(MULTI) EGOCENTERED COMMUNITIES

Jean-Loup Guillaume - jean-loup.guillaume@univ-lr.fr

Joint work with M. Danisch – B. Le Grand

Supported by CODDDE ANR-13-CORD-0017-01 and

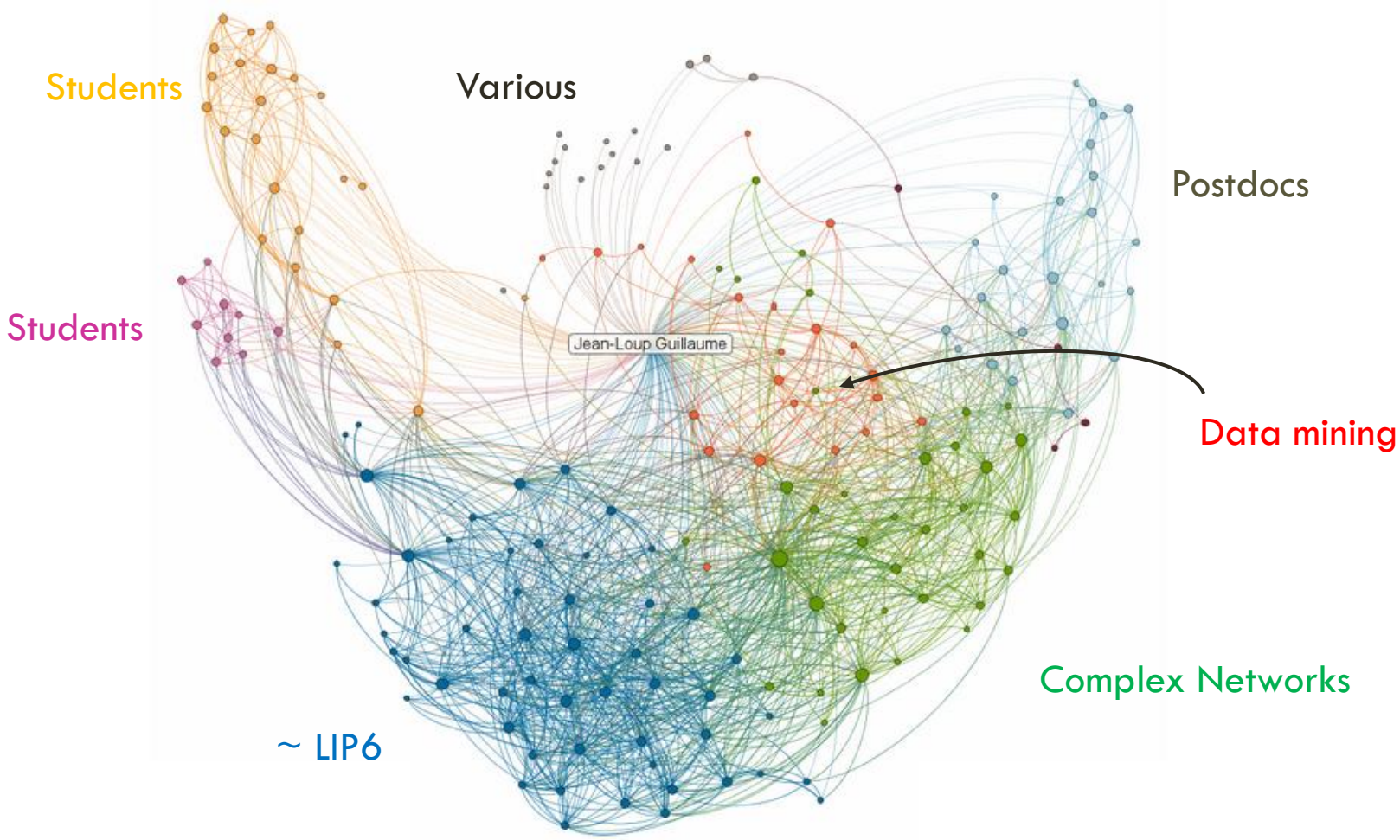
REQUEST projet Investissement d'avenir, 2014-2017

Laboratoire Informatique Image Interaction (L3I)

Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France

Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

LINKEDIN/INMAPS - 04/2014



COMPLEX NETWORKS

Relational data modeled using graphs:

- Computer science: web, the Internet, email, P2P, …
- Social sciences: friendships, collaborations, phone calls…
- Biology: neurons, proteins interactions, ethology, …
- Linguistics, transportation, …

Many common topological properties:

- Low average distance / small world effect
- Heterogeneous degrees / scale free networks
- Clustering / variation of density and communities
- Frequent motifs / triangles or more complex subgraphs

COMMUNITY DETECTION - APPLICATIONS

(online) Social networks:

- Automatic identification of groups
- Classification of "unknown" persons

Biology / epidemiology :

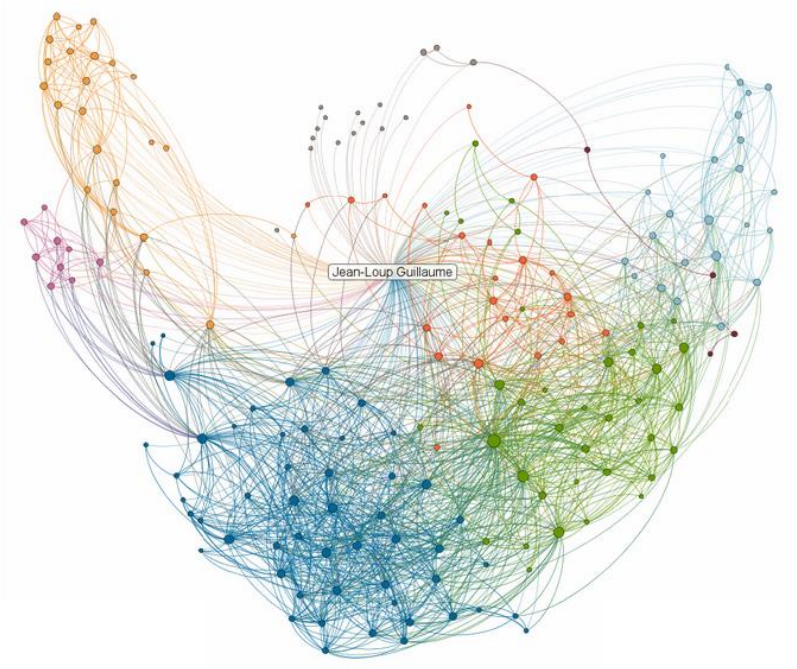
- Brain: identification of functional areas
- Proteins: prediction of the function of proteins

Graph visualization/navigation

Images/video segmentation

Hierarchical routing in networks

…



COMMUNITY DETECTION - APPLICATIONS

(online) Social networks:

- Automatic identification of groups
- Classification of "unknown" persons

Biology

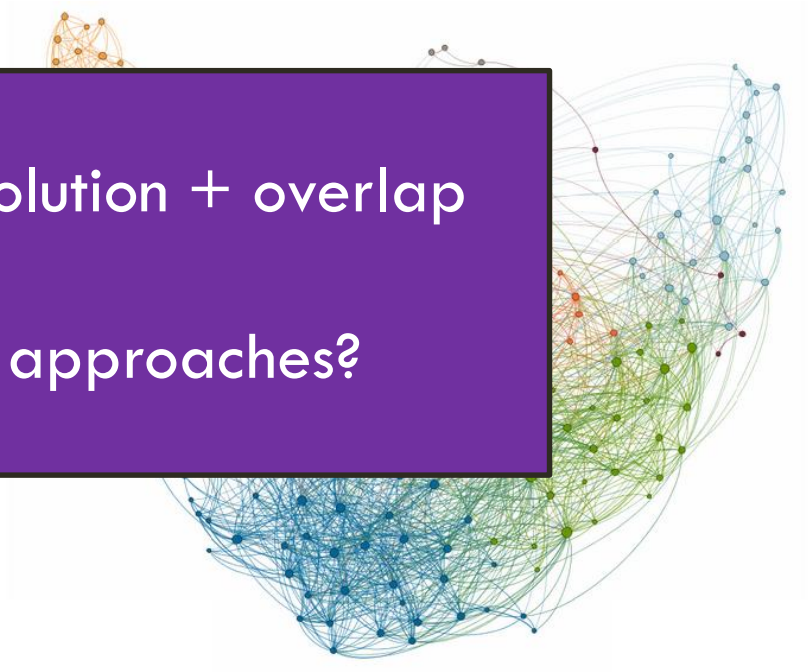
- Bra
  are
- Prot
  of p

Graph visualization/navigation

Images/video segmentation

Hierarchical routing in networks

…



Massive datasets + evolution + overlap

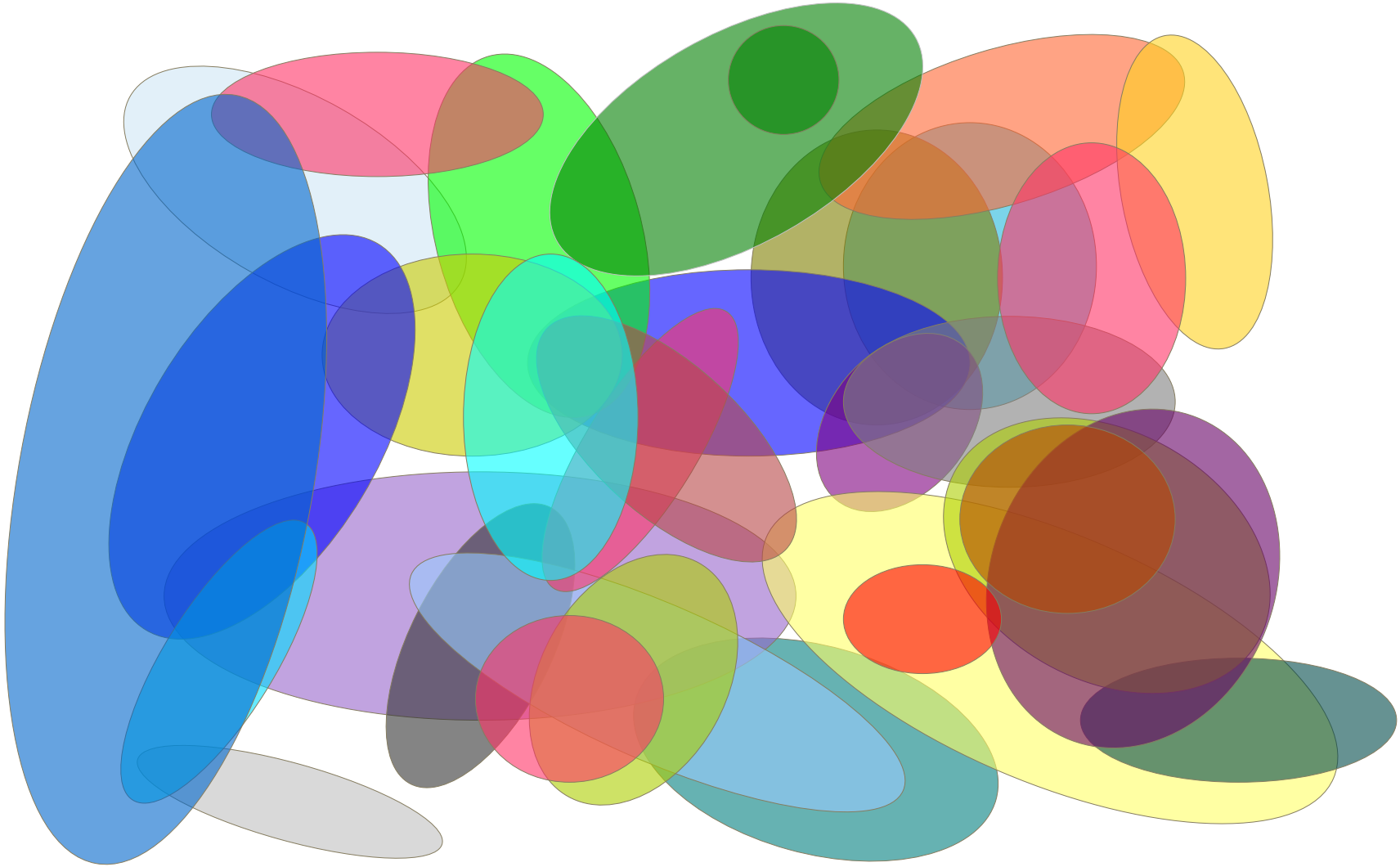
↓

Egocentered/local approaches?

