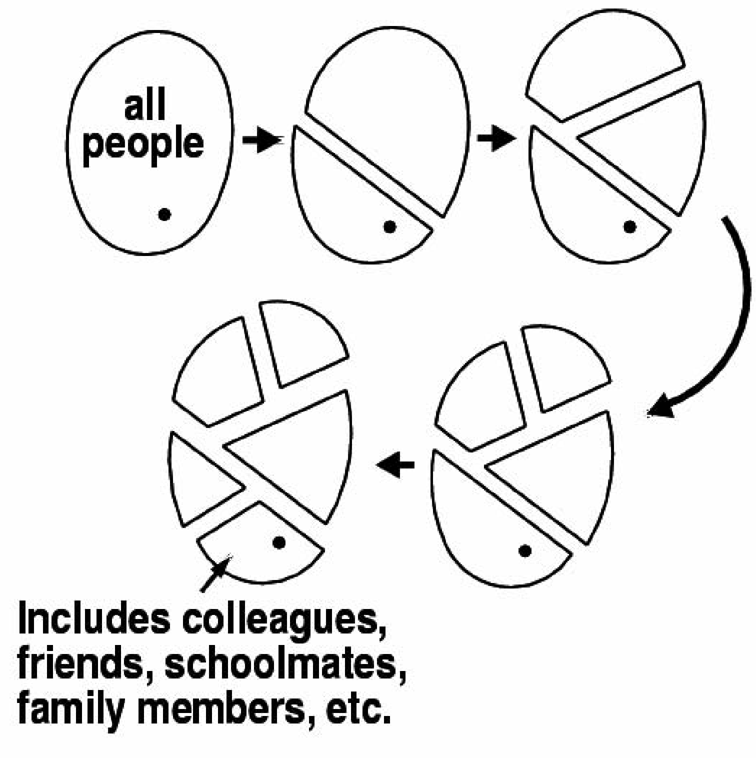
Two nodes belong to the same community if they can be connected through adjacent *k*-cliques

Finding communities

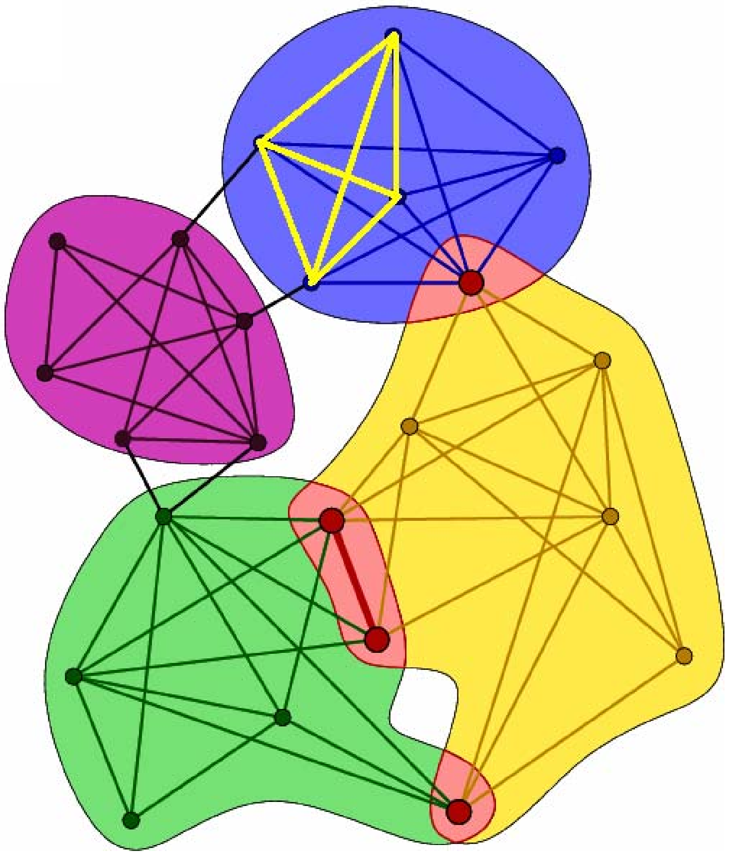
a 4-clique



Hierarchical methods



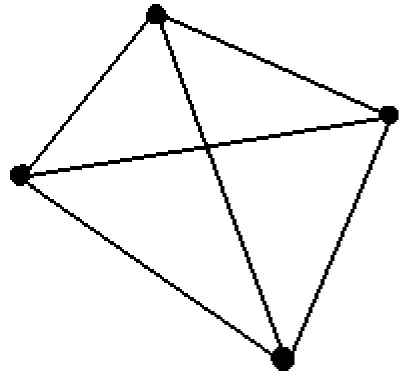
k-clique template rolling



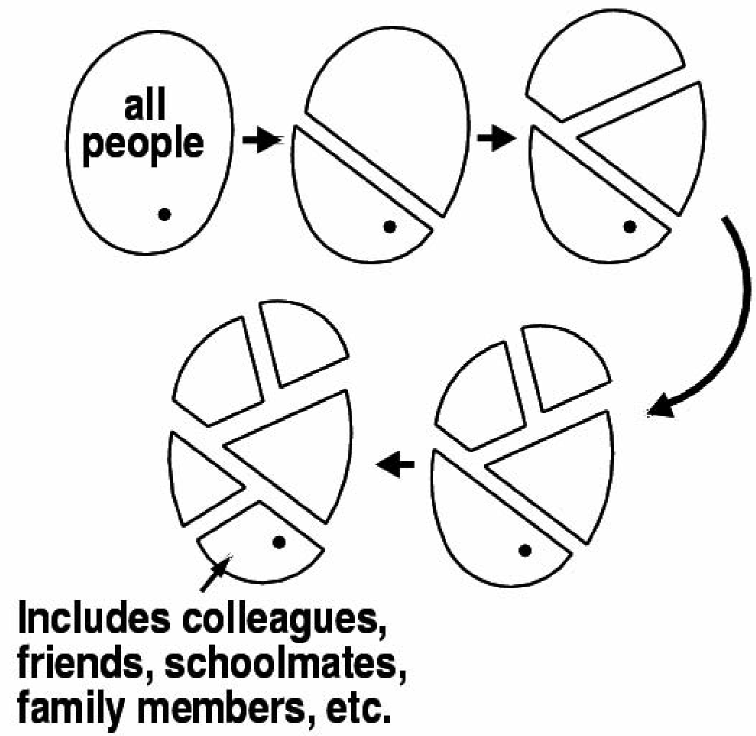
Two nodes belong to the same community if they can be connected through adjacent *k*-cliques

Finding communities

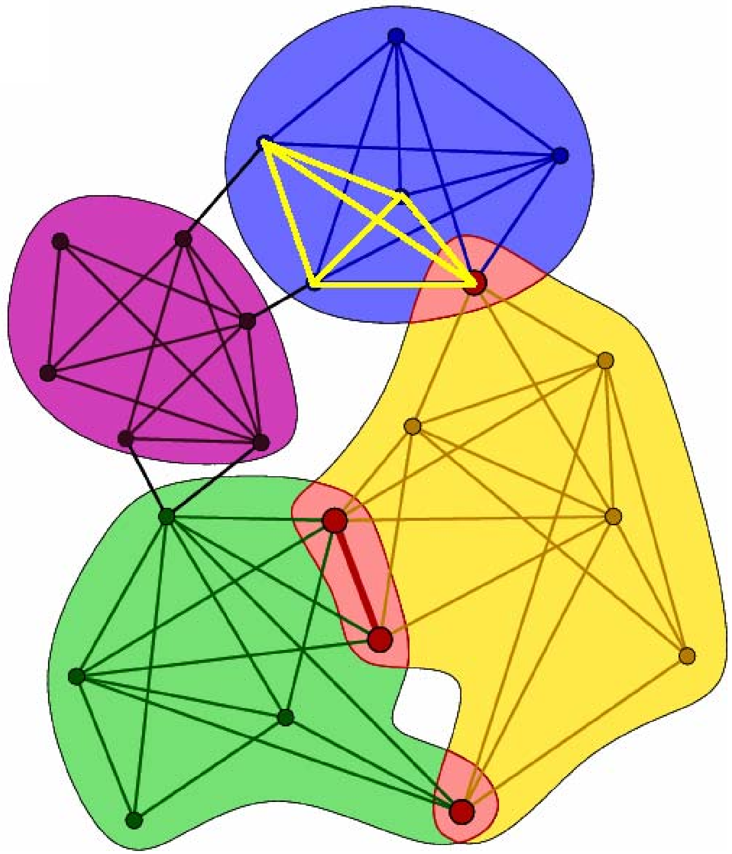
a 4-clique



Hierarchical methods



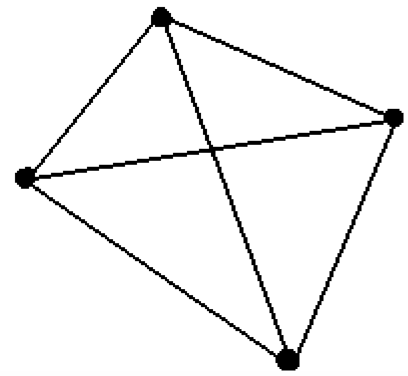
k-clique template rolling



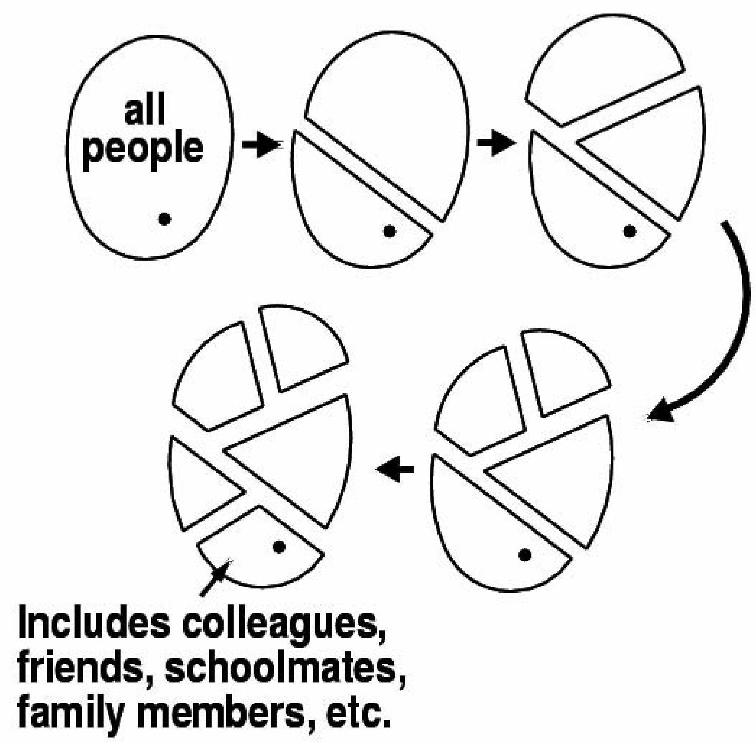
Two nodes belong to the same community if they can be connected through adjacent *k*-cliques

Finding communities

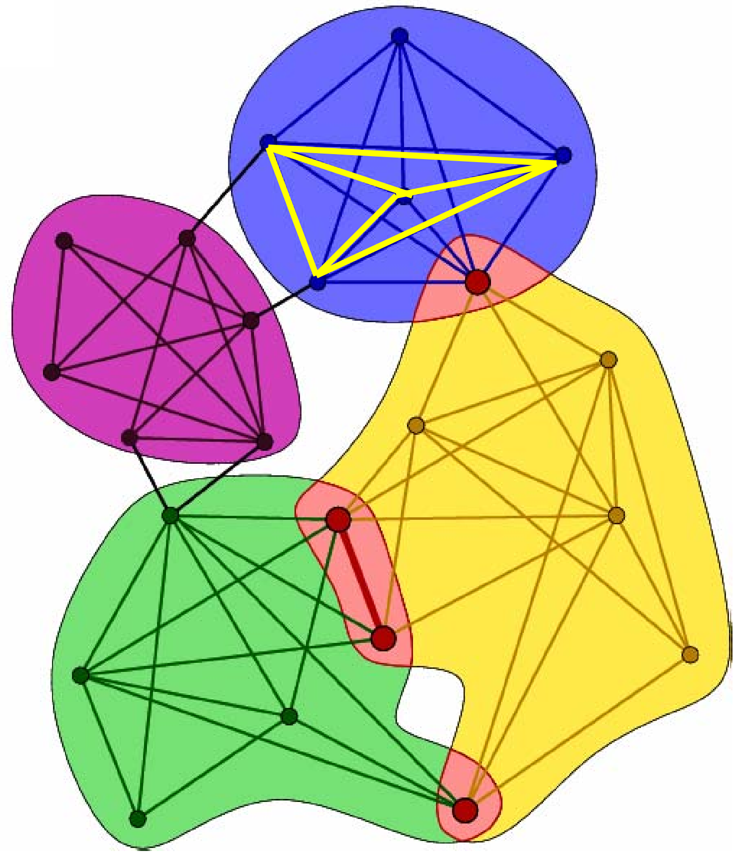
a 4-clique



Hierarchical methods



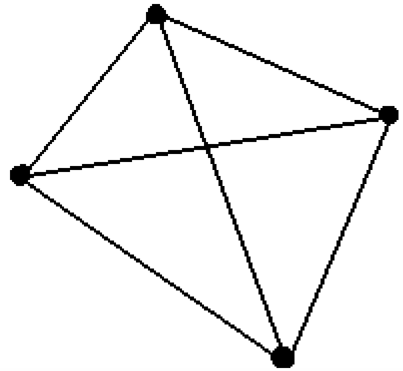
k-clique template rolling



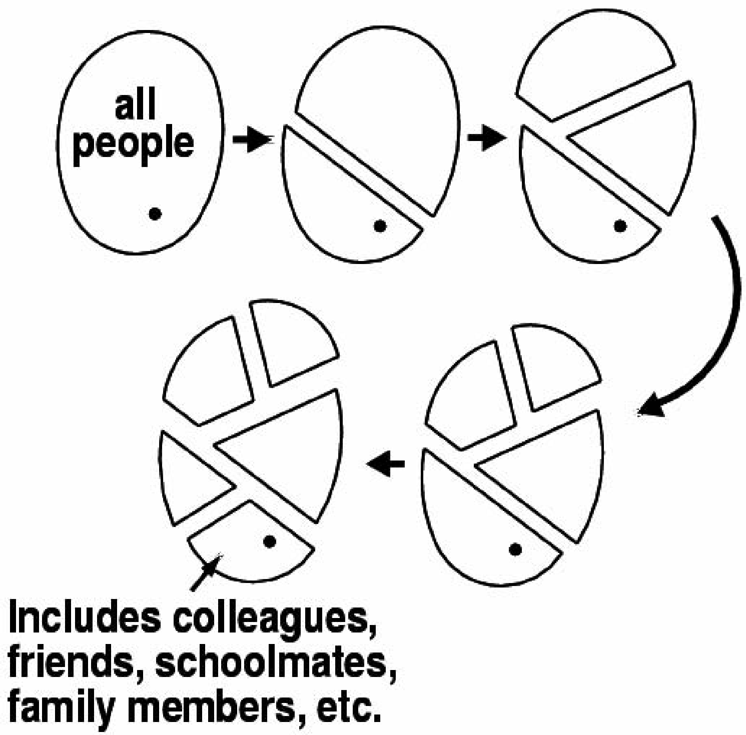
Two nodes belong to the same community if they can be connected through adjacent *k*-cliques

Finding communities

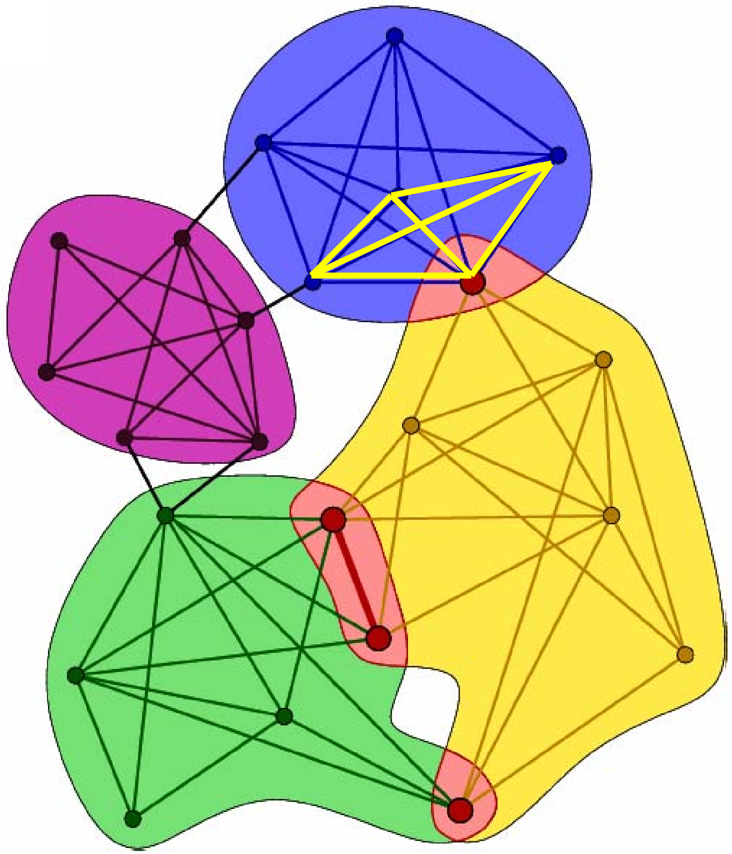
a 4-clique



Hierarchical methods



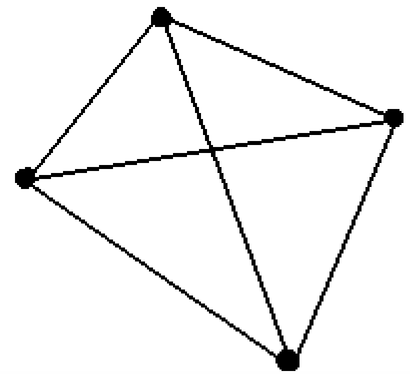
k-clique template rolling



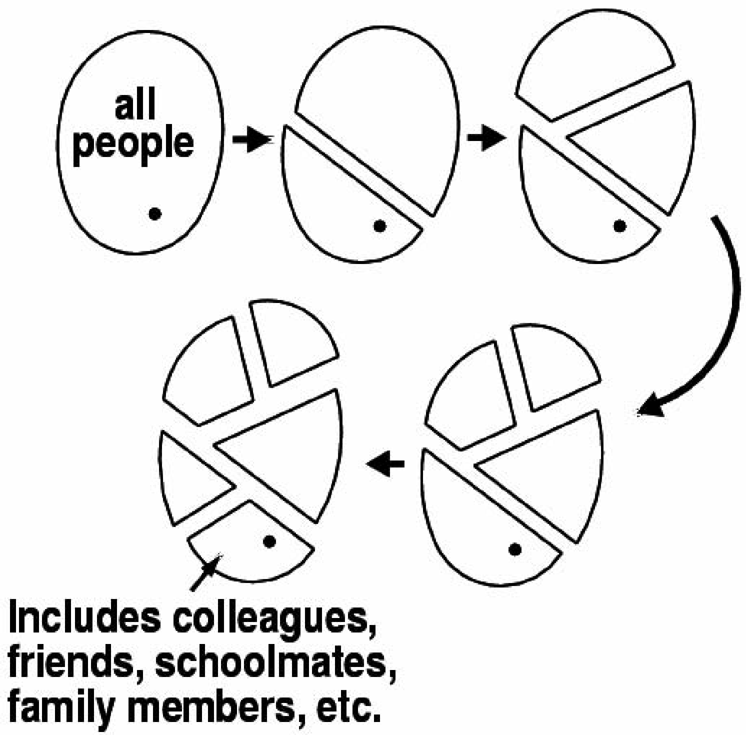
Two nodes belong to the same community if they can be connected through adjacent *k*-cliques

