Network Analysis

Dino Pedreschi, Fosca Giannotti
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

Master MAINS 2018

SUMMARY

- Network everywhere
- Discoverying the fabric of networks: communities
    - Discoverying Mobility Borders
    - Estimating active services of skype
- Forms of information spreading & Innovators

MODULE Outline

- **Lesson 4 Complex network**
  - **Community Discovey Homophily with Demon**
  - **Innovators and forms of spreading of innovation**
  - **Economic Complexity**
  - **Measuring Success in sport**
  - **The BigData ICT scenario**
  - **SoBigData**

Complex

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

–adjective

1.

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

2.

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

3.

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

Source: Dictionary.com

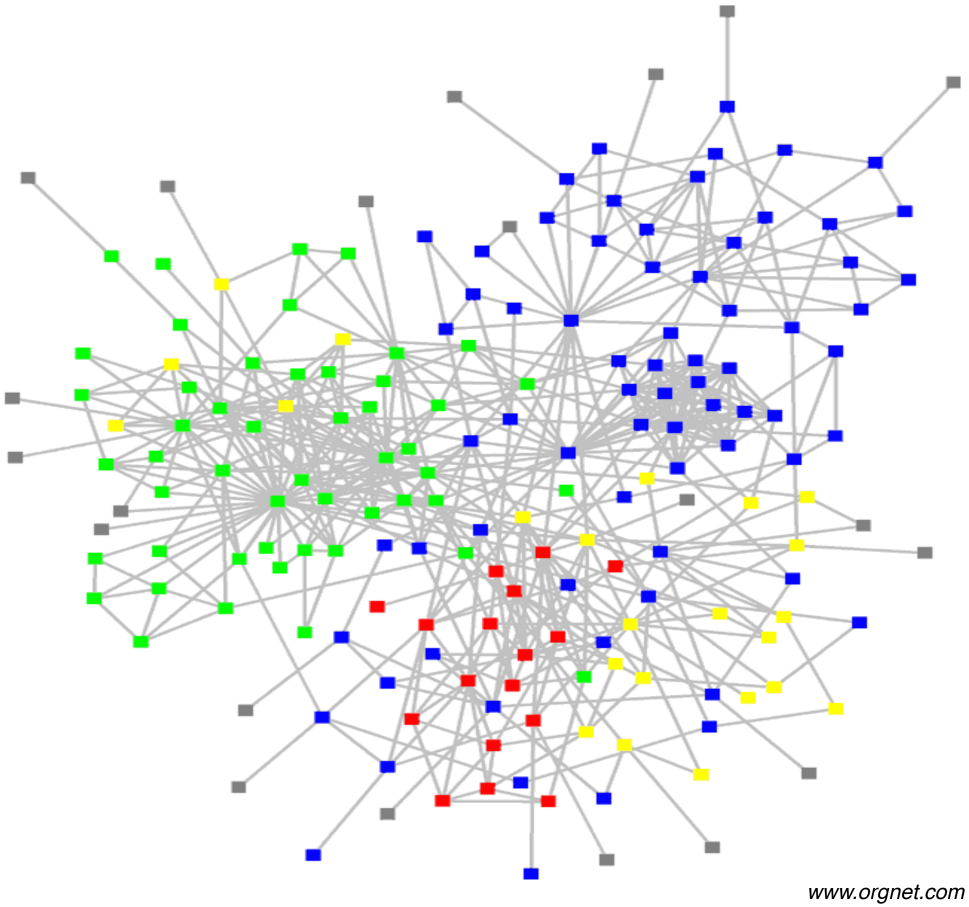
Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving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

Source: John L. Casti, Encyclopædia Britannica

Complexity

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

STRUCTURE OF AN ORGANIZATION



: departments

: consultants

: external experts

BUSINESS TIES IN US BIOTECH-INDUSTRY

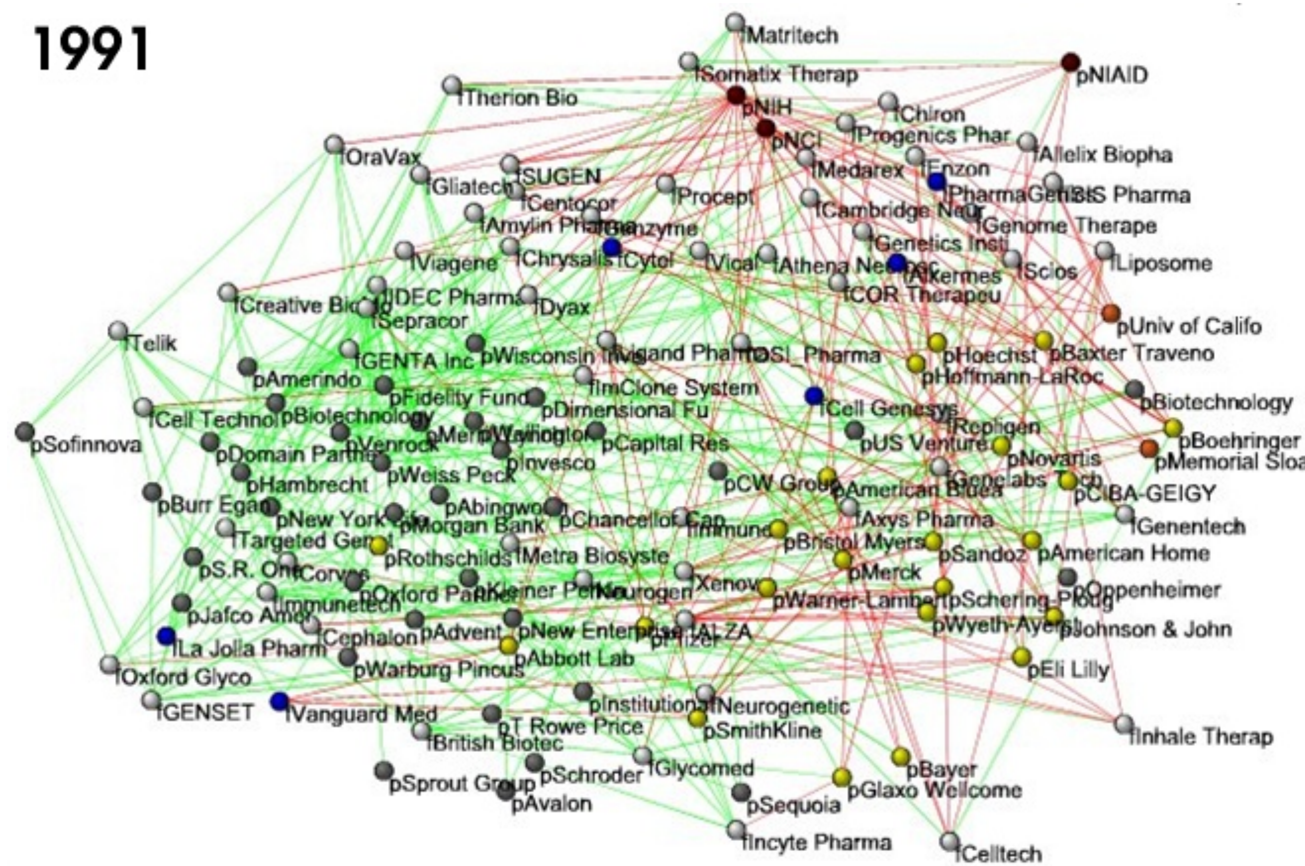
Nodes:

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

Links:

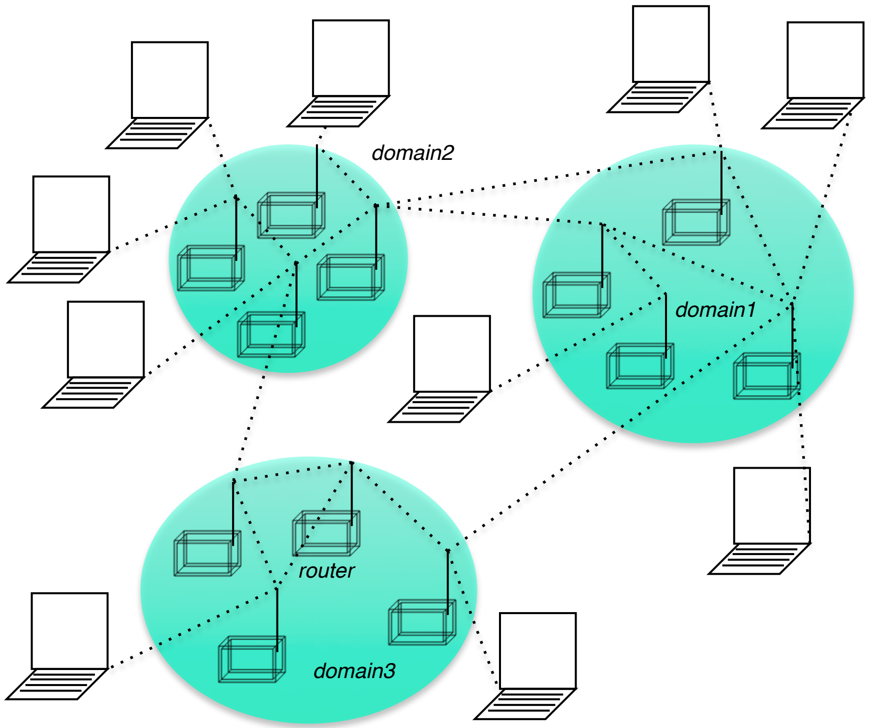
- Collaborations
- Financial
- R&D

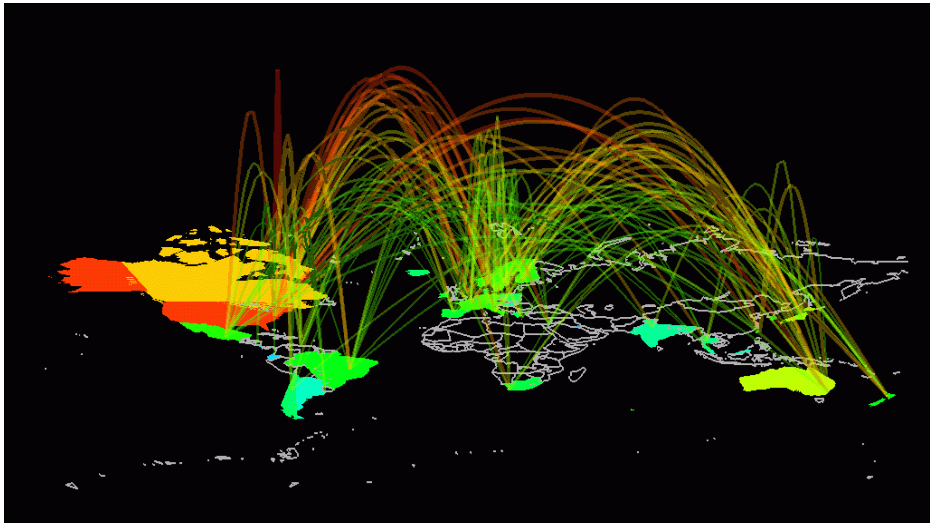
1991

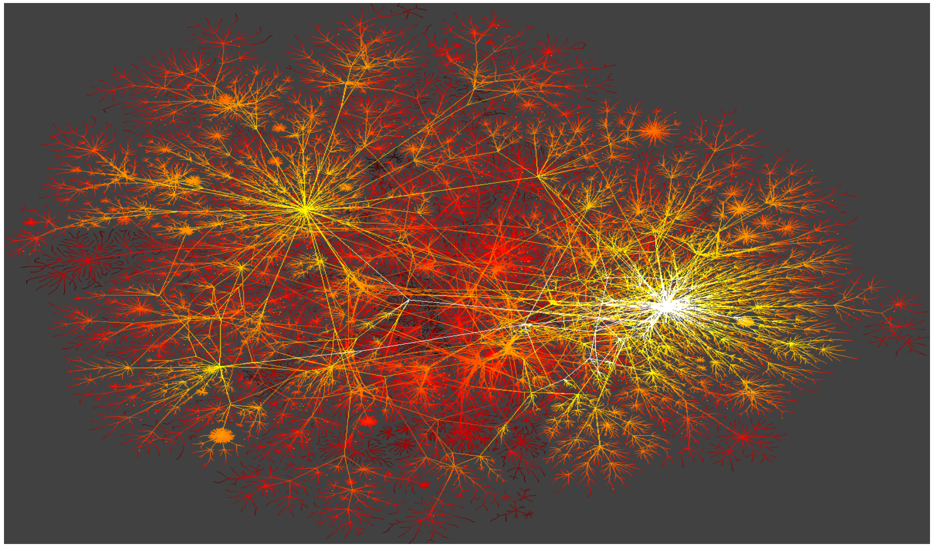


http://ecclectic.ss.uci.edu/~drwhite/Movie

INTERNET







Society

Nodes: individuals

Links: social relationship (family/work/friendship/etc.)