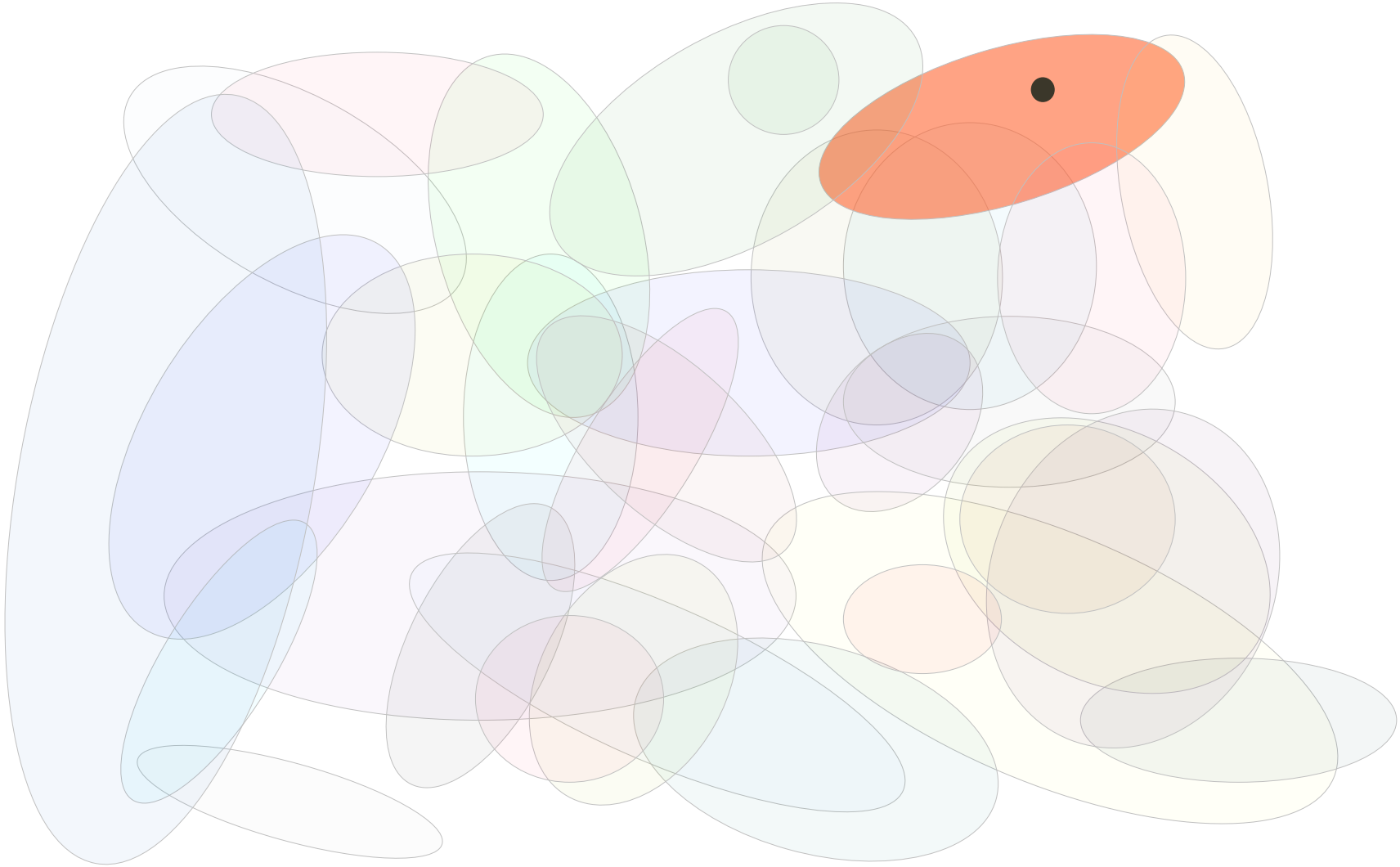
LOCAL / EGOCENTERED APPROACHES

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

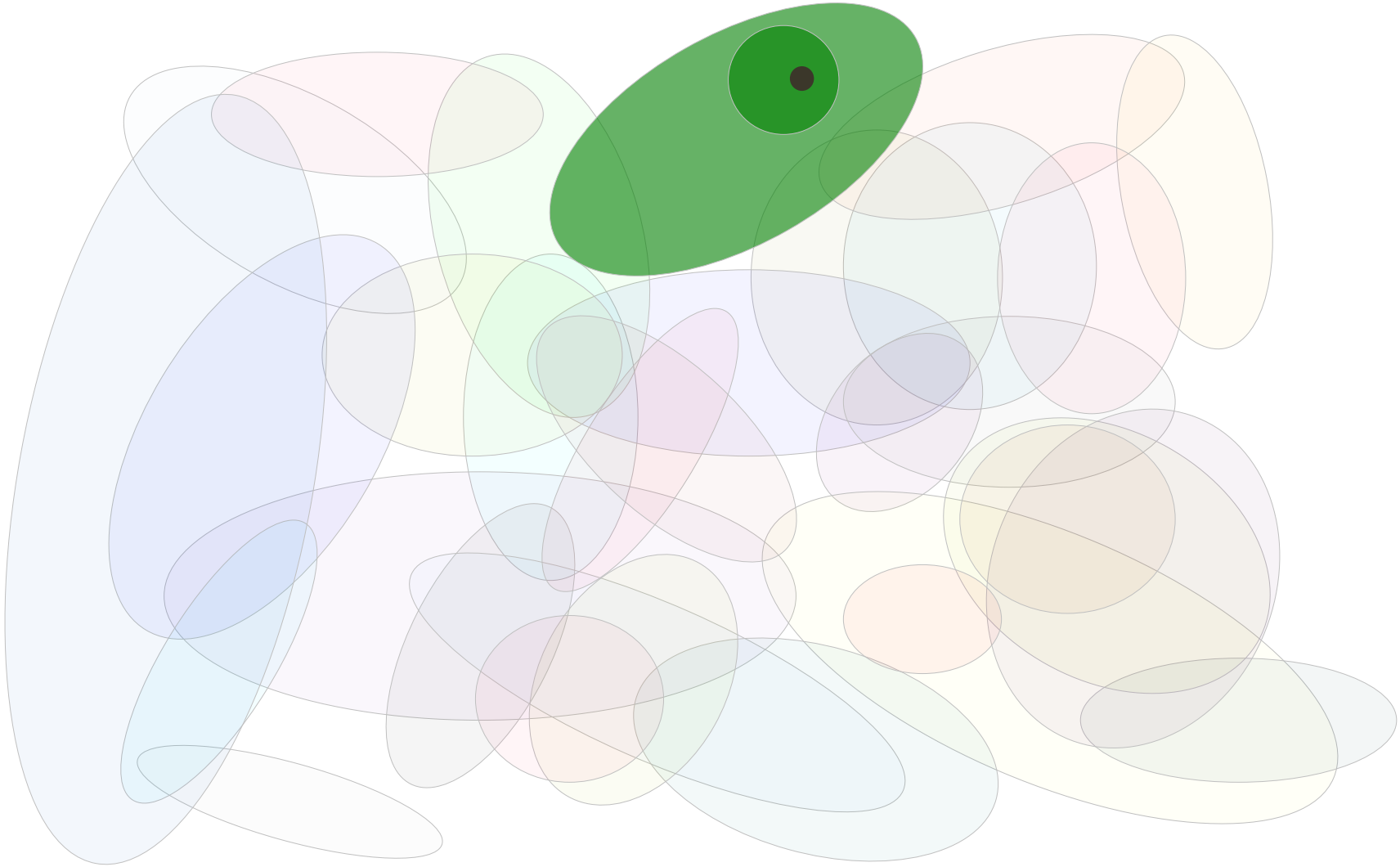
OVERLAPPING COMMUNITY STRUCTURE



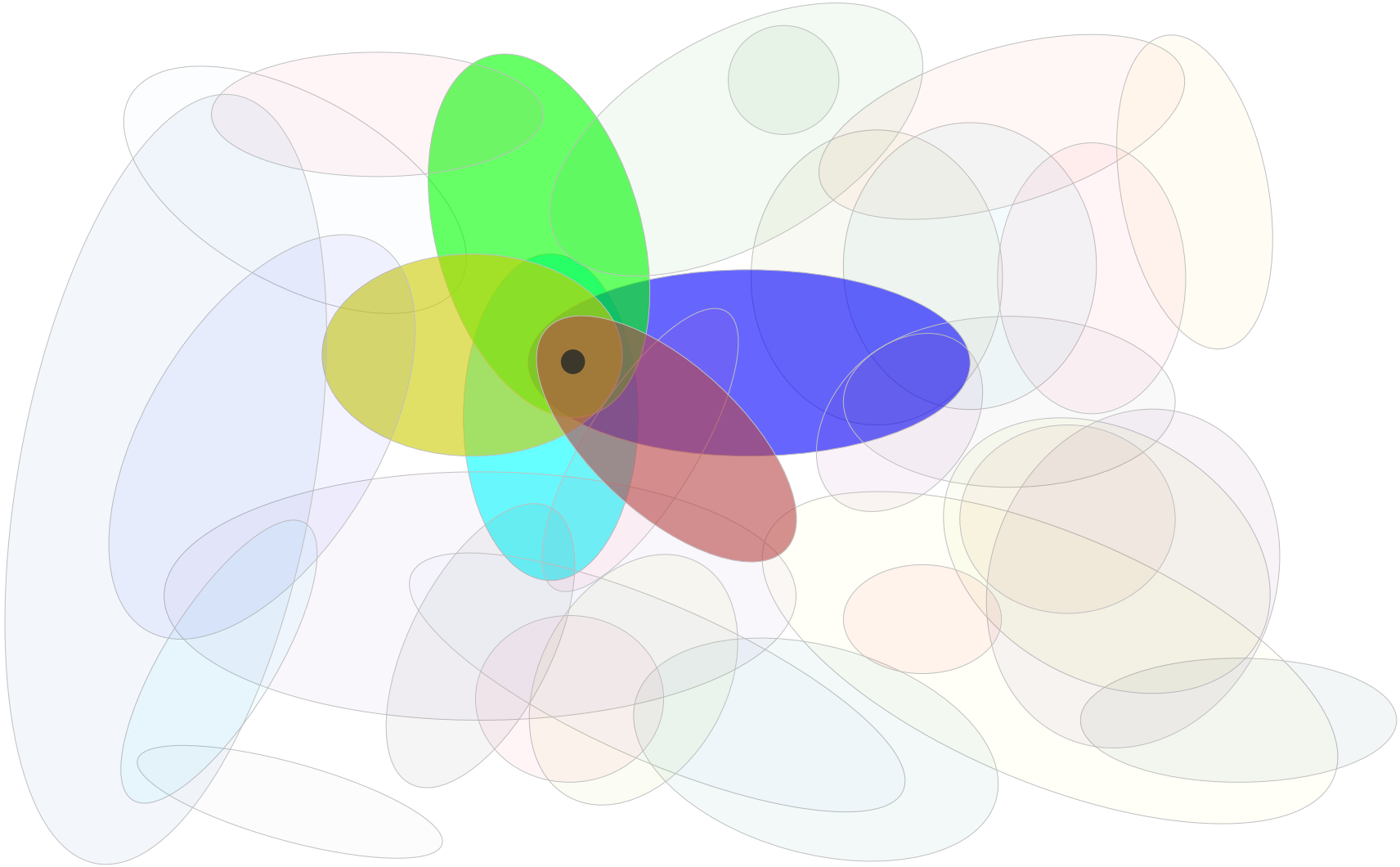
LOCAL / EGOCENTERED COMMUNITIES



LOCAL / EGOCENTERED COMMUNITIES



LOCAL / EGOCENTERED COMMUNITIES



QUALITY OR PROXIMITY?

Quality functions:

- Tell whether a set S is a good community or not
- Generally based on the links inside S vs. going outside S

Proximity measures:

- Given two nodes, tell how close they are

Given a node u:

- Quality functions find “a community” of u
  - Generally done in a greedy fashion starting from u
- Proximity measures identify nodes close to u (rank nodes by proximity)
  - Good measures should clearly indicate the border of the community

QUALITY - CONDUCTANCE

For a (small) set of nodes S

[Shi and Malik, 2000; Andersen FOCS 2006]

$$\phi_{approx}(S) = \frac{degree_{out}(S)}{degree(S)} = \frac{degree_{out}(S)}{degree_{in}(S) + degree_{out}(S)}$$

If S is large

$$\phi(S) = \frac{degree_{out}(S)}{\min(degree(S), degree(\bar{S}))}$$

- Exact minimization is a hard problem
- Very used for local communities

QUALITY - CONDUCTANCE LIKE FUNCTIONS

Relative density or local modularity

[Clauset Phys Rev E 2005; Luo, Wang and Promislow 2008]

$$rd(S) = \frac{degree_{in}(S)}{degree_{out}(S)}, rd2(S) = \frac{degree_{\text{in}}(S)}{degree_{in}(S) + degree_{out}(S)}$$

Controlling the size of the community

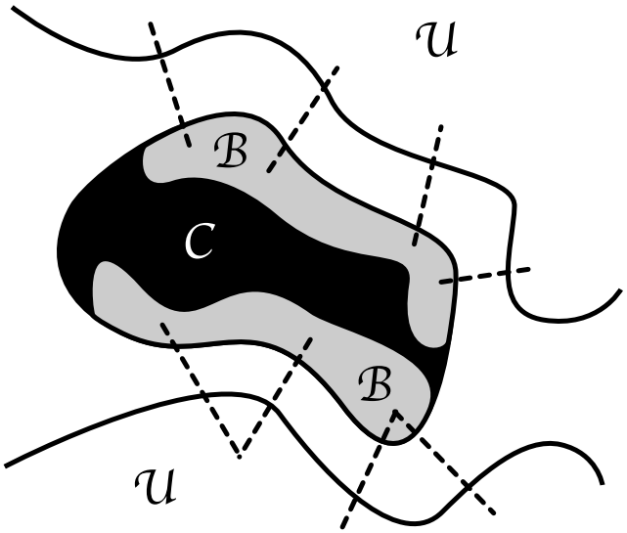
[Lancichinetti et al. 2009]

$$\phi_\alpha(S) = \frac{degree_{out}(S)}{(degree_{in}(S) + degree_{out}(S))^\alpha}$$

QUALITY - INSIDE/BORDER AND OUTSIDE

Some quality function restrict to the study of the border

[Clauset 2005]



$$R = \frac{links((B \cup C) \leftrightarrow B)}{degrees(B)}$$

Further improvements

[Chen, ASONAM 2009; Ngonmang et al. PPL 2012; …]

Image taken from Clauset 2005

TRIANGLE BASED APPROACH

Count in and out-triangles rather than in and out-links:

[Friggeri et al. SocialCom 2011]

$$C(S) = \frac{\Delta_{in}(S)}{\binom{|S|}{3}} \times \frac{\Delta_{in}(S)}{\Delta_{in}(S) + \Delta_{out}(S)}$$

- First term: triangle density inside S
- Second term: triangle isolation of S (an out-triangle has one node outside)

Pros:

- Triangles are more likely composed of 3 links of same nature
- Not penalized by outgoing links

OPTIMIZATION OF QUALITY FUNCTIONS

Most heuristics are greedy like:

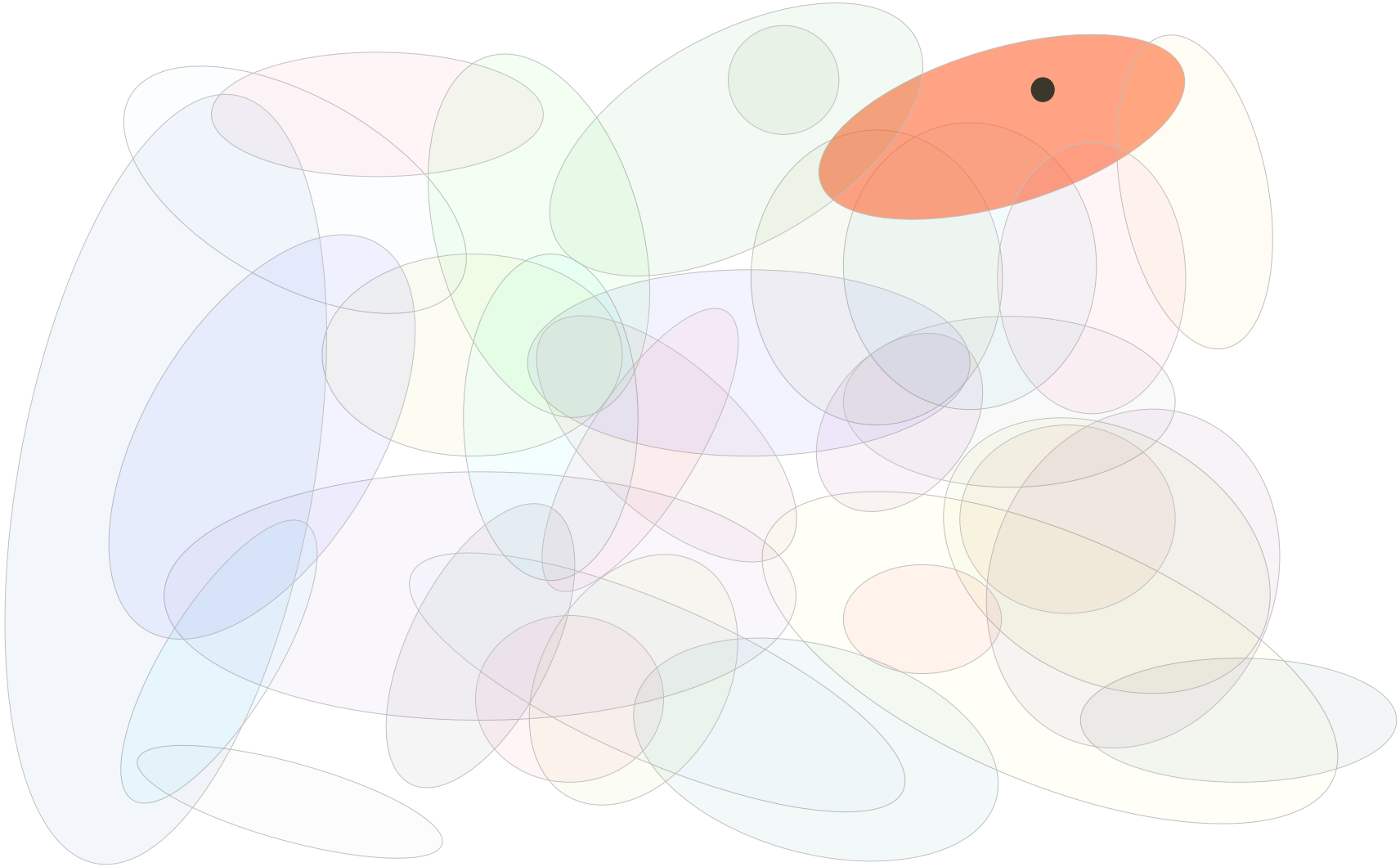
- Start with a community containing only one node of interest
- At each step add the (neighbor) node so as to maximize the gain
- Repeat until no further improvement can be obtained

