Community detection in networks

Santo Fortunato

Aalto University

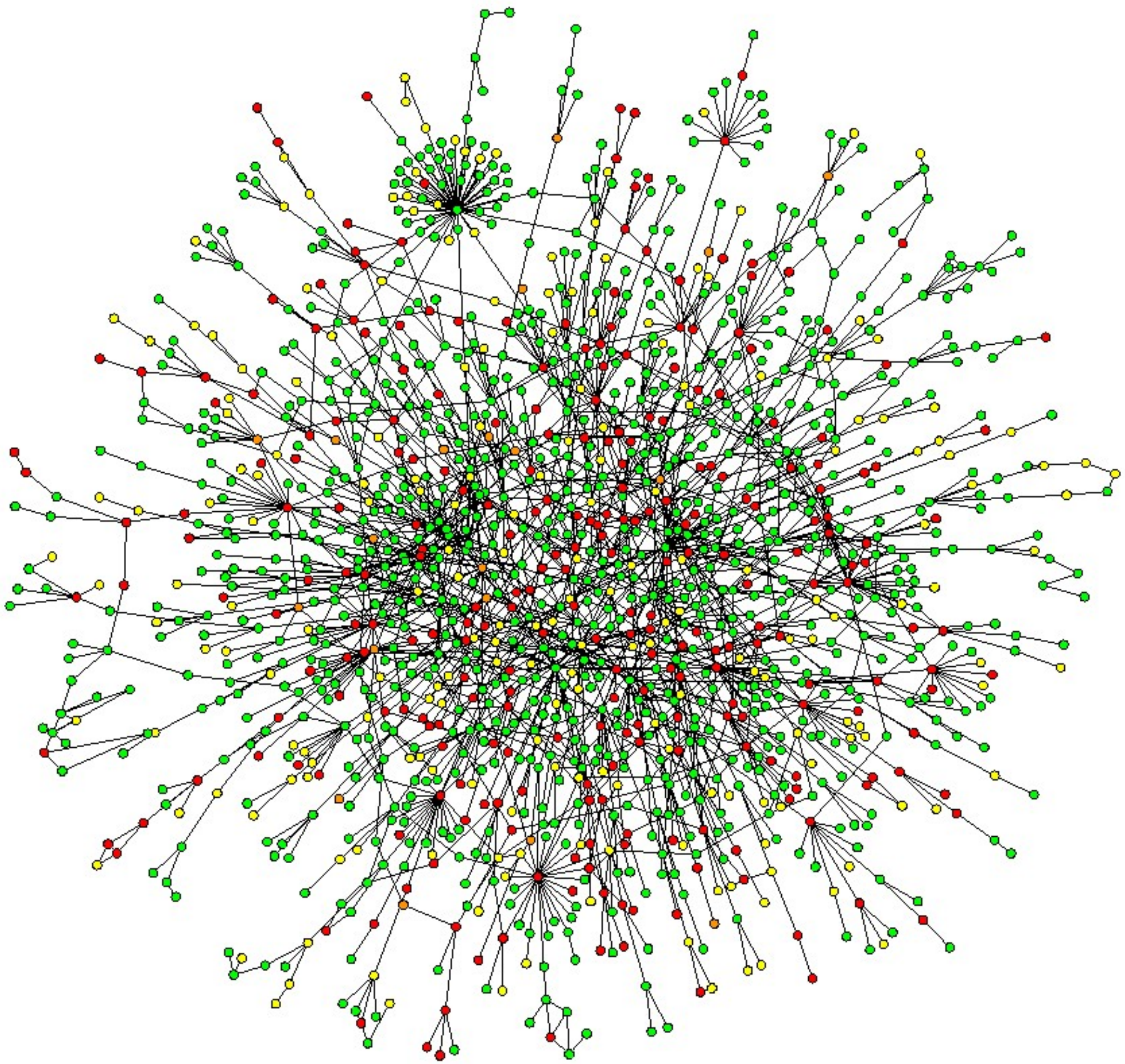
Outline

1) Introduction
2) Global optimization techniques: limits
3) Local techniques: OSLOM
4) Testing algorithms
5) Summary