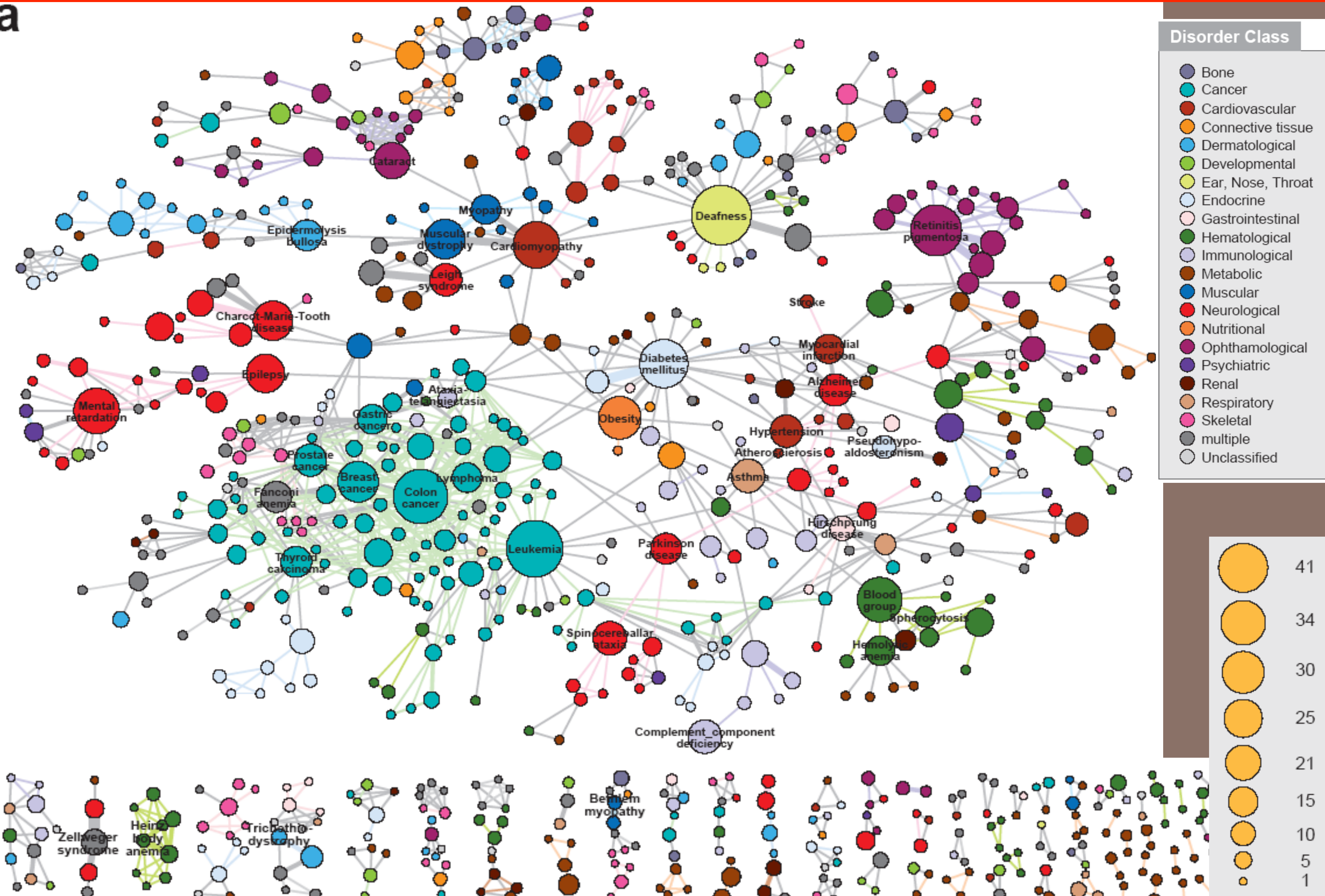
S. Milgram (1967)

John Guare

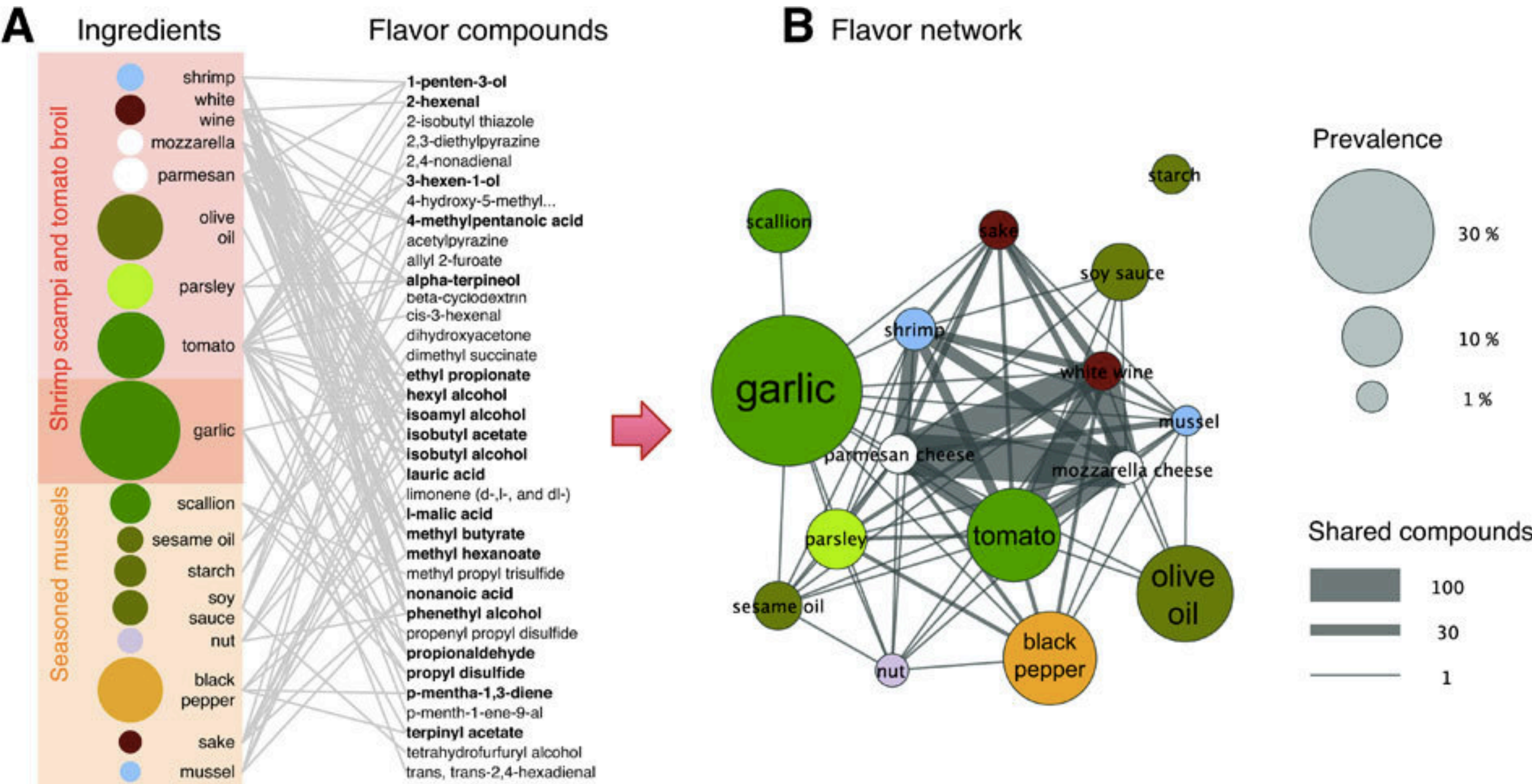
Six Degrees of Separation

Social networks: Many individuals with diverse social interactions between them.

HUMAN DISEASE NETWORK

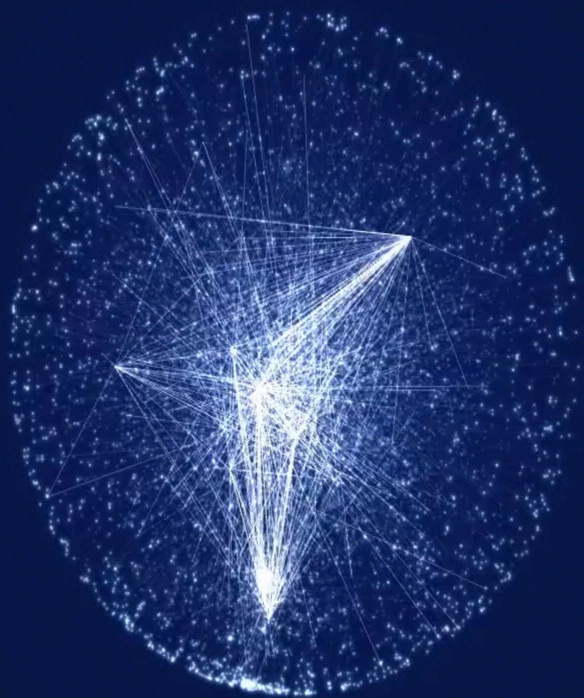


Ingredient-Flavor Bipartite Network



Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási
Flavor network and the principles of food pairing , Scientific Reports 196, (2011).

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Undirected network

N=2,018 proteins as nodes

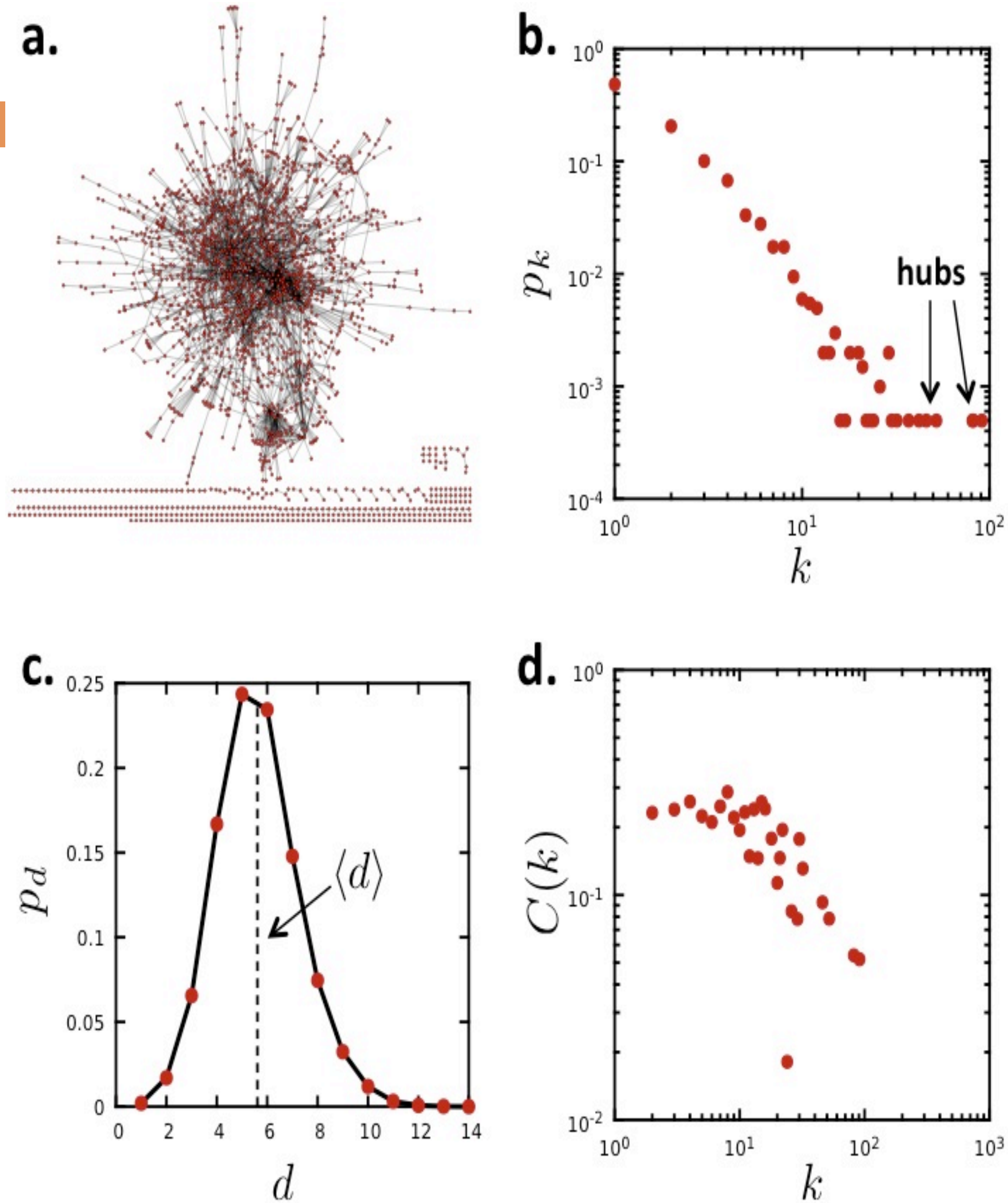
L=2,930 binding interactions as links.

Average degree <k>=2.90.

Not connected: 185 components

the largest (giant component) 1,647 nodes

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Undirected network

N=2,018 proteins as nodes

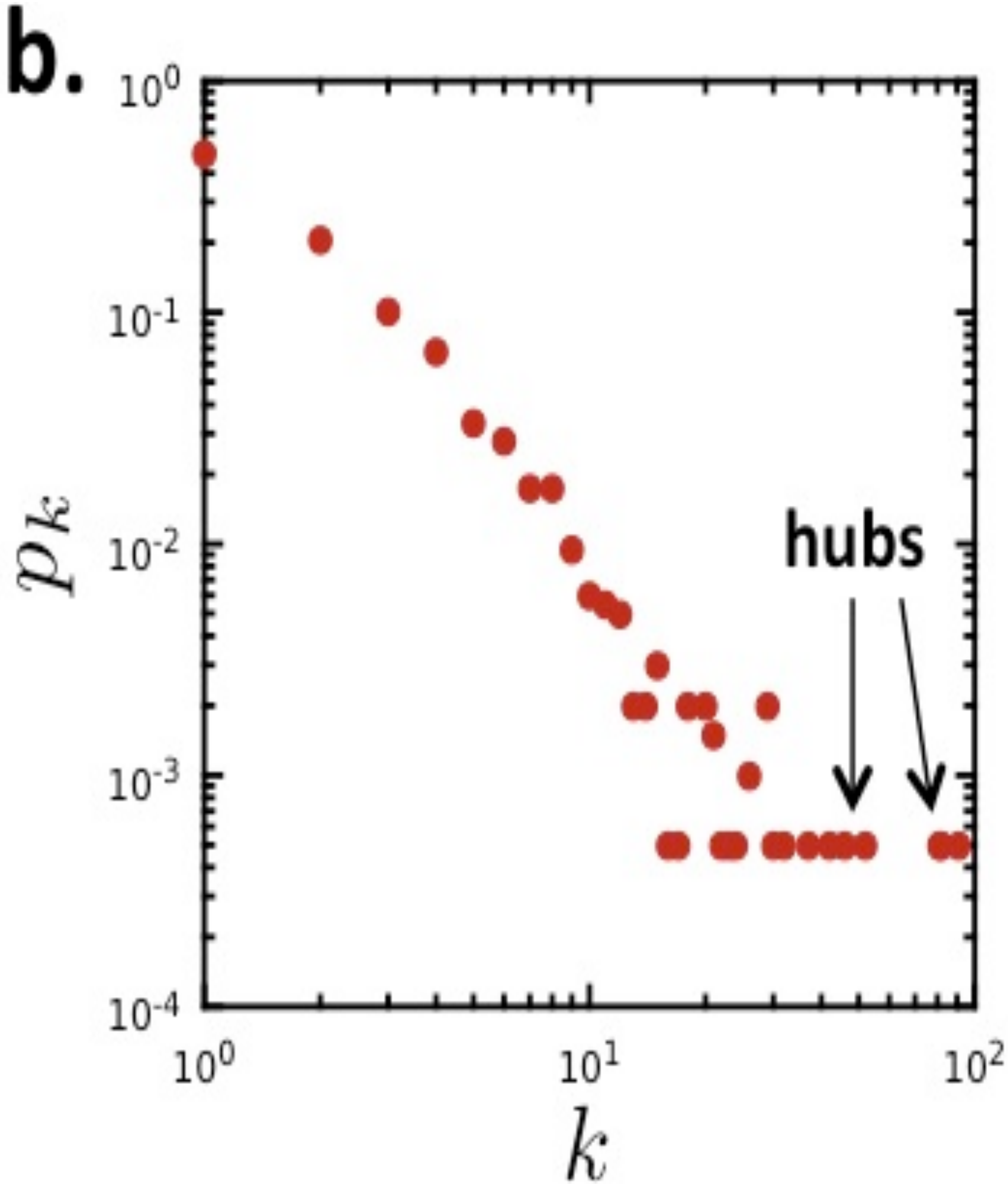
L=2,930 binding interactions as links.