Potential modifications/improvements

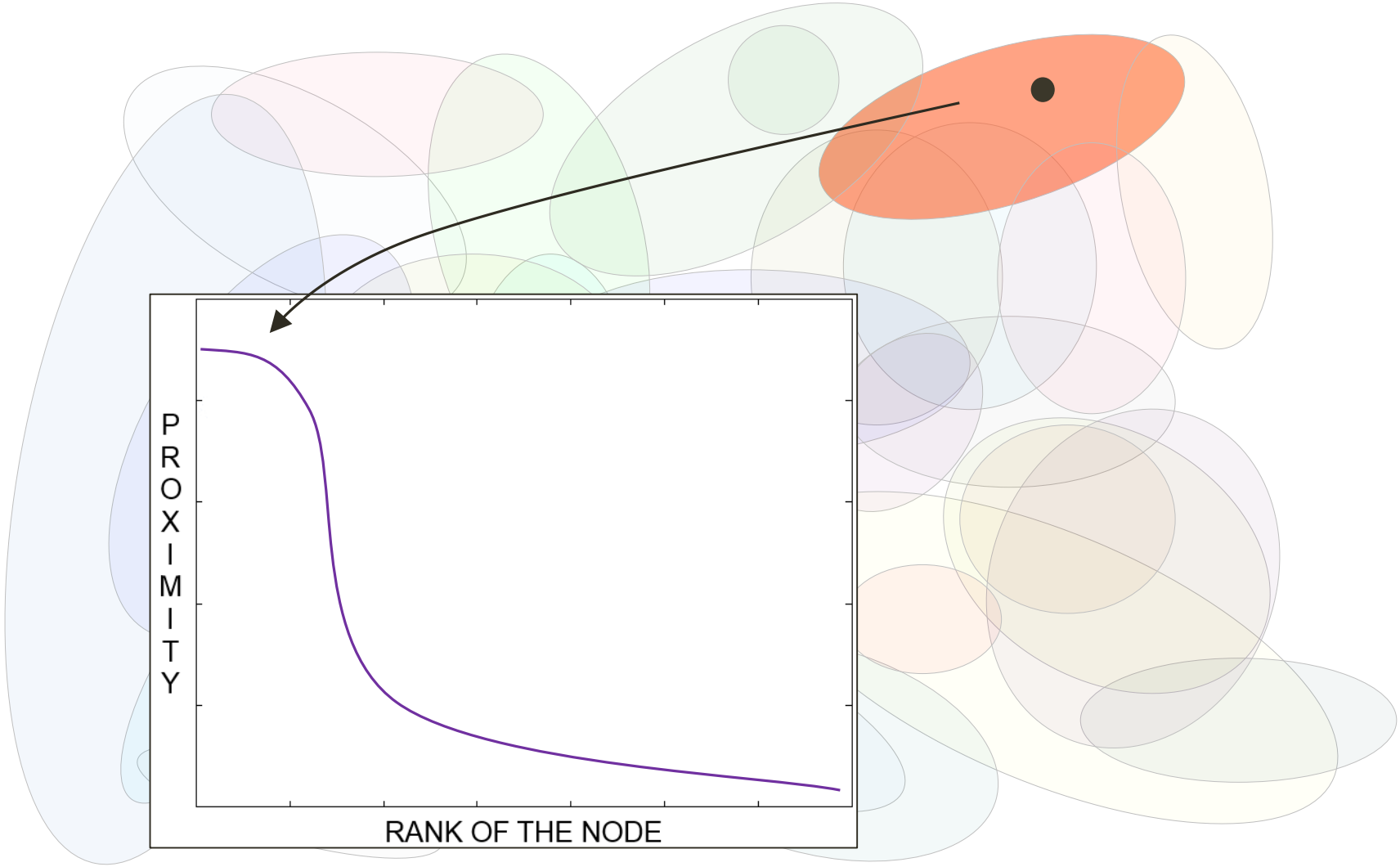
- Step 1:
  - Start with more than one node or with all nodes and remove
- Step 2:
  - Pick a “quality increase” node at random rather than the best one
  - Add simultaneously all “quality increase” nodes rather than one
- Step 3:
  - Add nodes even if the quality decrease (might re-increase later)

Many other optimization techniques

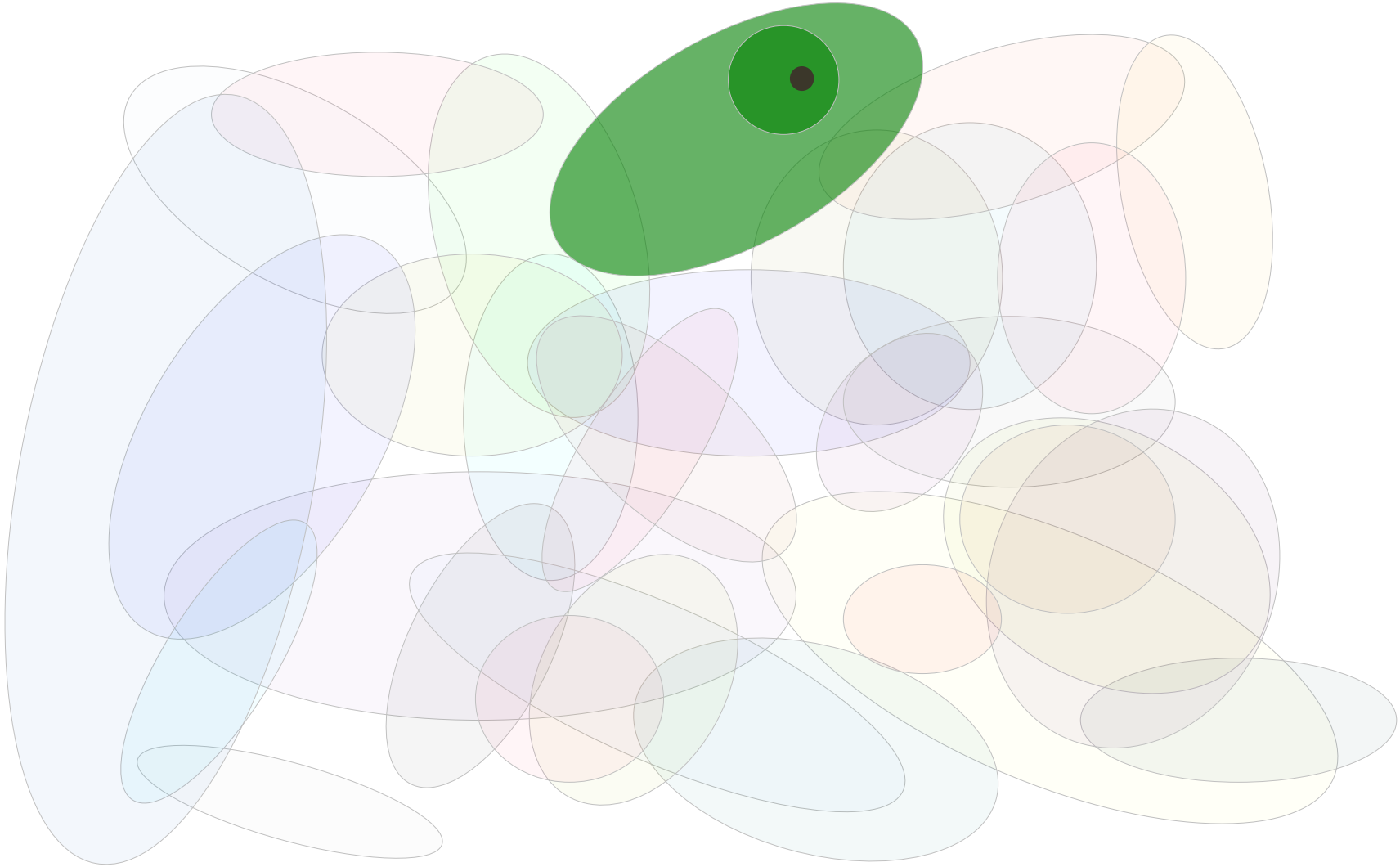
PROXIMITY



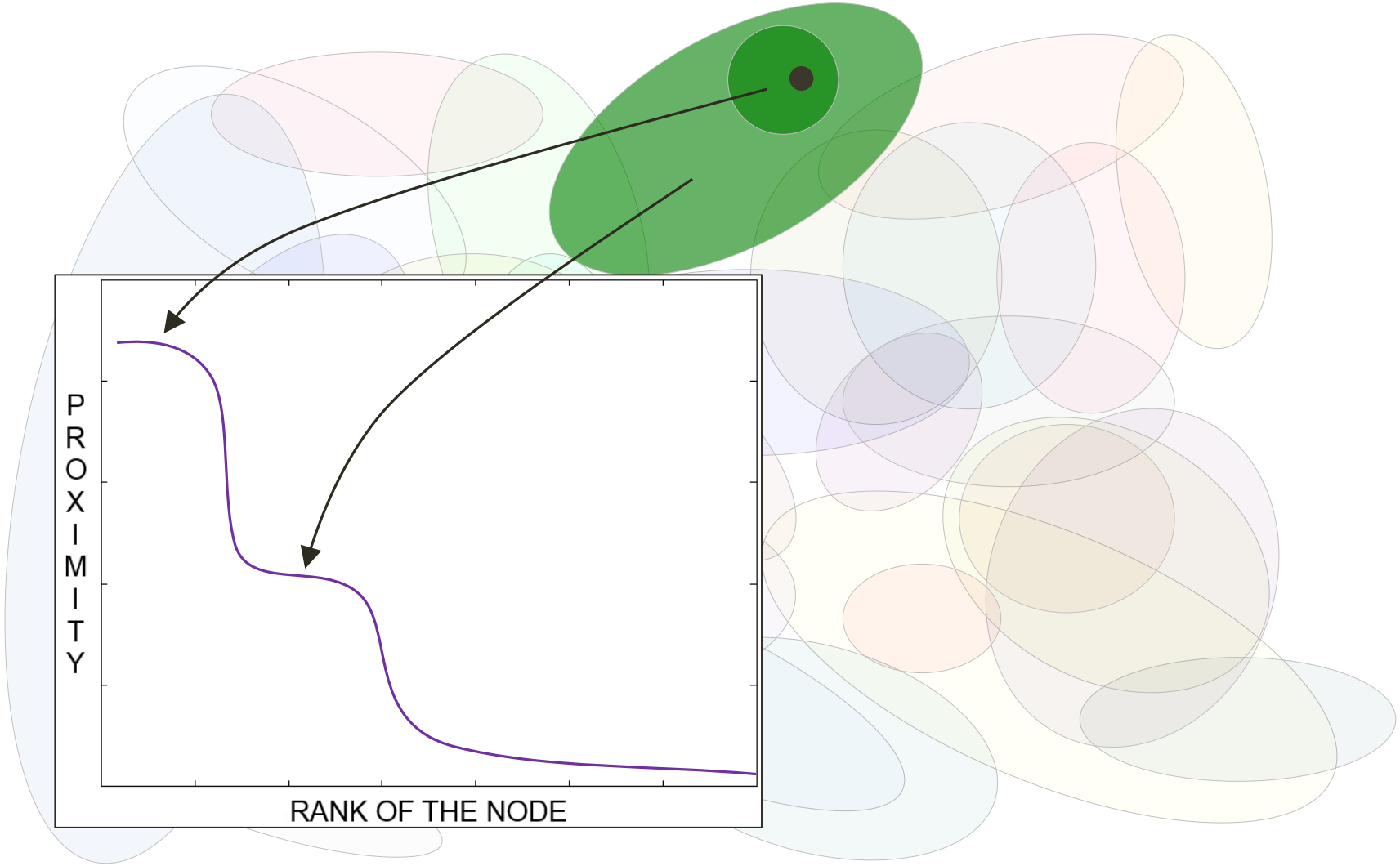
PROXIMITY



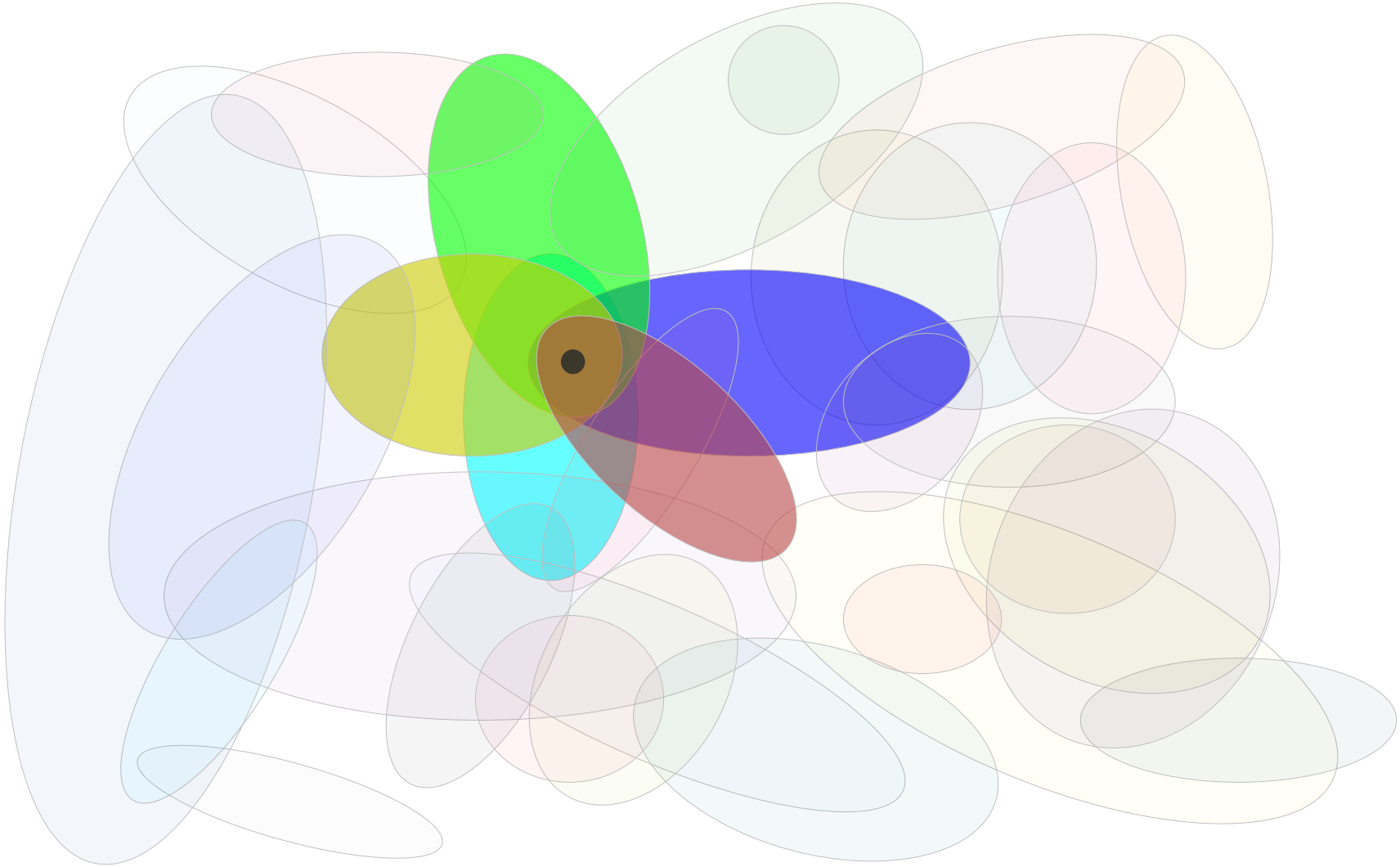
PROXIMITY



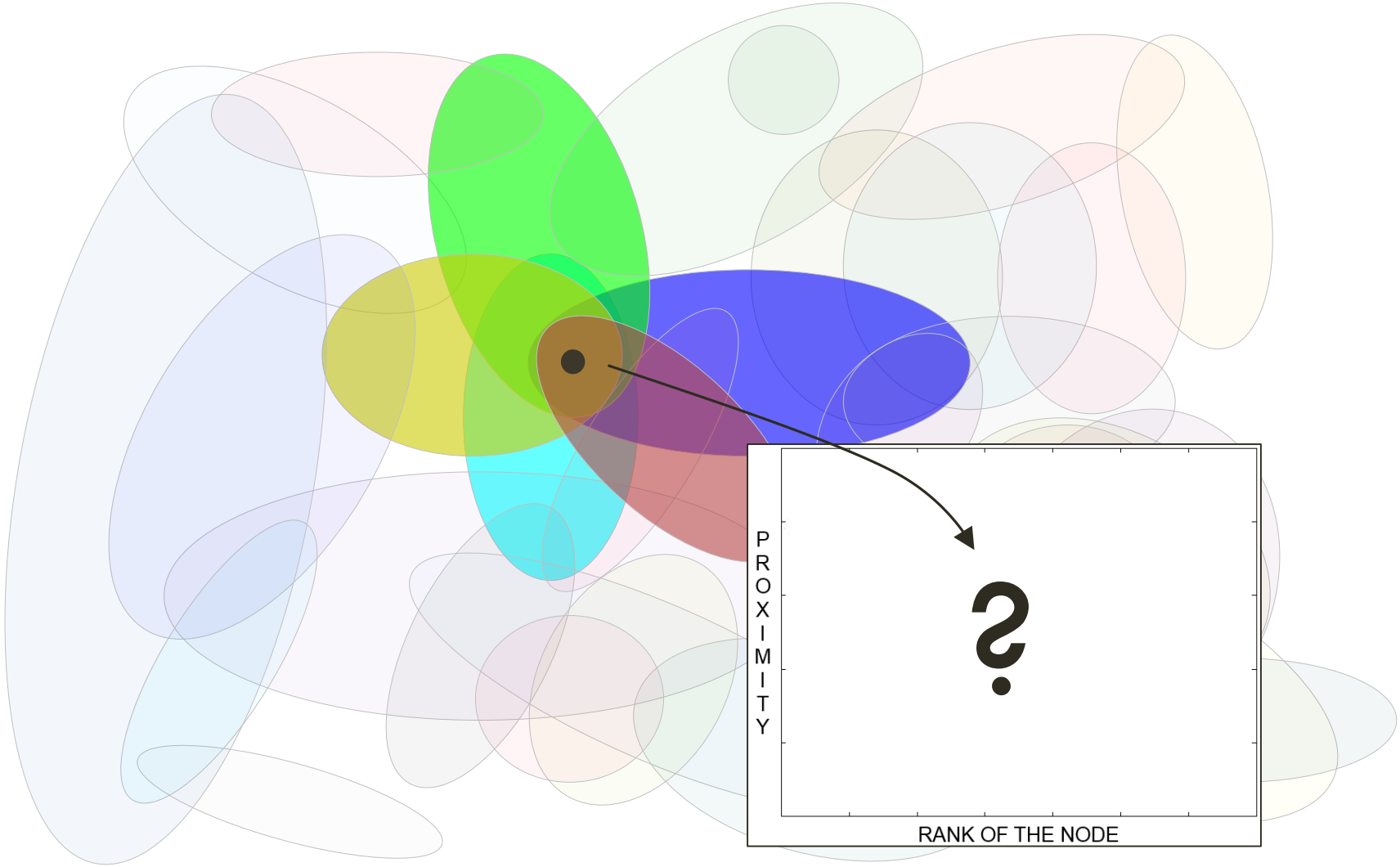
PROXIMITY



PROXIMITY



PROXIMITY



EGOCENTERED COMMUNITIES PARAMETER FREE MEASURE

IJWBC 2013

CompleNet 2013

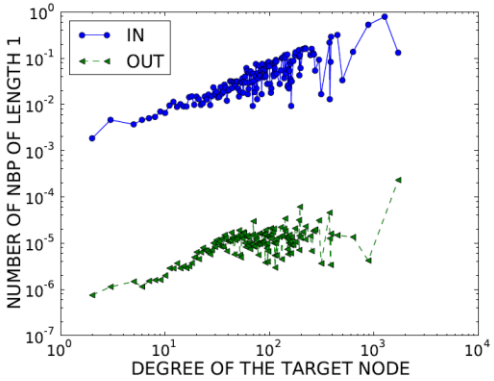
SNAM 2014

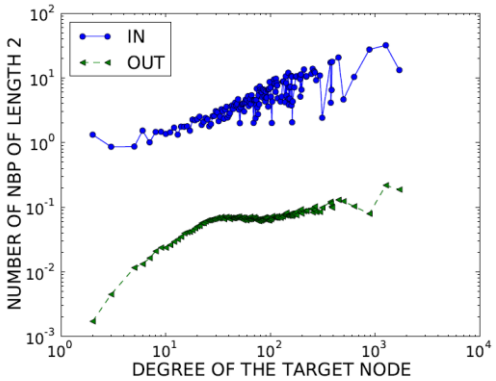
Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