Networks

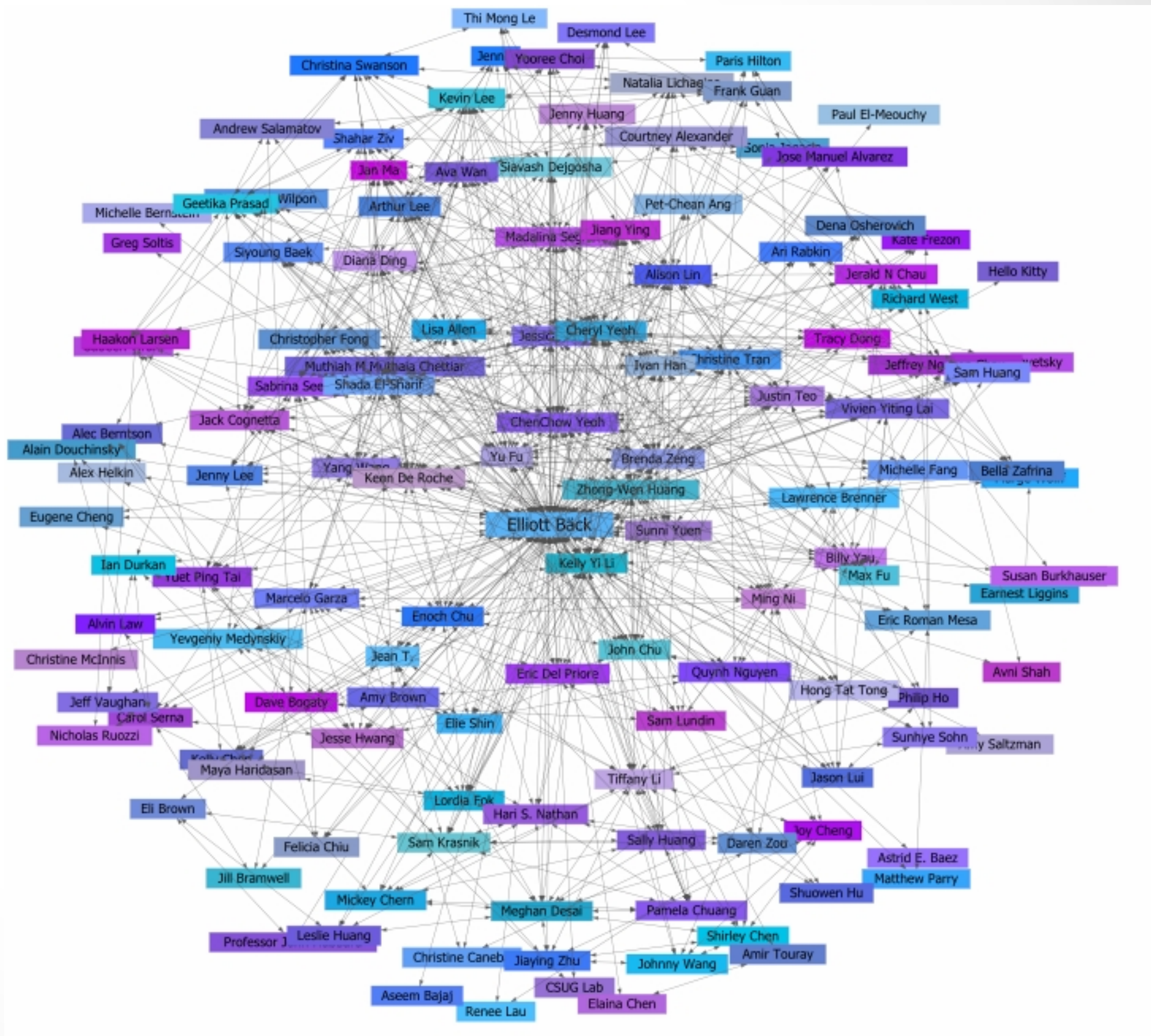
Protein-protein interaction networks

Network: simplest representation of a complex system



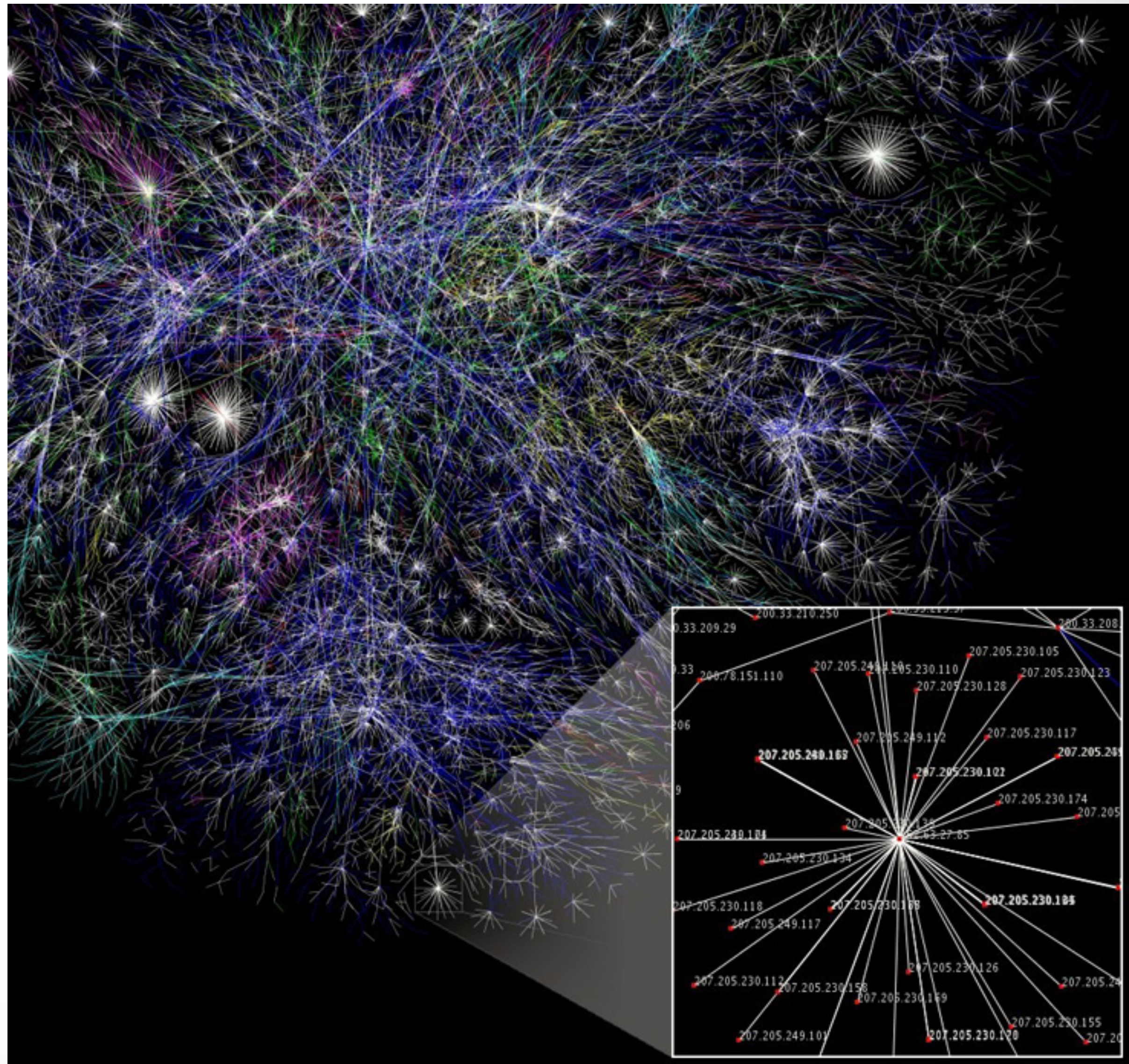
Networks

Social networks



Networks

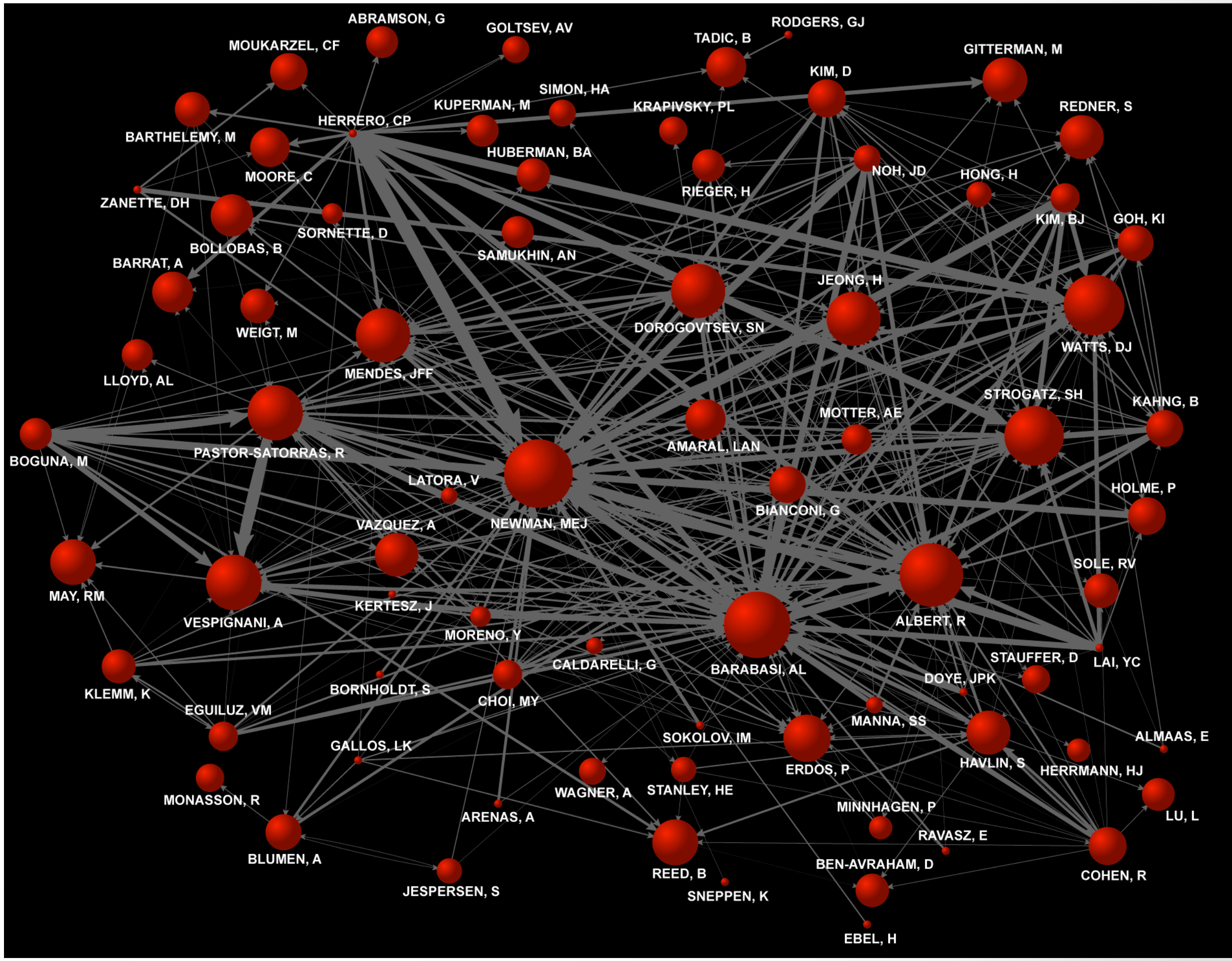
Internet



Networks

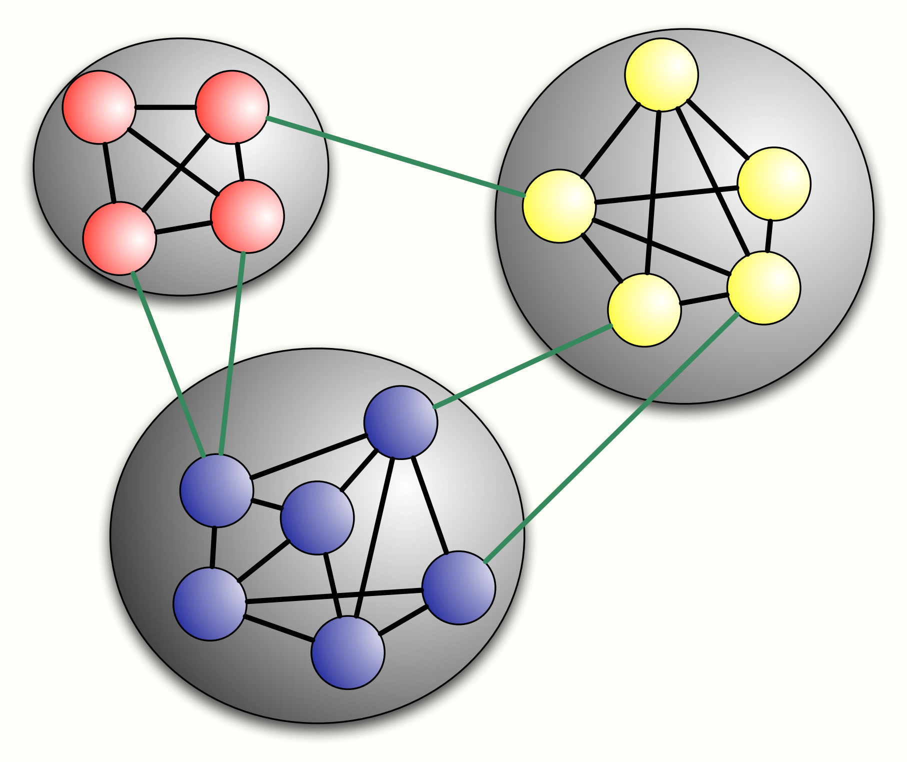
Citation networks