Finding communities

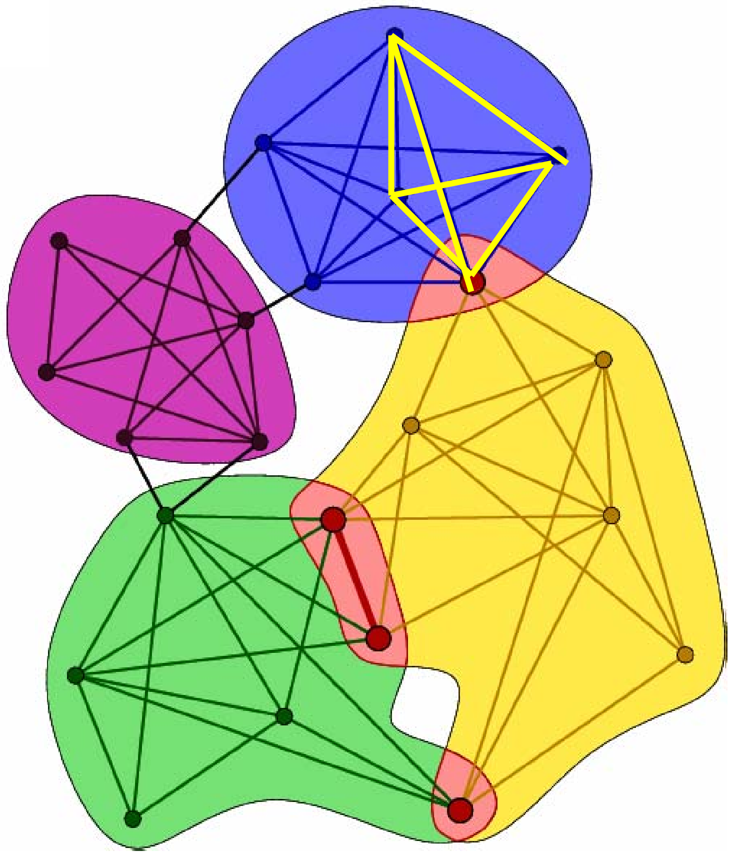
a 4-clique



Hierarchical methods



***k*-clique template rolling**



Two nodes belong to the same community if they can be connected through adjacent *k*-cliques

Hierarchical versus template rolling clustering

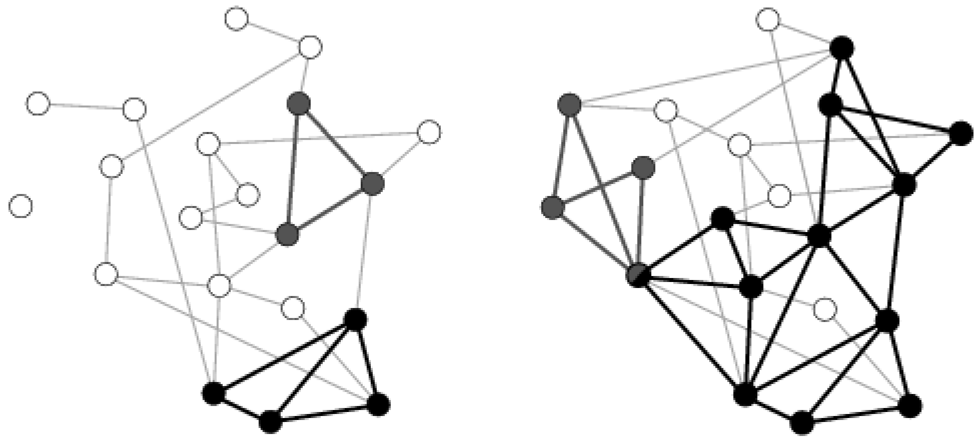
Common clustering methods lead to a partitioning in which someone (a node) can belong to a single community at a time only.

For example, I can be located as a member of the community “physicists”, but not, at the same time, be found as a member of my community “family” or “friends”, etc.

k-clique template rolling allows large scale, systematic (deterministic) analysis of the network of overlapping communities

k-CLIQUE PERCOLATION

with I. Derényi and G. Palla



Definitions

***k*-clique**: complete subgraph of *k* vertices

***k*-clique adjacency**: two *k*-cliques share a *k*-1 – clique

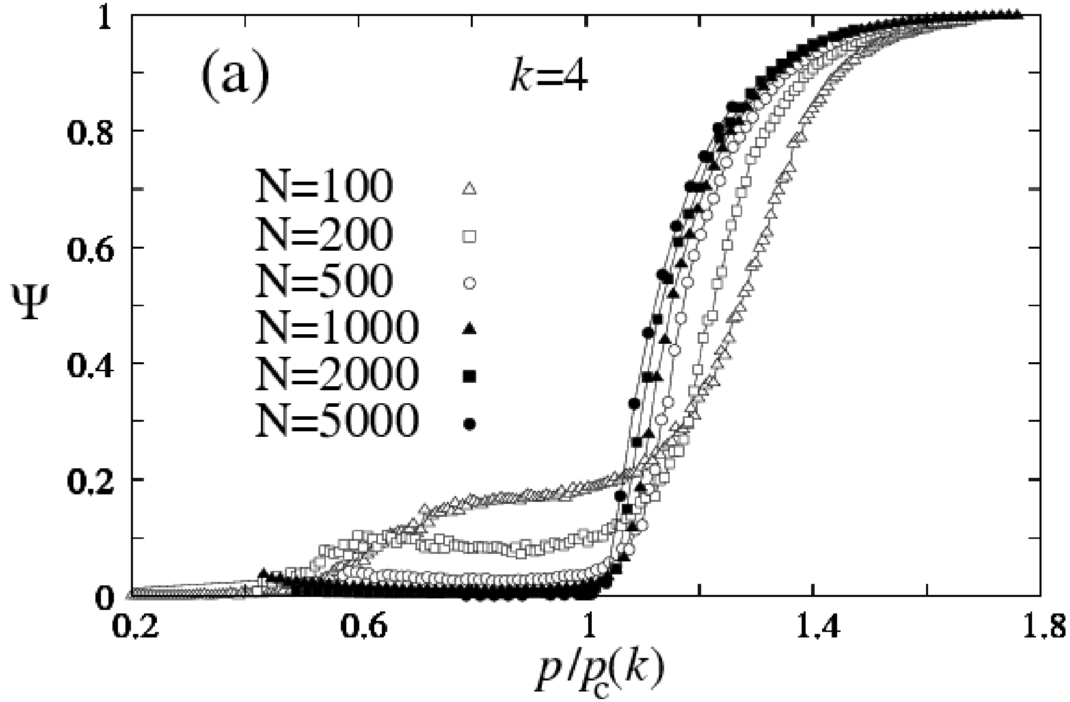
***k*-clique walk**: series of steps to adjacent *k*-cliques

***k*-clique cluster**: set of vertices of all *k*-clique walks from a given *k*-clique

(E-R percolation is the *k=2* case)

Details

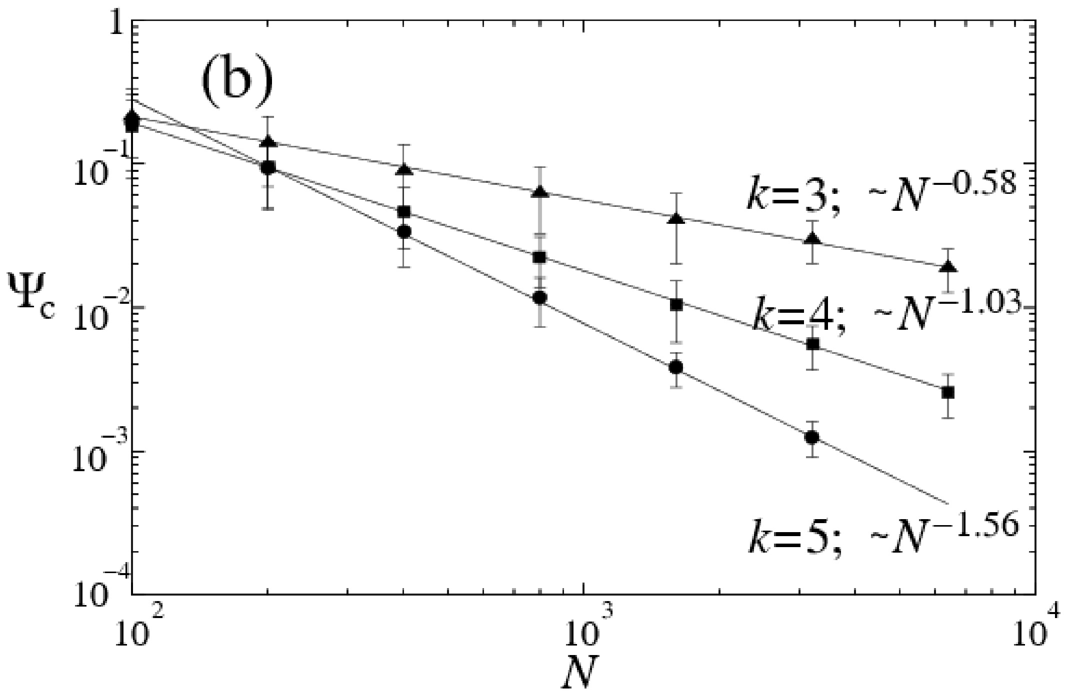
I. D, G. P. and T.V., *Phys. Rev. Lett*. 2005



Order parameter for clique percolation, k=4

Percolation threshold at

$p_c(k) = [(k\text{-}1)N]^{(-1/(k-1))}$



The scaling of the relative size of the giant cluster of *k*=3,4 and 5-cliques at p_c

For $k \leq 3$, $N_k^*/N_k(p_c) \sim N^{-k/6}$

For $k > 3$ $N_k^*/N_k(p_c) \sim N^{1-k/2}$

UNCOVERING THE OVERLAPPING COMMUNITY STRUCTURE OF COMPLEX NETWORKS IN NATURE AND SOCIETY

with G. Palla, I. Derényi, and I. Farkas

Definitions

An order *k* community is a *k*-clique percolation cluster

Such communities/clusters obviously can overlap

This is why a lot of new interesting questions can be posed

New fundamental quantities (cumulative distributions) defined:

$P(d^{com})$	community degree distribution
$P(m)$	membership number distribution
$P(s^{ov})$	community overlap distribution
$P(s)$	community size distribution (not new)

G.P,I.D,I.F,T.V *Nature* 2005

DATA

cond-mat (electronic preprints, about 30,000 authors)

protein-protein (DIP database, yeast, 2,600 nodes)

word association (sets of words associated with given words, questionnaire, 10,600 words)

mobile phone (~ 4,000,000 users calling each other)

school friendship (84 schools from USA)

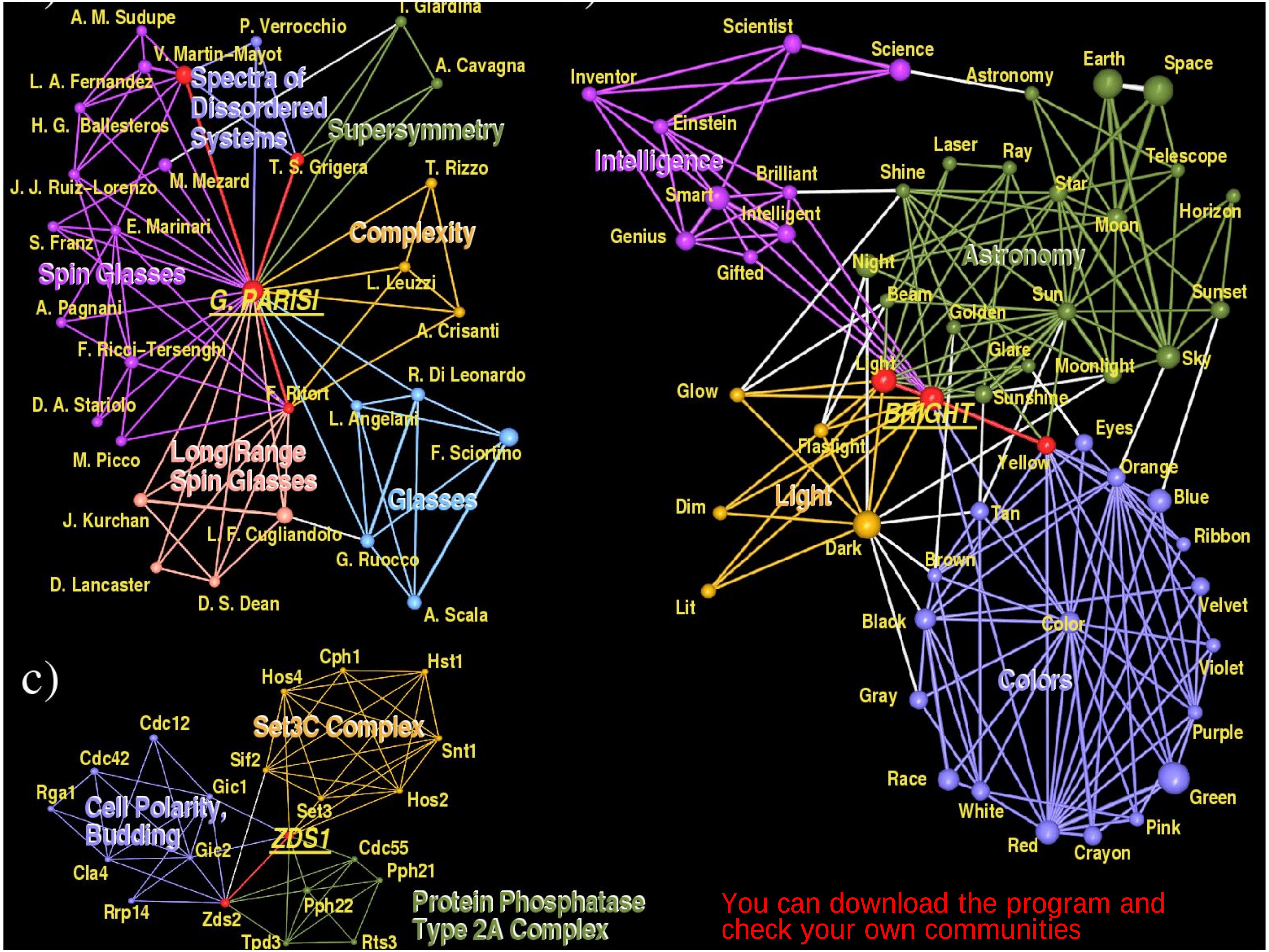
large data sets: efficient algorithm is needed! Our method is the fastest known to us for these type of data

Steps:

determine: cliques (not *k*-cliques!)
clique overlap matrix
components of the corresponding adjacency matrix

Do this for "optimal" *k* and *w*, where optimal corresponds to the "richest" (most widely distributed cluster sizes) community structure

Method


A. M. Sudupe
P. Verrocchio
I. Giardina
V. Martin-Mayot
L. A. Fernandez
Spectra of Dissordered Systems
A. Cavagna
H. G. Ballesteros
Supersymmetry
T. S. Grigera
T. Rizzo
J. J. Ruiz-Lorenzo
M. Mezard
E. Marinari
Complexity
S. Franz
Spin Glasses
L. Leuzzi
G. PARISI
A. Pagnani
A. Crisanti
F. Ricci-Tersenghi
R. Di Leonardo
F. Ritort
D. A. Stariolo
L. Angelani
M. Picco
Long Range Spin Glasses
F. Sciortino
Glasses
J. Kurchan
L. F. Cugliandolo
G. Ruocco
D. Lancaster
D. S. Dean
A. Scala
Scientist
Science
Inventor
Astronomy
Earth
Space
Einstein
Laser
Ray
Telescope
Intelligence
Brilliant
Shine
Star
Smart
Intelligent
Horizon
Moon
Genius
Night
Astronomy
Gifted
Beam
Sun
Sunset
Golden
Glare
Light
Moonlight
Sky
Glow
Sunshine
BRIGHT
Eyes
Flashlight
Yellow
Orange
Light
Blue
Dim
Tan
Dark
Ribbon
Brown
Lit
Velvet
Black
Color
Violet
Colors
Gray
Purple
Race
Green
White
Pink
Red
Crayon
c)
Cph1
Hos4
Hst1
Set3C Complex
Cdc12
Snt1
Cdc42
Sif2
Gic1
Rga1
Hos2
Cell Polarity, Budding
Set3
ZDS1
Gic2
Cdc55
Cla4
Pph21
Pph22
Protein Phosphatase Type 2A Complex
Rrp14
Zds2
Tpd3
Rts3
You can download the program and check your own communities

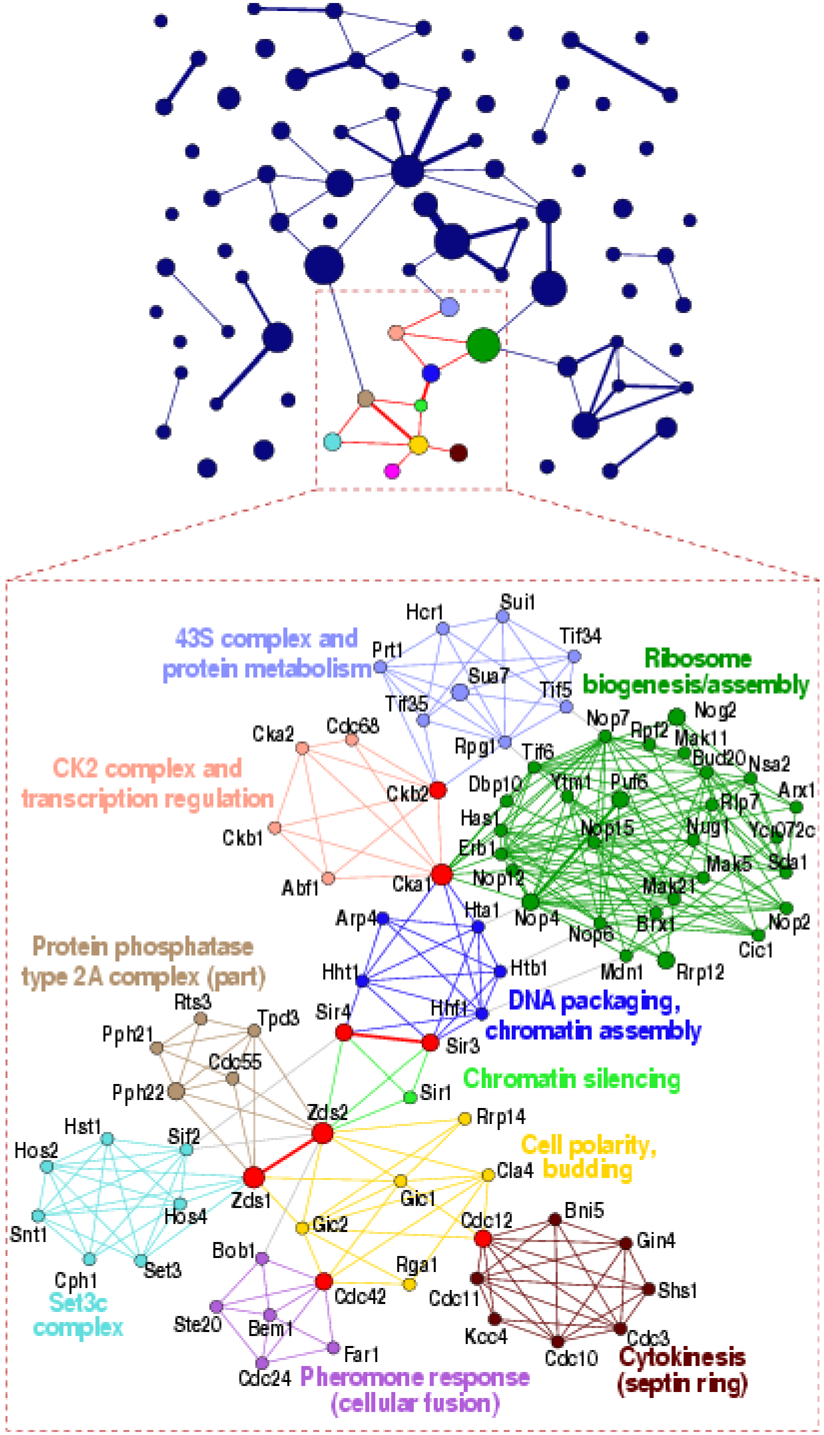
“Web of networks”

Each node is a community

Nodes are weighted for community size
Links are weighted for overlap size

DIP “core” data base of protein interactions (*S. cerevisiase*, a yeast)

The other networks we analysed are much larger!!