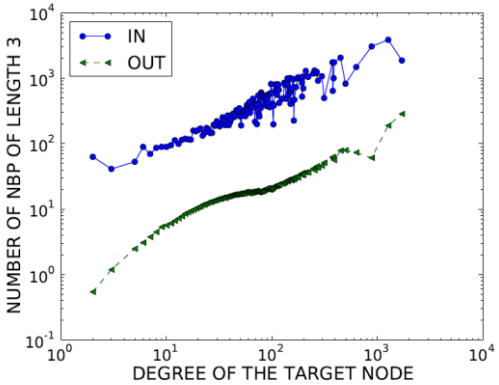
COMMUNITIES EXIST

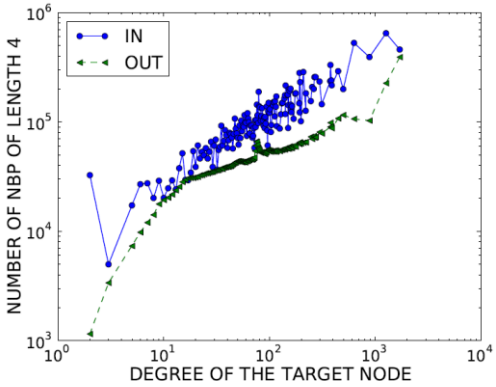
More short paths inside communities than outside

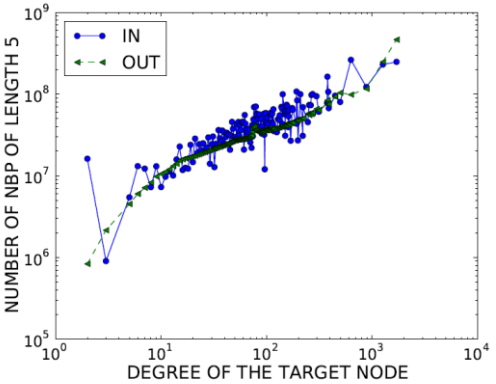
- For all pages from the Wikipedia “graph theory” category

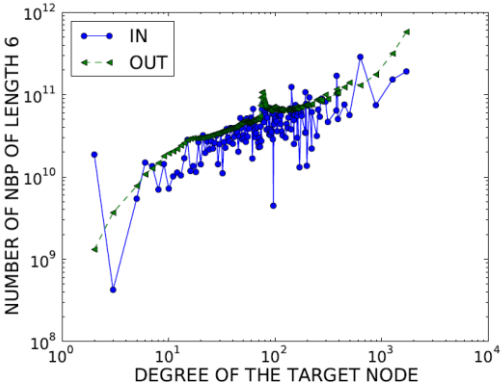












BASIC IDEA

Information may be trapped in communities

Proximity measure based on opinion dynamics

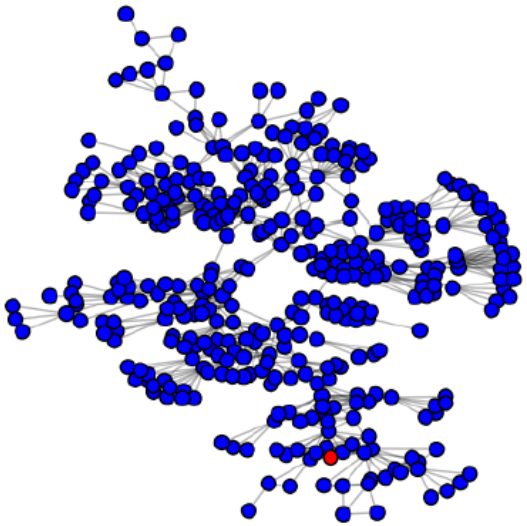
- Node of interest have a fixed opinion equal to 1
- Each node takes the average opinion of its neighbors
- Opinion is carried over from node to node

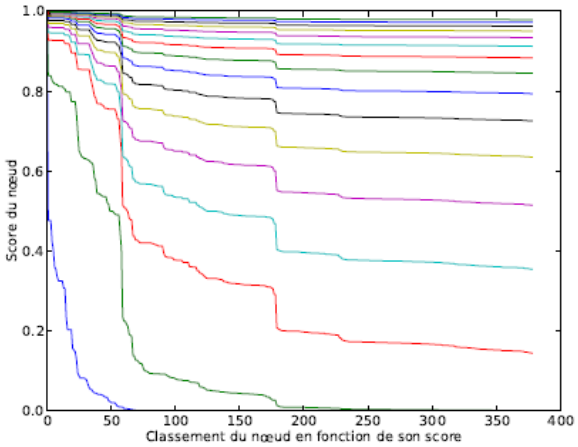
Close to random walks

DEFINITION / COMPUTATION

$$X_t = MX_{t-1}$$ AVERAGING

$$X_t^i = 1$$ RESETTING



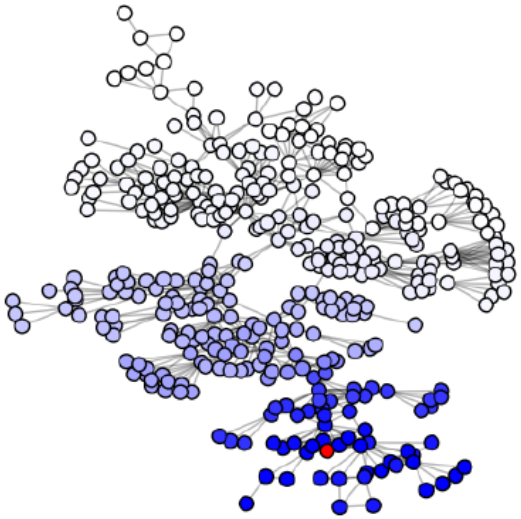


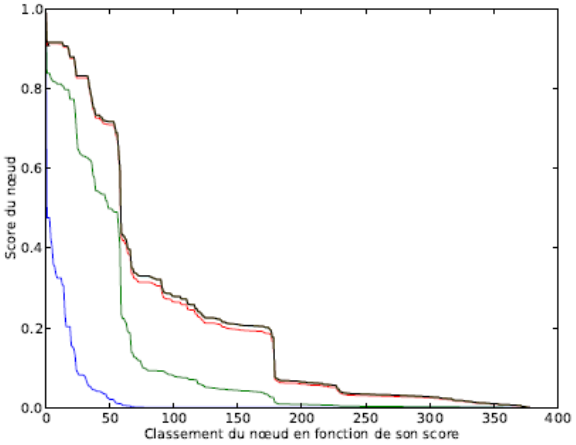
DEFINITION / COMPUTATION

$$X_t = MX_{t-1} \qquad \text{AVERAGING}$$

$$X_t = \frac{X_t - \min(X_t)}{1 - \min(X_t)} \qquad \text{RESCALING}$$

$$X_t^i = 1 \qquad \text{RESETTING}$$



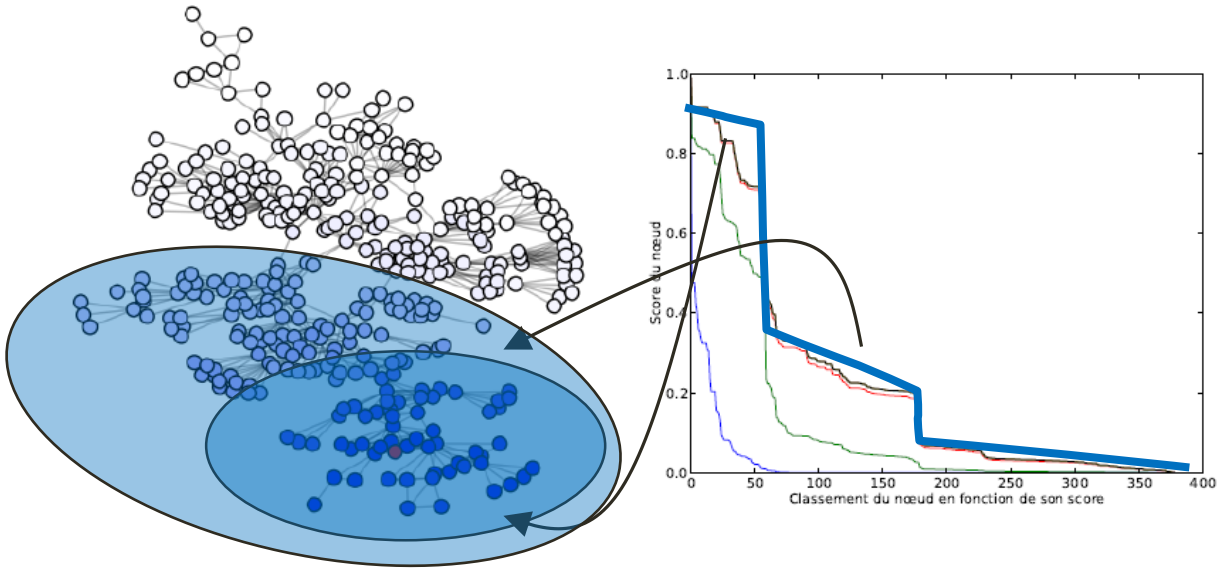


DEFINITION / COMPUTATION

$$X_t = MX_{t-1} \qquad \text{AVERAGING}$$

$$X_t = \frac{X_t - \min(X_t)}{1 - \min(X_t)} \qquad \text{RESCALING}$$

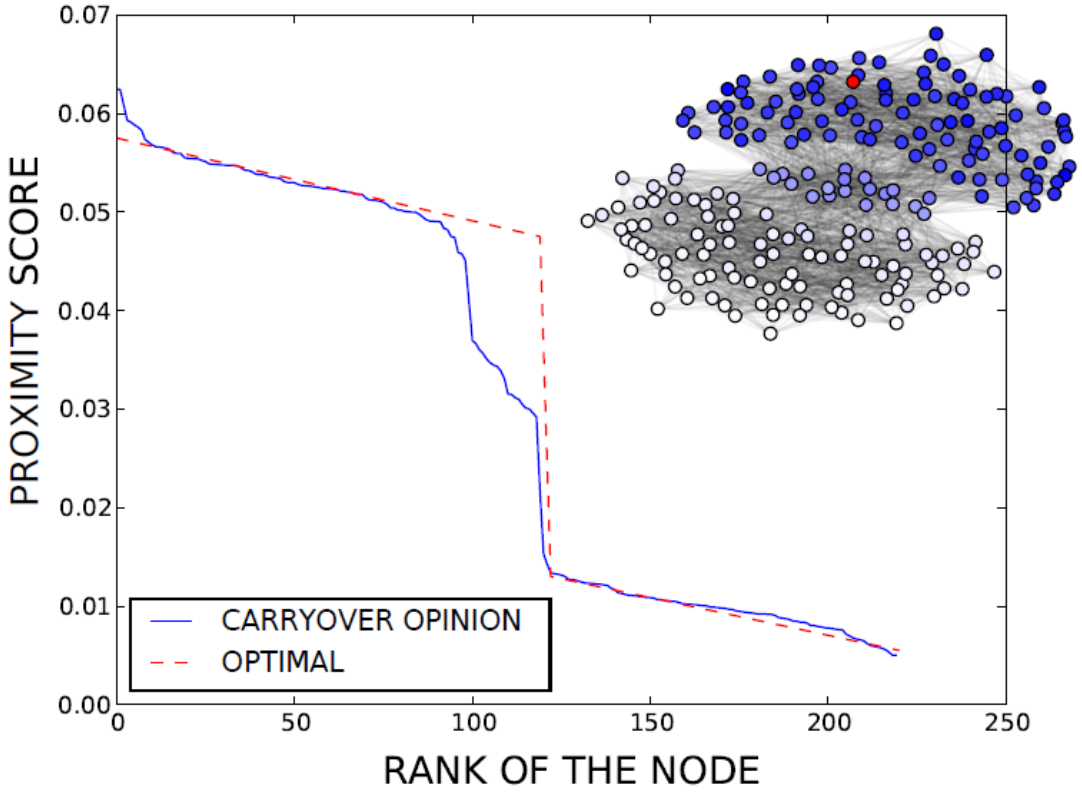
$$X_t^i = 1 \qquad \text{RESETTING}$$



CARRYOVER OPINION - LIMITATIONS

What if a node belong to two communities?

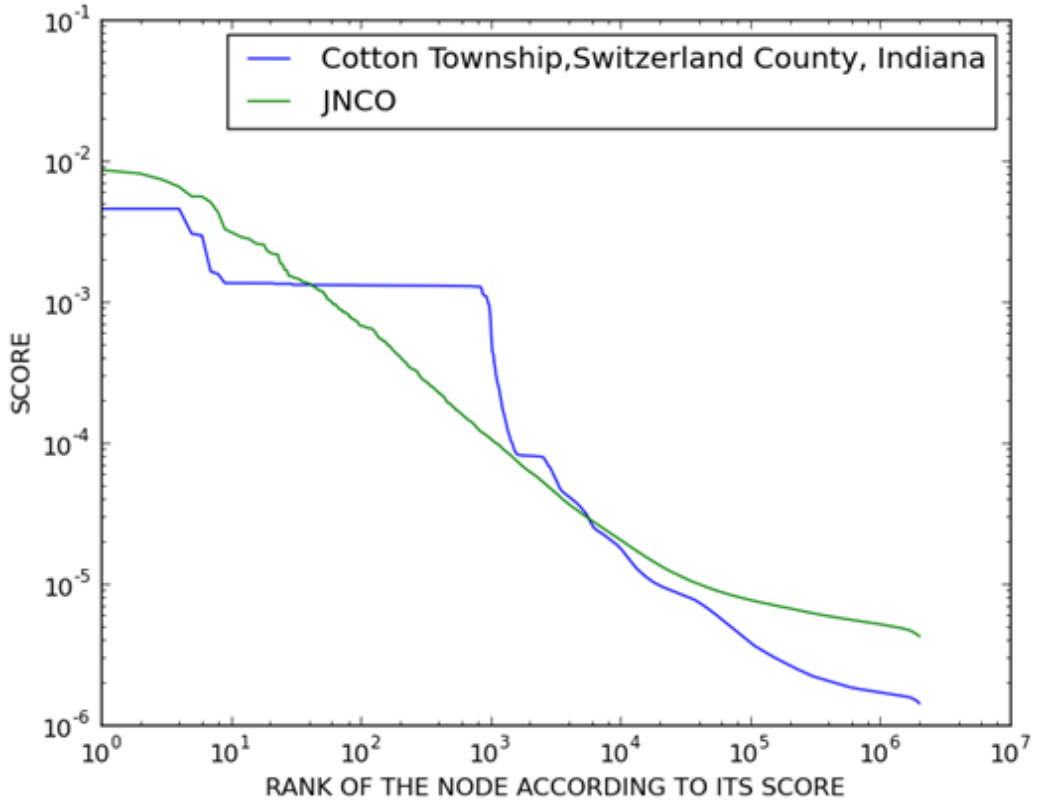
- Expected result?