Important features of a system and its dynamics from purely structural information



Community structure

Communities: sets of tightly connected nodes

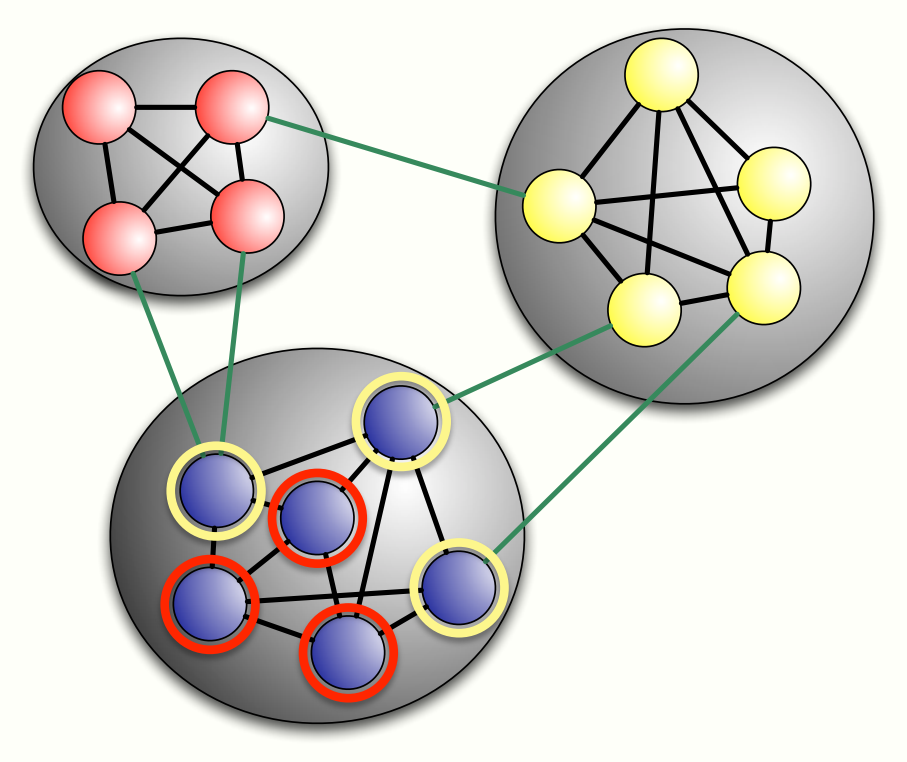


- People with common interests
- Scholars working on the same field
- Proteins with equal/similar functions
- Papers on the same/related topics
- …