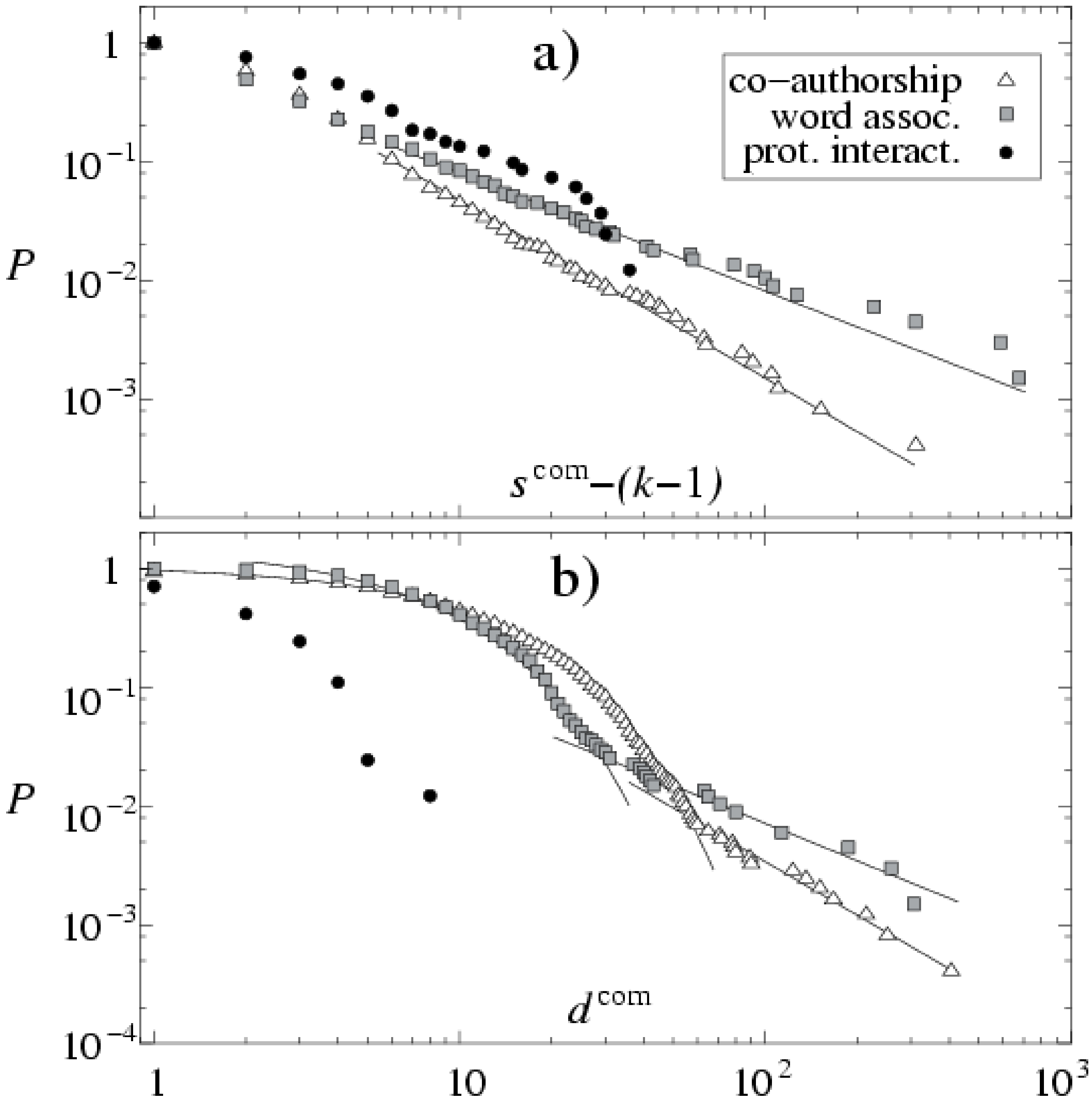


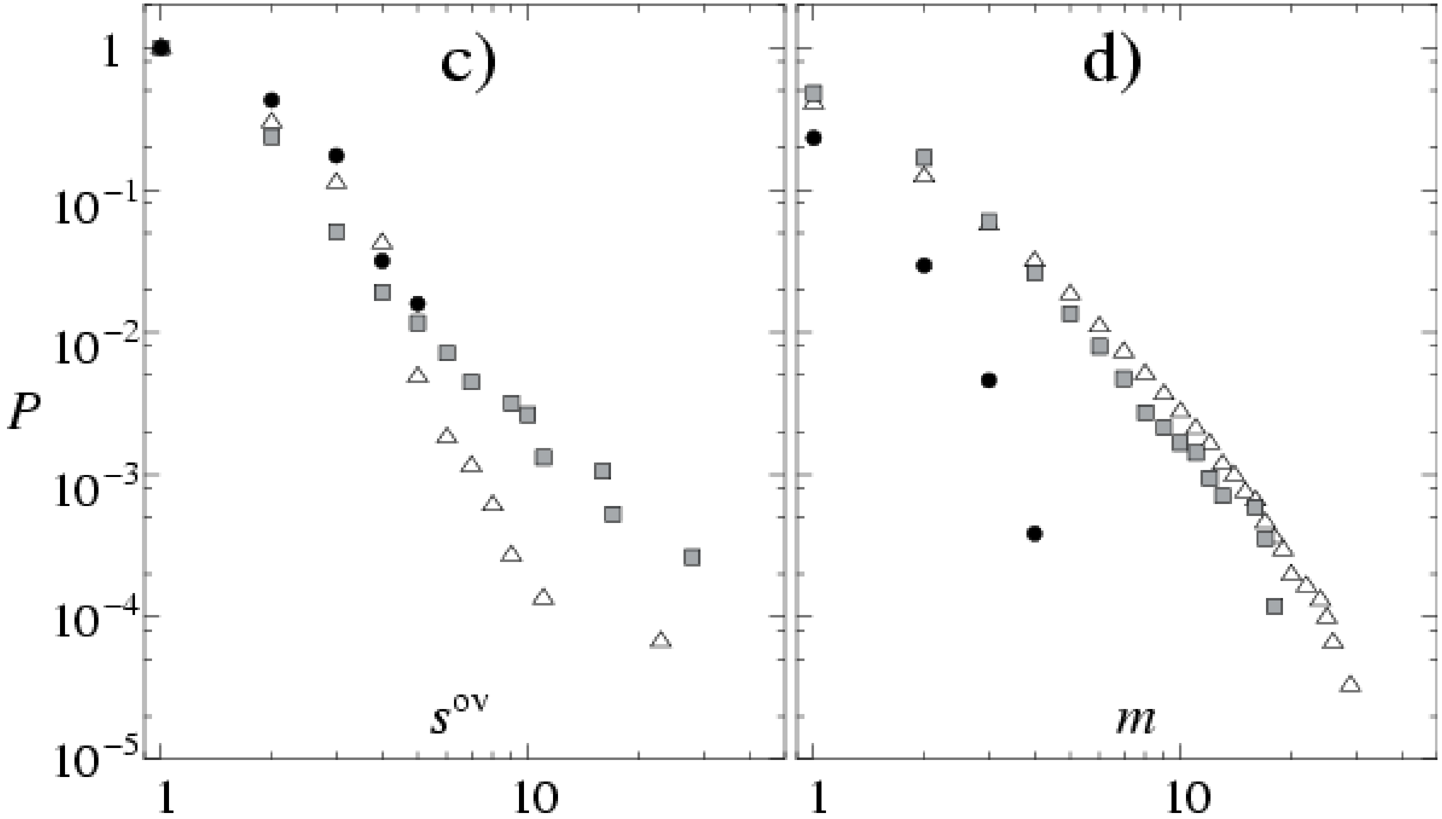
Community size distribution

Community degree distribution

Combination of exponential and power law!

Emergence of a new feature as going "up" to the next level





Community overlap size **membership number**

Case studies + dynamics

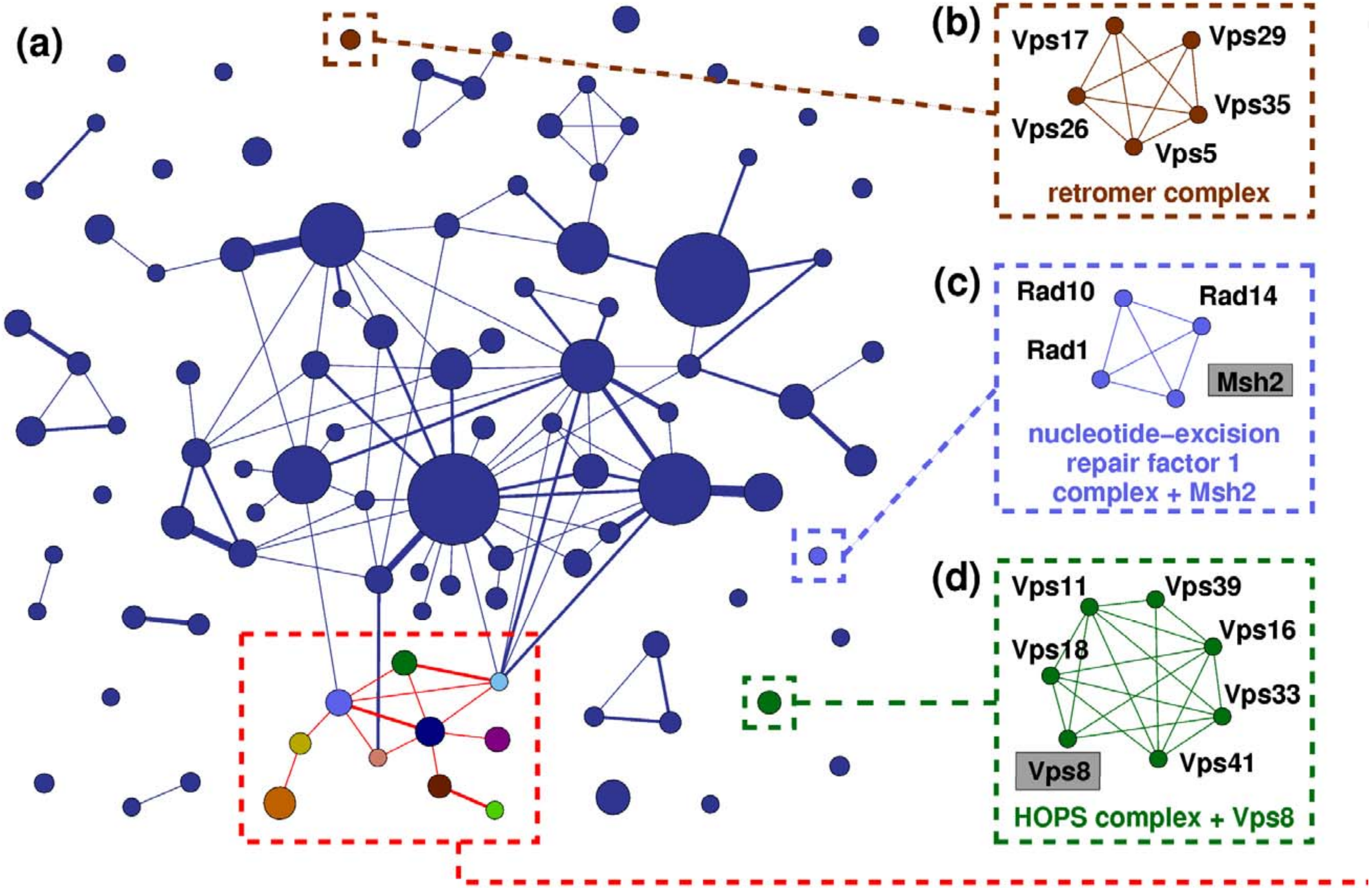
Protein interaction (prediction of function)

School friendship (disassortativity of communities, role of races)

Social group evolution in a co-authorship and
a mobile phone network

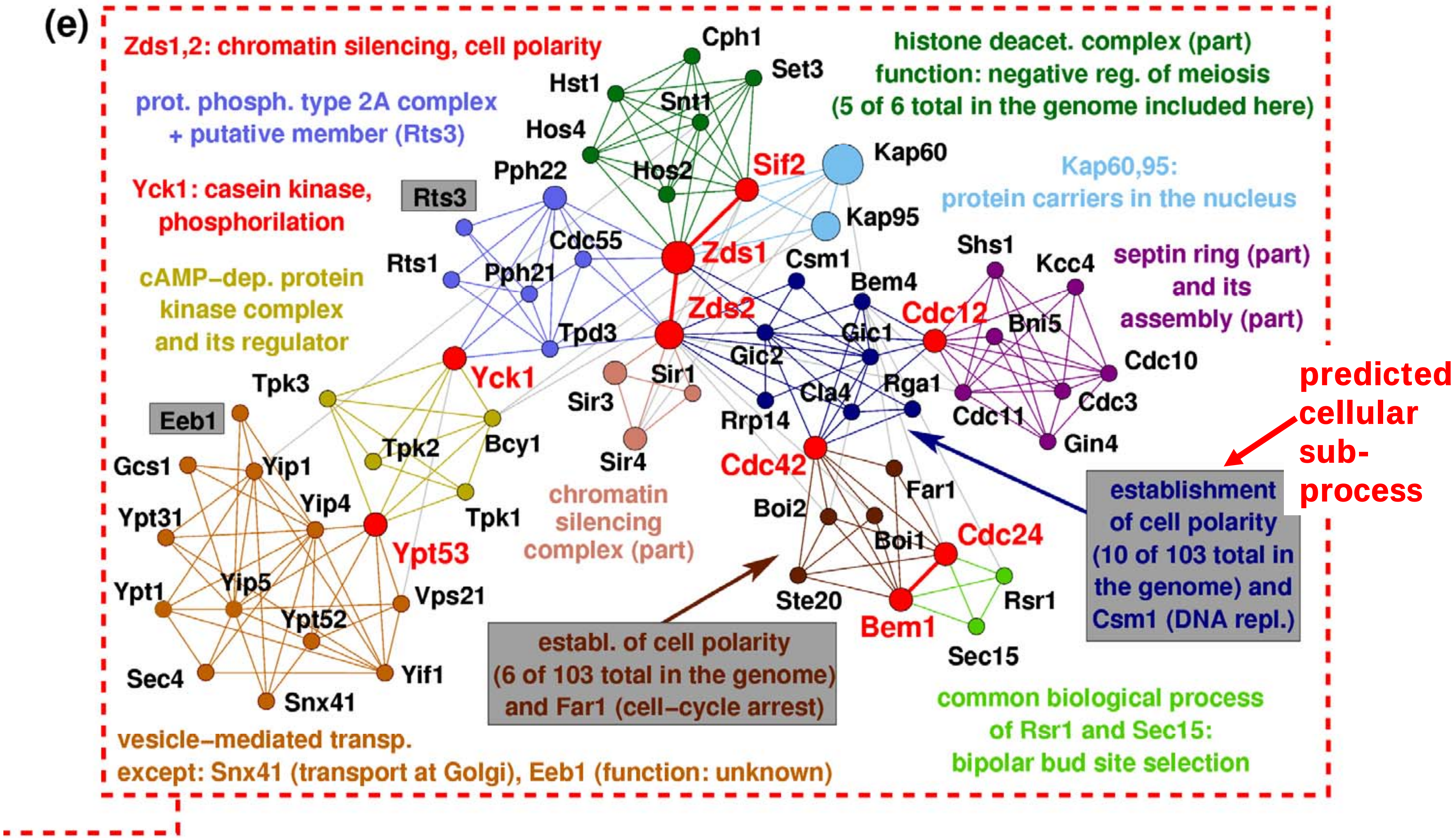
network of yeast PPI modules

node: module of proteins, link: overlap of modules



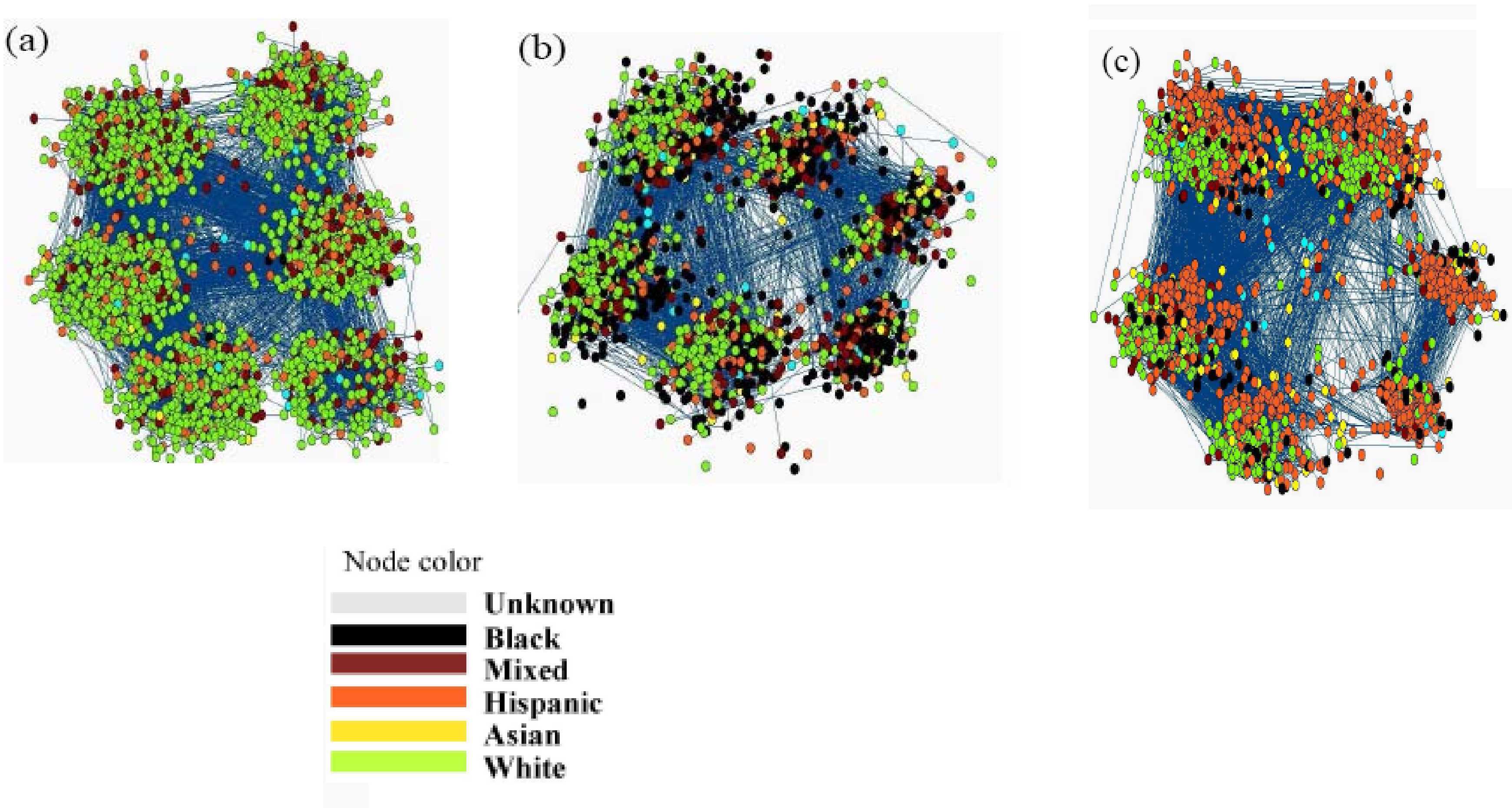
B.A,G.P,I.D,I.F,T.V.: *BIOINFROMATICS* 2006 enlarged portions of the network of modules **Marked:**

single proteins (function prediction) and groups (anticipated new modules)