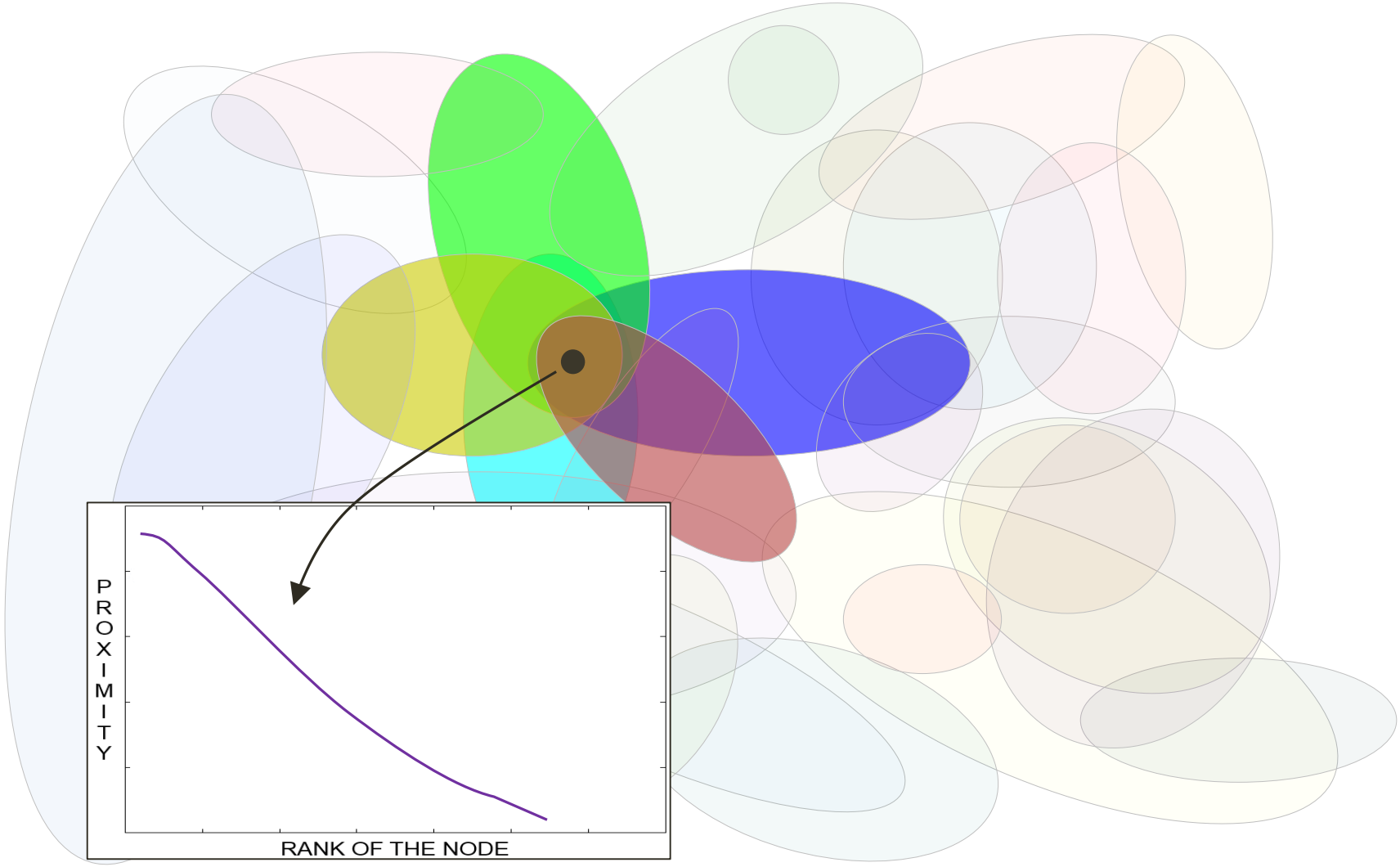
CARRYOVER OPINION - LIMITATIONS

What if a node belong to many communities?

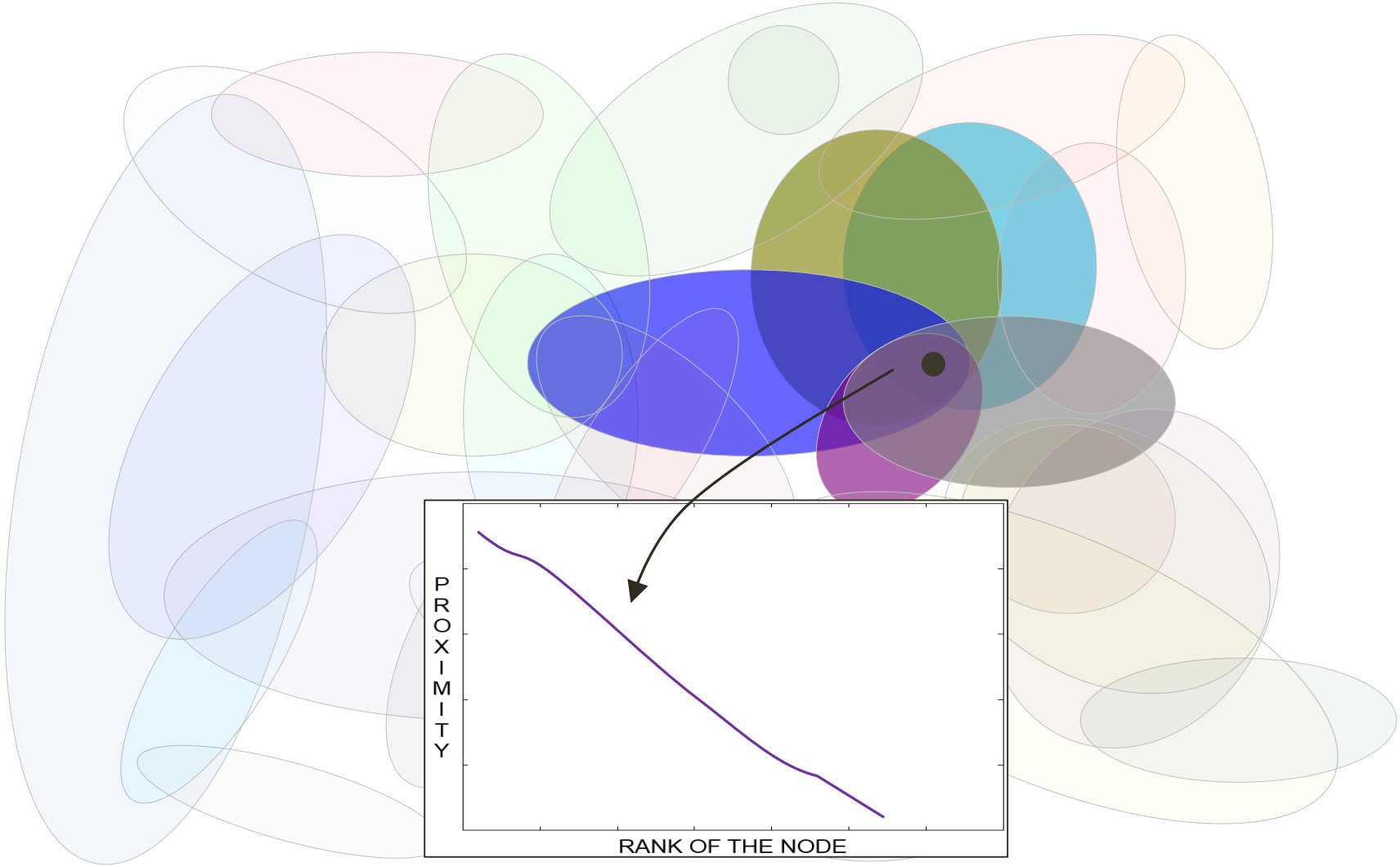
- Of different sizes / not well defined / overlapping