Community detection

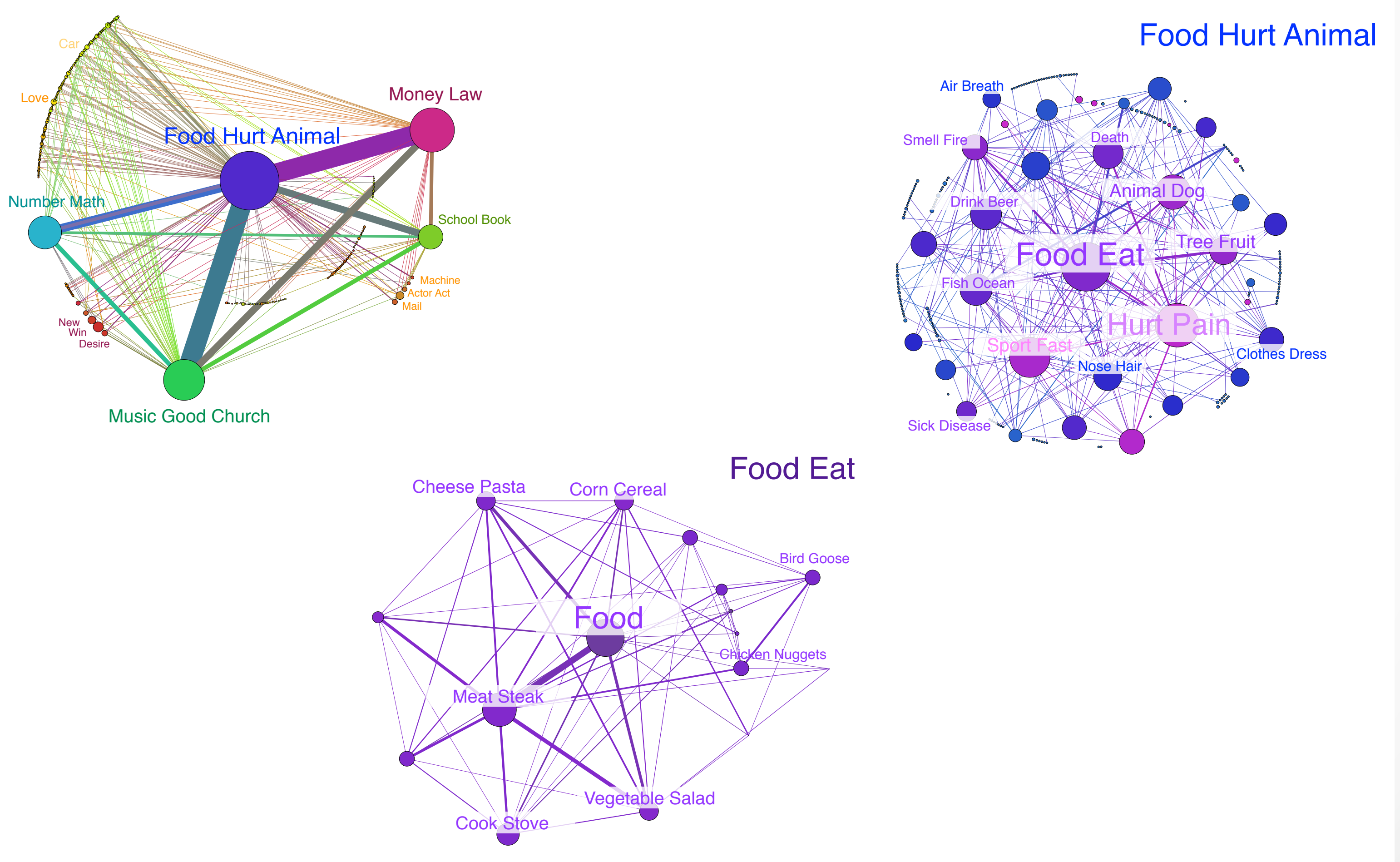
Theoretical reasons

- Organization
- Node features
- Node classification
- Missing links



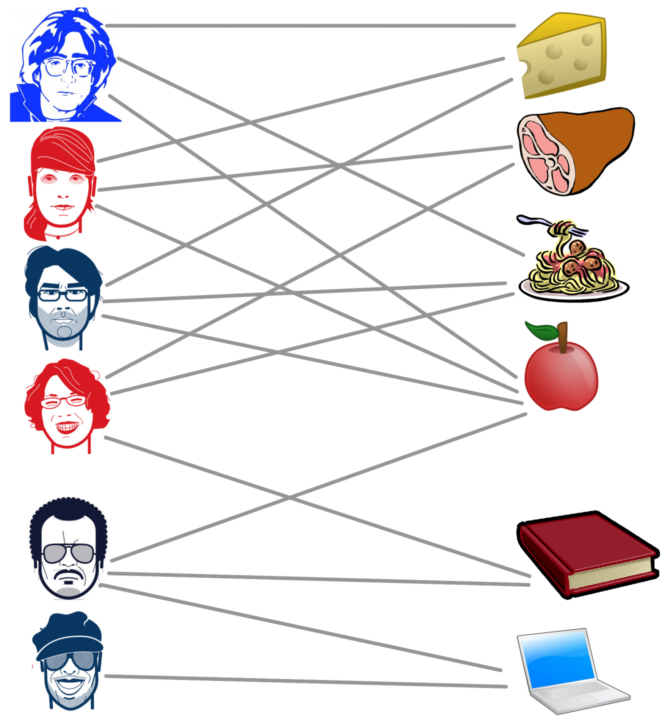
Community detection

Graph visualization



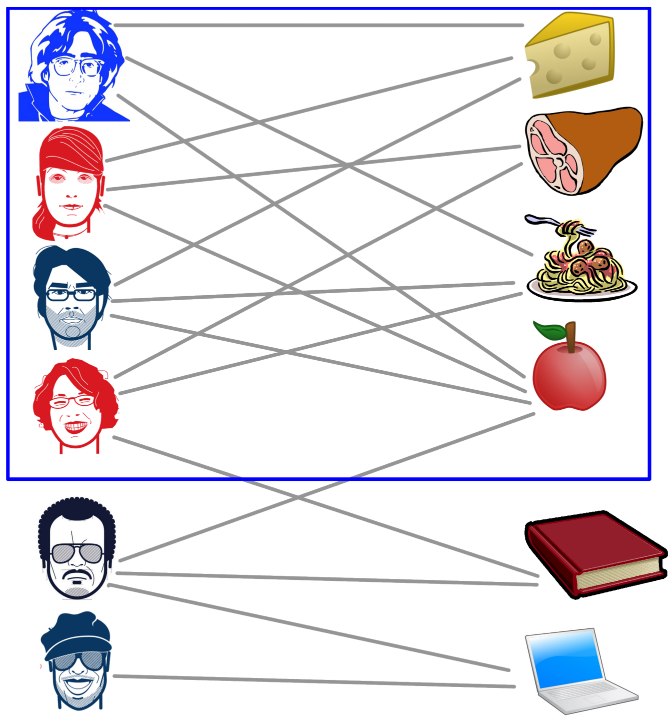
Community detection

Practical reasons: recommendation systems



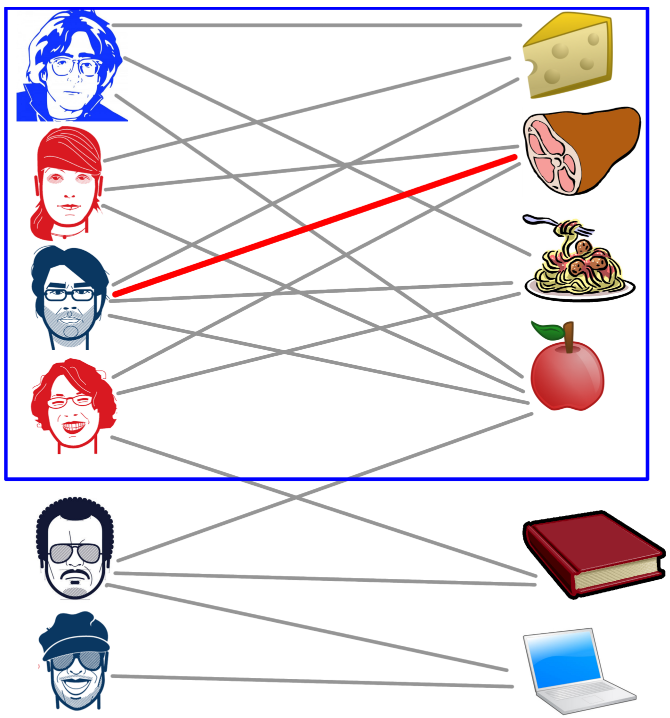
Community detection

Practical reasons: recommendation systems



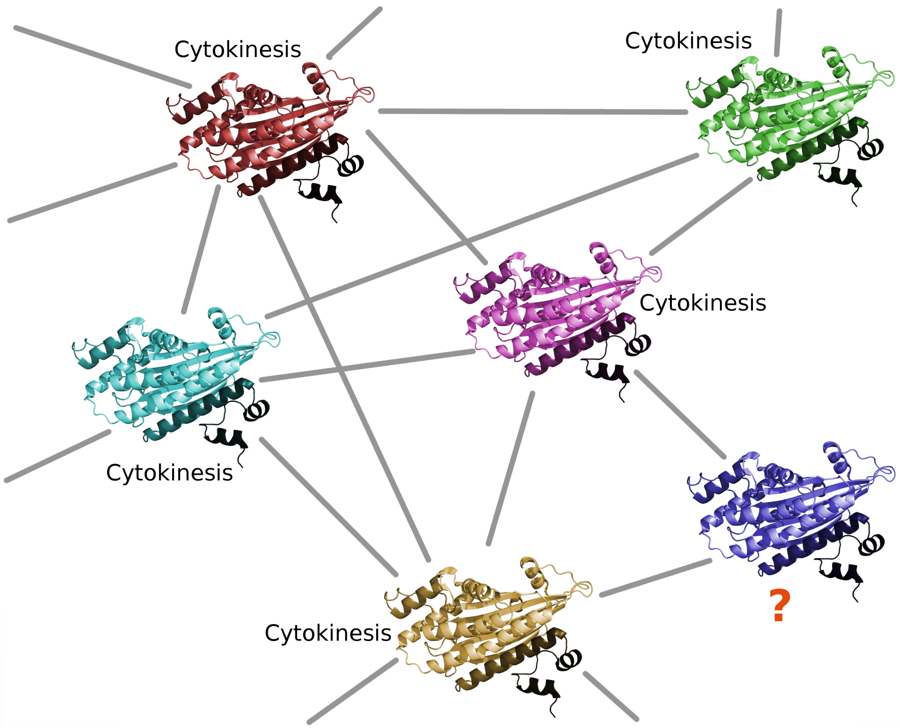
Community detection

Practical reasons: recommendation systems



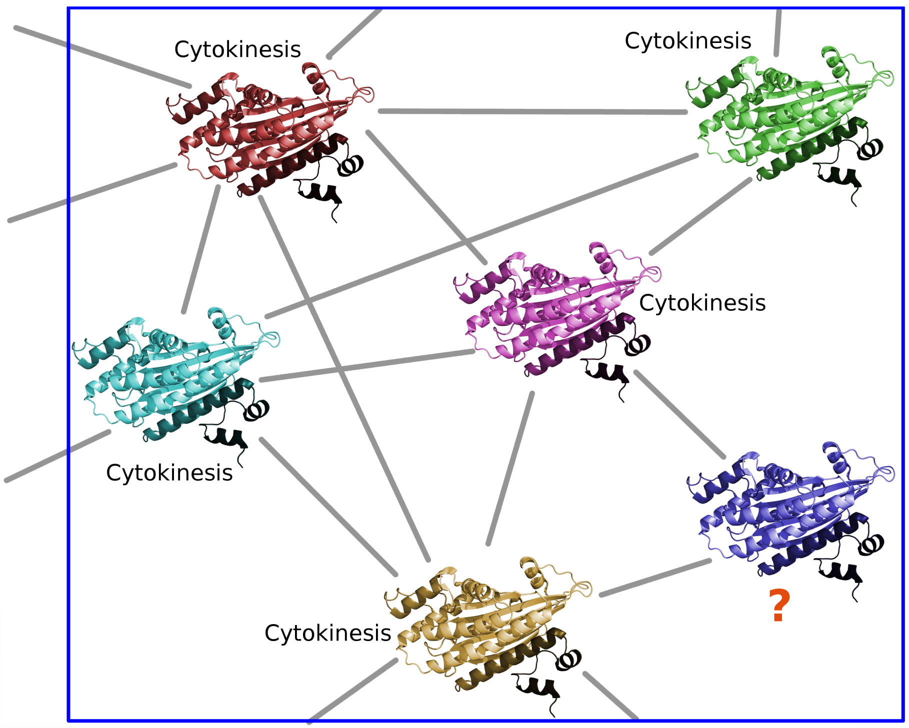
Community detection

Practical reasons: unknown protein functions



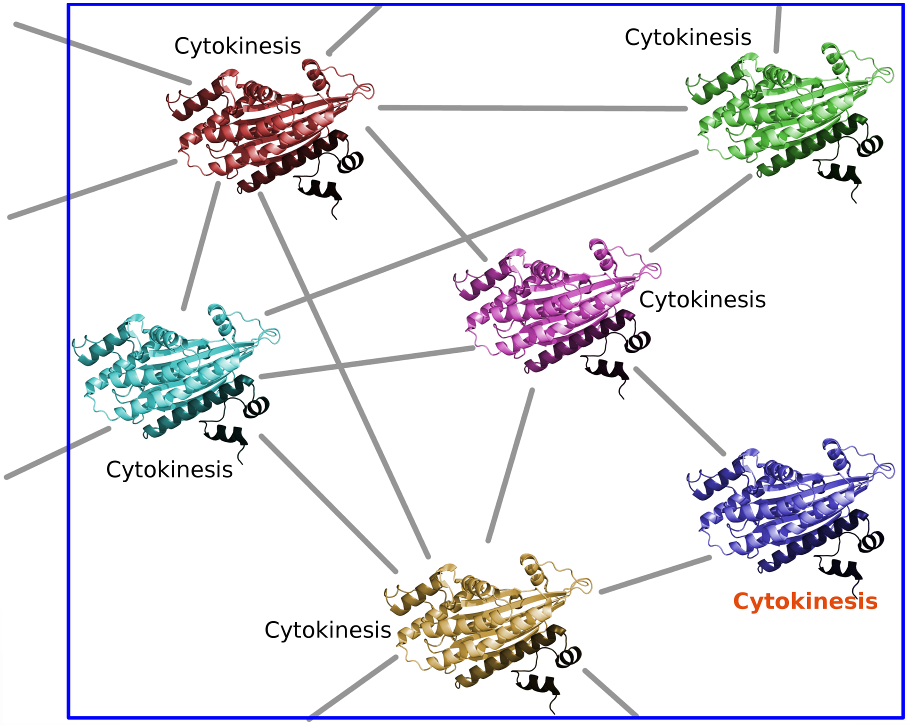
Community detection

Practical reasons: unknown protein functions

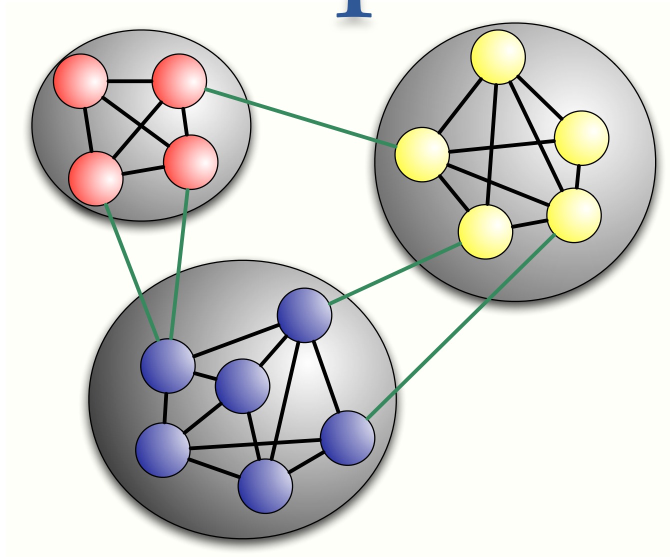


Community detection

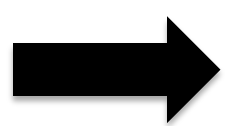
Practical reasons: unknown protein functions



Difficult problem!