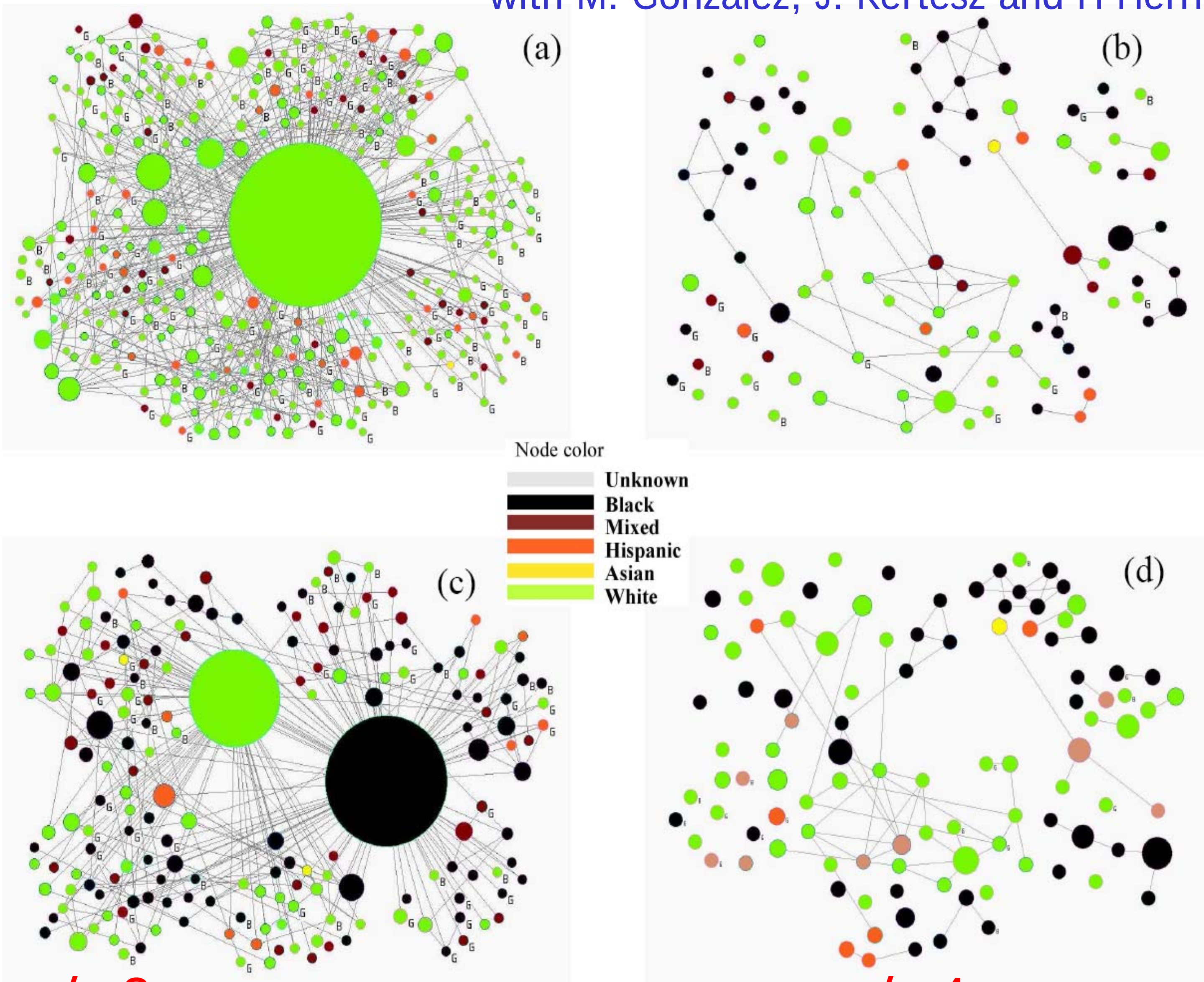
Three schools from the Add-Health school friendship data set

Grades 7-12



Network of school friendship communities

with M. Gonzalez, J. Kertész and H Herrmann

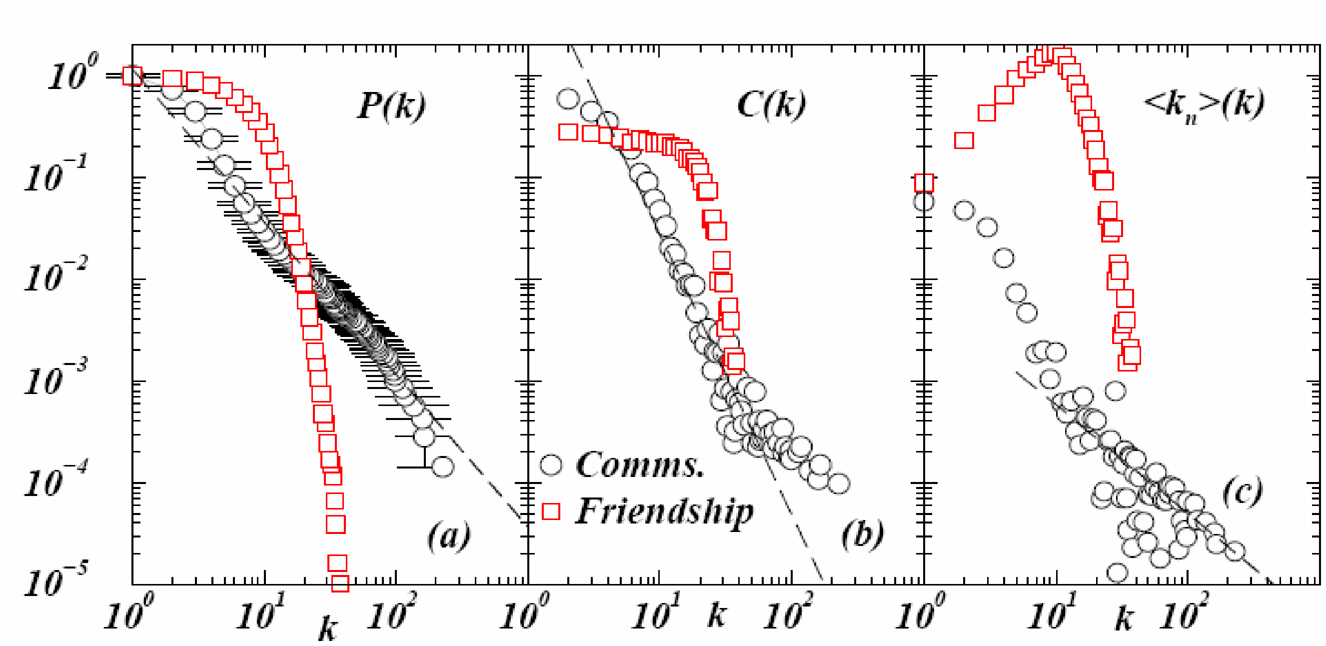


k=3 (looser) *k*=4 (more dense)

Minorities tend to form more densely interconnected groups

Distribution functions (for k=3)

○ communities □ individuals



P(k) – degree distribution

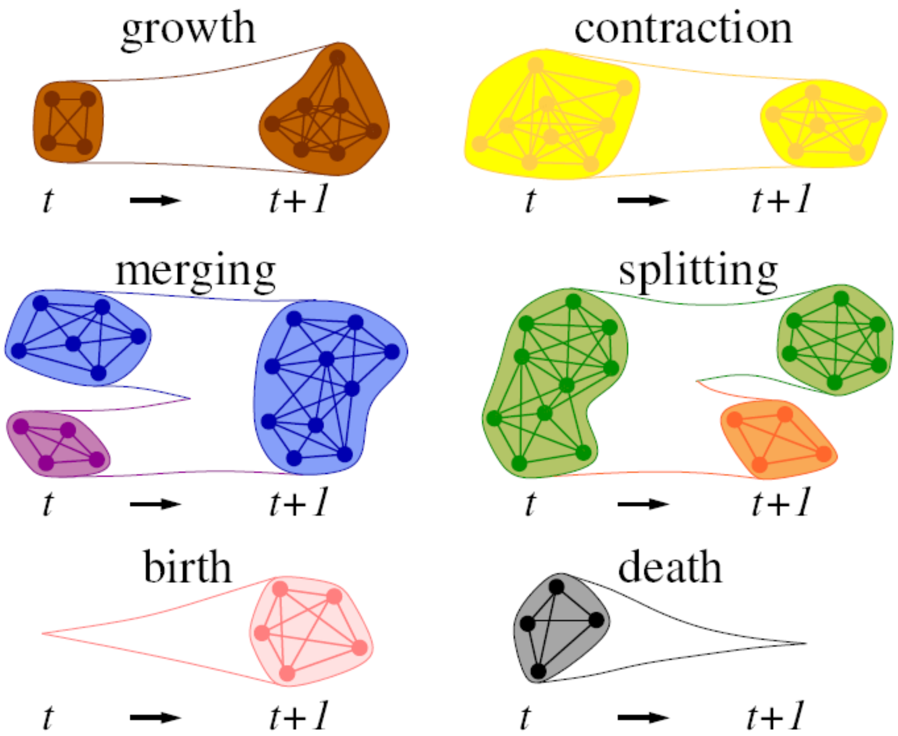
C(k) – clustering coefficient

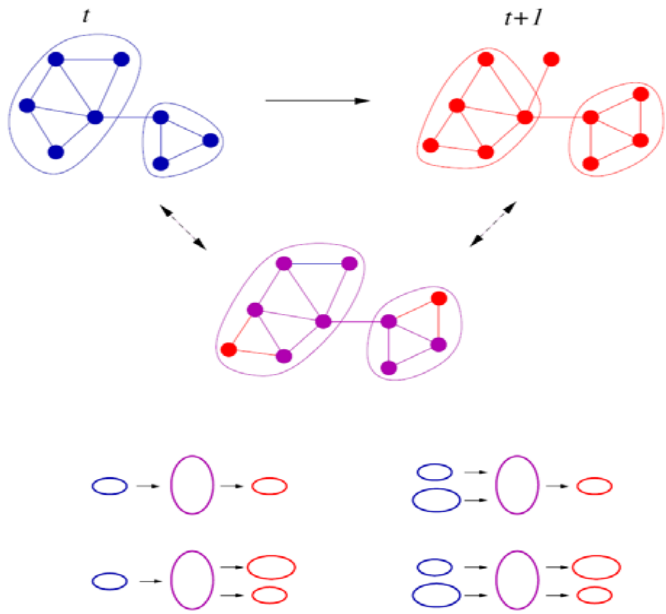
$<k_n>(k)$ – degree of neighbour (individuals: assortative

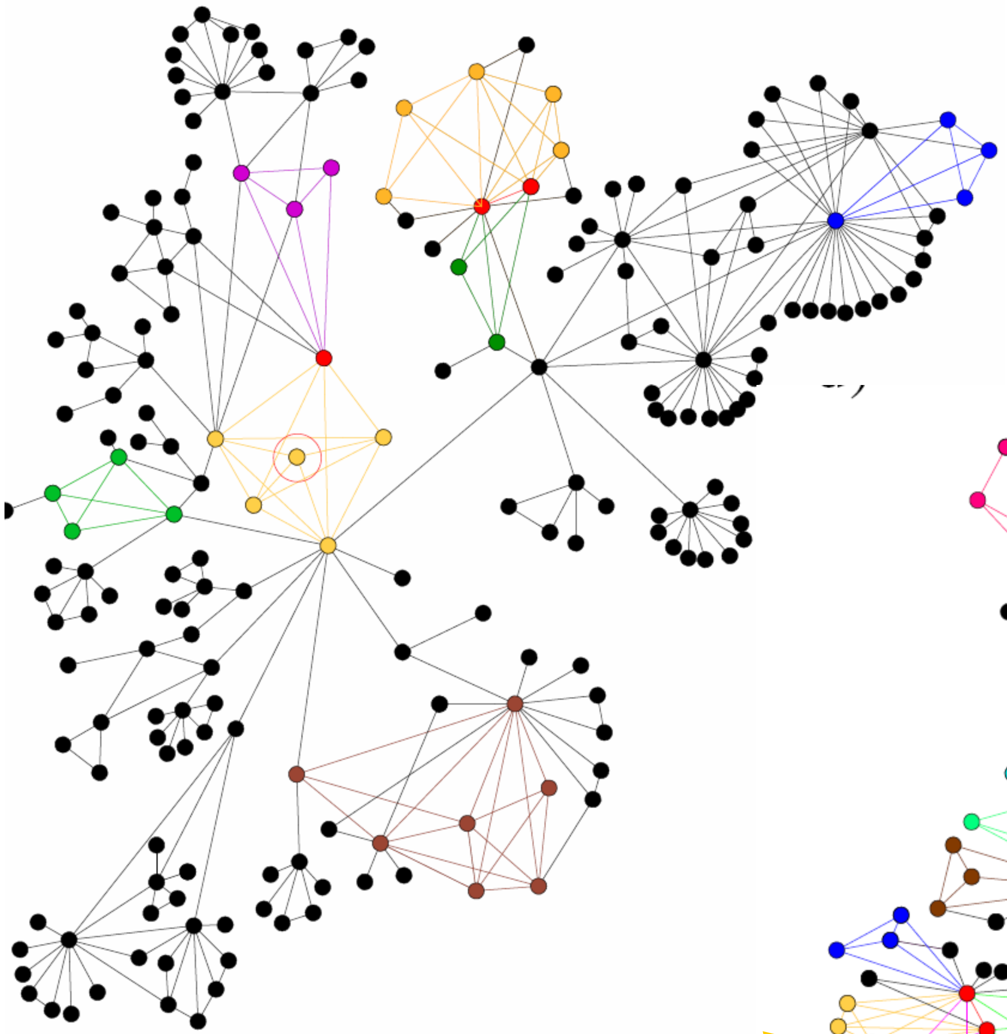
communities: diassortative)

Quantifying social group evolution

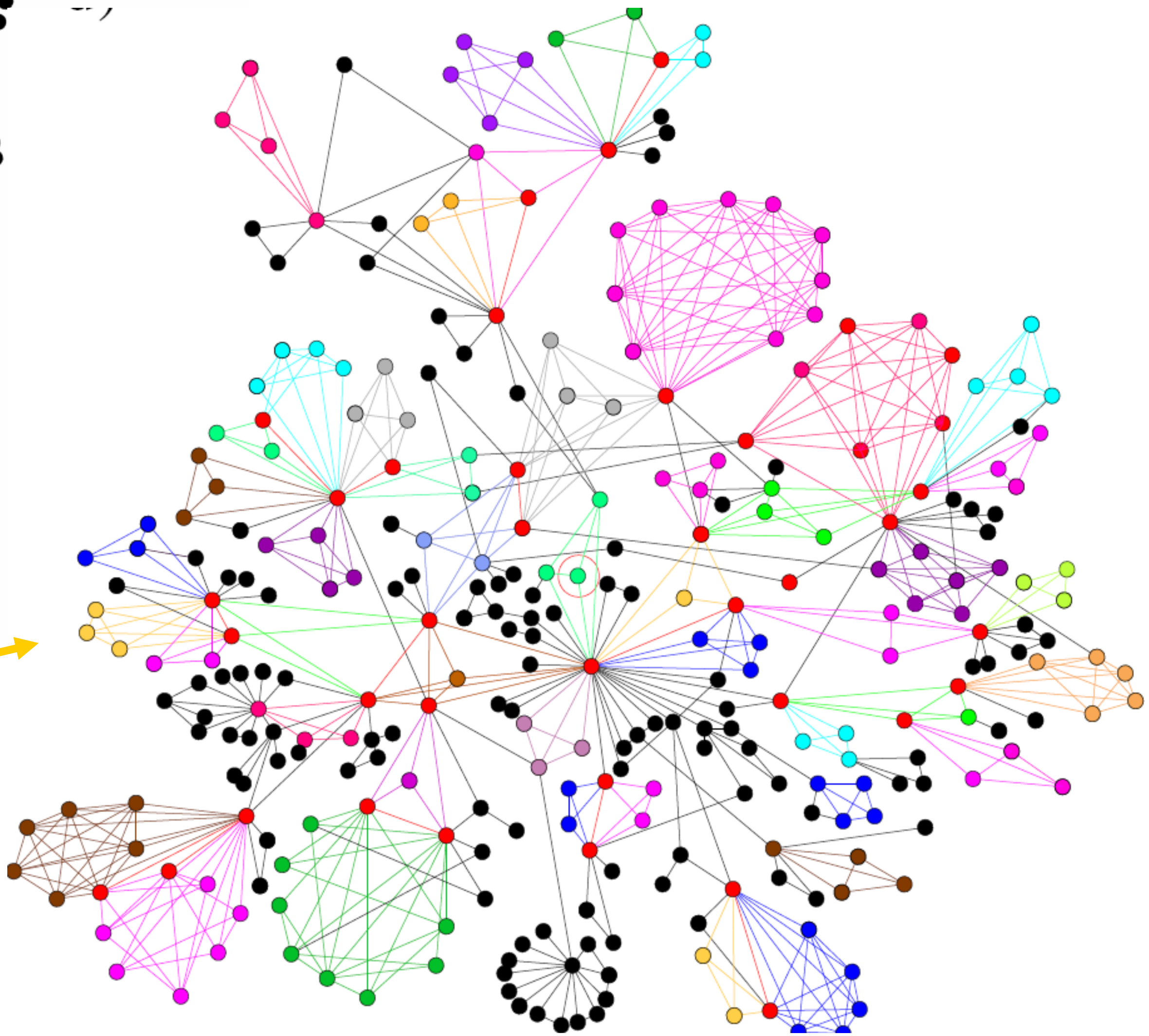
with G. Palla and A-L Barabási (*Nature*, April 2007)





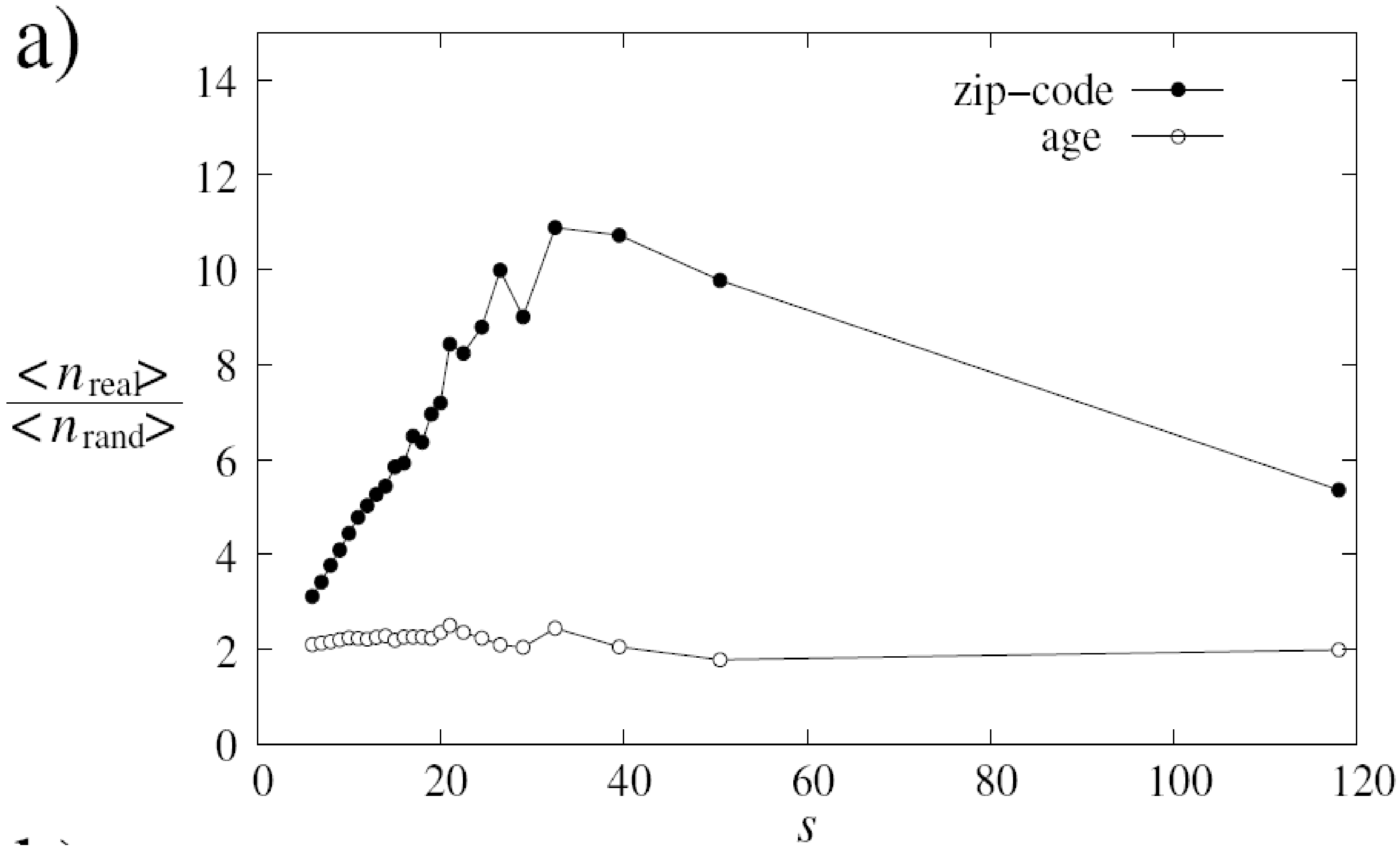


Small part of the phone call network (surrounding the circled yellow node up to the fourth neighbour)