Average degree $\langle k \rangle$=2.90.

Not connected: 185 components

the largest (giant component) 1,647 nodes

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

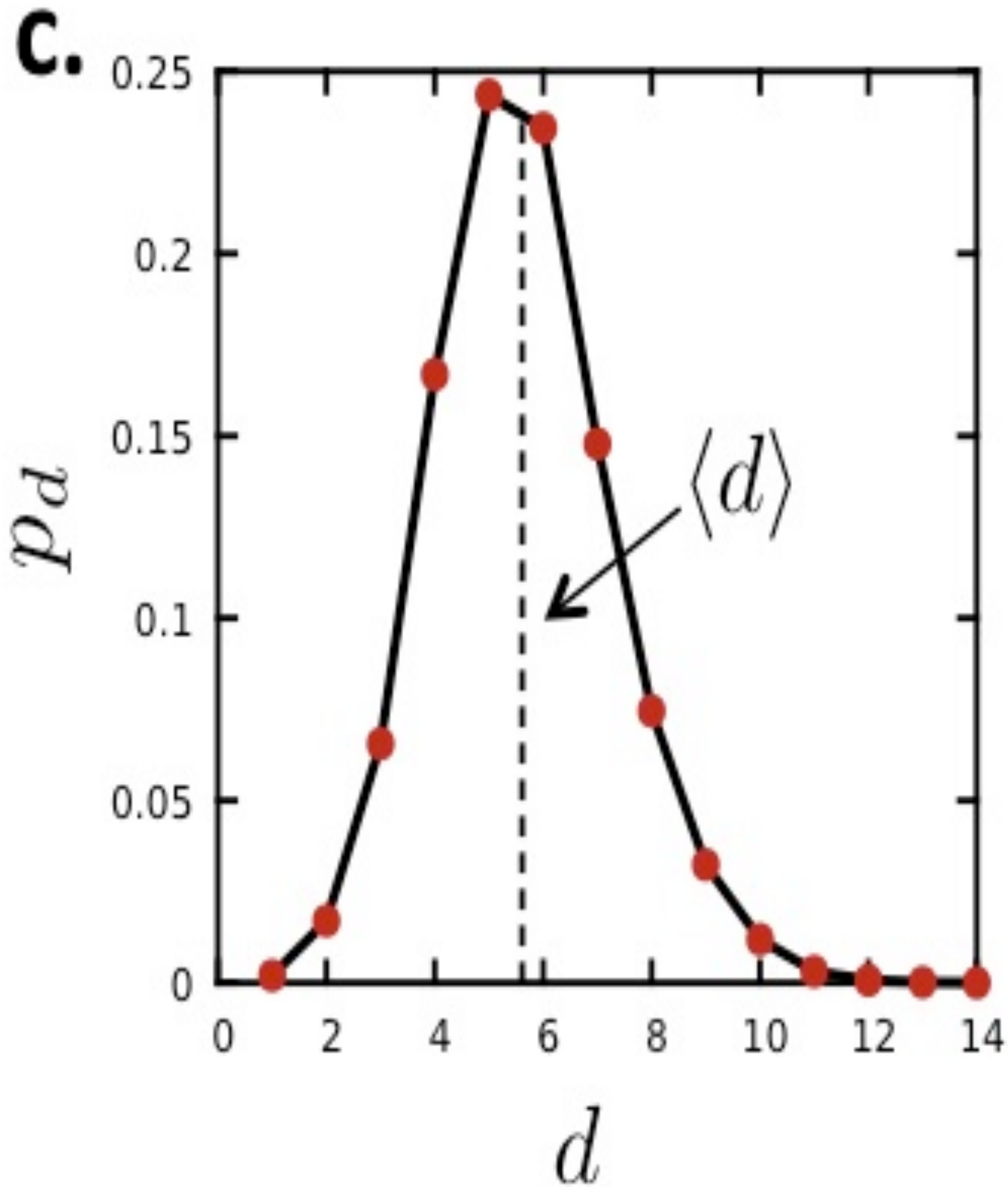


p_k is the probability that a node has degree **k**.

N_k = # nodes with degree k

$p_k = N_k / N$

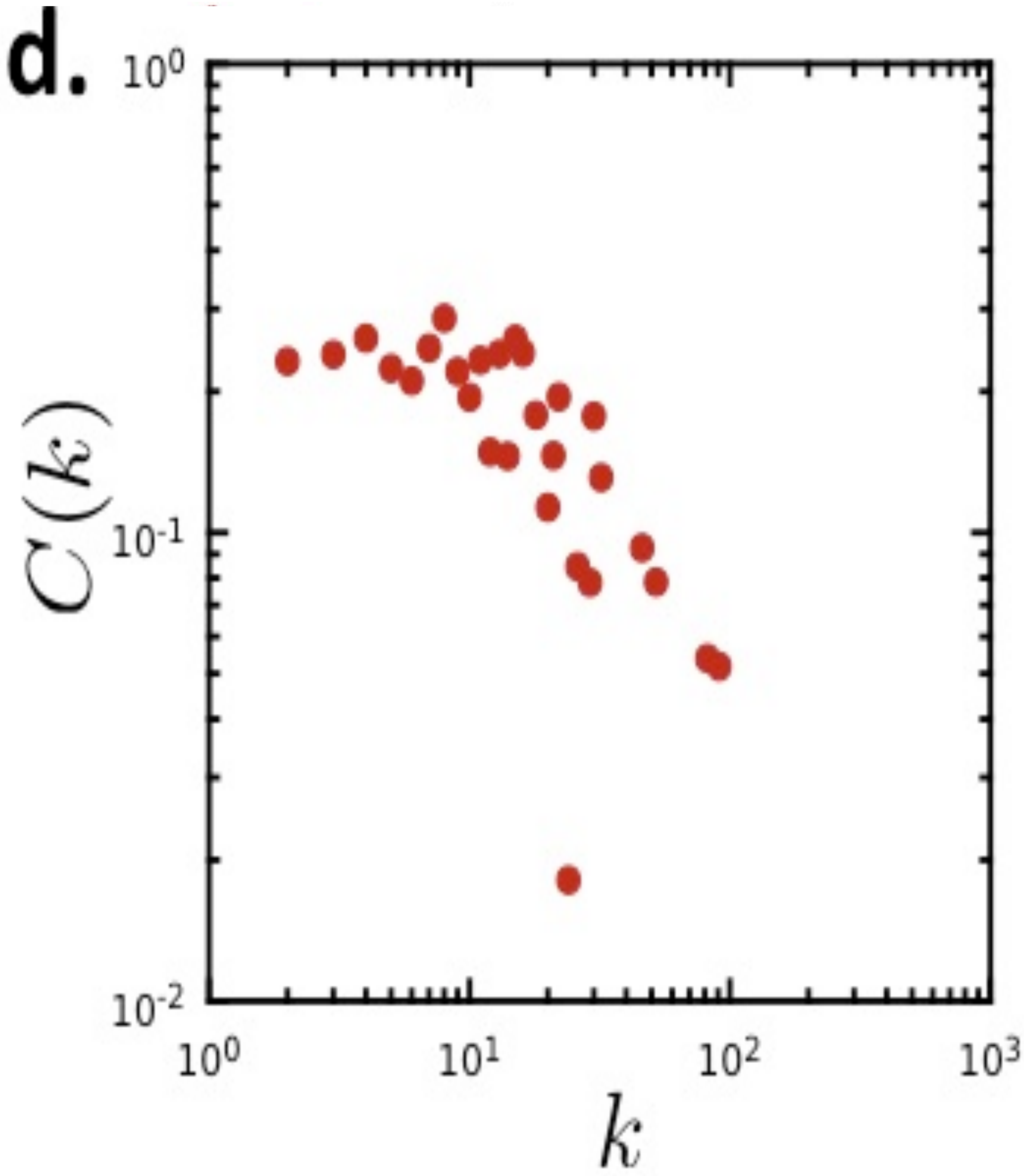
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