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1								1		1	1			1	1
2						1		1						1	
3	1											1			1
4				1						1		1			1
5							1	1					1	1	
6		1						1							1
7			1						1			1	1		
8		1				1	1							1	
9			1				1					1	1		
10		1						1						1	1
11							1			1		1	1		
12			1				1		1						1
13				1			1	1				1			
14		1				1	1	1							
15	1			1	1							1			

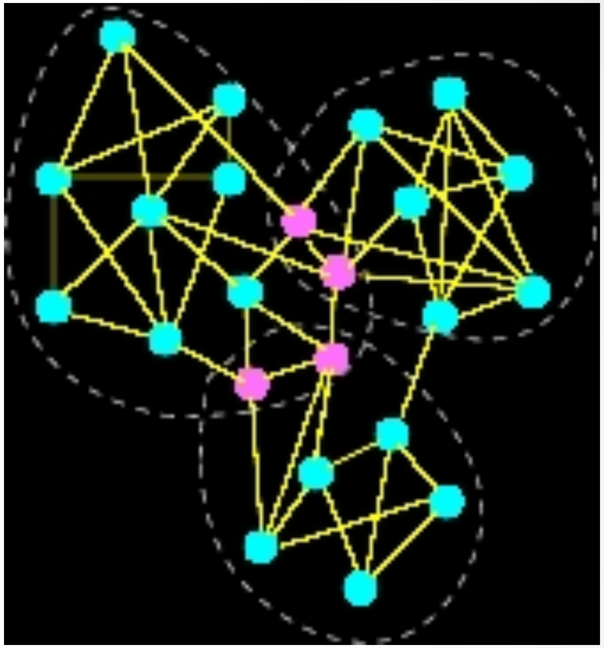


	6	2	8	14	5	7	13	11	9	12	10	4	1	3	15
6		1	1	1											
2	1		1	1							1				
8	1	1		1	1										
14	1	1	1		1										
5			1	1		1	1			1					
7					1		1	1	1						
13					1	1		1		1					
11						1	1		1	1					
9						1		1		1			1		
12					1		1	1	1			1			
10		1										1	1	1	1
4										1	1		1		1
1									1		1	1		1	1
3											1		1		1
15											1	1	1	1	

Difficult problem

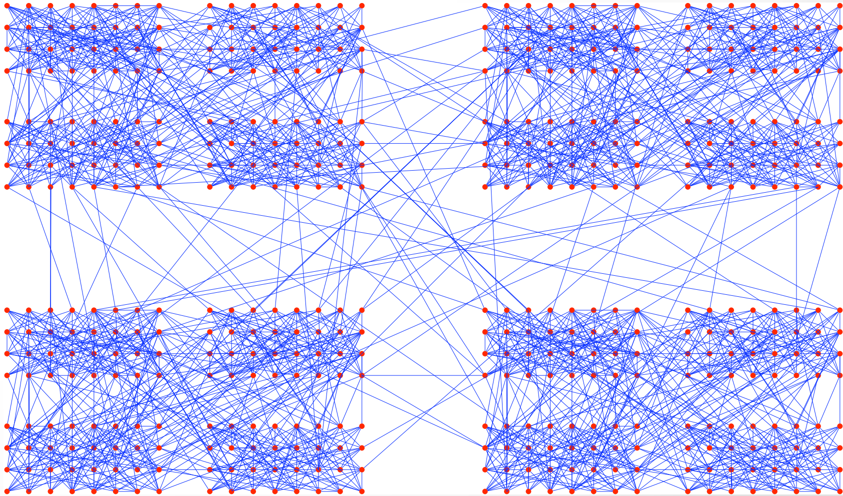
Ill-defined problem:

- What is a community/partition?
- What is a *good* community/partition?