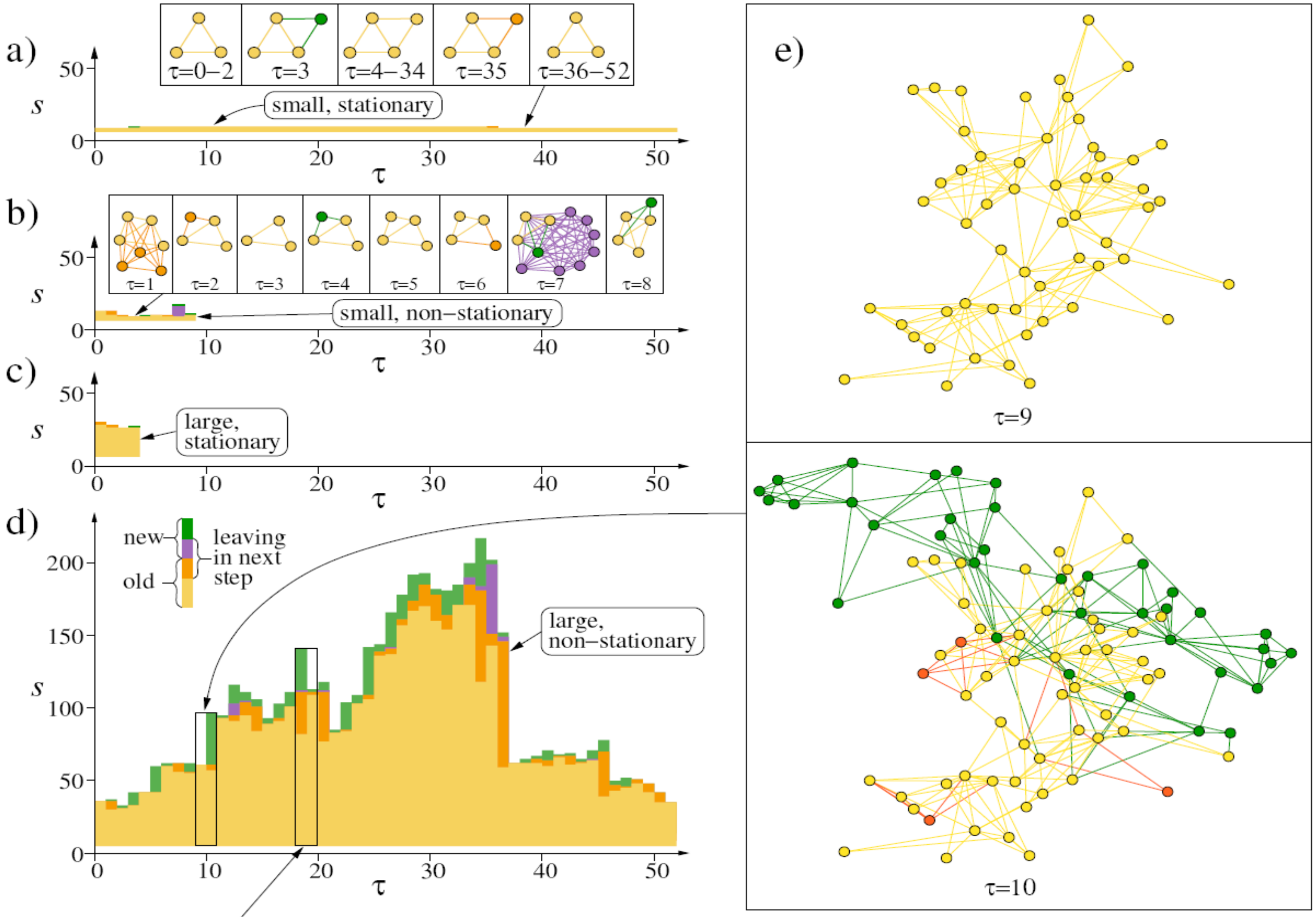


Small part of the collaboration network (surrounding the circled green node up to the fourth neighbour)

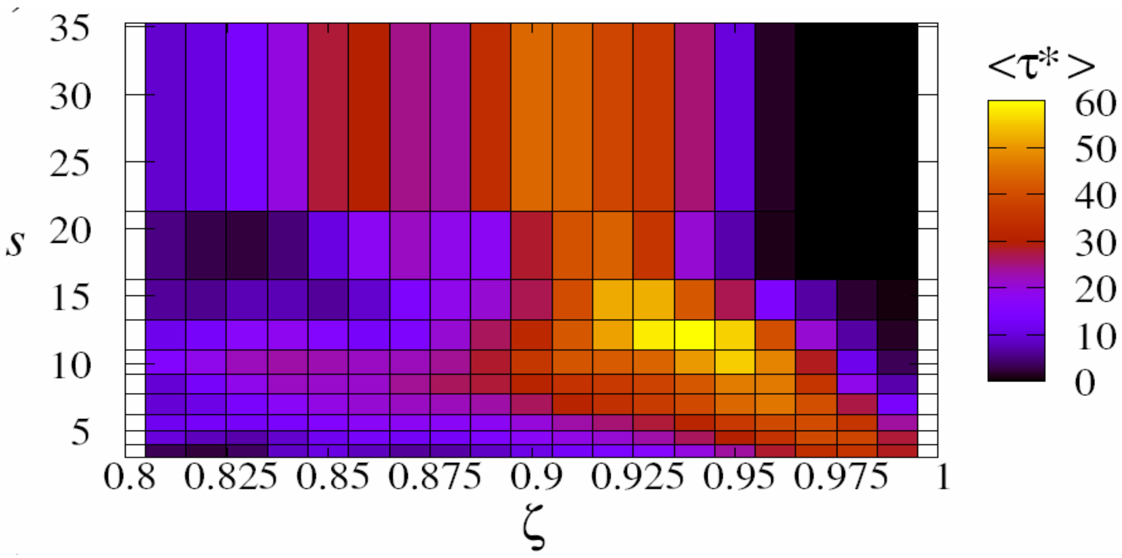
Callers with the same zip code or age
are over-represented in the communities we find



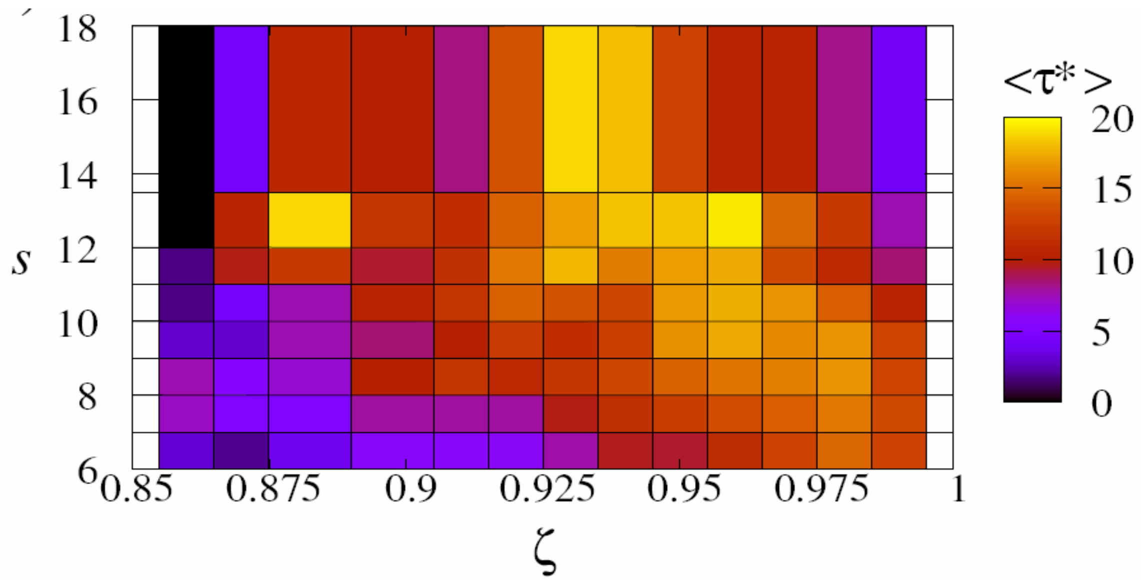
Examples for tracking individual communities.



Lifetime (τ) of a social group as a function of stability (steadiness, ζ) and size (s)



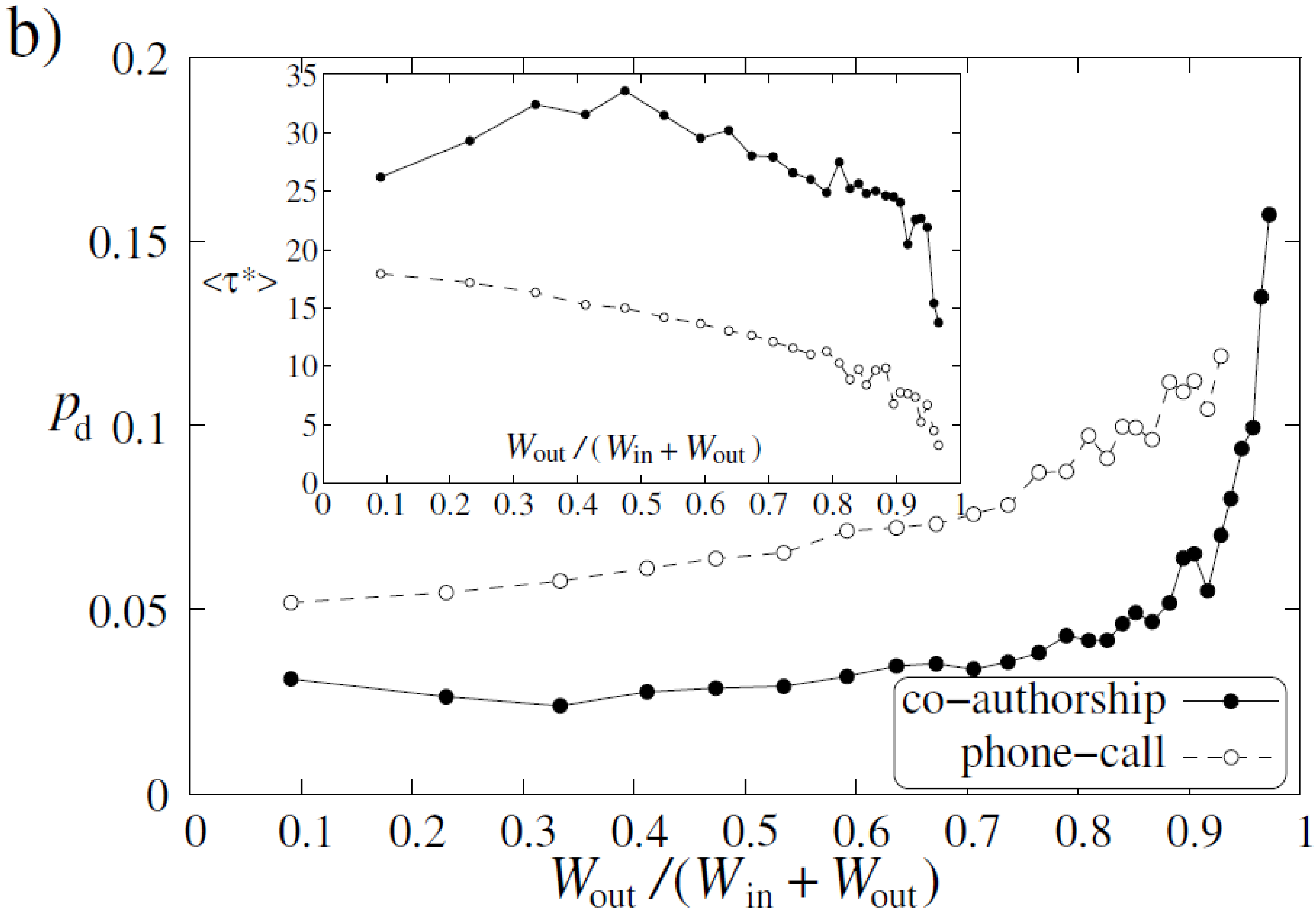
Cond-mat collaboration network



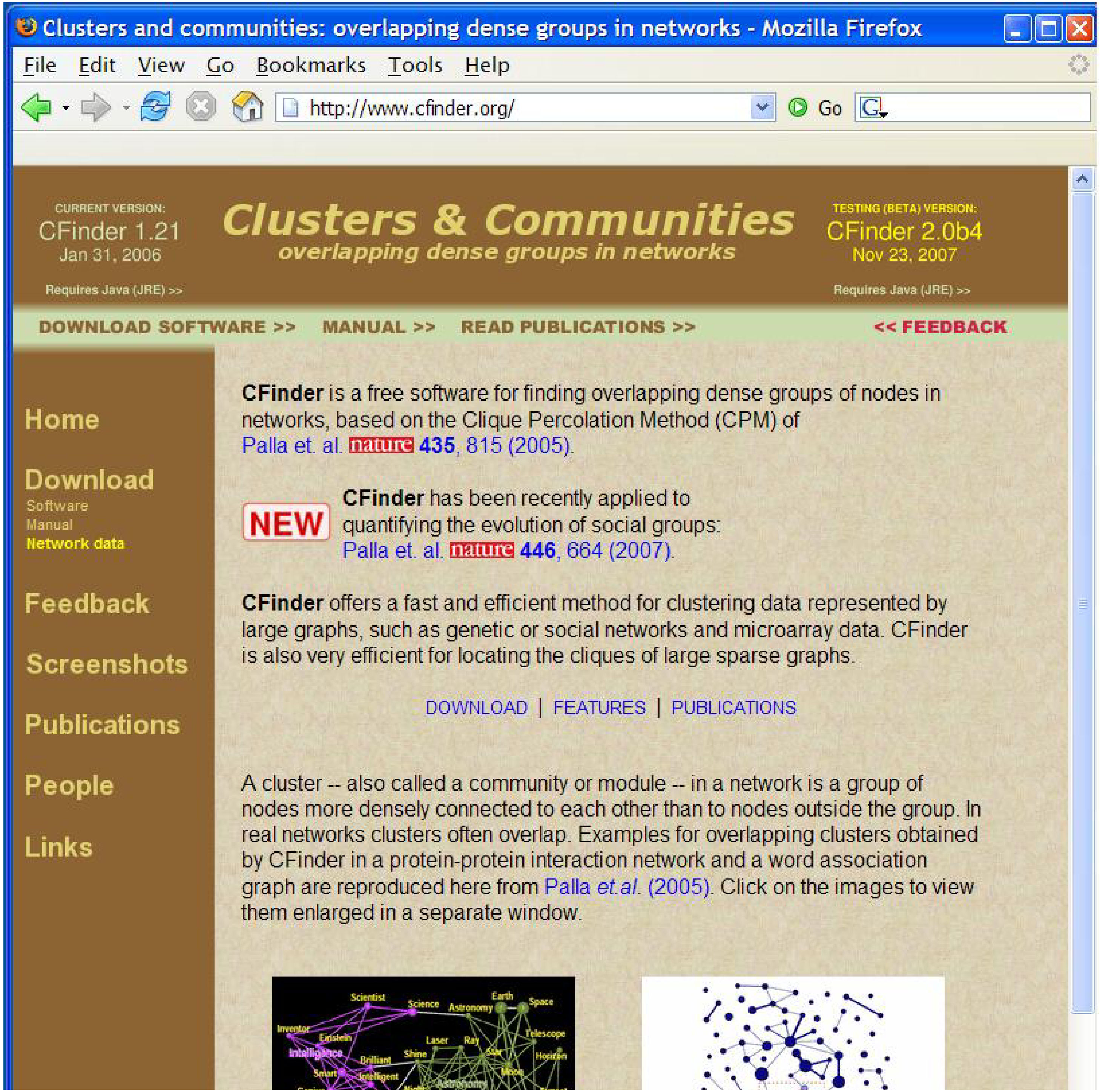
Phone call network

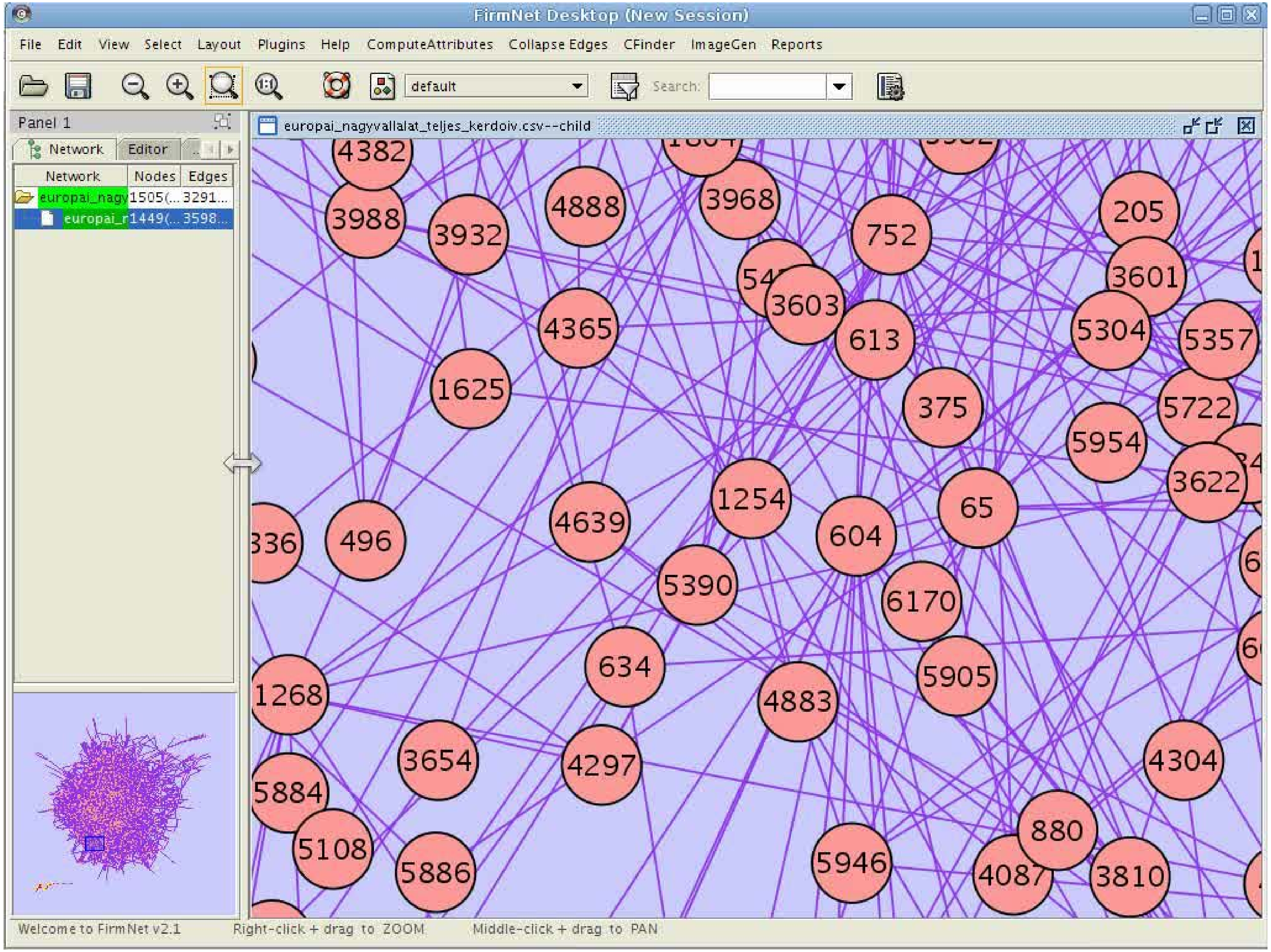
Thus, a large group is around longer if it is less steady (and the opposite is true for small groups)