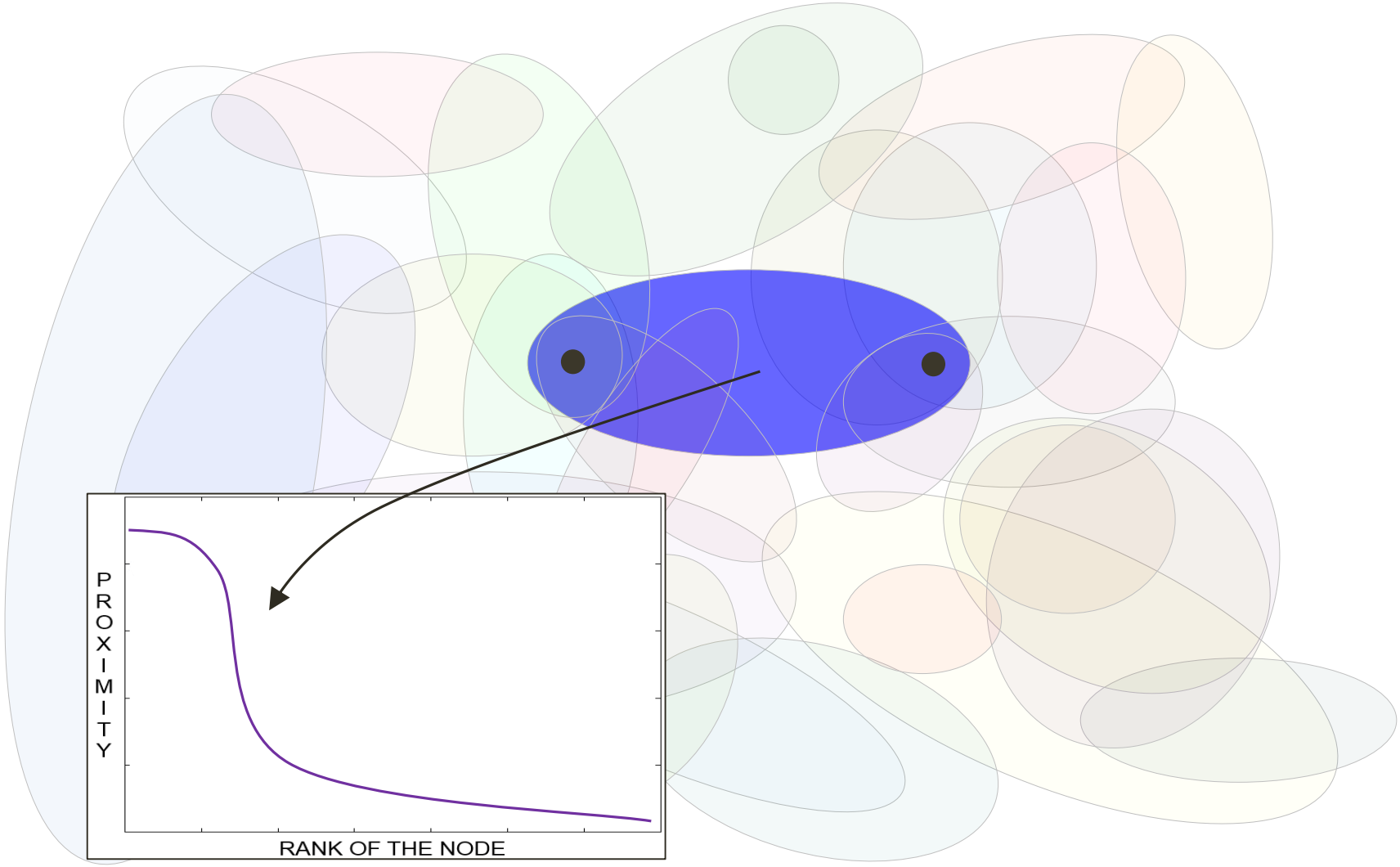
EGOCENTERED COMMUNITIES



EGOCENTERED COMMUNITIES

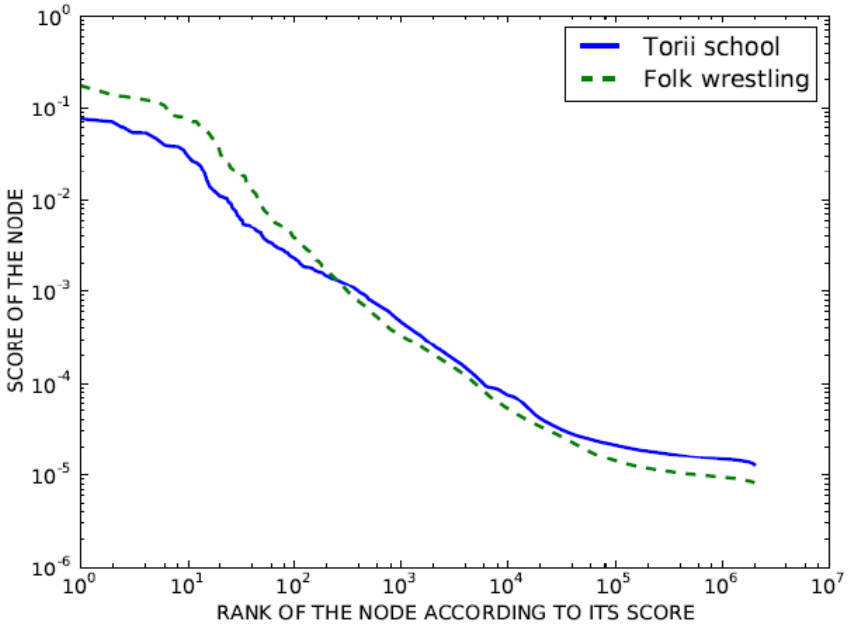


BI-EGOCENTERED COMMUNITIES



BI-EGOCENTERED COMMUNITIES

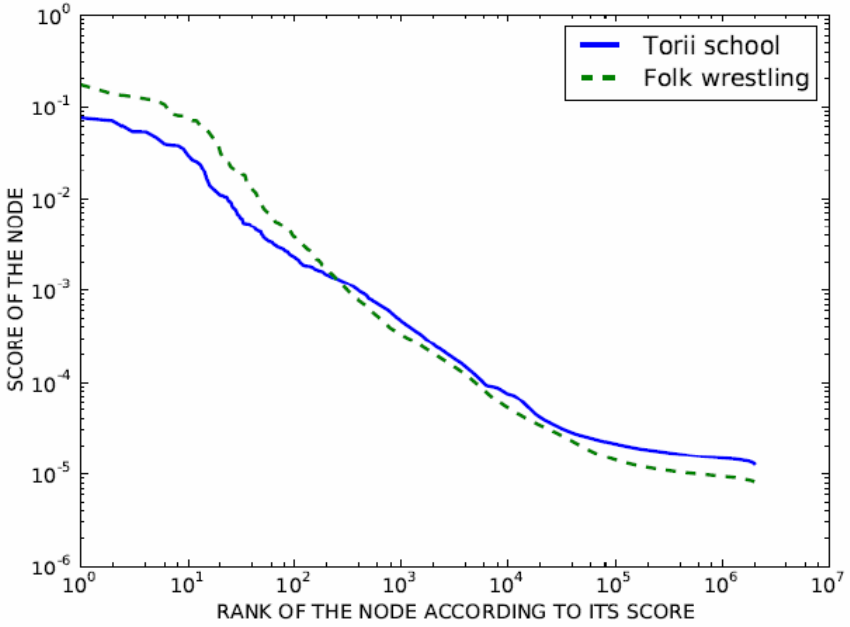
Torii school + Folk wrestling

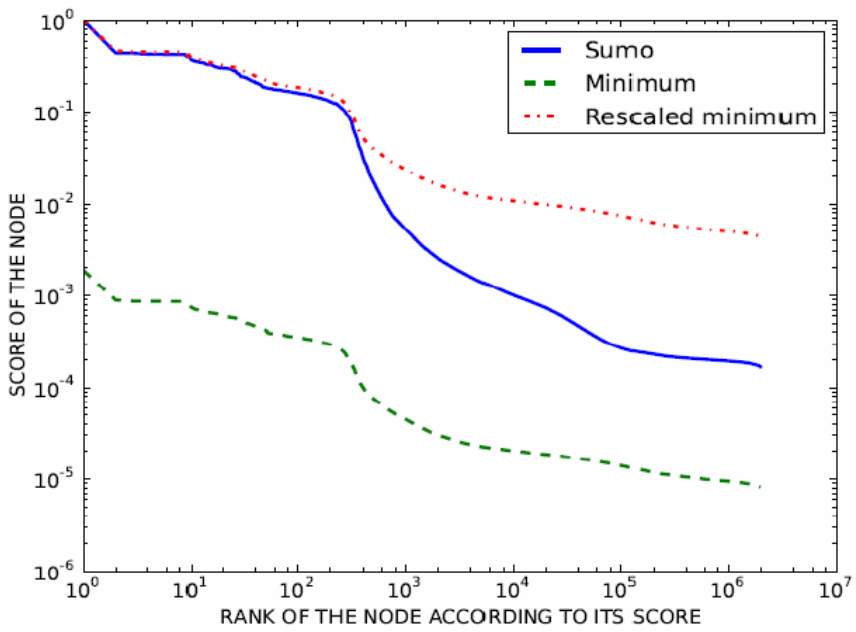


BI-EGOCENTERED COMMUNITIES

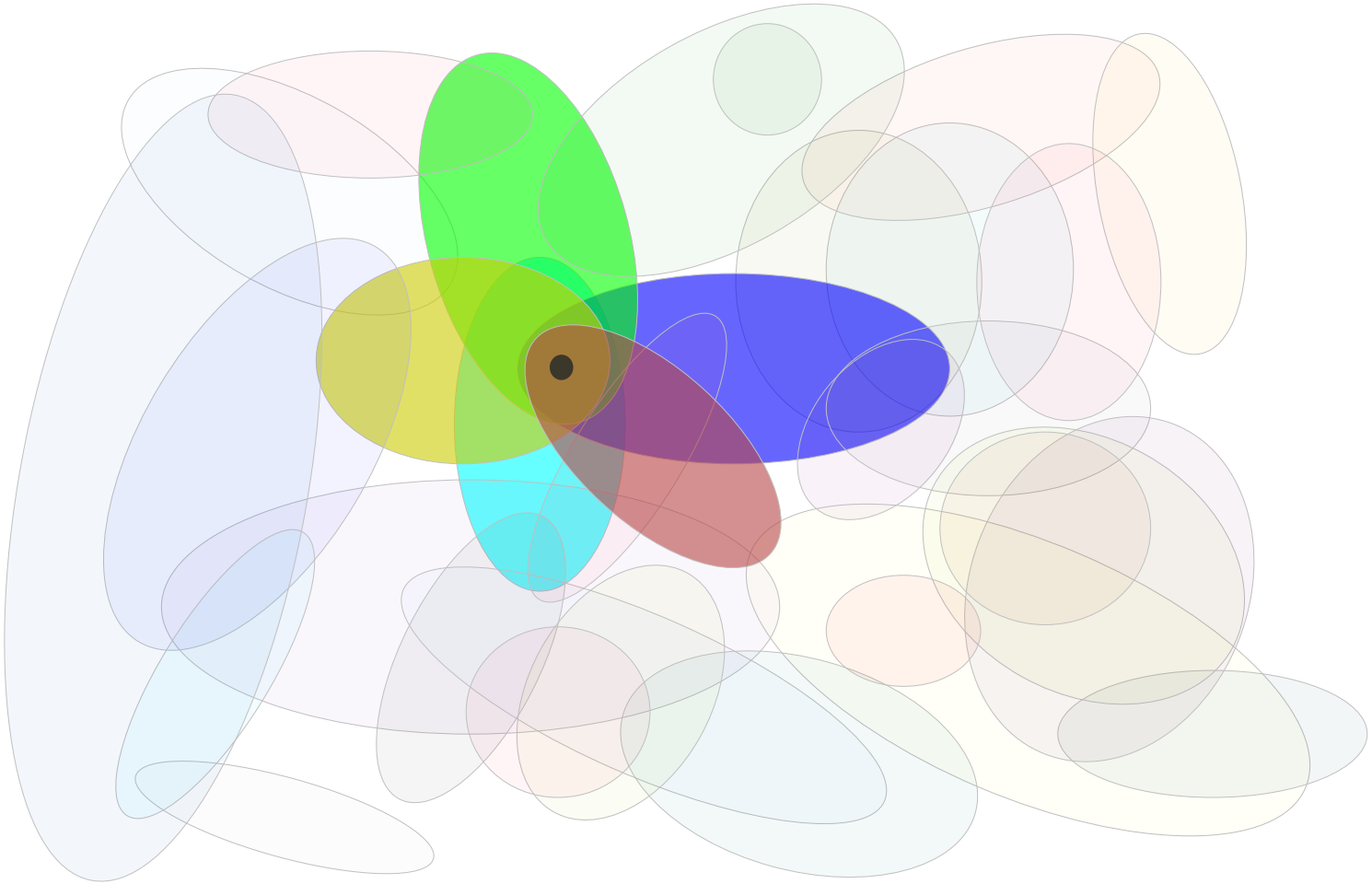
Torii school + Folk wrestling = Sumo

(350 first nodes of sumo contain 337 of the minimum)



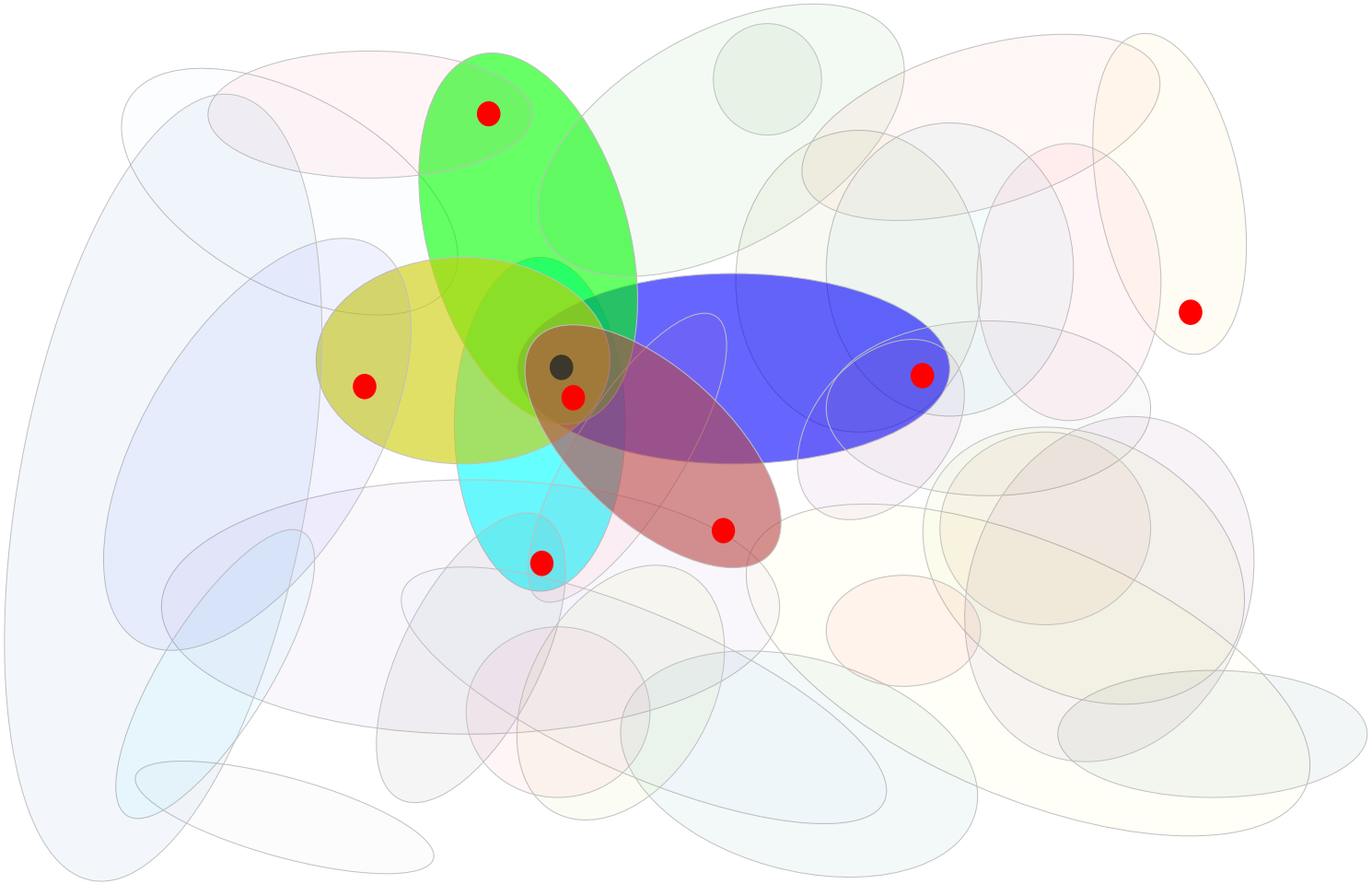


METHODOLOGY TO FIND ALL COMMUNITIES



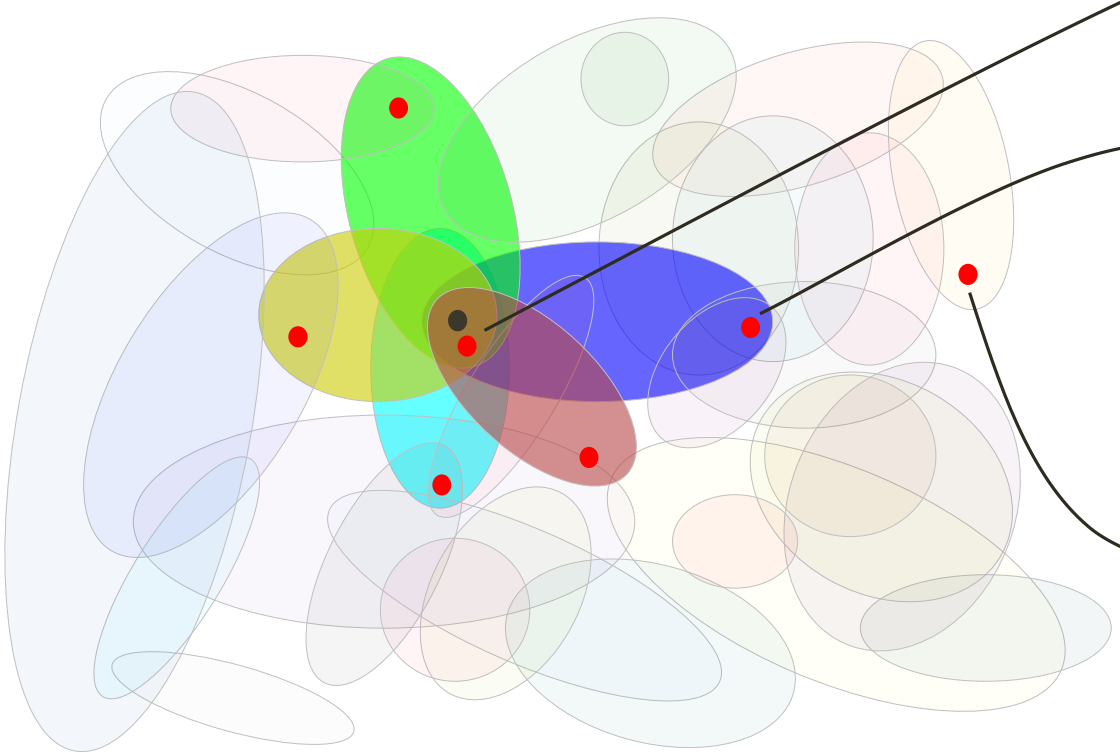
METHODOLOGY TO FIND ALL COMMUNITIES

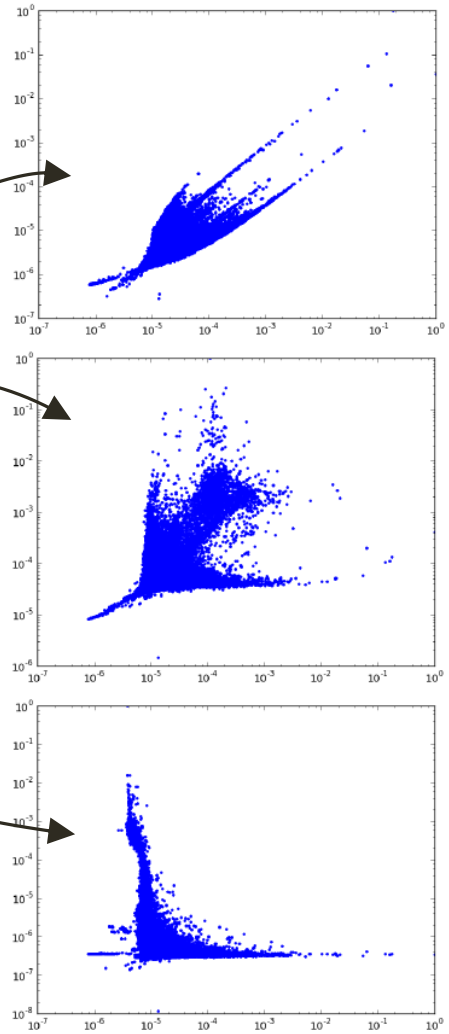
1. Select candidate nodes



METHODOLOGY TO FIND ALL COMMUNITIES

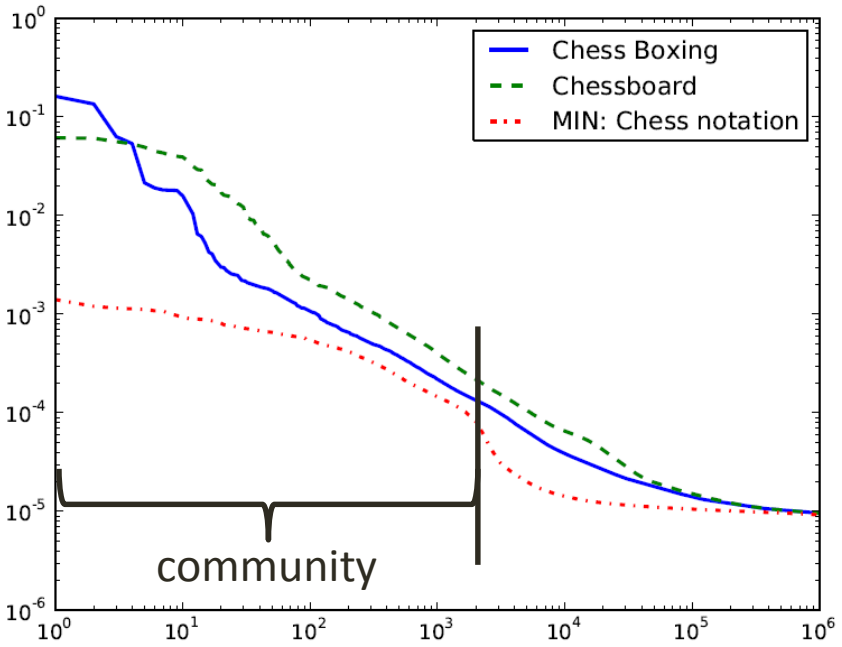
1. Select candidate nodes
   - All nodes? Random sample? Other heuristics?
   - Number of candidates vs time?