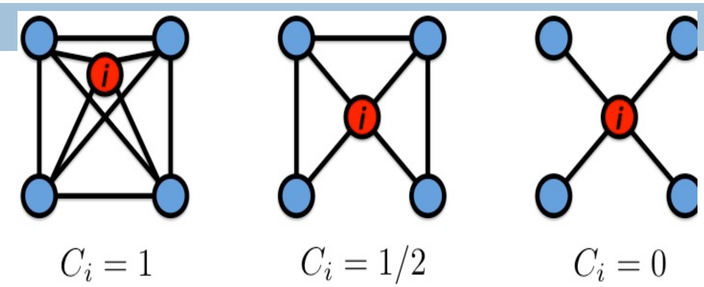


d_{max}=14

<d>=5.61

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



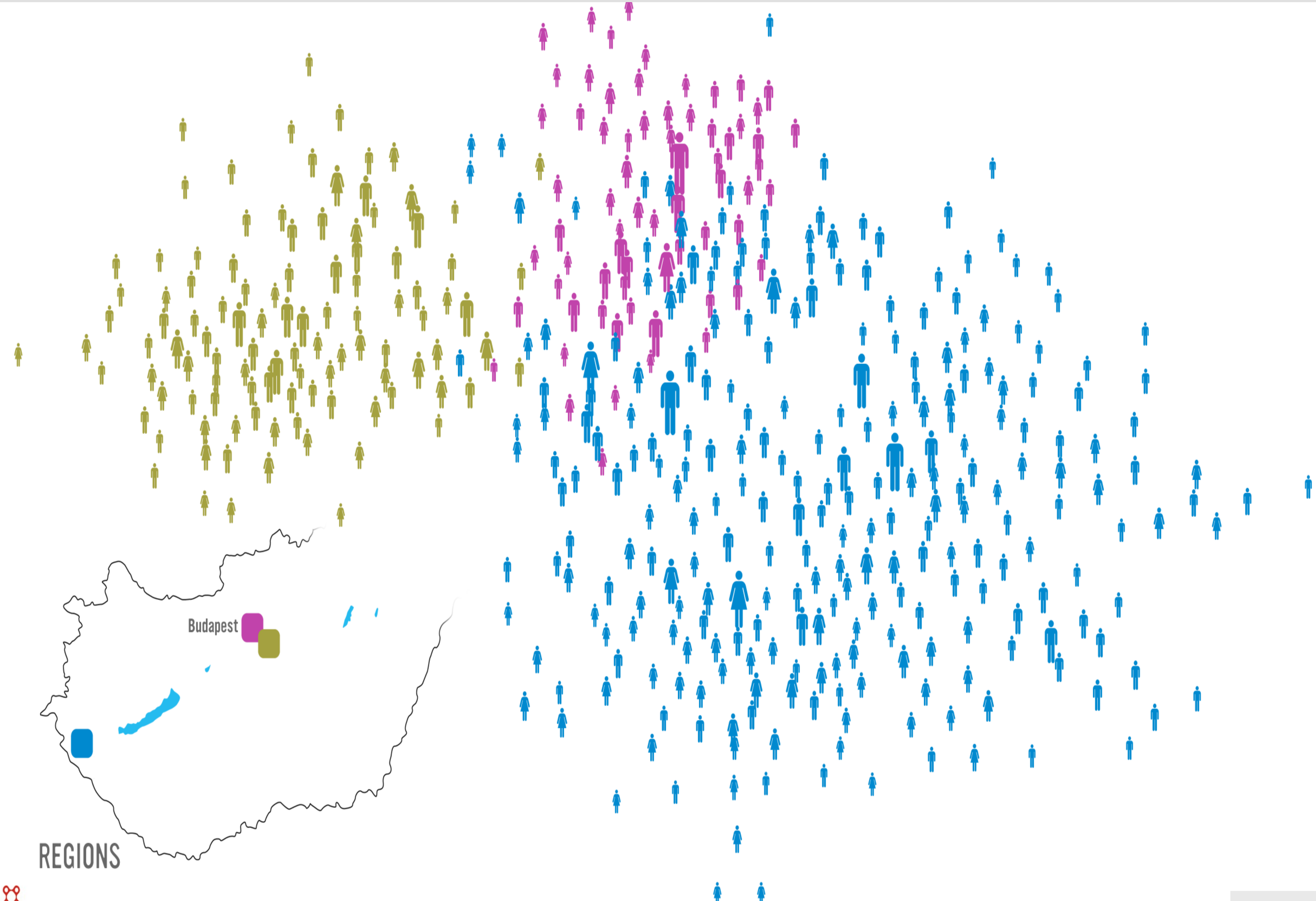


$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

<C>=0.12

INFORMATION DIFFUSION IN SOCIAL NETWORK

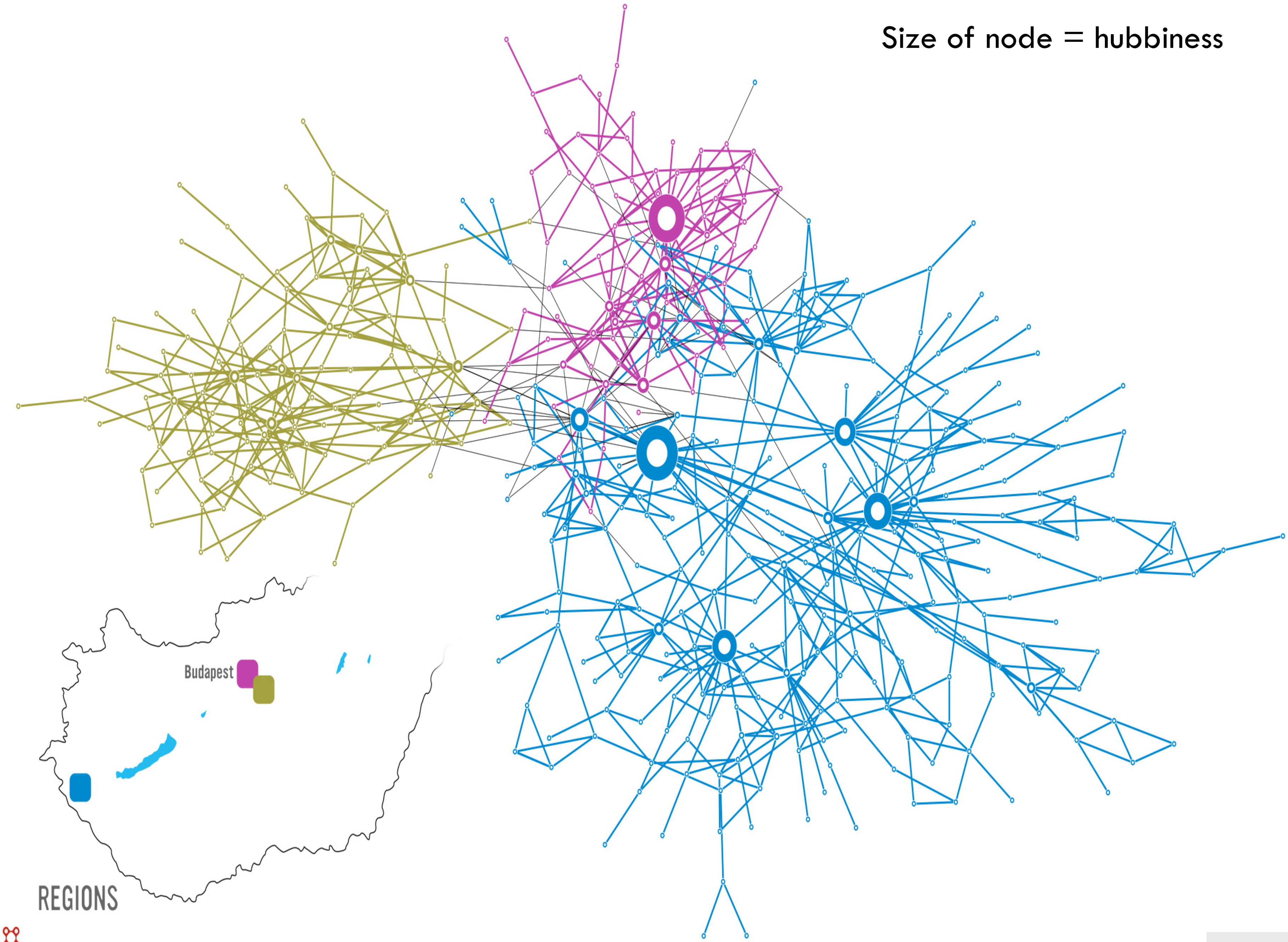
Mapping Organizations



maven7
connecting knowledge

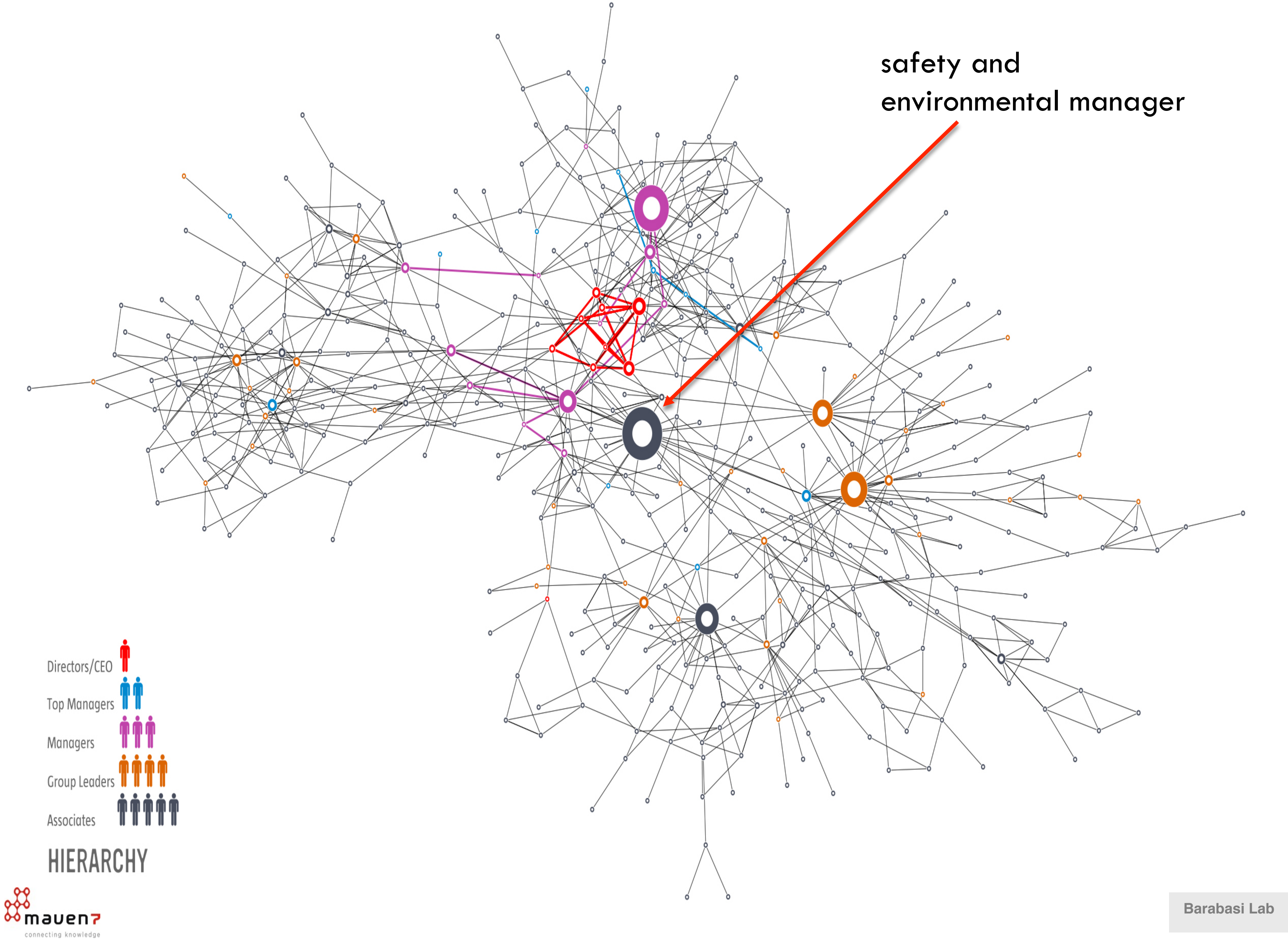
Barabasi Lab

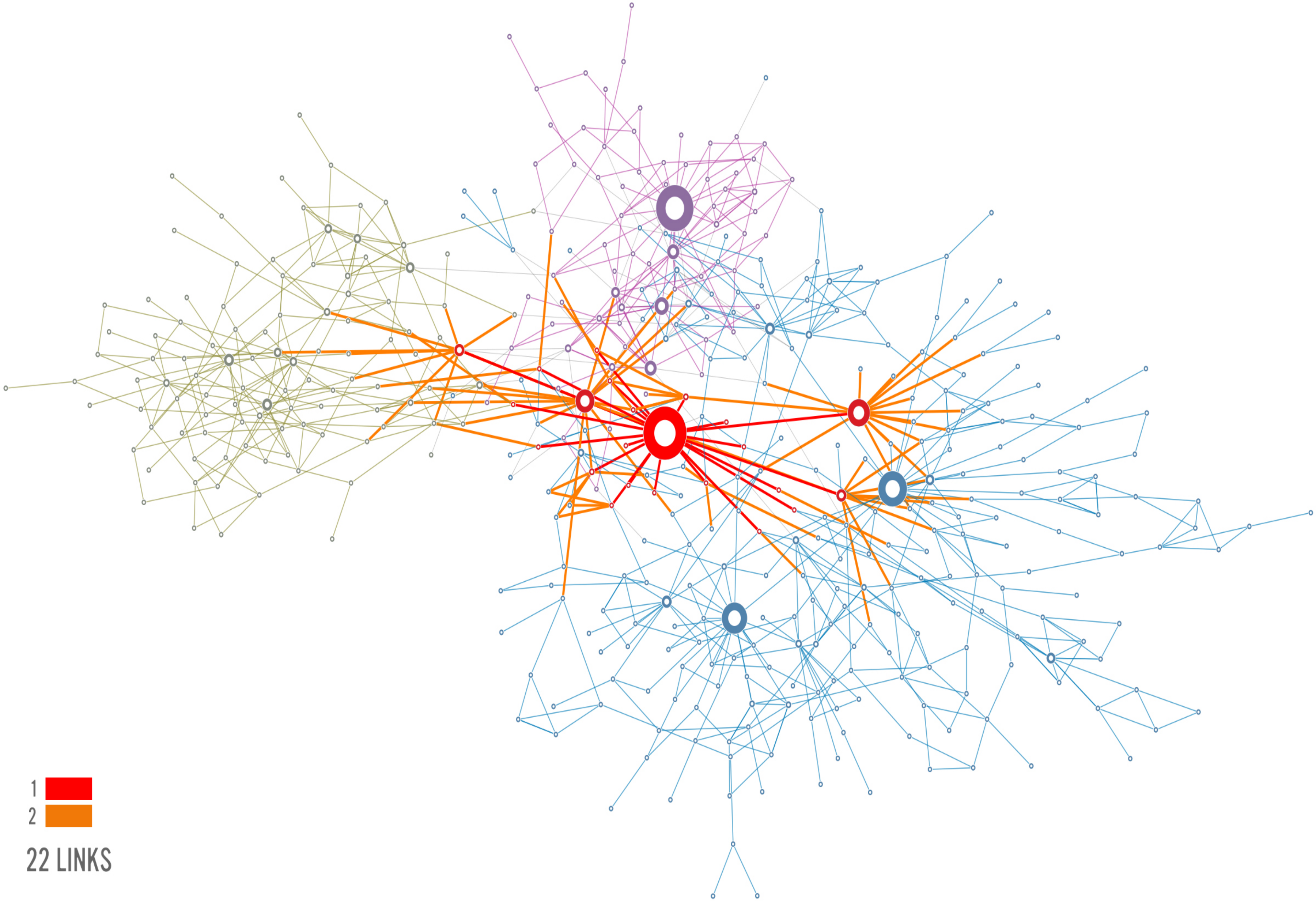
Size of node = hubbiness



maven7
connecting knowledge

Barabasi Lab


safety and
environmental manager
Directors/CEO
Top Managers
Managers
Group Leaders
Associates
HIERARCHY
maven7
connecting knowledge
Barabasi Lab


1
2
22 LINKS

SOCIAL NETWORK MINING
COMMUNITY DISCOVERY

How to highlight the modular structure of a network?

Skype Data: a first glance

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

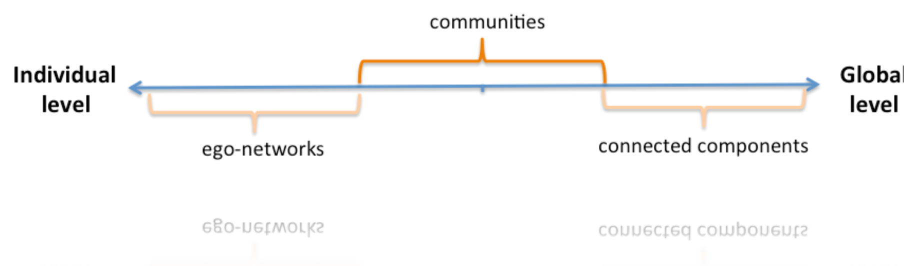
Given an online platform we often we need to *estimate* how its services (i.e., Skype Audio\Video call) are used by the registered users.
In particular we can be asked to answer the following questions:

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?

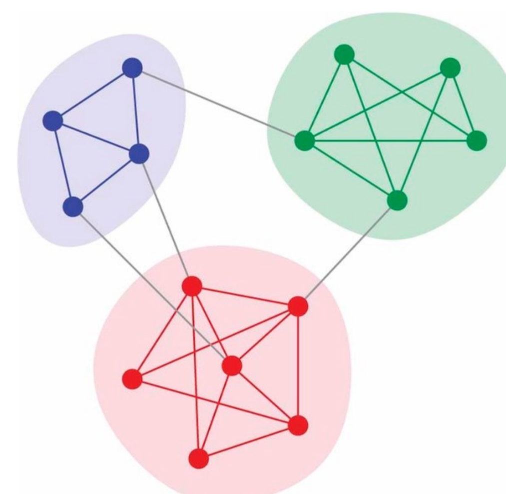
Observation Scale?

- **Problem:**
  Given the size of the dataset (several hundred millions of users) an individual level analysis can be redundant;



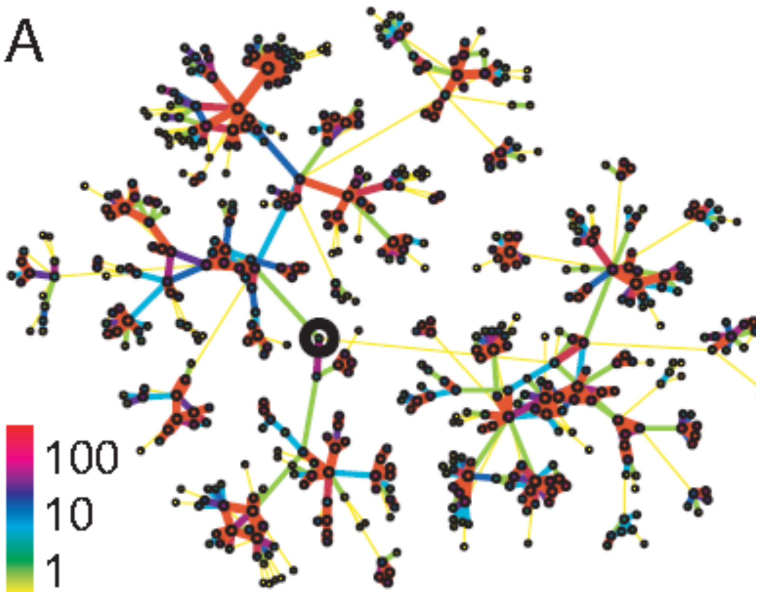
- **Idea:**
  Homophily has been proven to hold on several social context:

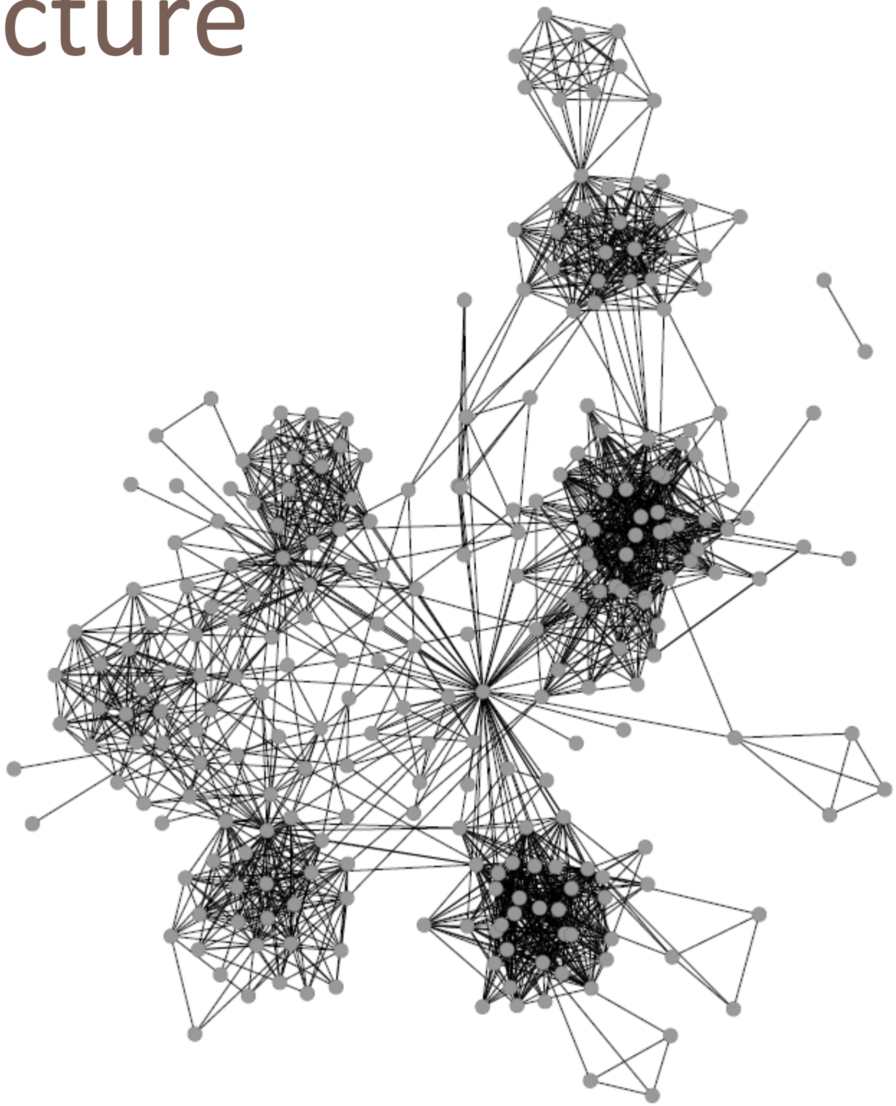
Identifying tight groups of "similar" users we can reduce the problem space