Probability of disintegrating (p_d) and the lifetime (τ^*) of a community whose members have a total amount of “commitments” to other communities equal to W_{out}

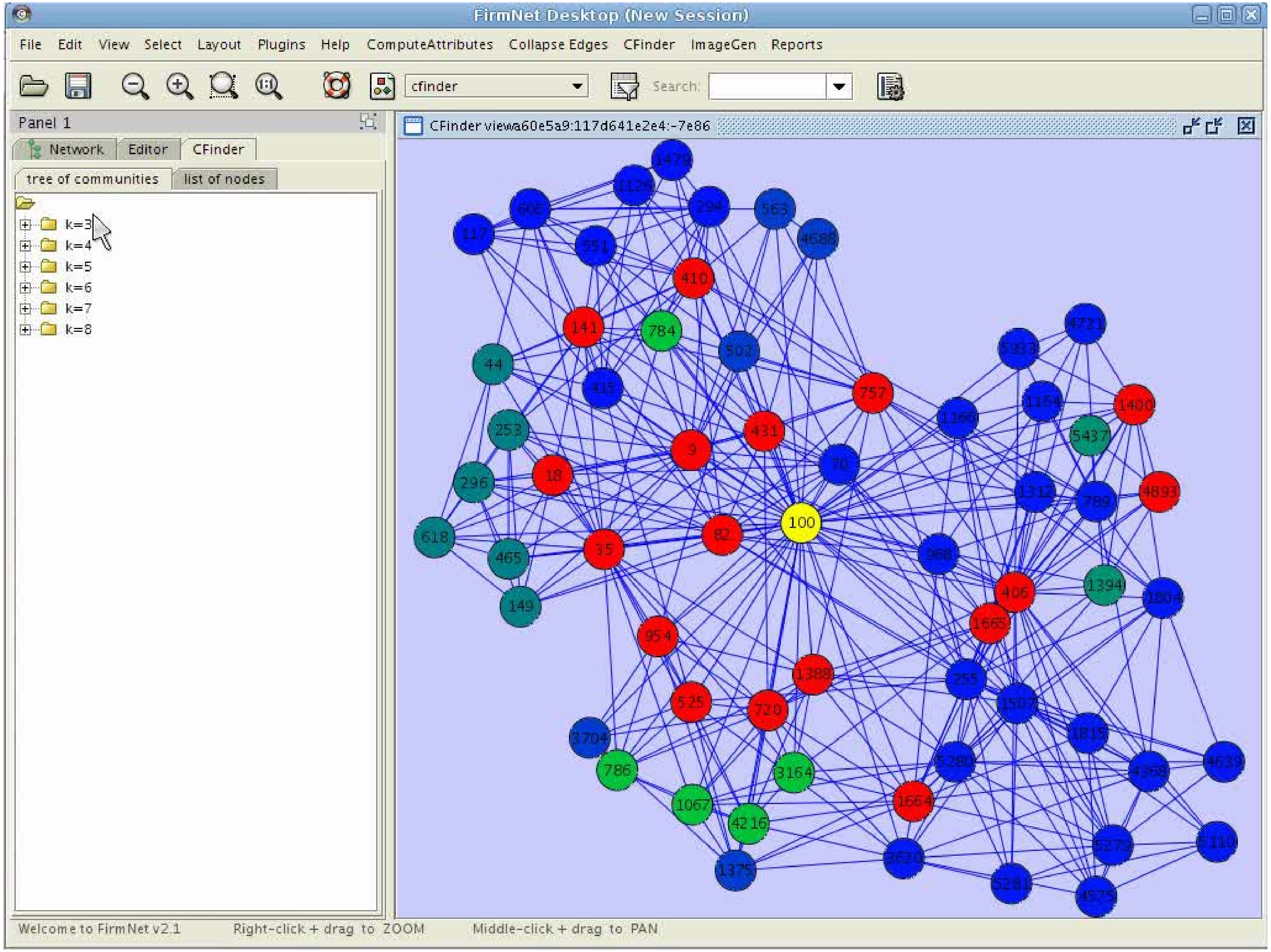


Home page of CFinder





Social network of the 3000 employees of an European company determined from an on-line survey. Visualization of the betweenness centrality



Visualization of the communities for the same company shown here using an adaptation of our CFinder-Firmnet software.

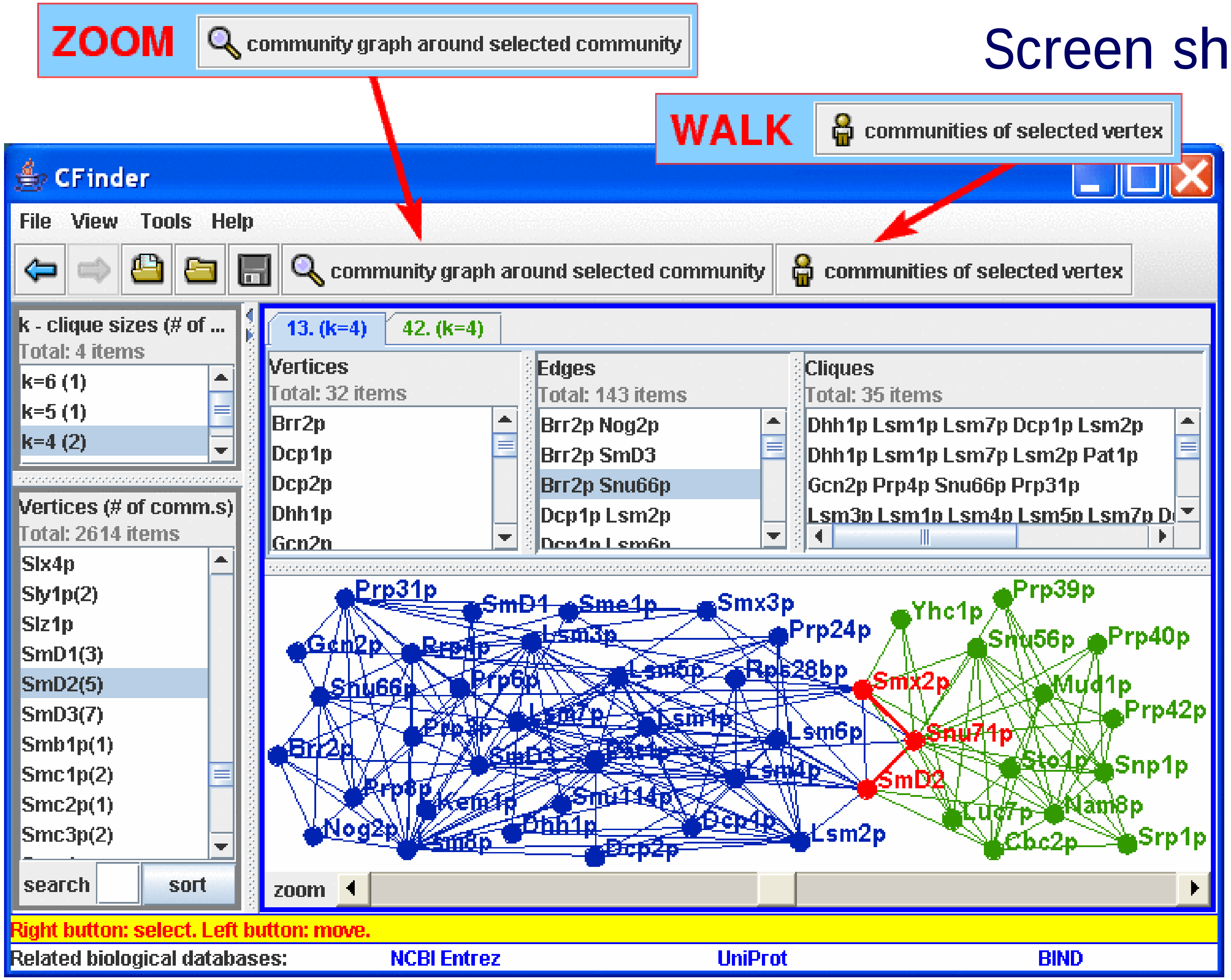
Theridion provides organizational development services based on network analysis.

Outlook:

Networks of networks

- hierarchical aspects
- correlations, clustering, etc.,
  i.e., everything you can do for vertices
- applications, such as protein function
  prediction or organizational development

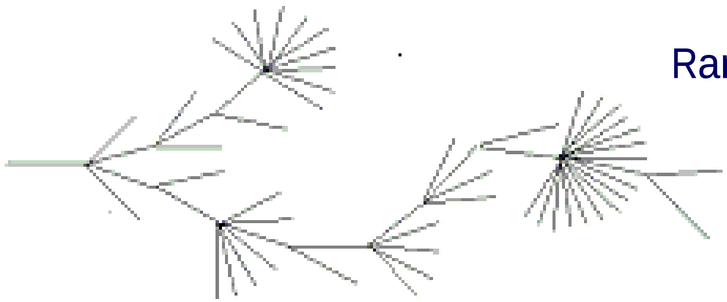
Screen shot of CFinder



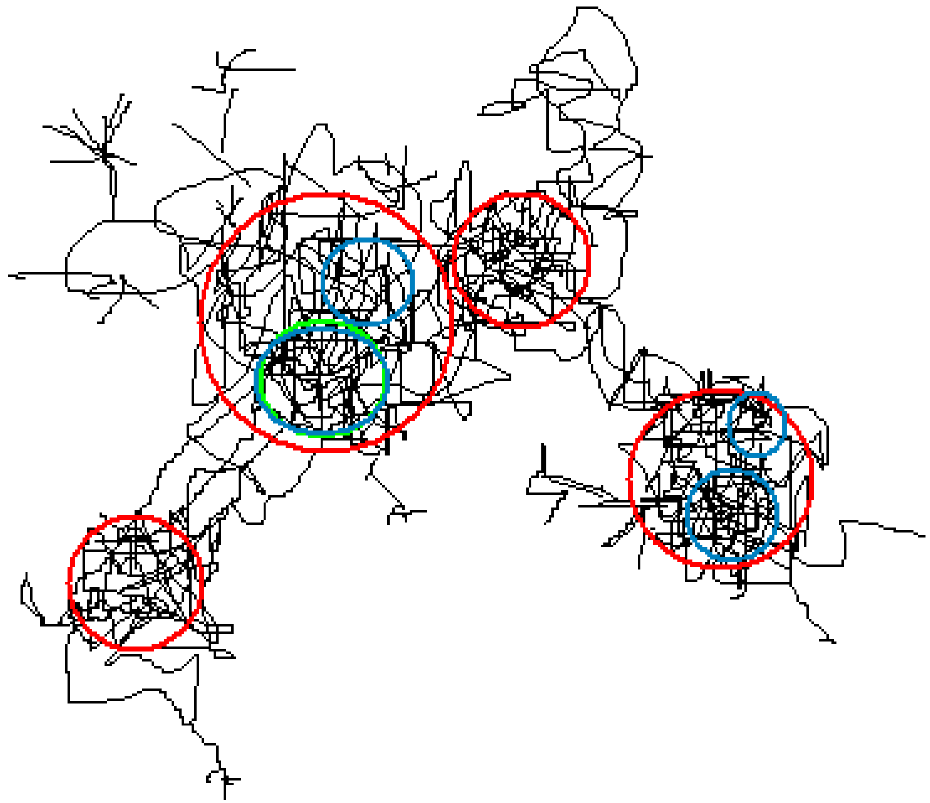
This will also become a commercial product by **Firmlinks** with **GORDIO**, a Budapest based HR company

Internal organization of large complex networks in terms of their modular structure

- Research on modules/communities is a very active field (Amaral, Barabási, Newman
- How does a large complex network may look like?   + many further groups)

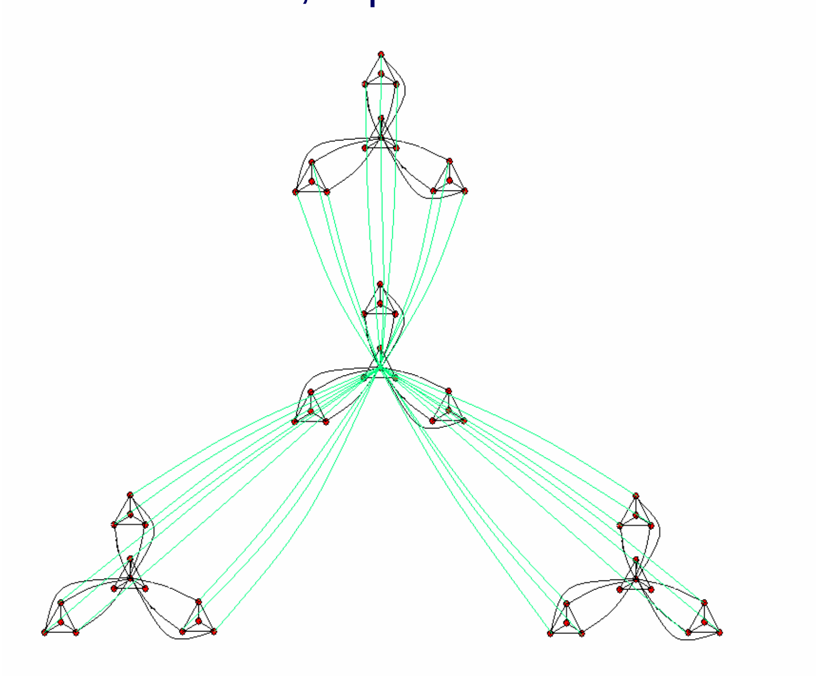


Random tree



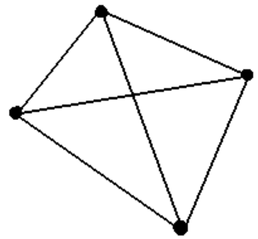
Random “blobs”

Deterministic,loops



To find overlapping communities we

consider: connected groups (clusters) of motifs e.g. a 4-clique



define: a cluster of adjacent complete subgraps (cliques) is a community (simple assumption)

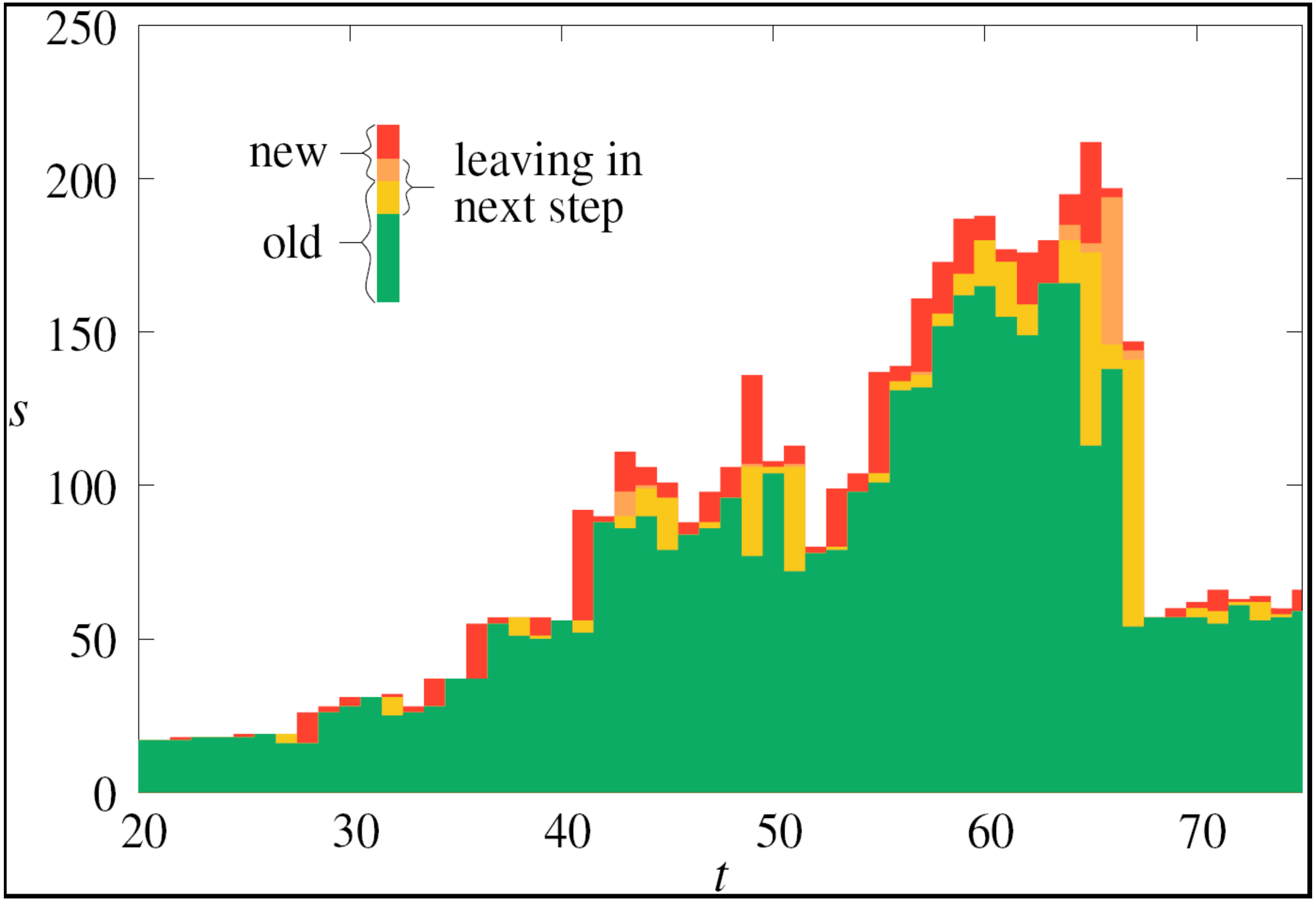
Two aspects

I) *k*-clique percolation

II) communities in large real networks: overlaps and their statistics

Evolution of a single large community of collaborators

s – size (number of authors), t – time (in months)