Complications:

- Link directions
- Link weights
- Overlapping communities
- Hierarchical structure



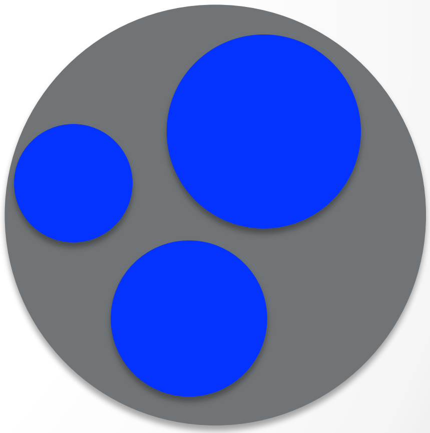
Global optimization

Principle:

- Function Q($\mathcal{P}$) that assigns a score to each partition
- Best partition of the network -> partition corresponding to the maximum/minimum of Q($\mathcal{P}$)

Problems:

- Good partition does not imply good clusters
- Answer depends on the whole graph -> it changes if one considers portions of it or if it is incomplete



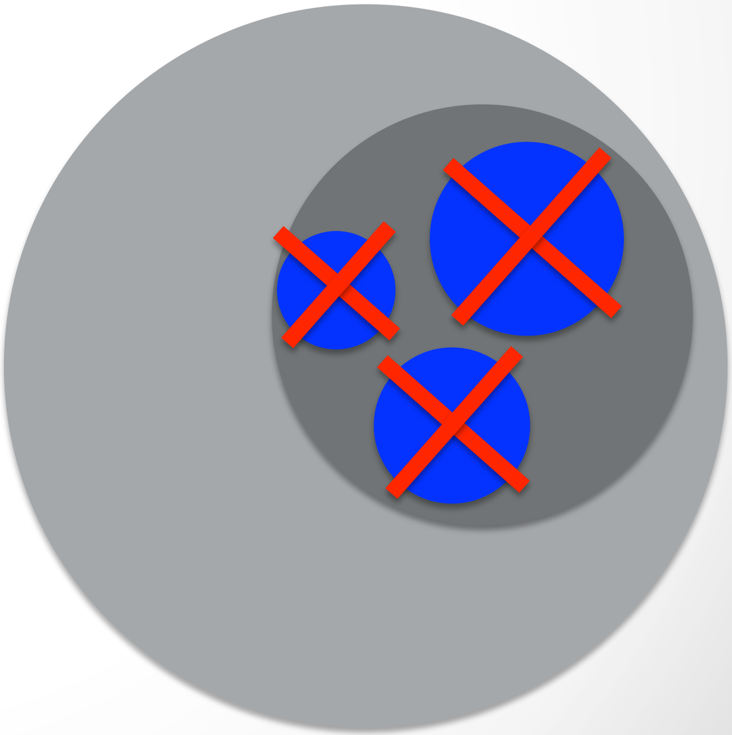
Global optimization

Principle:

- Function Q($\mathcal{P}$) that assigns a score to each partition
- Best partition of the network -> partition corresponding to the maximum/minimum of Q($\mathcal{P}$)

Problems:

- Good partition does not imply good clusters
- Answer depends on the whole graph -> it changes if one considers portions of it or if it is incomplete



Modularity optimization

$$Q = \frac{1}{m}\sum_{c=1}^{n_c}\left(l_c - \frac{d_c^2}{4m}\right)$$

M. E. J. Newman, M. Girvan, Phys. Rev. E 69, 026113 (2004)

M. E. J. Newman, Phys. Rev. E 69, 066133 (2004)

Goal: find the maximum of Q over all possible network partitions

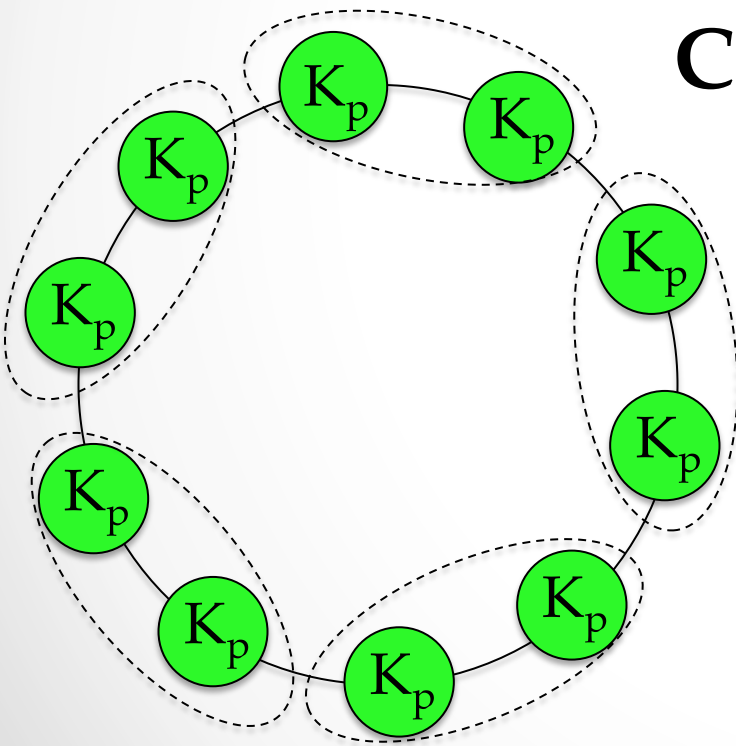
Problem: NP-complete (Brandes et al., 2007)!

Resolution limit

$$Q = \frac{1}{m}\sum_{c=1}^{n_c}\left[l_c - \frac{1}{4}\left(\frac{d_c}{\sqrt{m}}\right)^2\right]$$

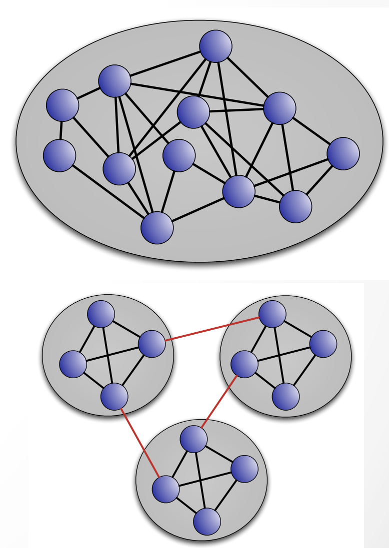
modularity's scale

Consequences



$d_c < \sqrt{m}$ ⟶

$d_c < \sqrt{m}$ ⟶



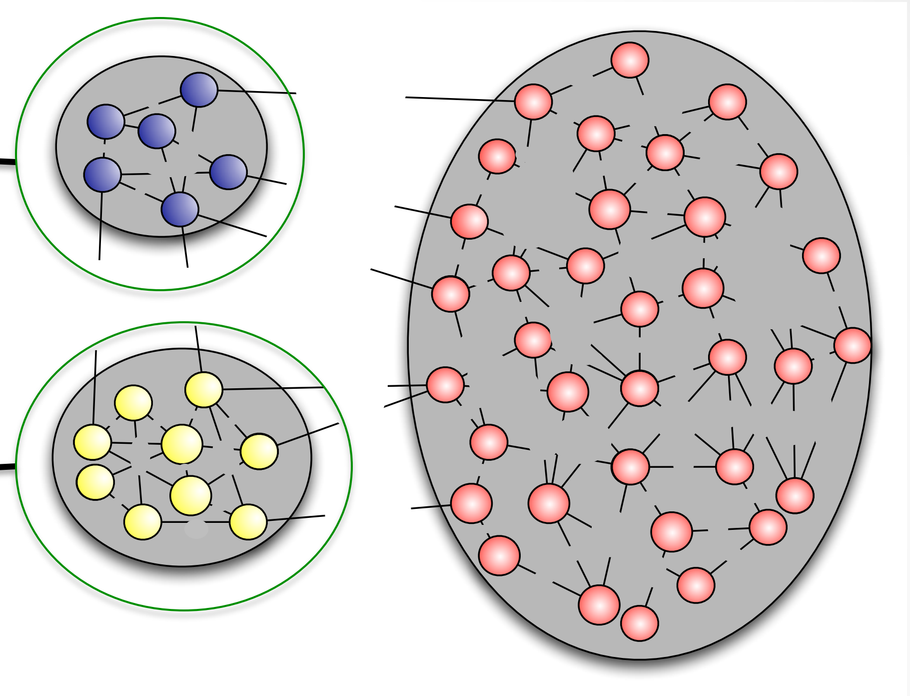
?