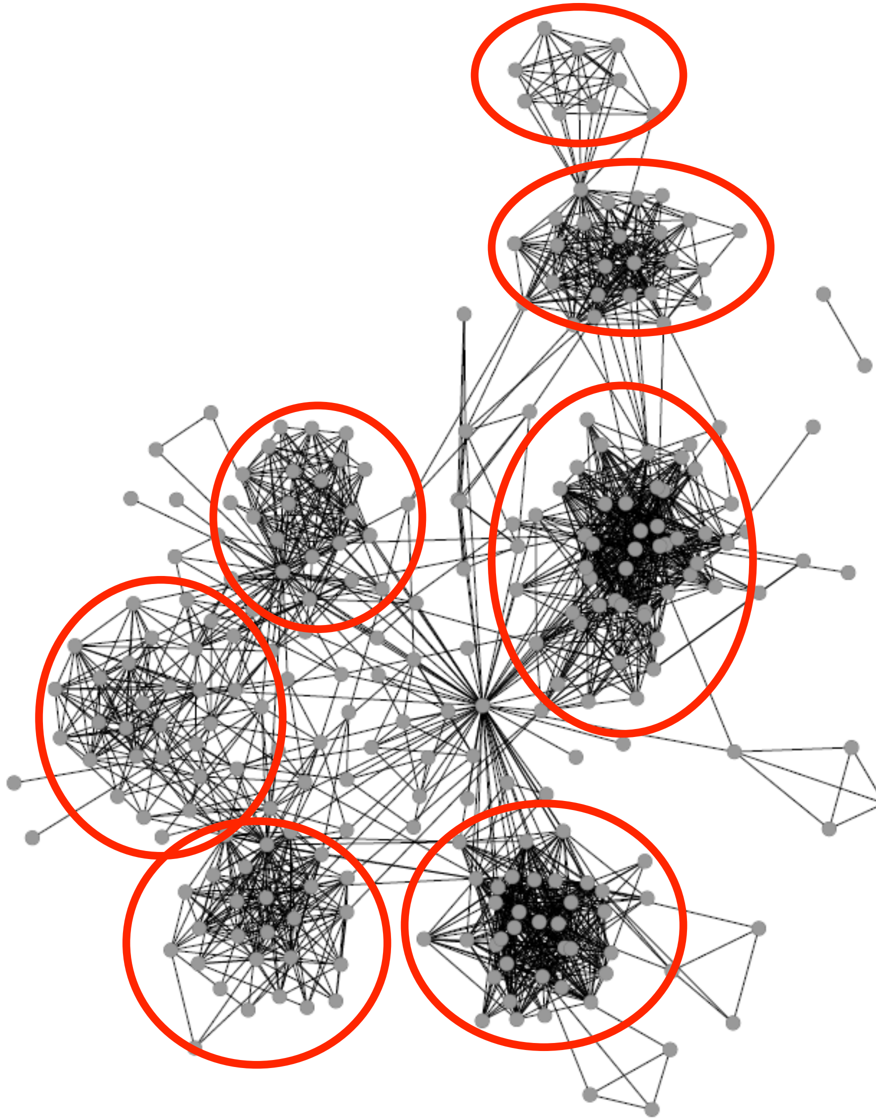
Community Discovery

Community structure

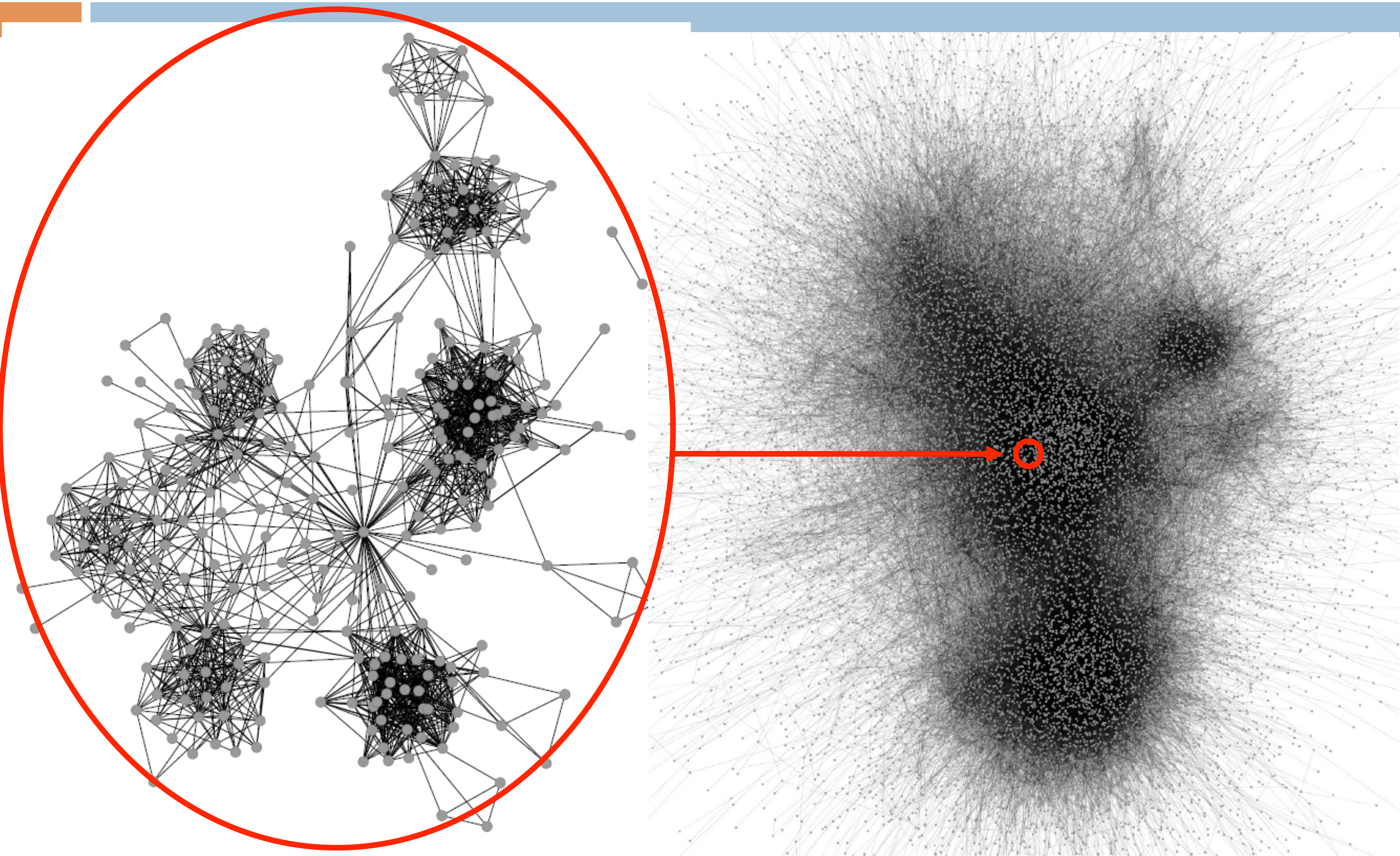




Communities



Lost in the crowd



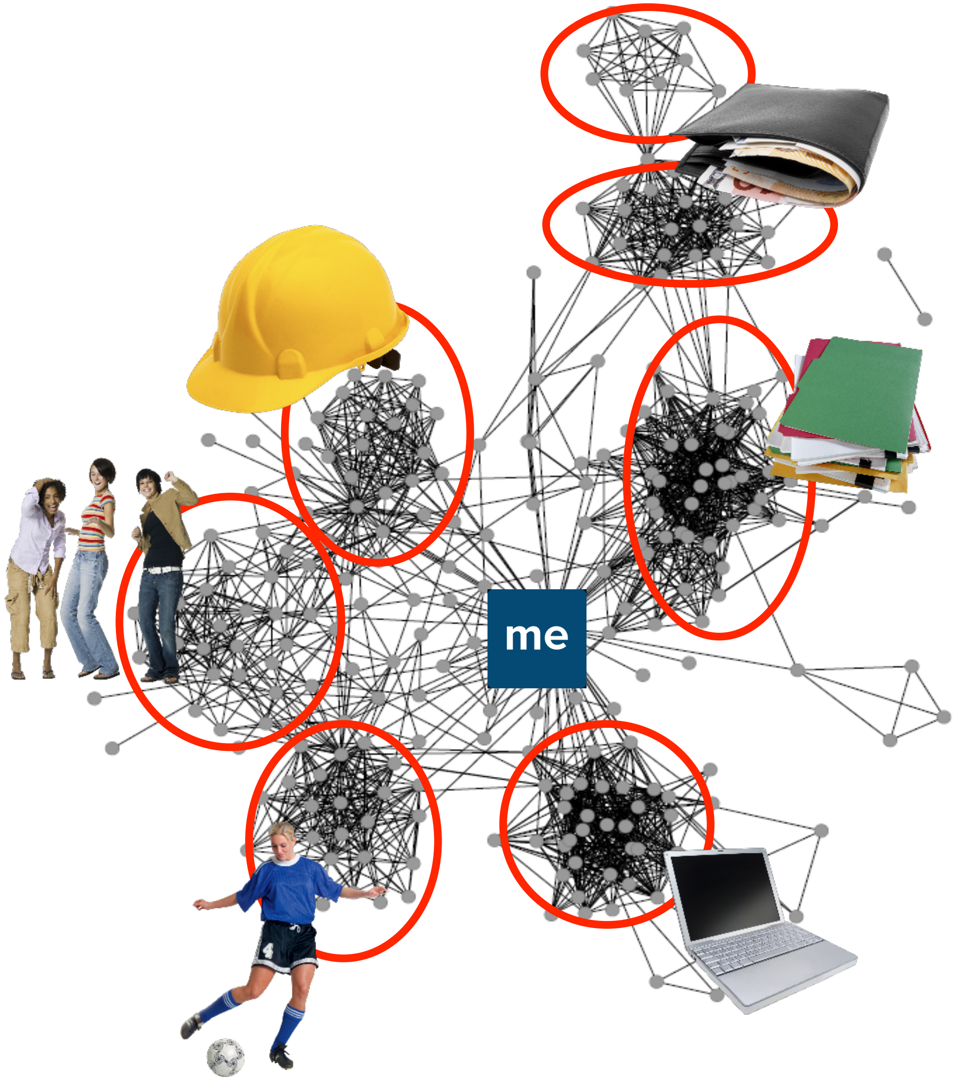
Reducing the complexity

Real Networks are Complex Objects

Can we make them “simpler”?

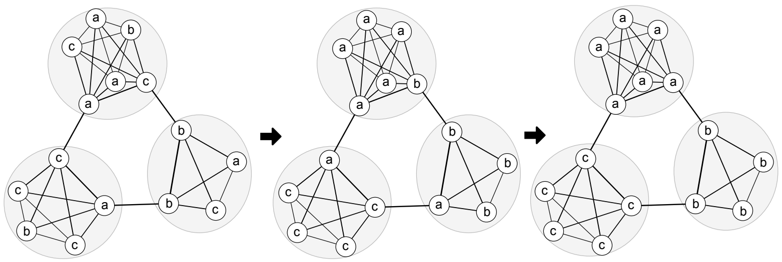
Ego-Networks

networks built upon a focal node , the "ego”, and the nodes to whom ego is directly connected to, including the ties, if any, among the alters



DEMON Algorithm

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



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

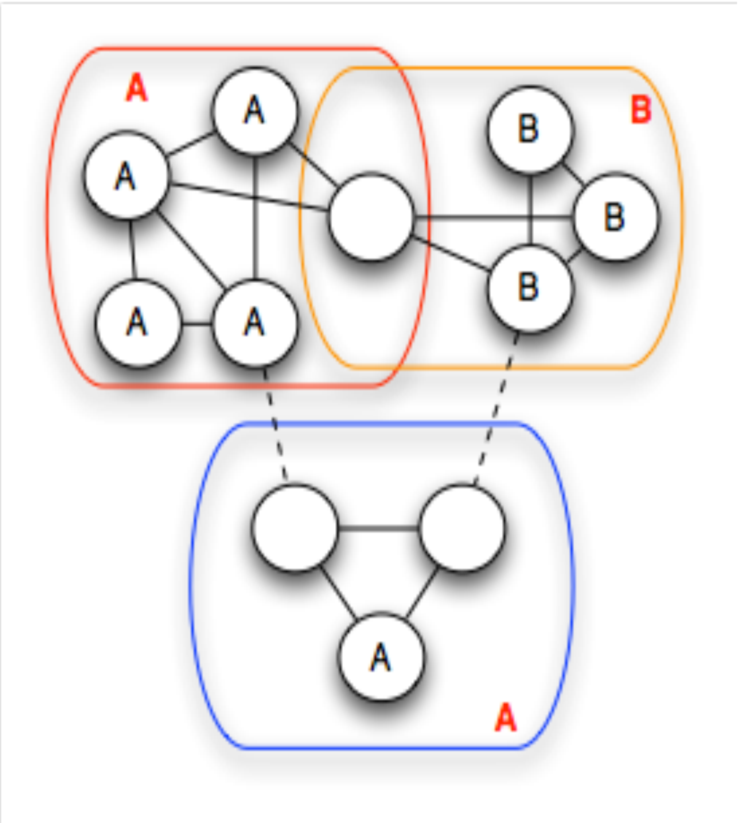
DEMON A Local-first Discovery Method For Overlapping Communities, Giulio Rossetti1,2 ,Michele Coscia3, Fosca Giannotti2, Dino Pedreschi1,2