Dedicated home page (software, papers, data)

http://angel.elte.hu/clustering/

Home

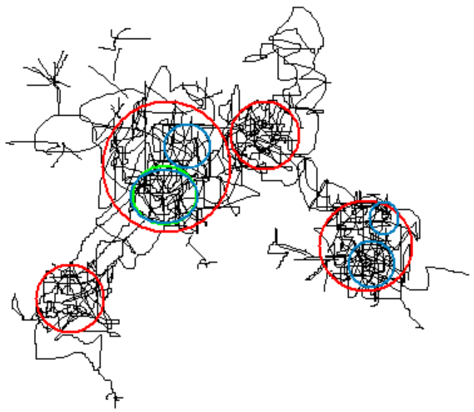
Screen shots

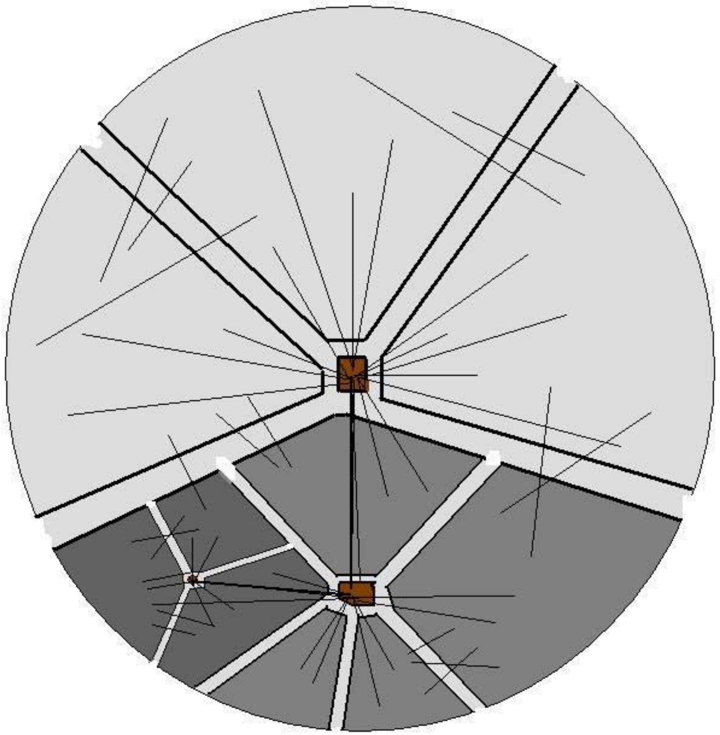
Basic observations:

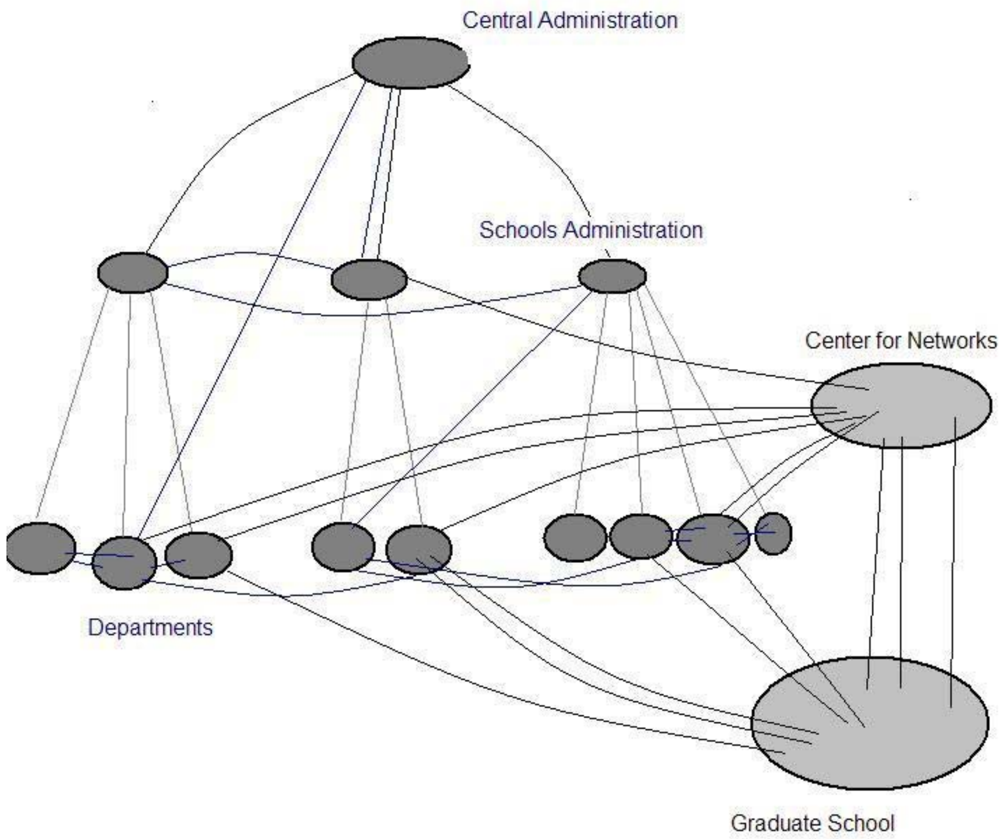
A large complex network is bounded to be highly structured
(has modules; function follows from structure)

The internal organization is typically hierarchical
(i.e., displays some sort of self-similarity of the structure)

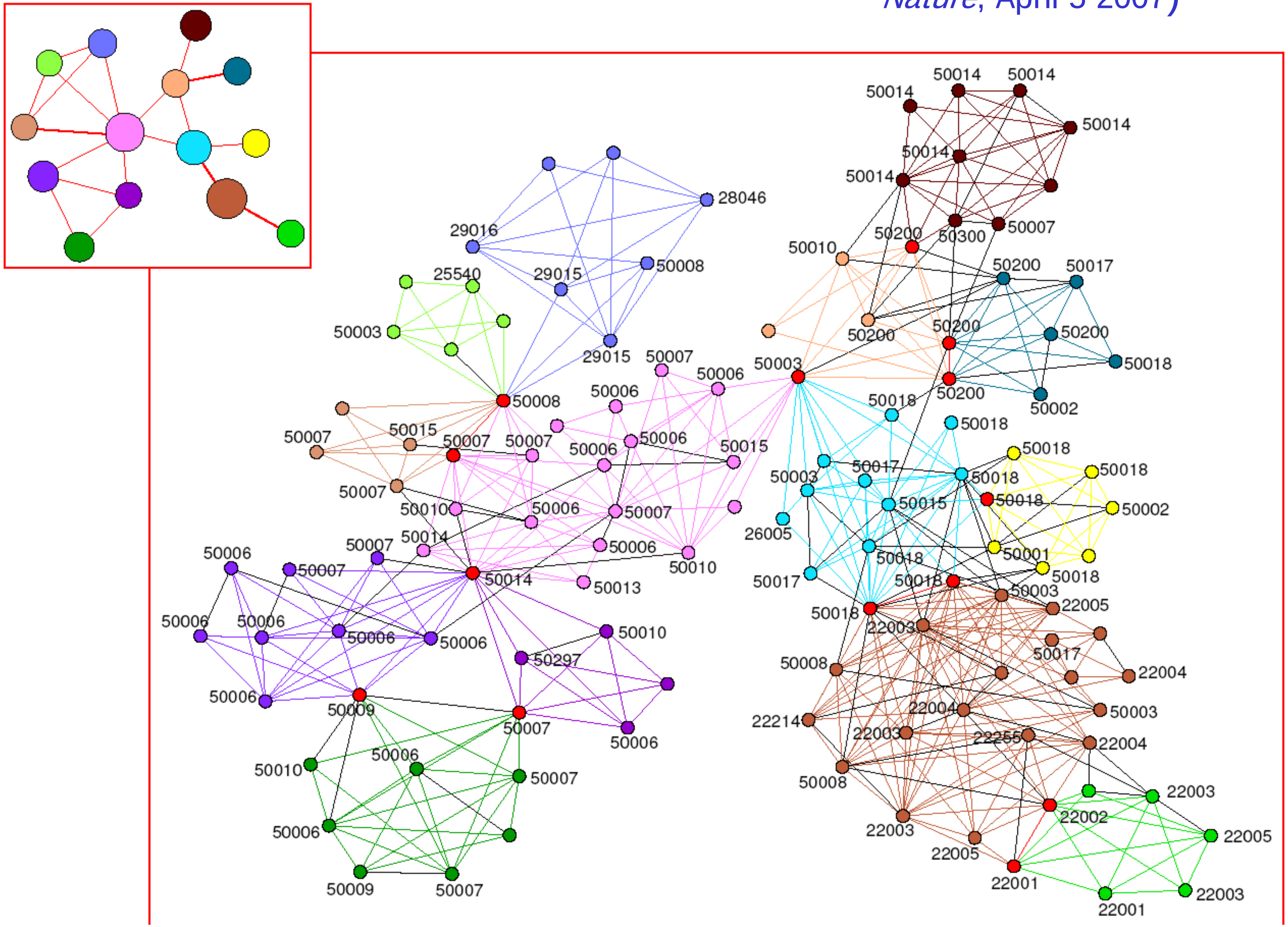
An important new aspect: Overlaps of modules are essential

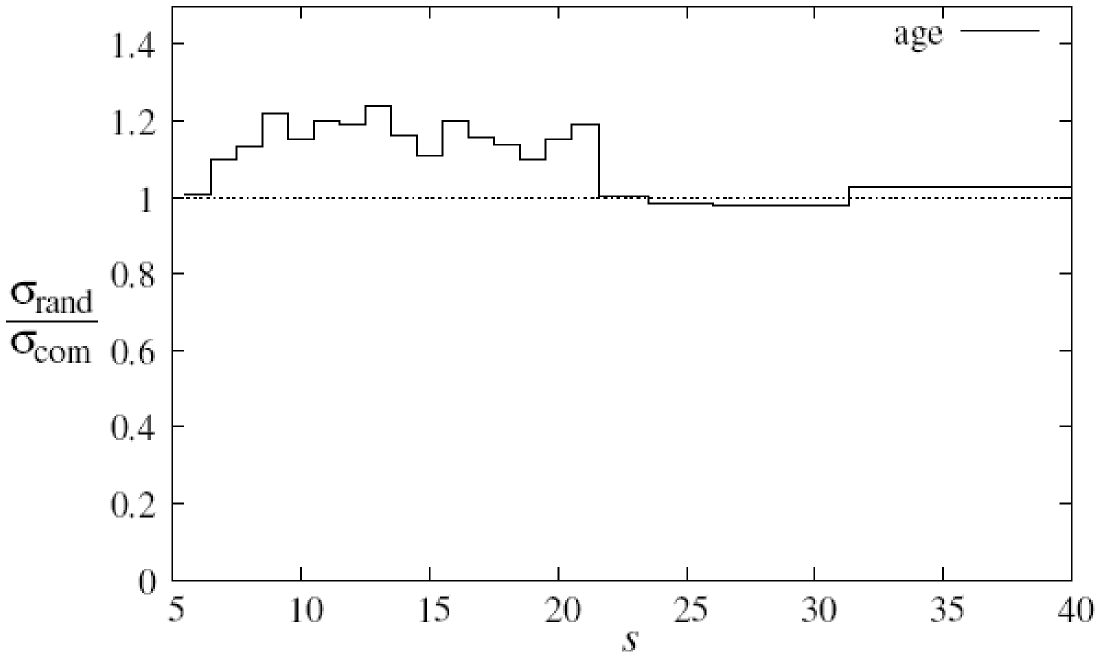




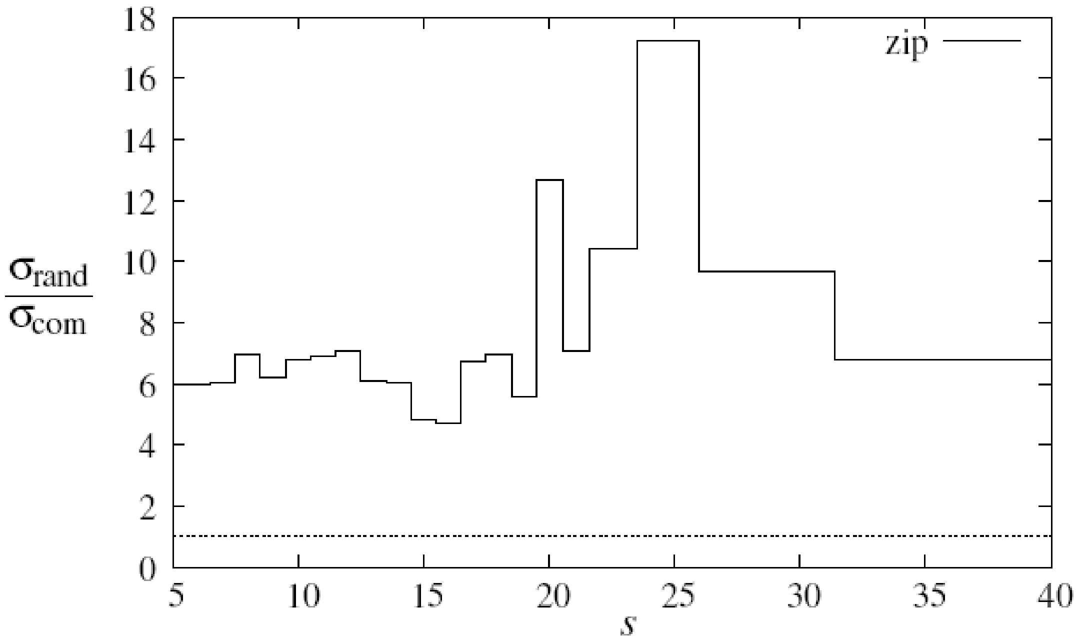


Communities in a “tiny” part of a phone calls network of 4 million users (with A-L Barabási and G. Palla *Nature*, April 5 2007)



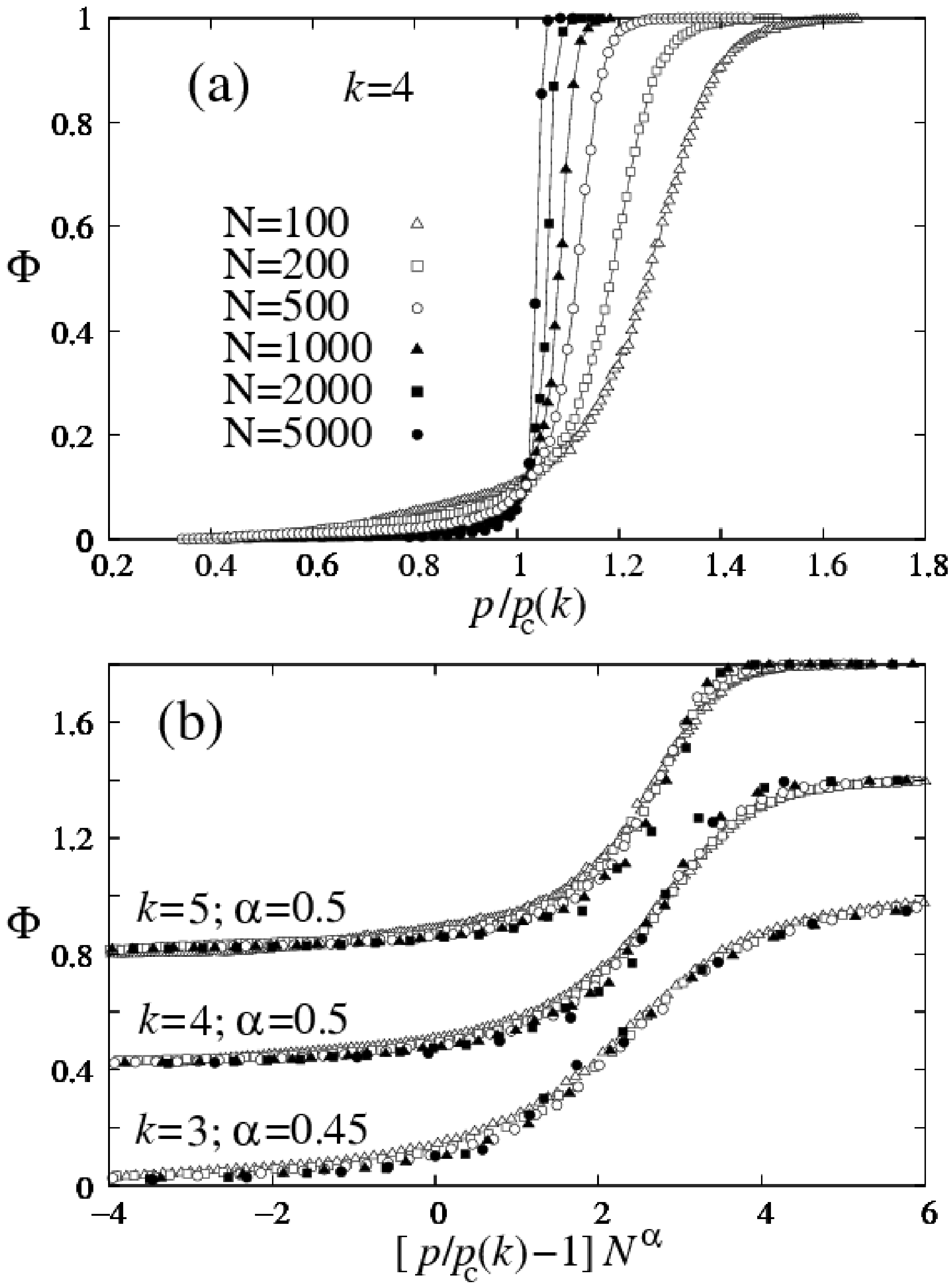


Information about the age distribution of users in communities of size *s*
(Ratio of the standard deviation in a randomized set over actual)



Information about the Zip code (spatial) distribution of users in communities of size *s*

(Ratio of the standard deviation in a randomized set over actual)



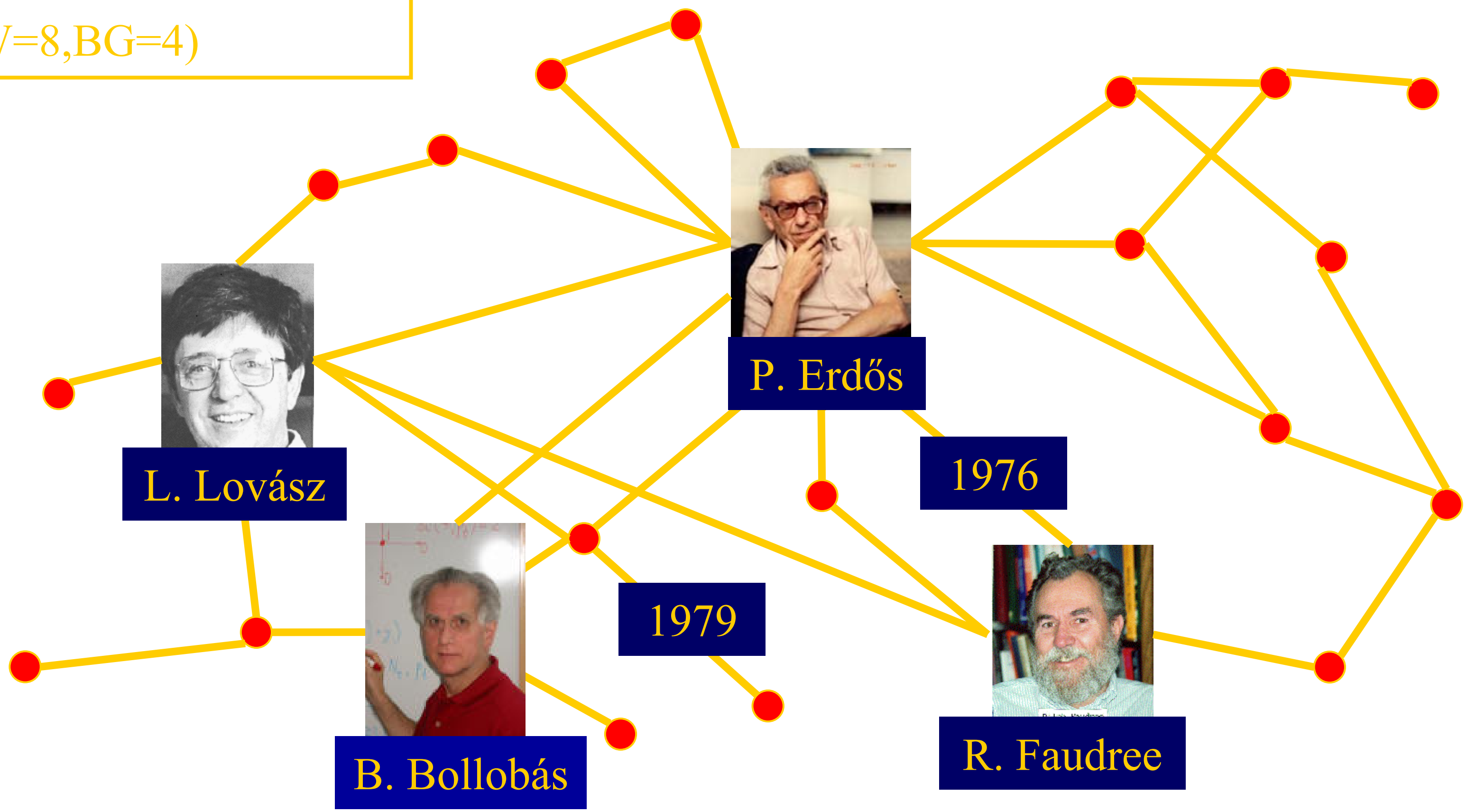
The number of vertices in the largest component