S. F. & M. Barthélemy, PNAS 104, 36-41 (2007)

Resolution limit

Subgraph 1, degree k_1

Subgraph 2, degree k_2



Expected number of edges between the two subgraphs in modularity's null model:

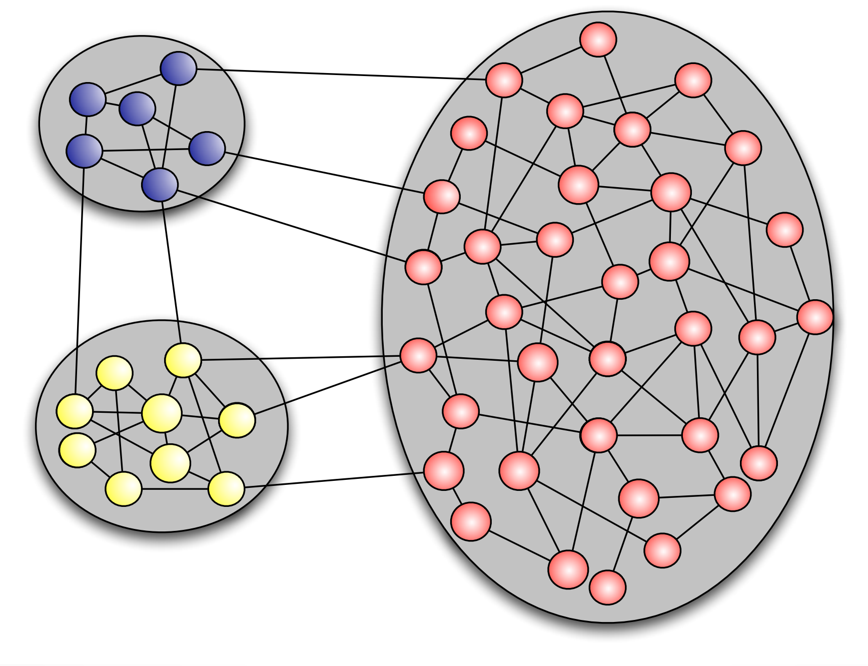
$$m\left(2 \cdot \frac{k_1}{2m} \cdot \frac{k_2}{2m}\right) = \frac{k_1 k_2}{2m}$$

$$\text{if } k_1 = k_2 = d_c \rightarrow \frac{d_c^2}{2m}$$

Resolution limit

Question: What is the origin of the resolution limit?

Answer: global null model is unrealistic!

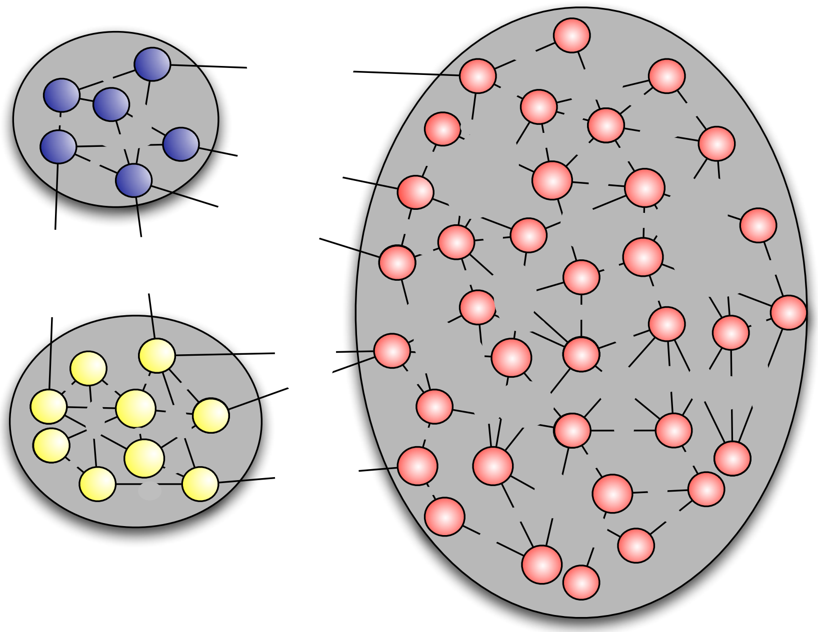


S. F. & M. Barthélemy, PNAS 104, 36-41 (2007)

Resolution limit

Question: What is the origin of the resolution limit?

Answer: global null model is unrealistic!



S. F. & M. Barthélemy, PNAS 104, 36-41 (2007)

Multi-resolution methods?

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[l_c - \gamma \frac{d_c^2}{4m} \right]$$

Double trouble:

1) Small clusters are merged
2) Large clusters are split

Hard to find values of resolution parameter that eliminate both problems!

A. Lancichinetti, S. F., Phys. Rev. E 84, 066122 (2011)

Local optimization

Principle:
- Communities are local structures
- Local exploration of the network, involving the subgraph and its neighborhood

Advantages:
- Conceptual advantage: communities are "local"
- Absence of global scales -> no resolution limit
- One can analyze only parts of the network