[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

DEMON Algorithm

- Discovers Overlapping communities
- Microscopic
- High homophily

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



Classifying communities of users

- Classification through Stochastic Gradient Descent
- 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

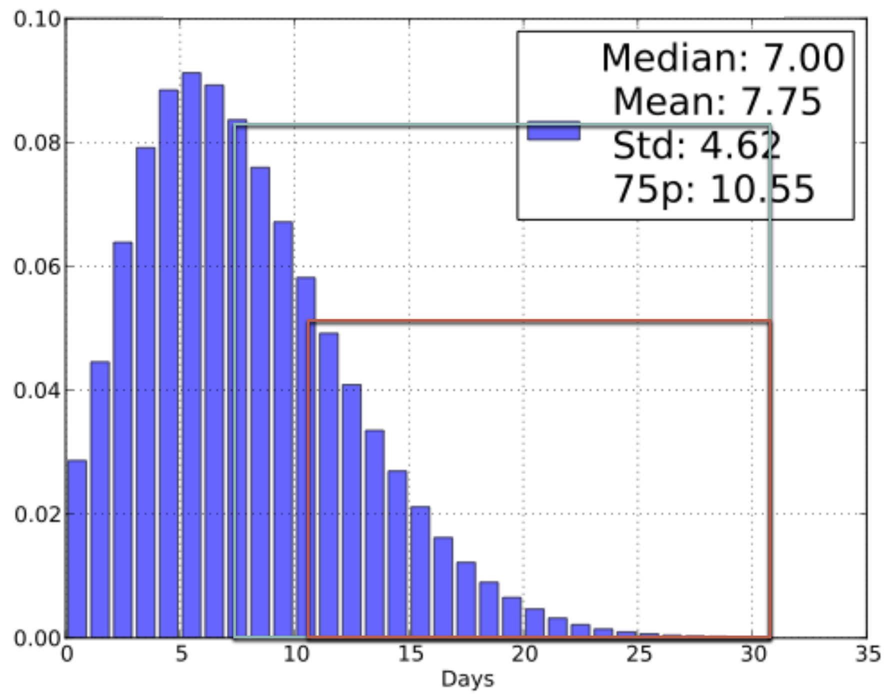
Video	mean number of days of video
Chat	mean number of days of chat

Target Class (for each service)

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

Two scenarios:

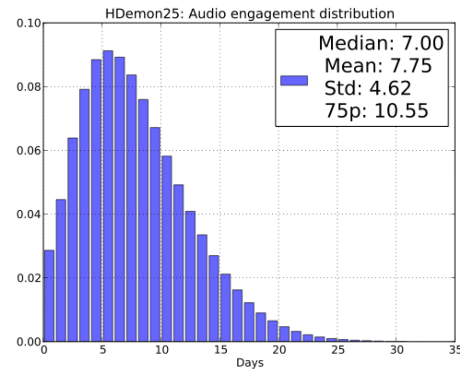
1. Low/High activity is identified by the median of the distribution
2. High activity communities are the one above the 75th percentile

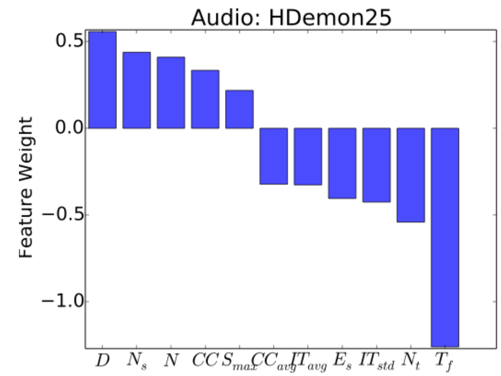


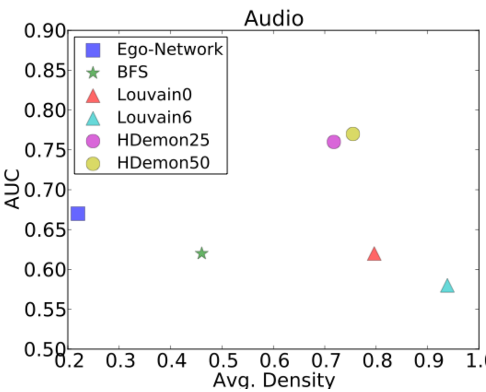
“Social Engagement” : Skype social graph

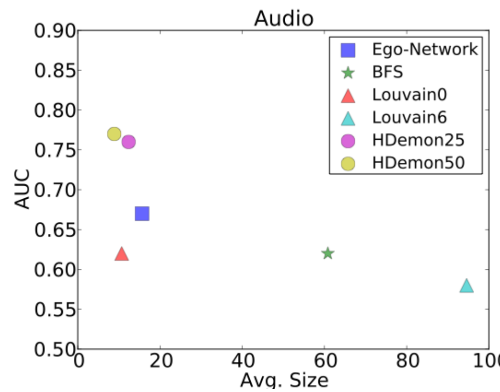
- **Problem:**
  Given the Skype social graph and its user information (i.e., location…) predict average level of community activity for the Audio\Video services.

- **Main Results:**
  - Smaller and denser communities are easier to classify correctly
  - Topological, Temporal and Geographical features of communities are valuable activity level predictors







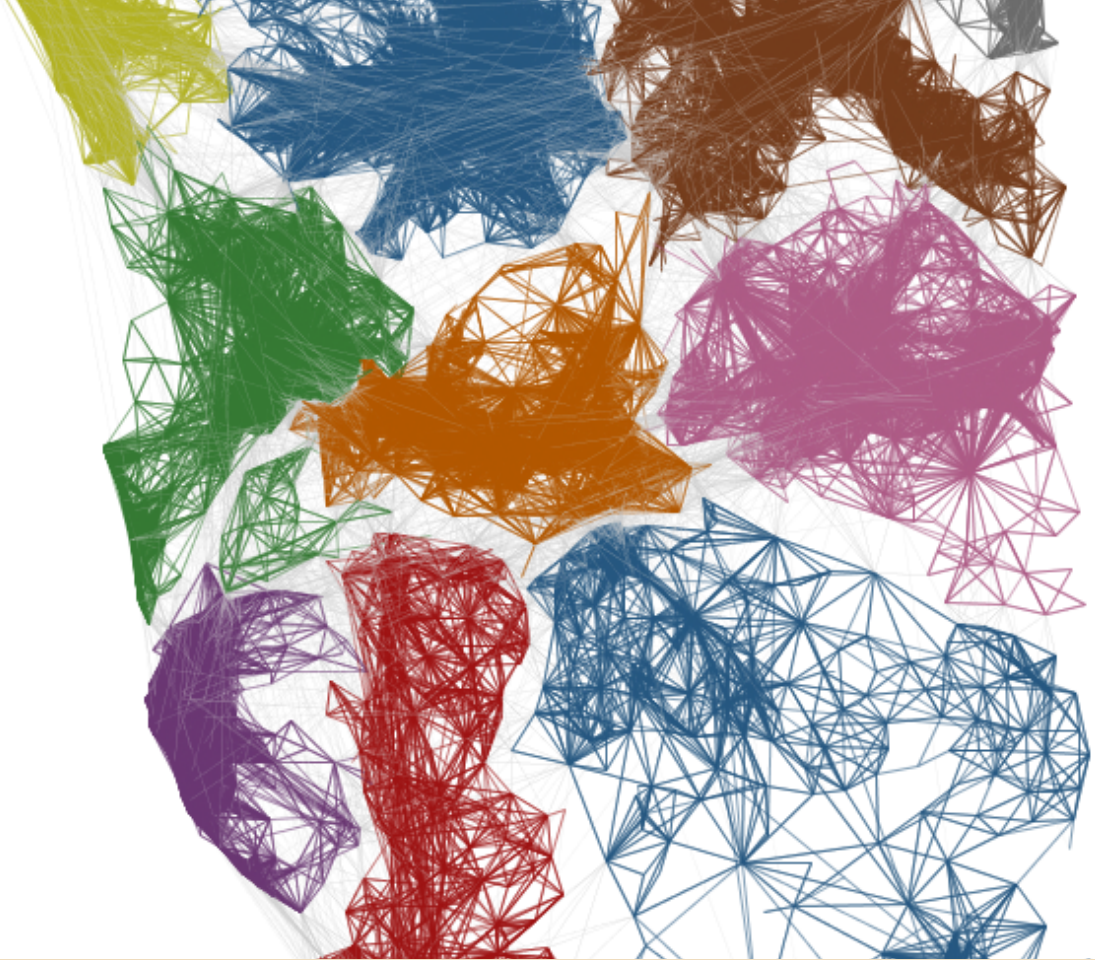


G. Rossetti, L. Pappalardo, R. Kikas, F. Giannotti, D. Pedreschi, M. Dumas
Community-centric analysis of service en- gagement in Skype social networks.
IEEE ASONAM 2015, France (Accepted)

Community Description

Looking at the weight assigned to each feature we can identify some common characteristics of Highly active communities

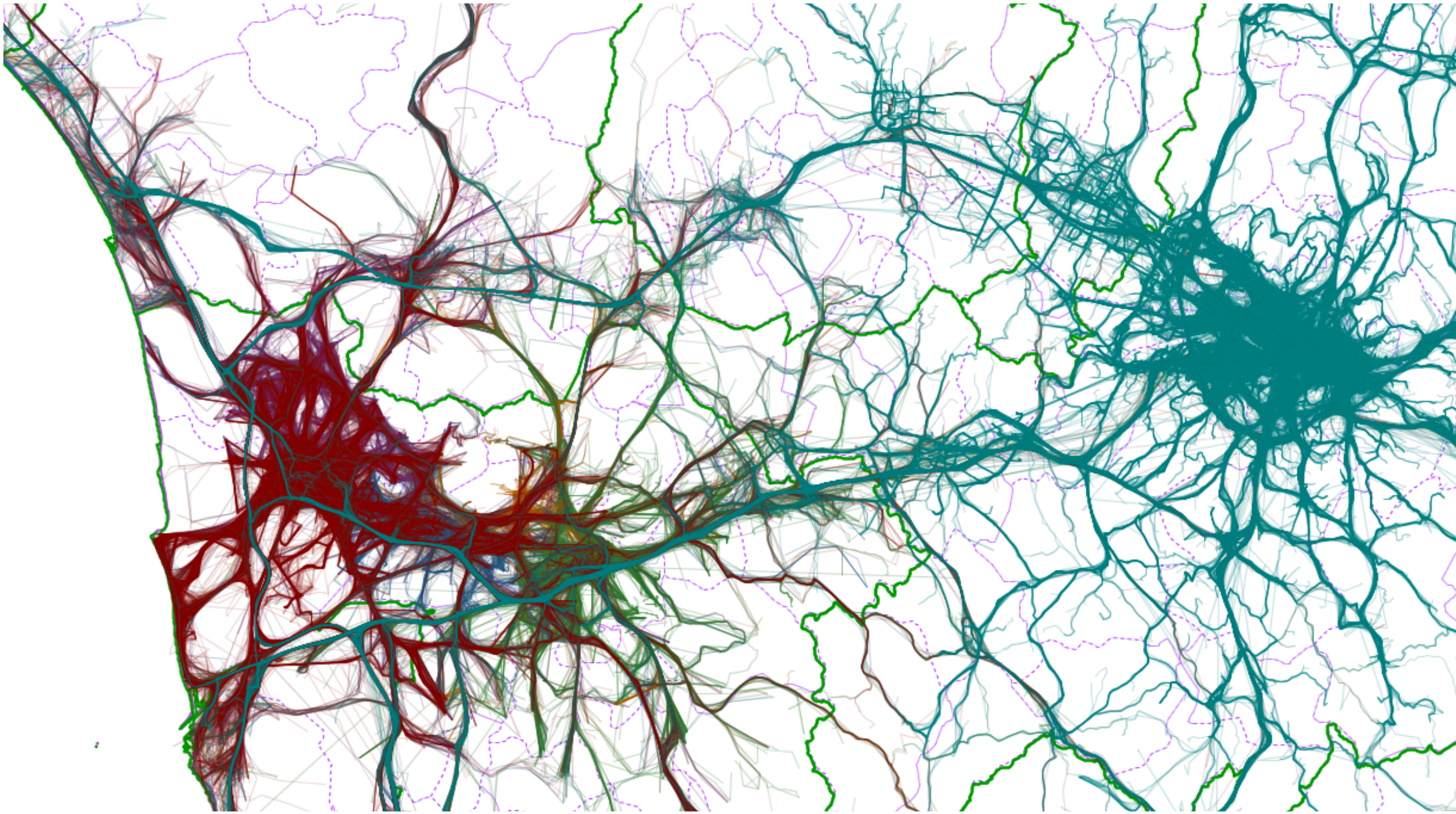
(e.g., for Audio\Chat, low clustering coefficient, reduced size, geographical compactness)