As *N* grows the width of the quickly growing region decays as $1/N^{1/2}$

A.-L. B., H.J, Z.N., E.R., A. S., T. V. (Physica A, 2002)

The Erdős graph and the Erdős number

(Ei=2,W=8,BG=4)



Data: collaboration graphs in (M) Mathematics and (NS) Neuroscience

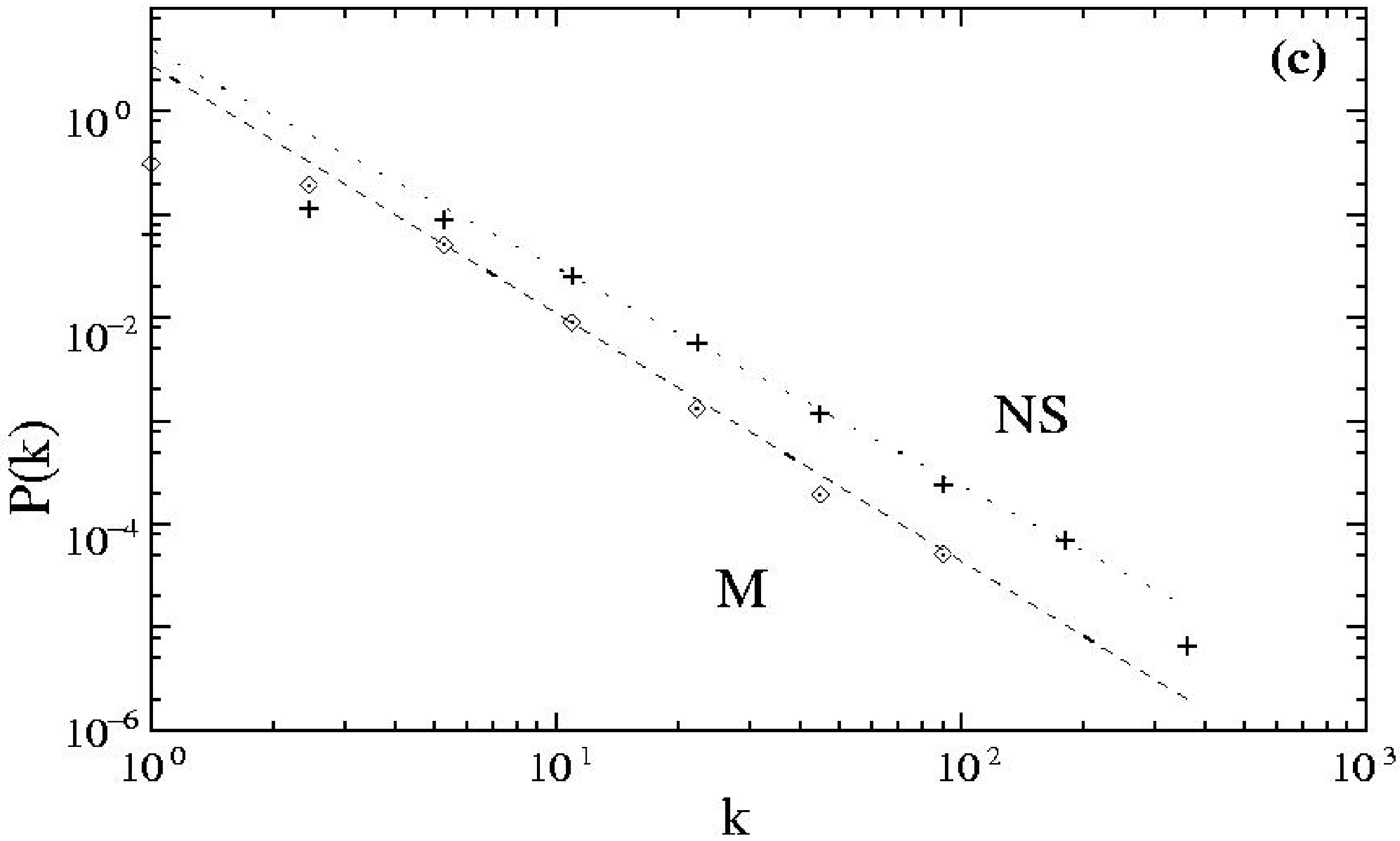
Cumulative data, 1991 - 98

Degree distribution:

power-law with

$$\gamma_{M} = 2.1, \ \gamma_{NS} = 2.4$$

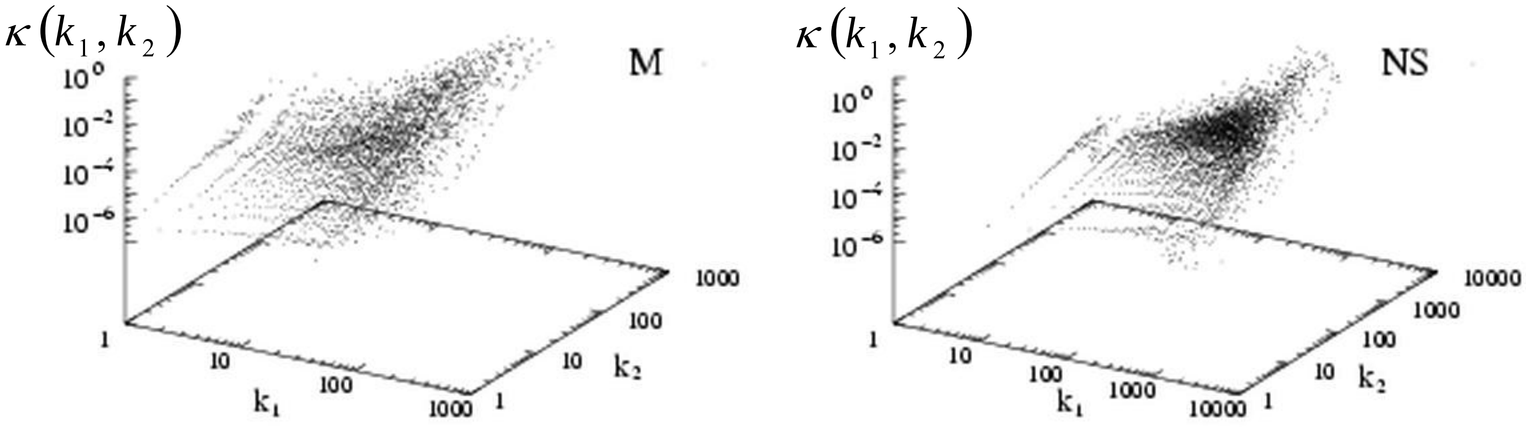
due to growth and preferential attachment



Internal preferential attachment:

Collaboration network

cumulative attachment rate: $\kappa(k_1,k_2) = \int_1^{k_1 k_2} \Pi(k_1,k_2)\, d(k_1,k_2)$

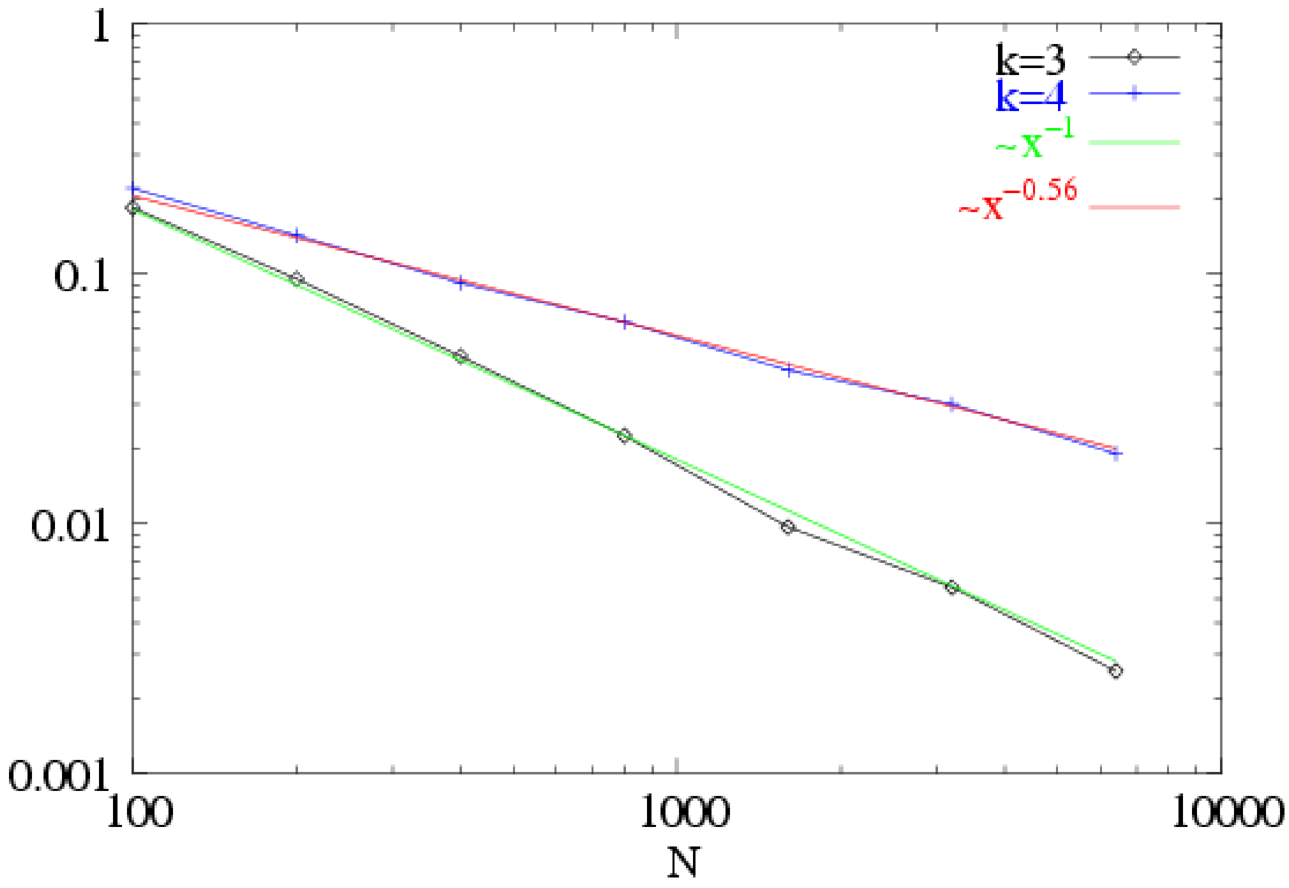


Measured data shows: $\kappa(k_1,k_2)$ is quadratic in $k_1 k_2$

Attachment rate $\Pi(k_1,k_2)$ is linear in $k_1 k_2$

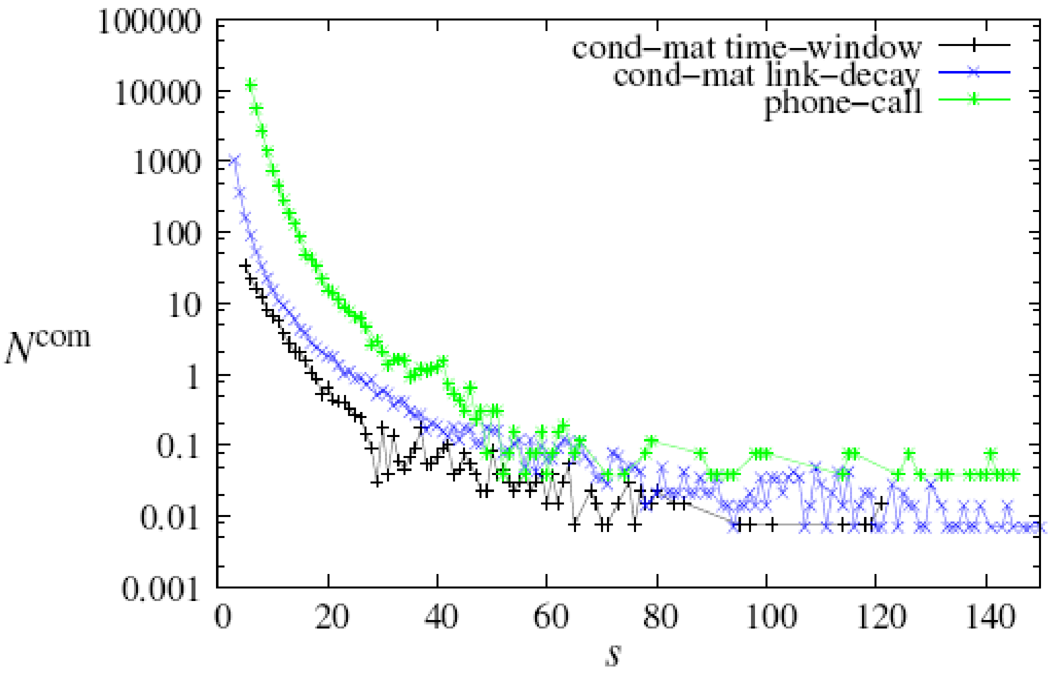
communities of collaborators are formed

The scaling of the relative size of the giant cluster of *k*-cliques at p_c

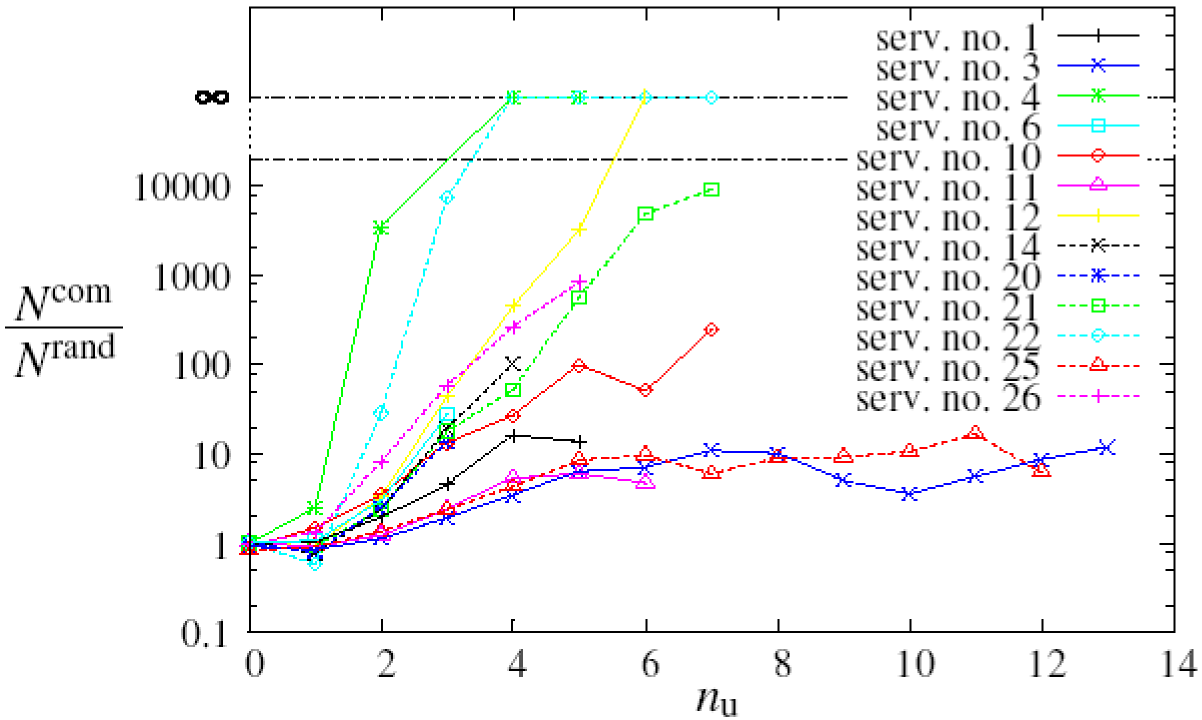


For $k \leq 3$, $N_k^*/N_k(p_c) \sim N^{-k/6}$

For $k > 3$ $N_k^*/N_k(p_c) \sim N^{1-k/2}$



Distribution of community sizes

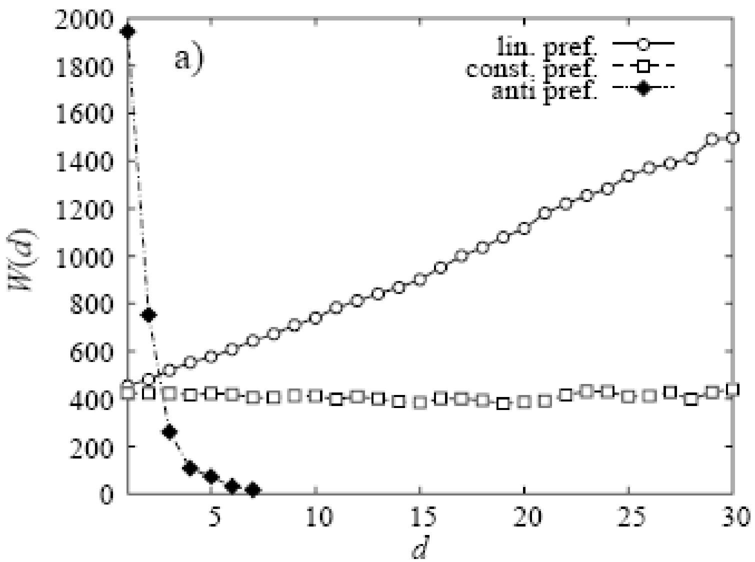


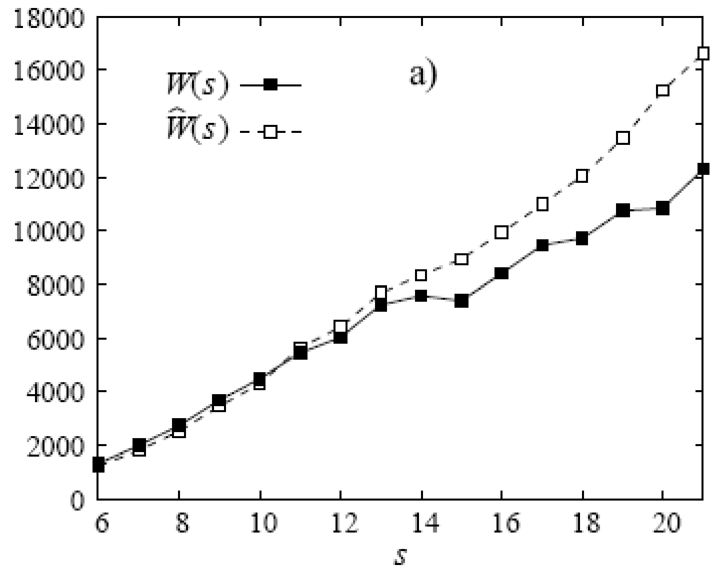
Over-representation of the usage of a given service as a function of the number of users in a community

Community dynamics

with P. Pollner and G. Palla

Dynamics of community growth: the preferential attachment principle applies on the level of communities as well





The probability that a previously unlinked community joins a community larger than *s* grows approximately linearly (for the cond-mat coauthorship network)

P.P,G.P,T.V Europhys Lett. 2006