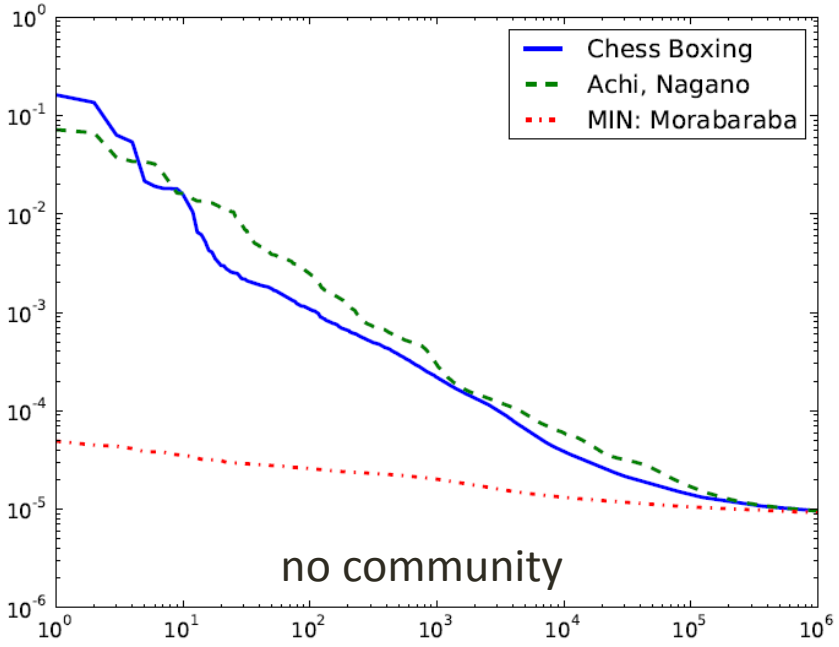




METHODOLOGY TO FIND ALL COMMUNITIES

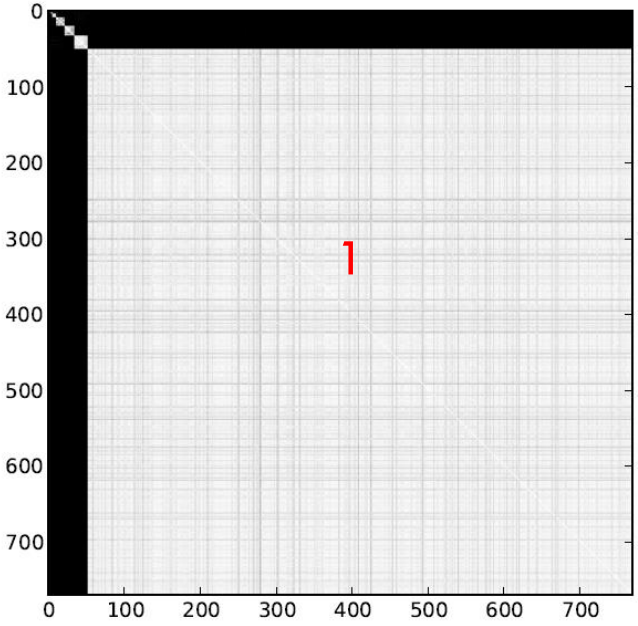
2. Compute bi-egocentered communities
   - Minimum of the two scores
   - Keep nodes before the "sharp decrease" (only if source node is before)
   - Ex: among 3000 candidates, 770 give a sharp decrease

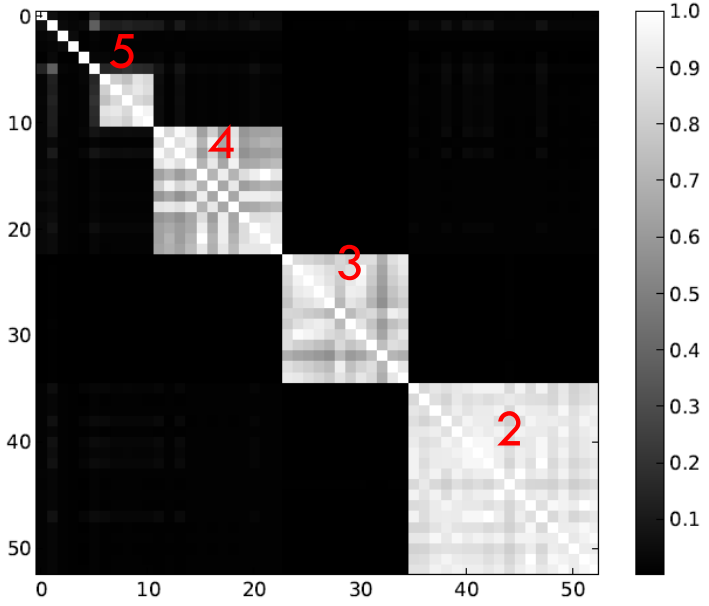




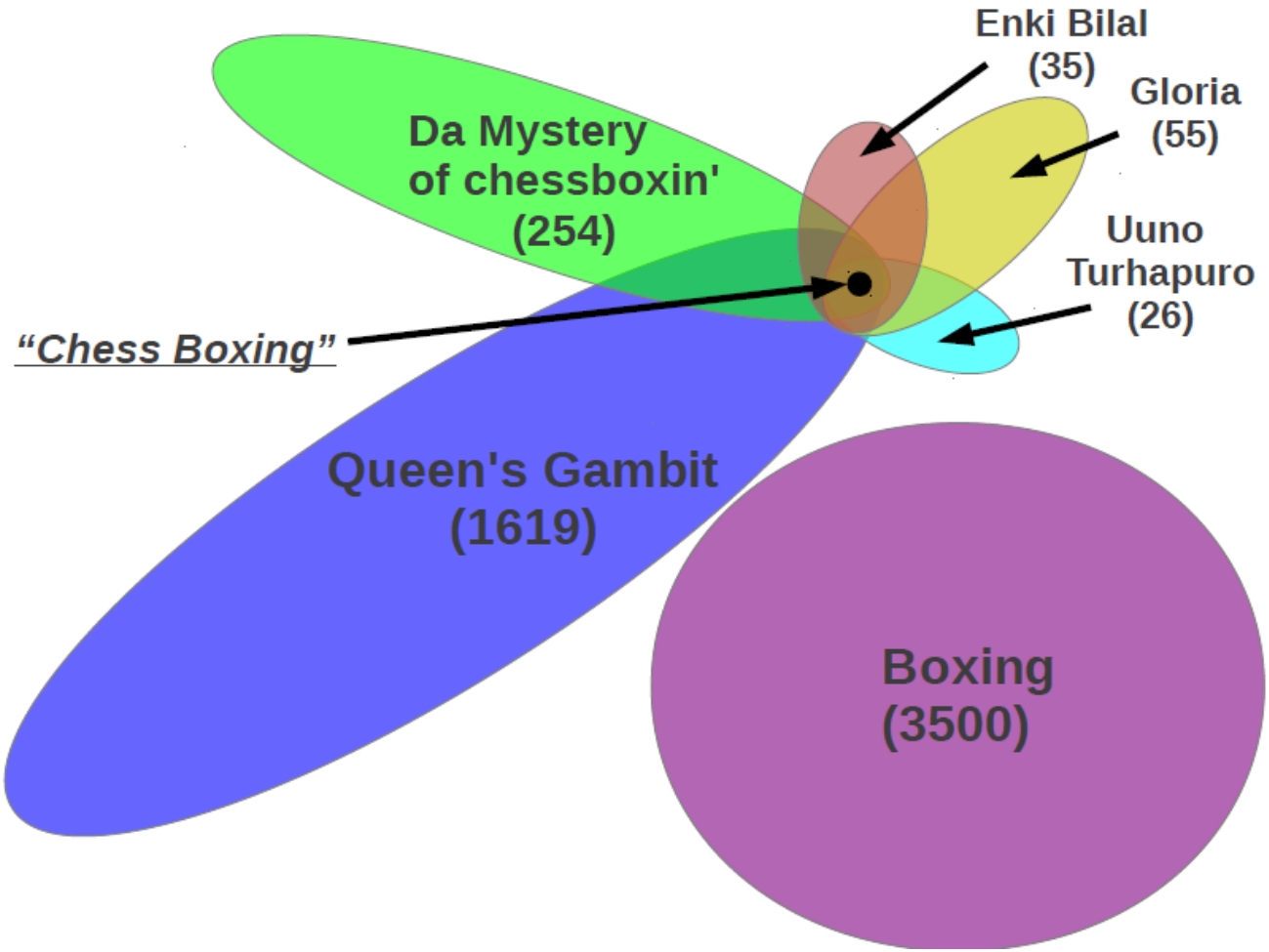
METHODOLOGY TO FIND ALL COMMUNITIES

3. Clean found communities
   - Some communities are found more than once: merge
   - Some communities are found only once (noise): remove
   - Ex: 3000 candidates, 770 communities, 5 remain





RESULT ON CHESS-BOXING



CONCLUSION / PERSPECTIVES

Method to find (multi) egocentered communities

- "Fast" to compute and parameter-free

Detection of irregularities detect only the sharpest decrease

- Other relevant irregularities?
- What about nodes on the decrease?

Limitation to bi-centered communities

- What about communities centered on 3 or more nodes?
- Computation time => fine selection of candidates

Communities cannot be found for popular pages!

EGOCENTERED COMMUNITIES PARAMETERIZED MEASURE

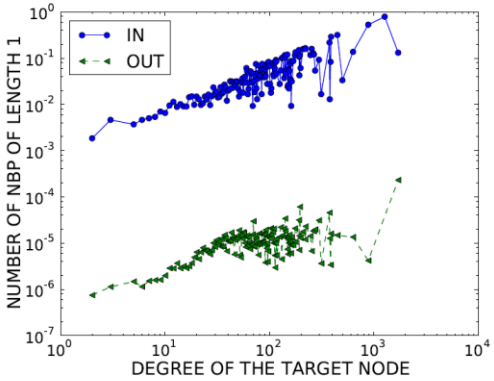
DSAA 2014

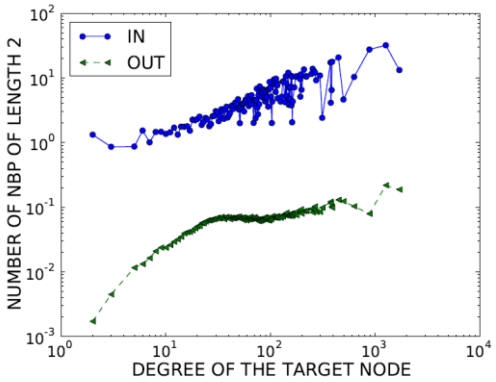
Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

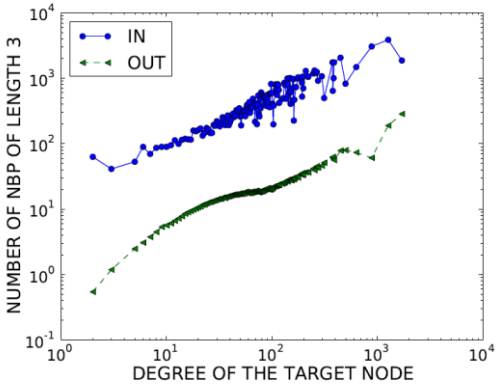
DESIGNING A PROXIMITY MEASURE

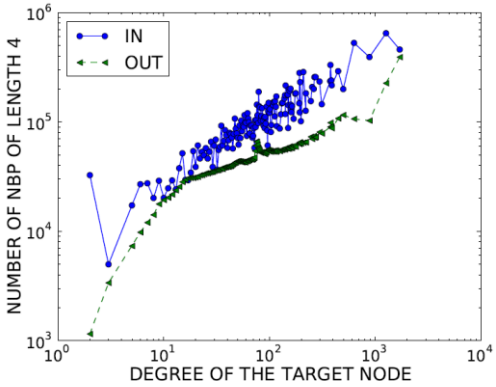
More short paths inside communities than outside

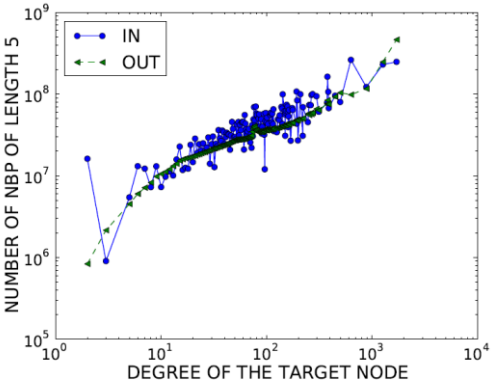
- For all pages from the Wikipedia “graph theory” category

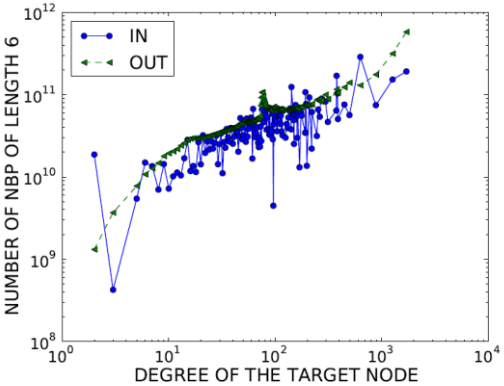








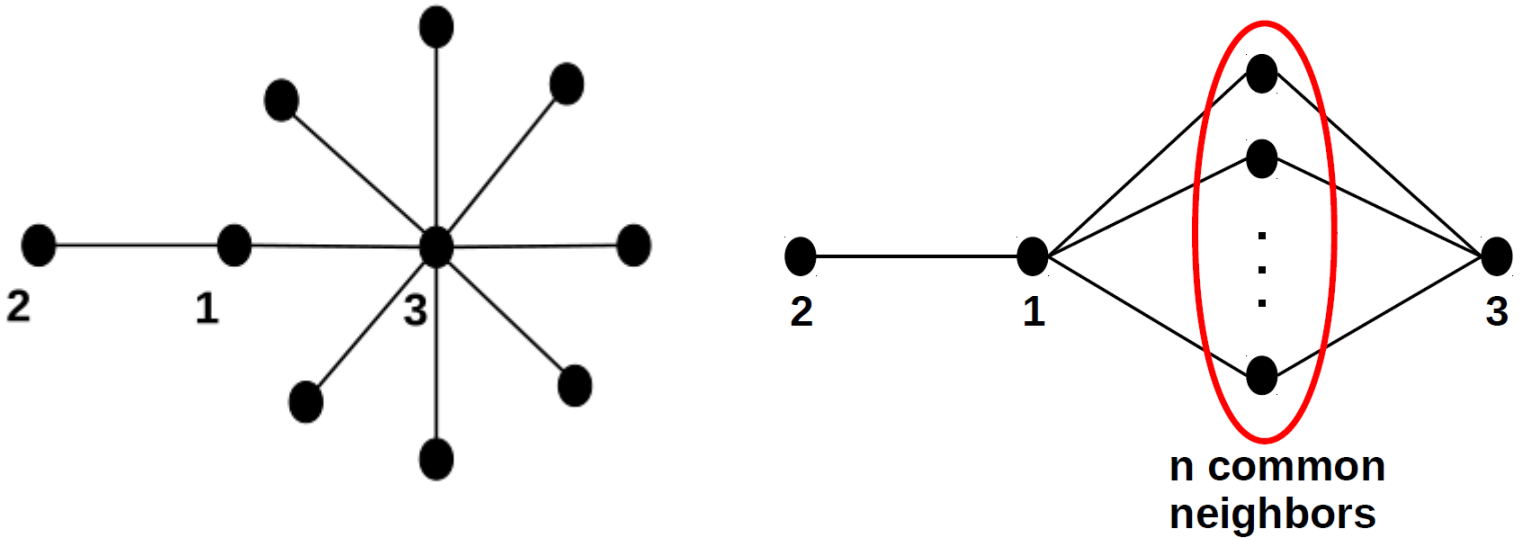




DESIGNING A PROXIMITY MEASURE

Popularity vs intimacy

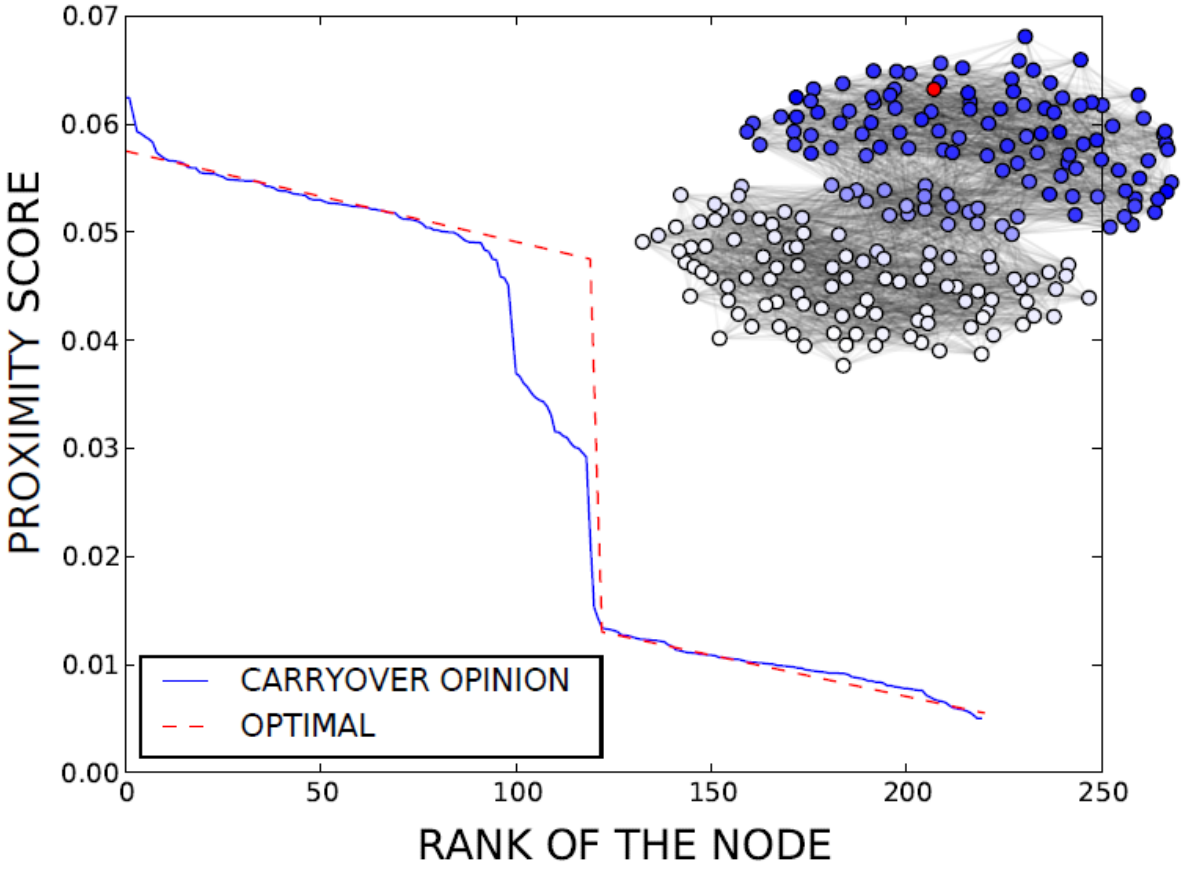
Distance vs redundancy



prox(1,2)<prox(1,3) or prox(1,2)>prox(1,3)?