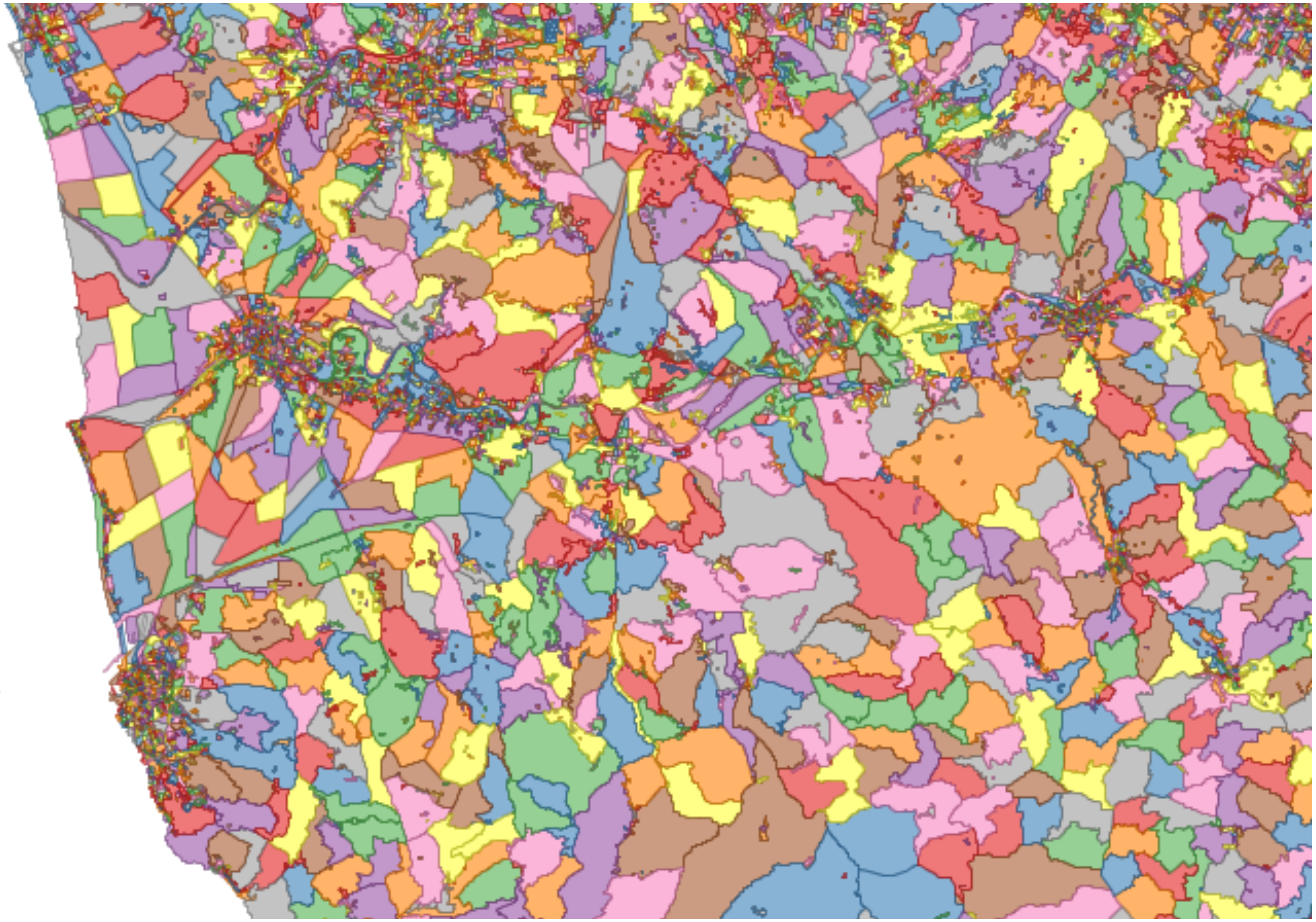
Mining Geographical Mobility Networks

S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, F. Giannotti

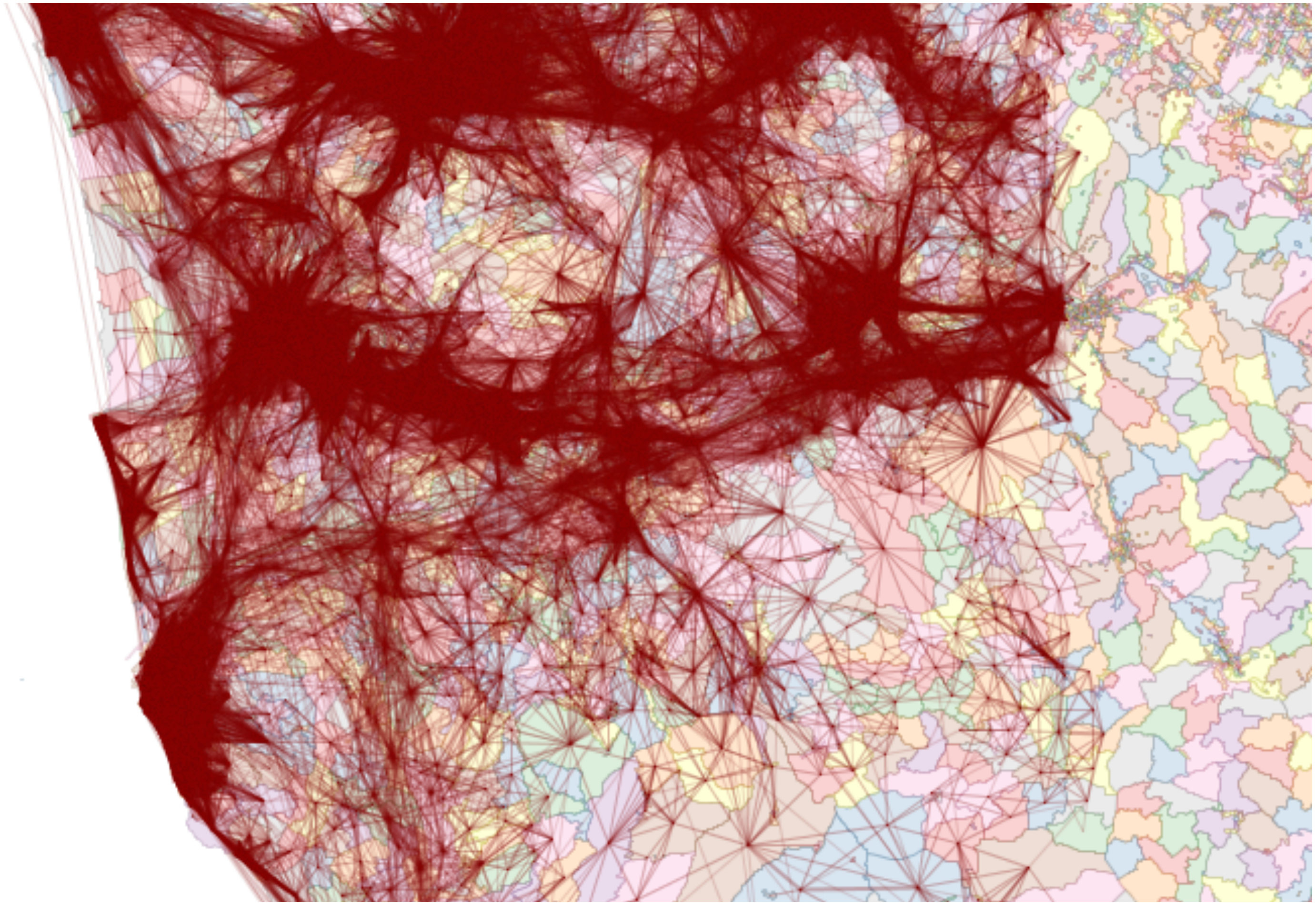
Discovering the Geographical Borders of Human Mobility. KI - Künstliche Intelligenz, 2012.



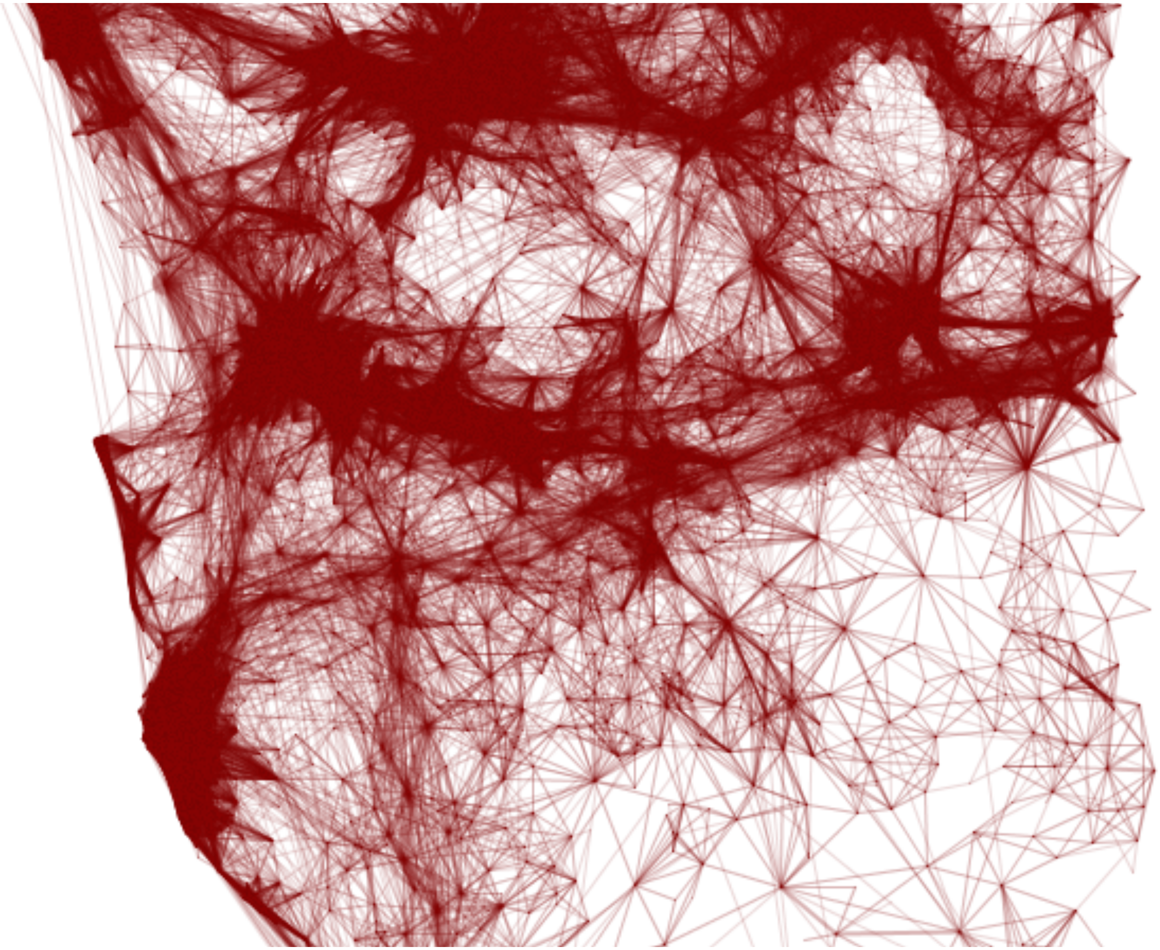
Step 1: spatial regions



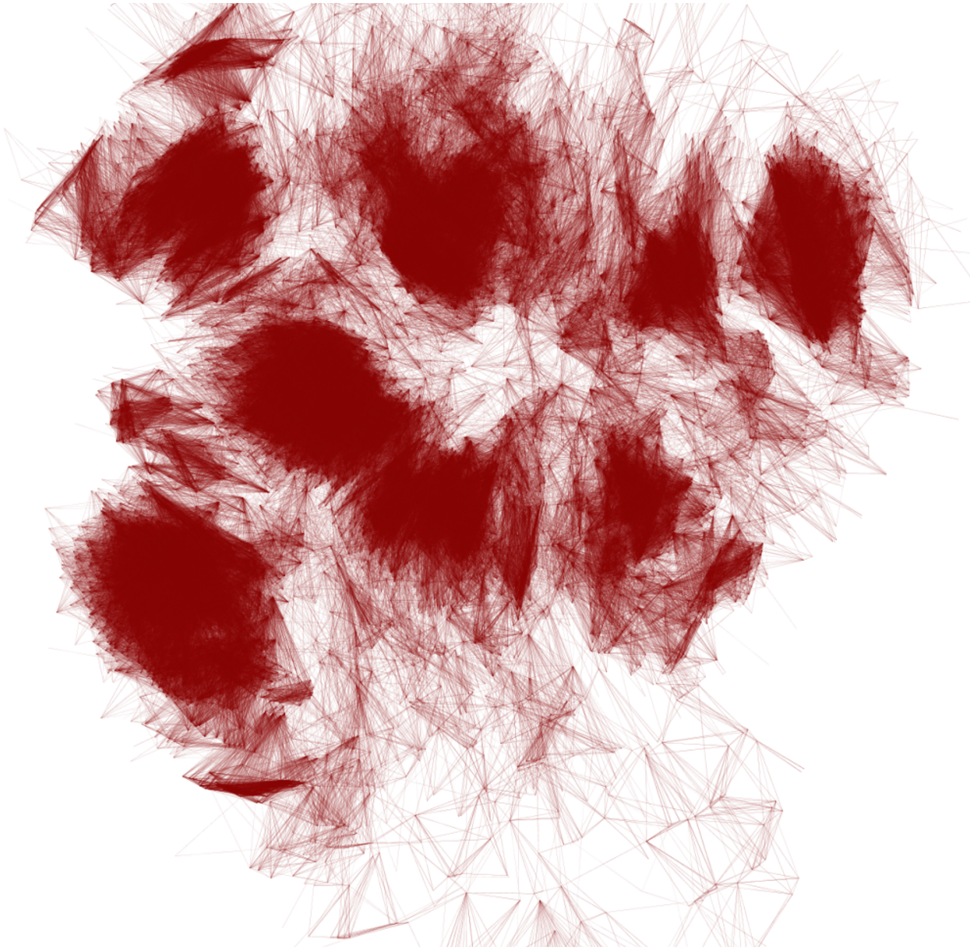
Step 2: evaluate flows among regions



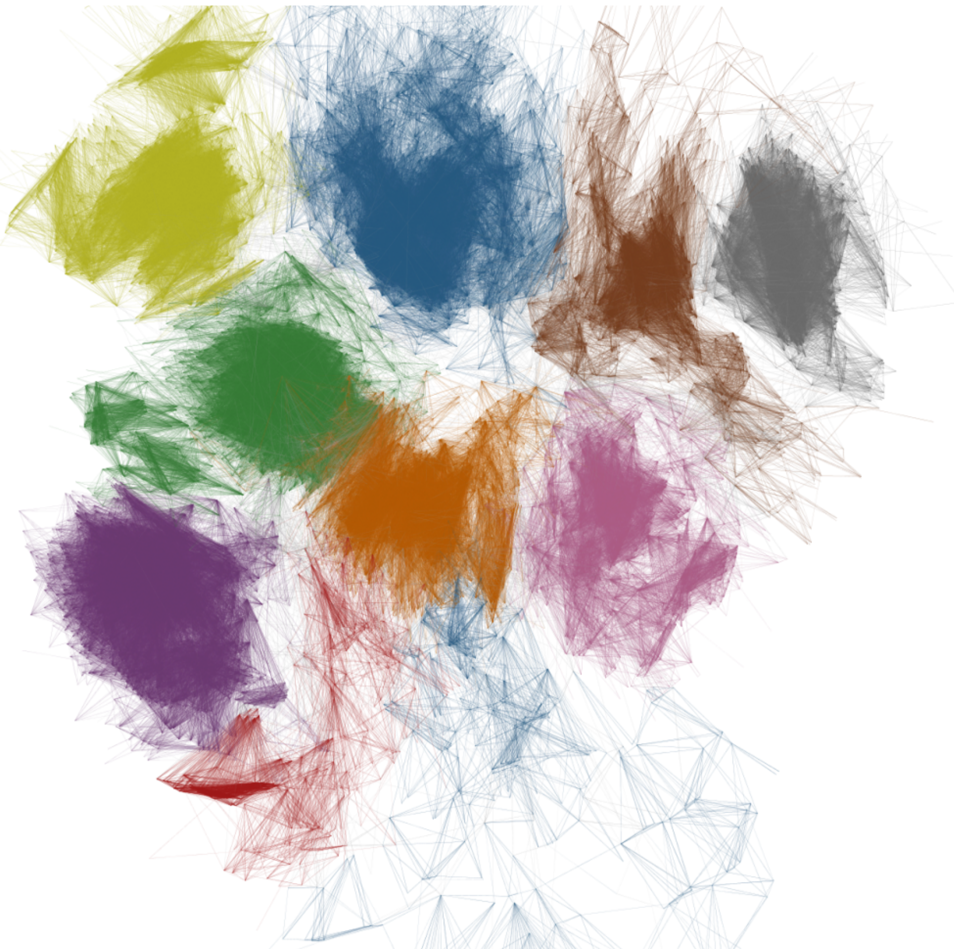
Step 3: forget geography



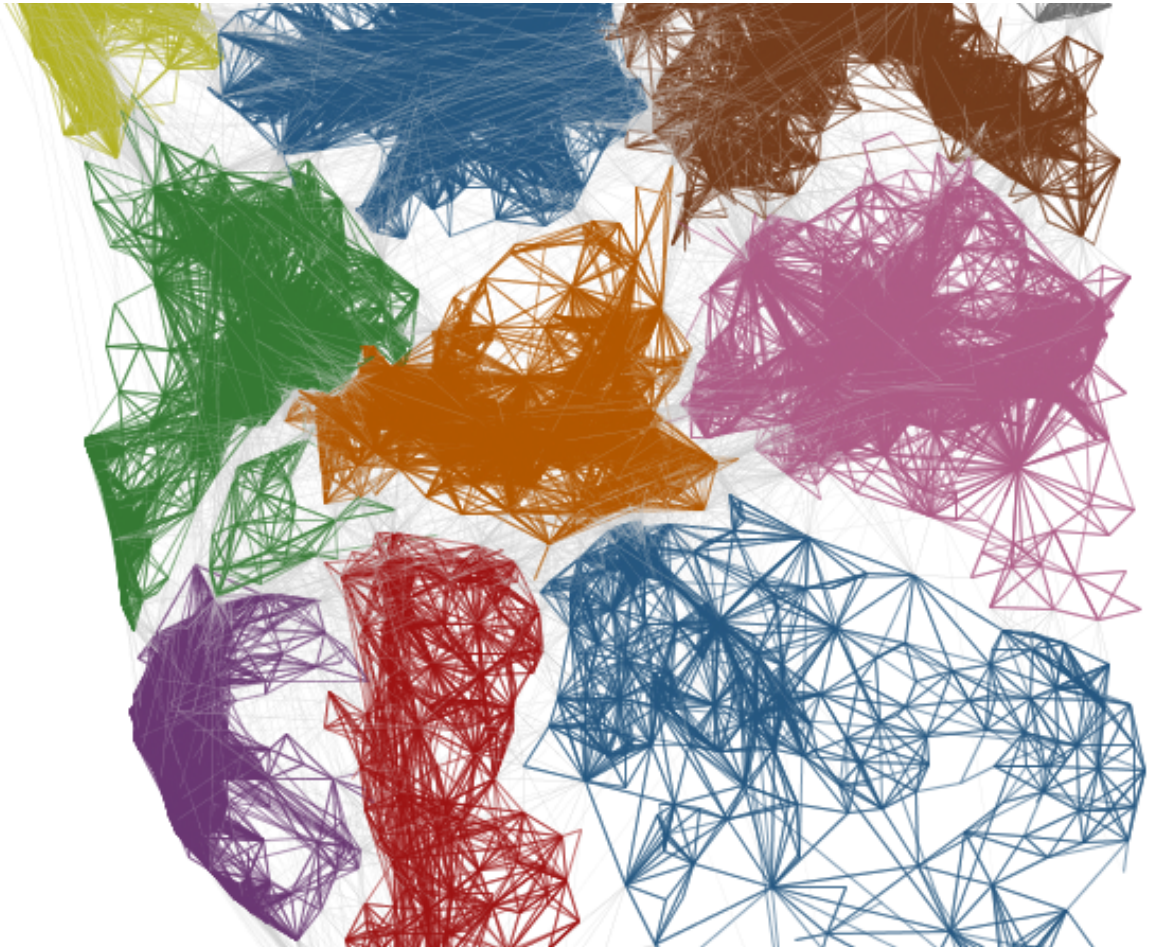
Step 4: perform community detection



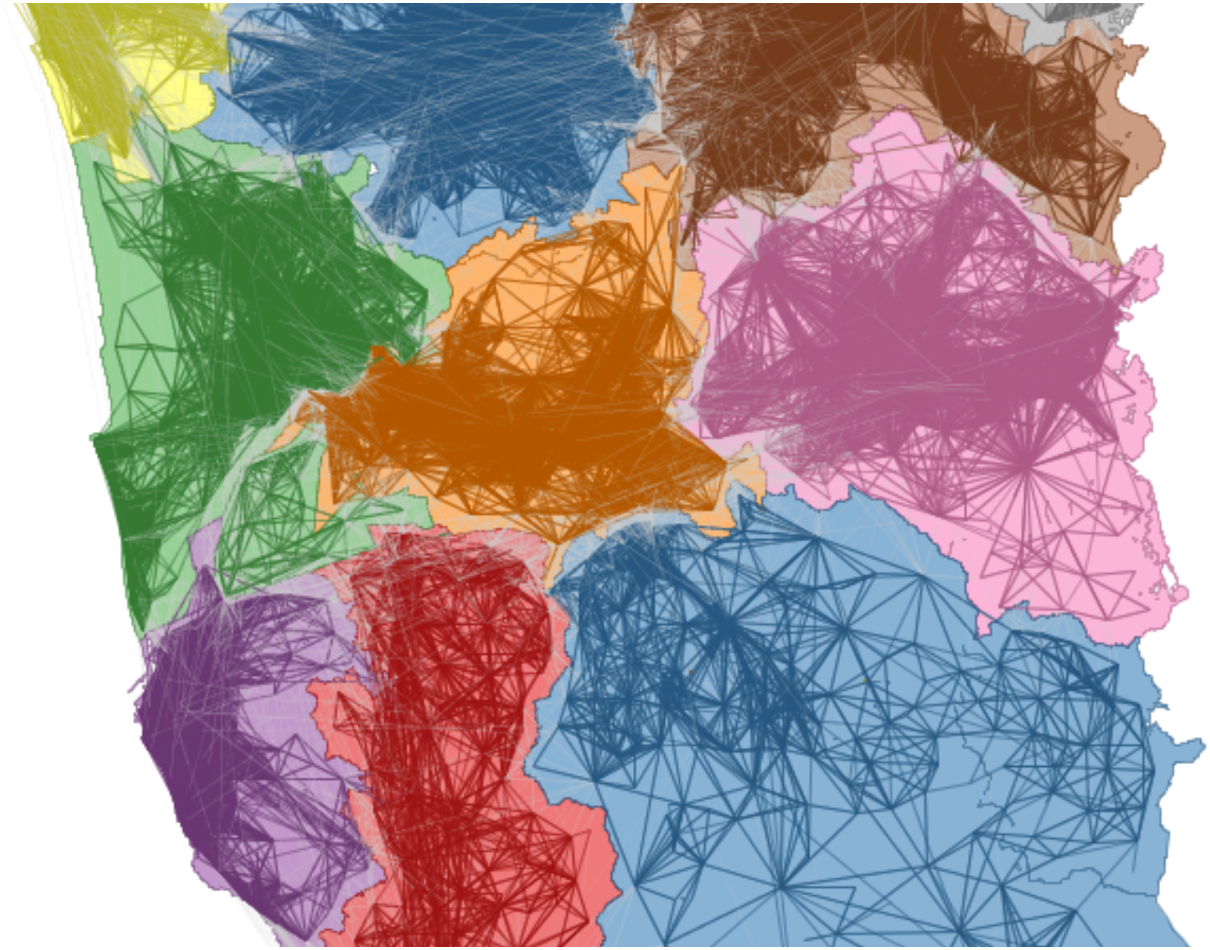
Step 4: perform community detection



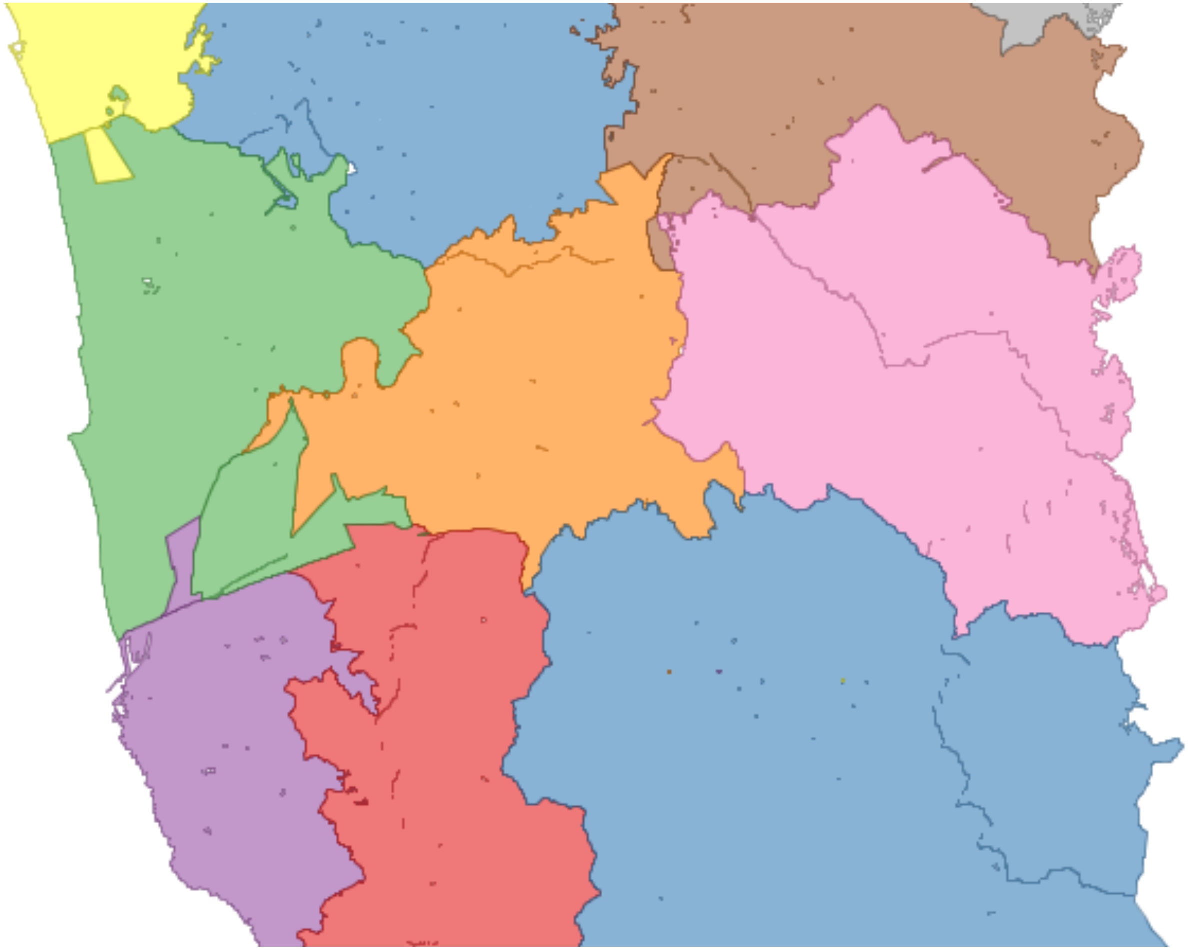
Step 5: map back to geography



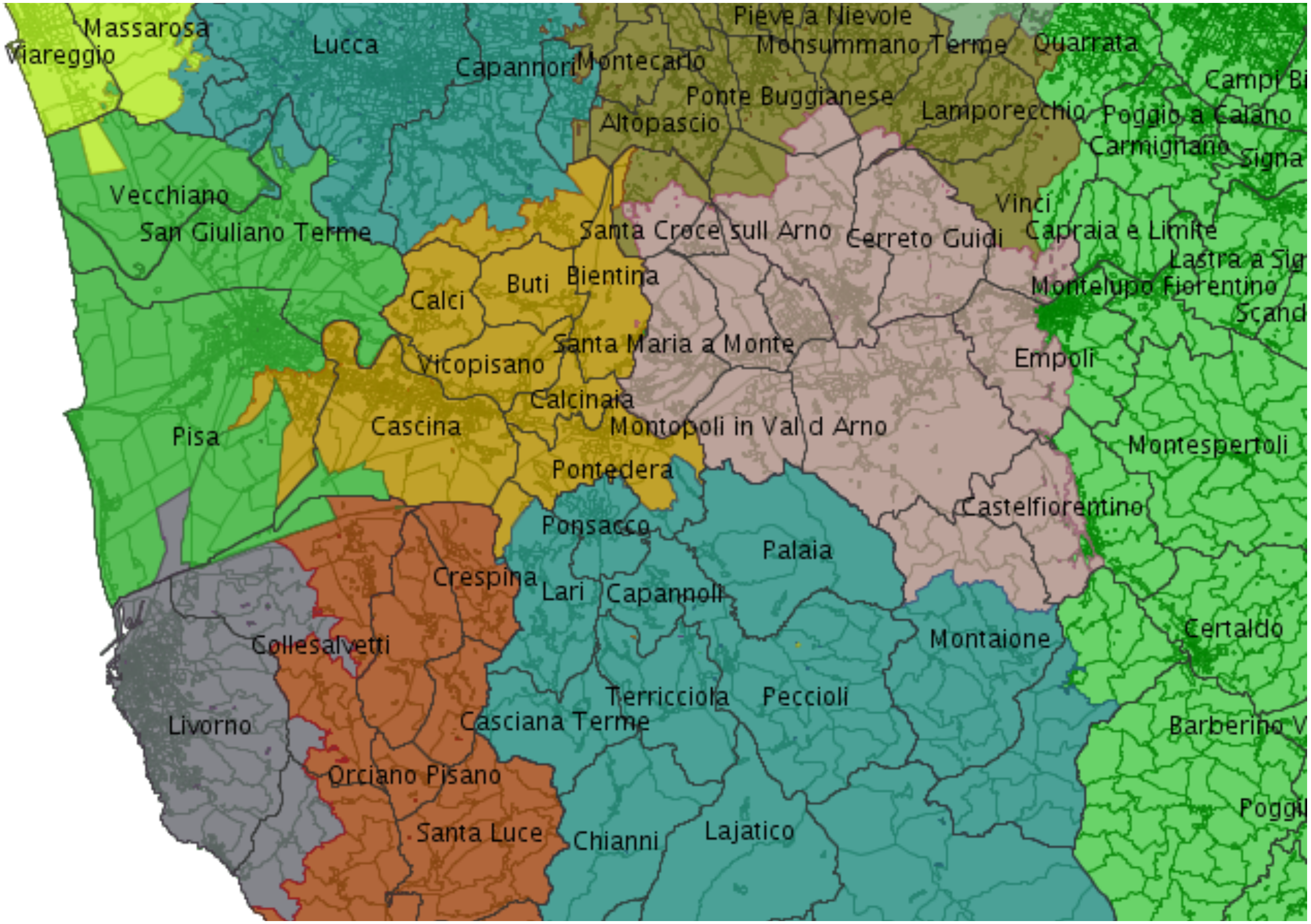
Step 6: draw borders



Final result



Final result vs. municipality borders



BREATH AND ASK

THANK YOU !

Questions?