DESIGNING A PROXIMITY MEASURE

Impact of overlapping communities



FINAL PROXIMITY MEASURE

Important features :

- Paths: number, length and maximum length
- Degree of target node

$$P(i,j) = \frac{1}{d_j^{\beta}} \sum_{l=0}^{\lambda} \alpha^l N_l^{\mathrm{NBP}}(i,j)$$

Close to Katz index

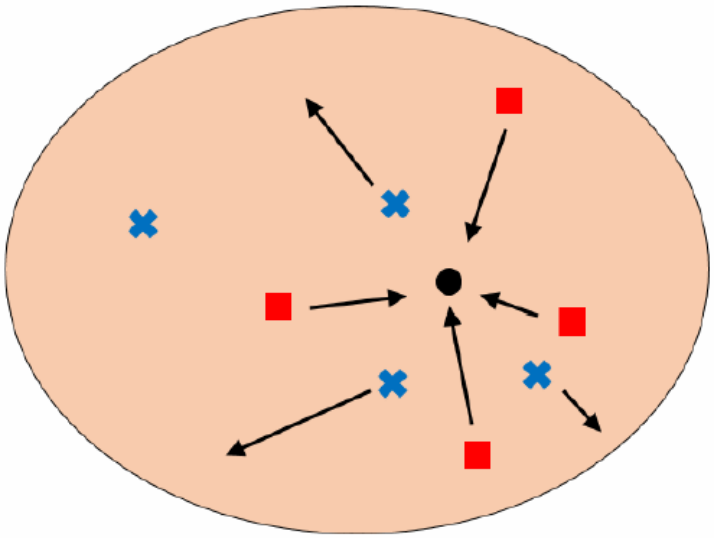
More parameters could be used:

- One for each path length ($\alpha(l)$ vs α^l)
- Degrees on the paths…

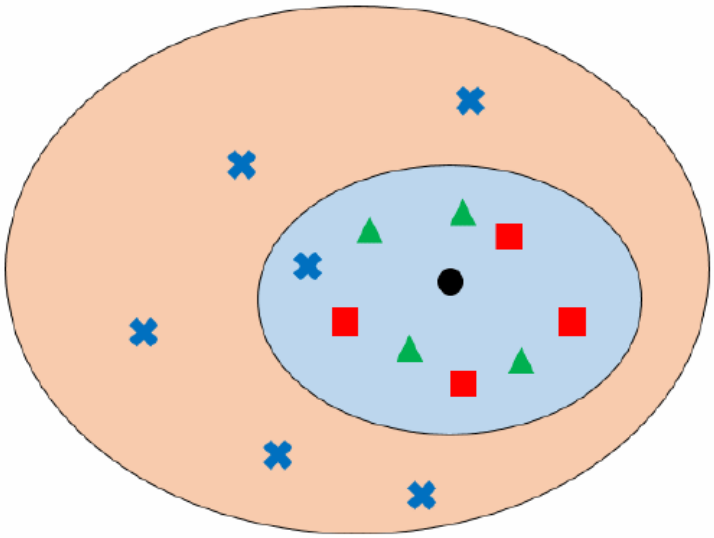
FINAL PROXIMITY MEASURE - LEARNING

Completing a community: parameters can be learnt

- Given a node of interest i and positive (negative) examples
- Find the parameters that maximize the proximity of positive nodes to i



We learn P(●, ■) > P (●, ✖)

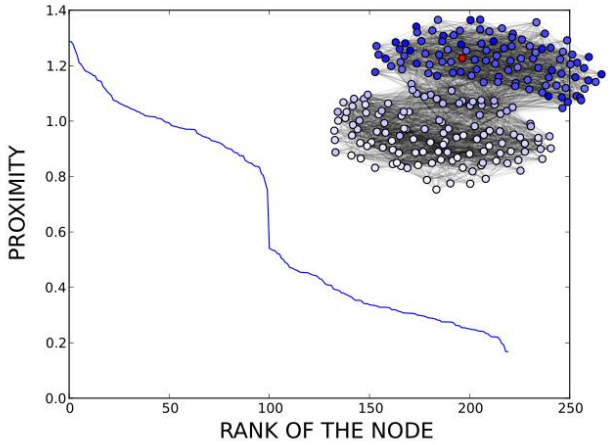


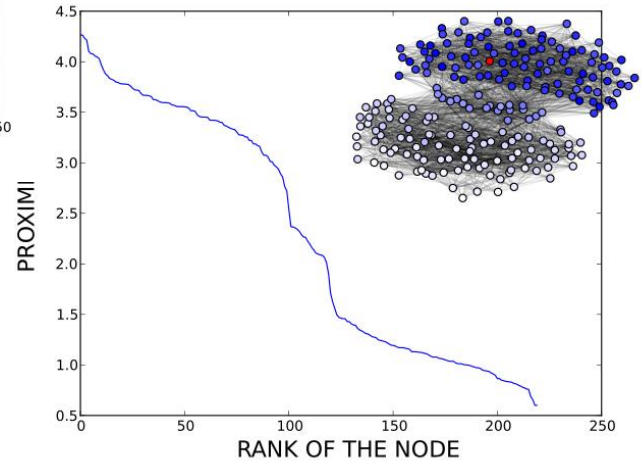
We hope P(●, ▲) > P(●, ✖)

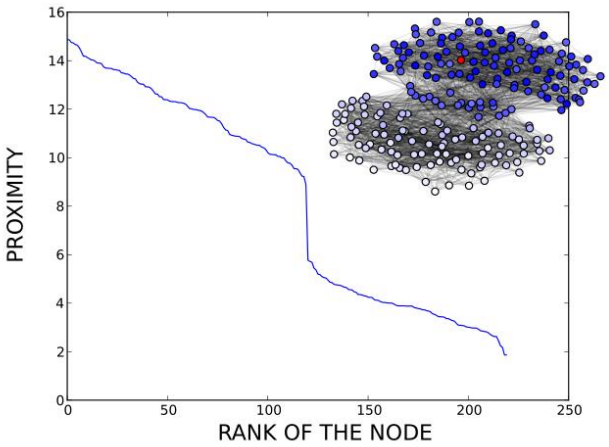
VALIDATION

Random graph with two overlapping communities

- We can choose what to do with in-between nodes



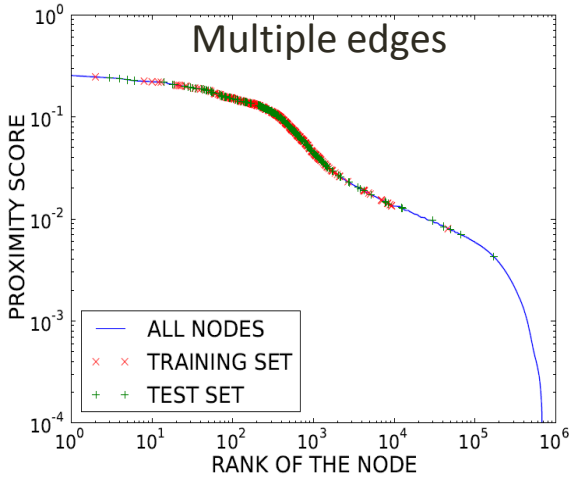


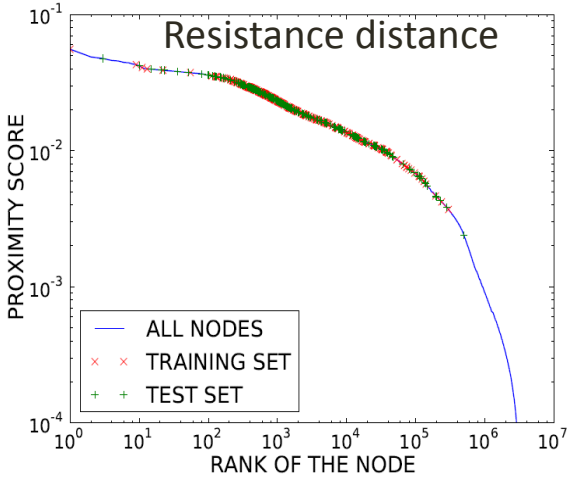


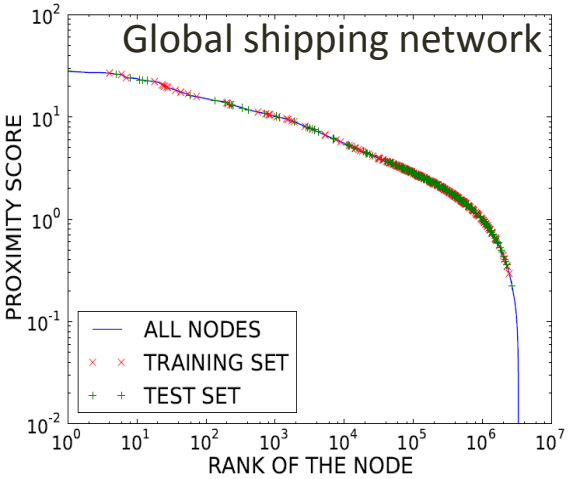
VALIDATION - GRAPH THEORY

Wikipedia "graph theory" category (and direct subcats):

- Manually categorized pages split in train and test sets
- For each node of interest, learn the parameters using train nodes







Questions:

- Can parameters be tuned? → Community exists?
- Are test nodes also close? → Ok or over fitting?
- Are there some train/test badly ranked? → Outside the category?
- Are there outside nodes well ranked? → Should be inside?

VALIDATION - GRAPH THEORY

Best ranked pages outside “graph theory”:

- Mostly belong to subcategories of graph theory
- Graphs vs Networks

Rank	Page	Category
3	Graphlets	Networks
6	Wall and Lines	G. Theory (added since)
8	Complete graph	Regular G.
9	Chang graphs	Regular G.
13	Local McLaughlin graph	Regular G.
14	Complete bipartite graph	Param. Families of G.
15	Quartic graph	Regular G.
23	Watkins snark	Regular G.
30	Brouwer-Haemers graph	Regular G.
33	Bipartite graph	G. families

CONCLUSION

Two approaches for egocentered communities

- Poor information: parameter-free method
- Rich information: parameters + learning techniques
- Both are “computationally effective”

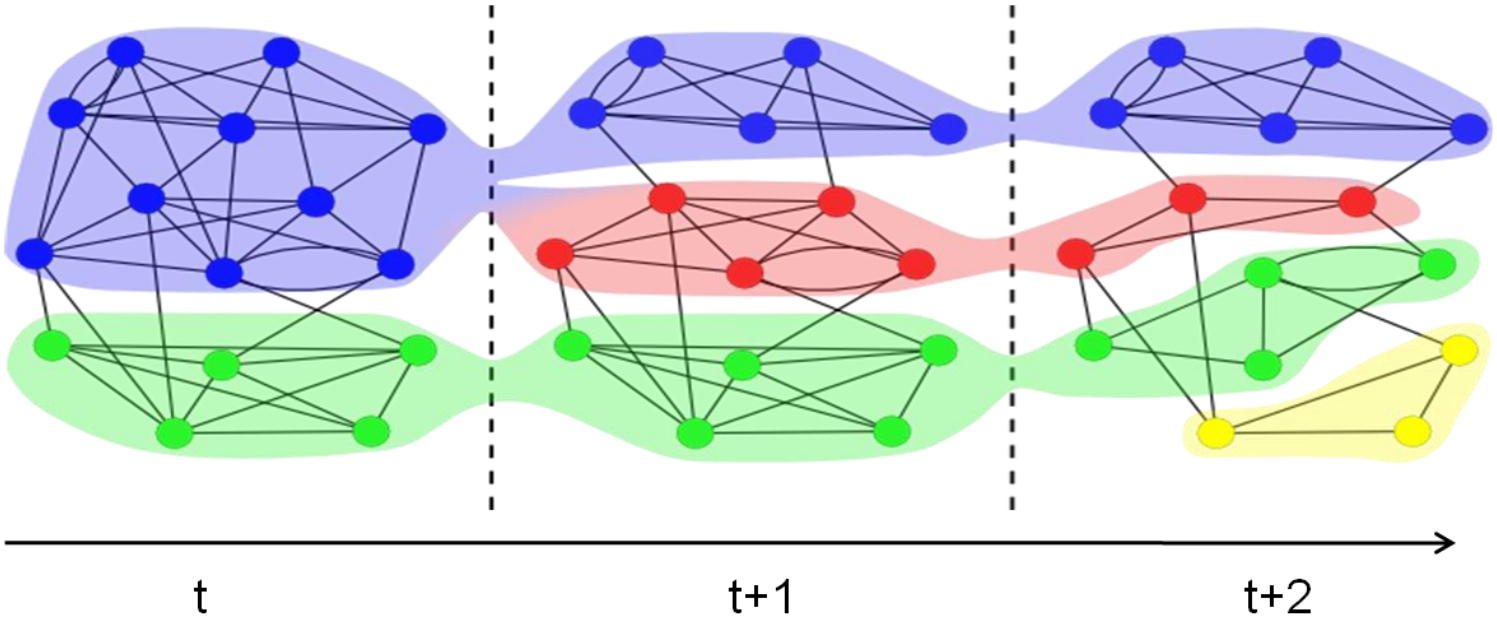
Notion of multi-egocentered community

EVOLVING OVERLAPPING COMMUNITIES

Egocentered communities → all communities

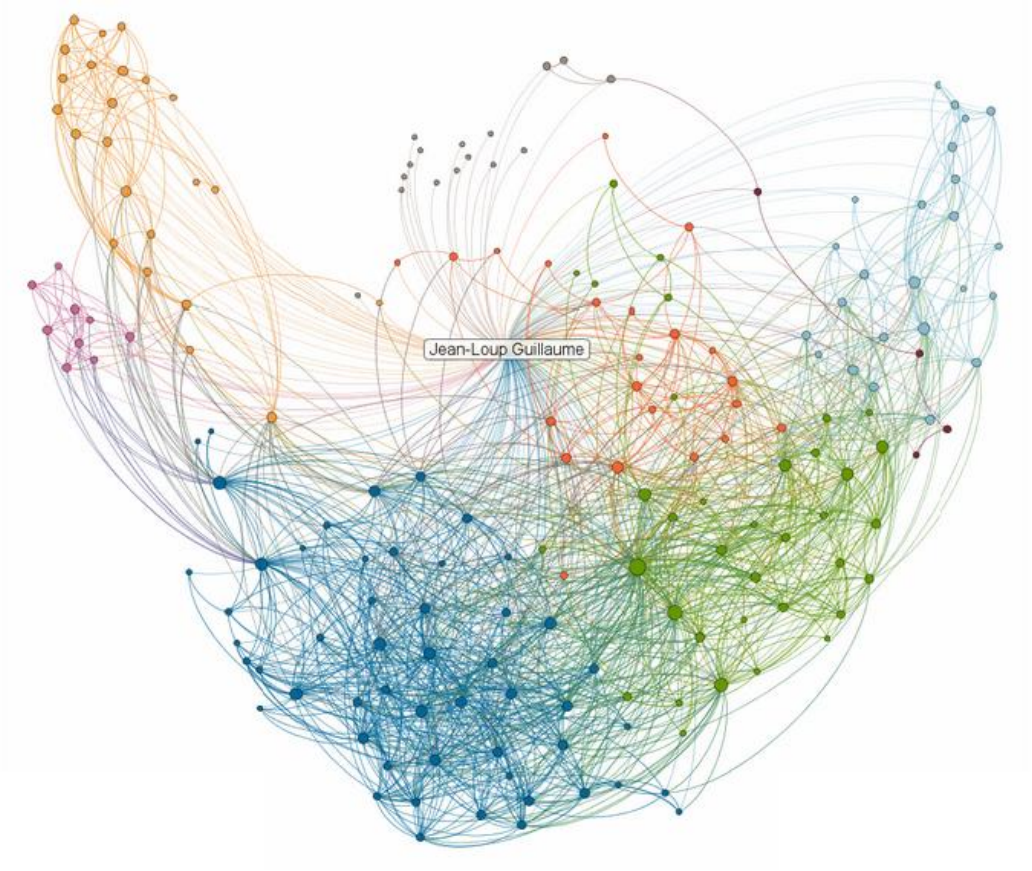
- Computation for every node

Study the evolution of communities



HYBRID / MULTIPLEX NETWORKS

Use more that the explicit interconnections?



MERCI

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/