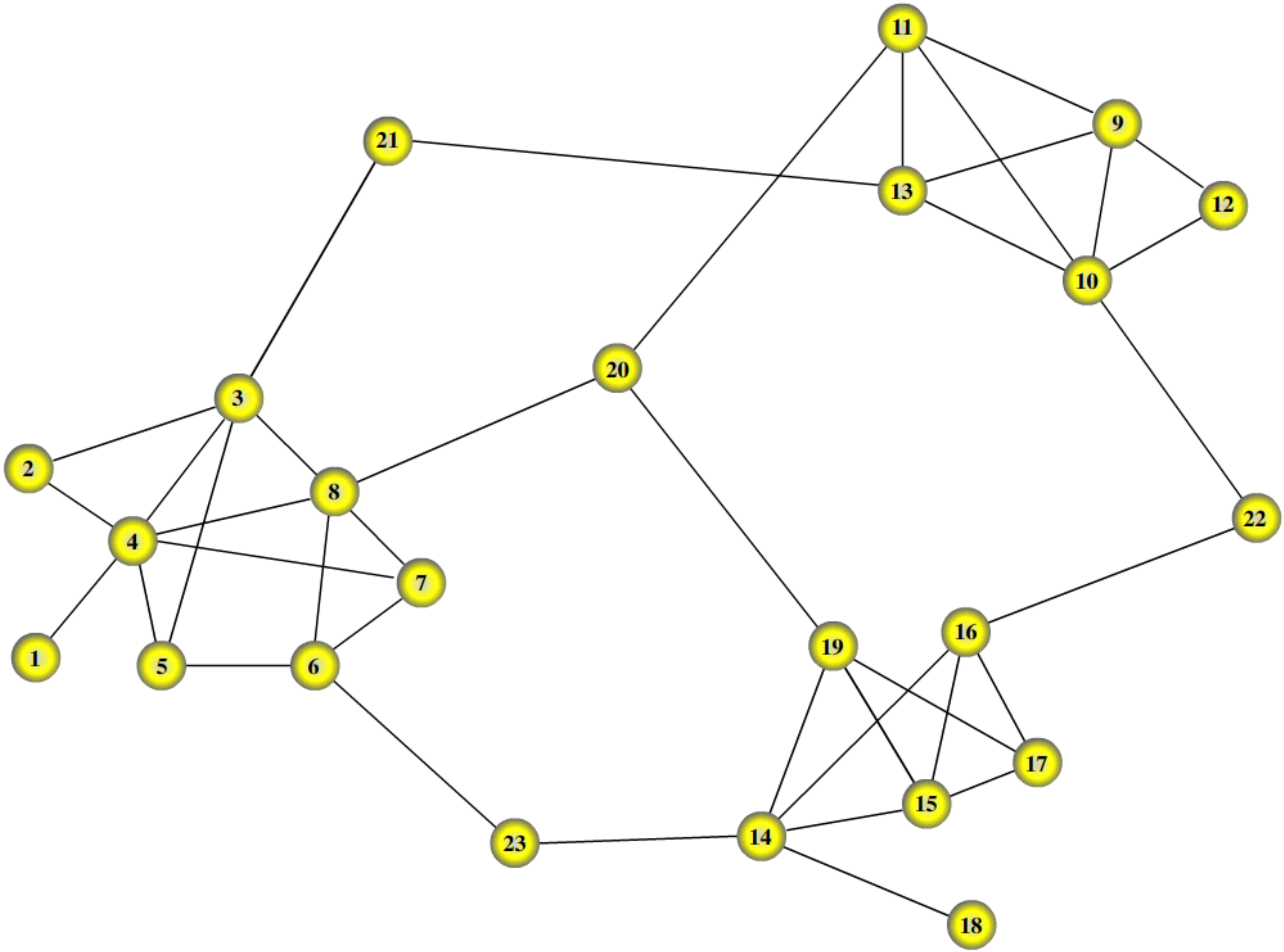
Local optimization

Implementation:

- Function Q(*C*) that assigns a score to each subgraph
- Best cluster -> cluster corresponding to the maximum/minimum of Q(*C*) over the set of subgraphs including a seed node

Example: Local Fitness Method (LFM)

A. Lancichinetti, S. F., J. Kertész, New. J. Phys. 11, 033015 (2009)



11
9
21
13
12
10
20
3
2
8
22
4
7
16
19
1
5
6
17
15
23
14
18

Local optimization: OSLOM

Basics:

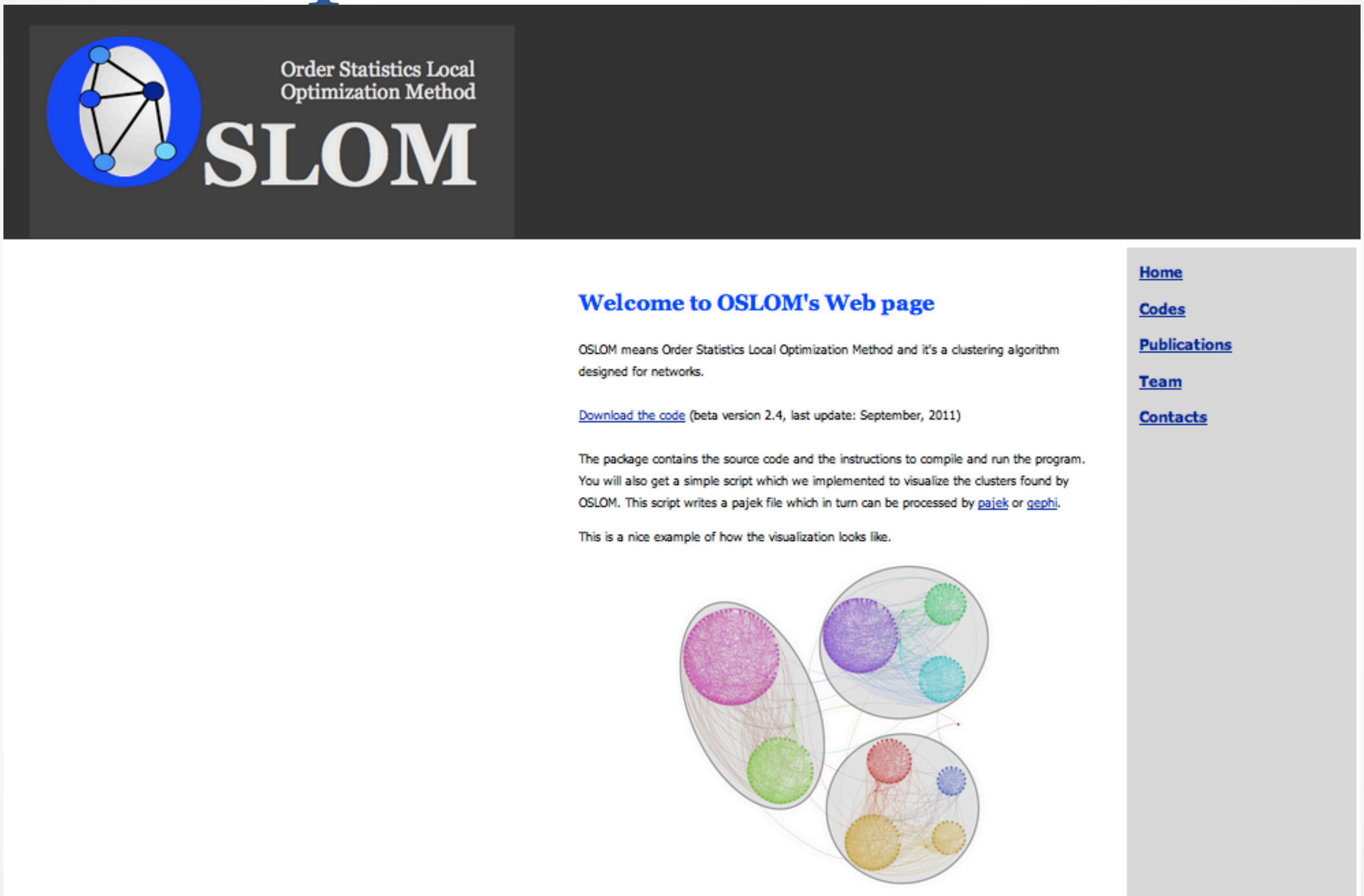
- LFM with fitness expressing the statistical significance of a cluster with respect to random fluctuations
- Statistical significance evaluated with Order Statistics

First multifunctional method:

- Link direction
- Link weight
- Overlapping clusters
- Hierarchy

A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. F., PLoS One 6, e18961 (2011)

Local optimization: OSLOM



http://www.oslom.org/

Testing clustering algorithms

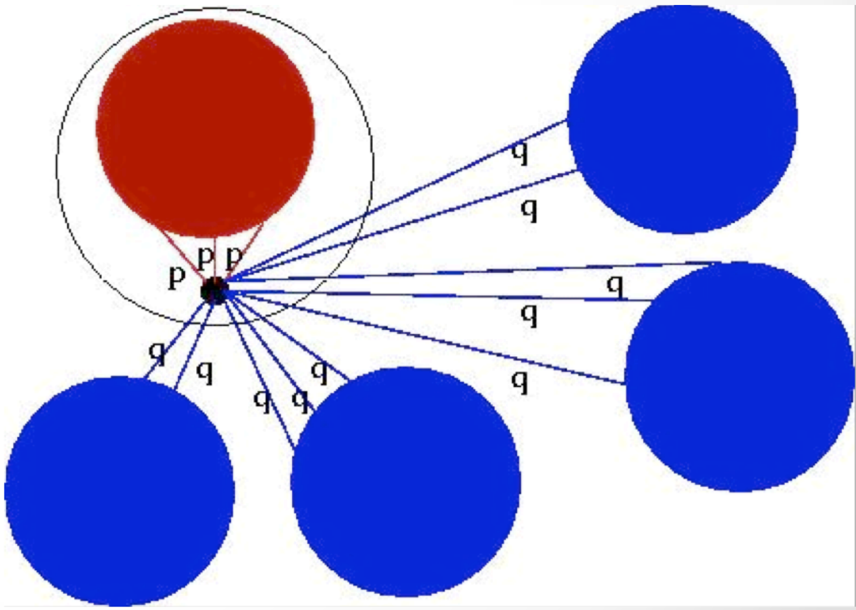
Question: how to test clustering algorithms?

Answer: checking whether they are able to recover the known community structure of benchmark graphs

Planted l-partition model (Condon & Karp, 1999)

Ingredients:

1) p=probability that vertices of the same cluster are joined
2) q=probability that vertices of different clusters are joined

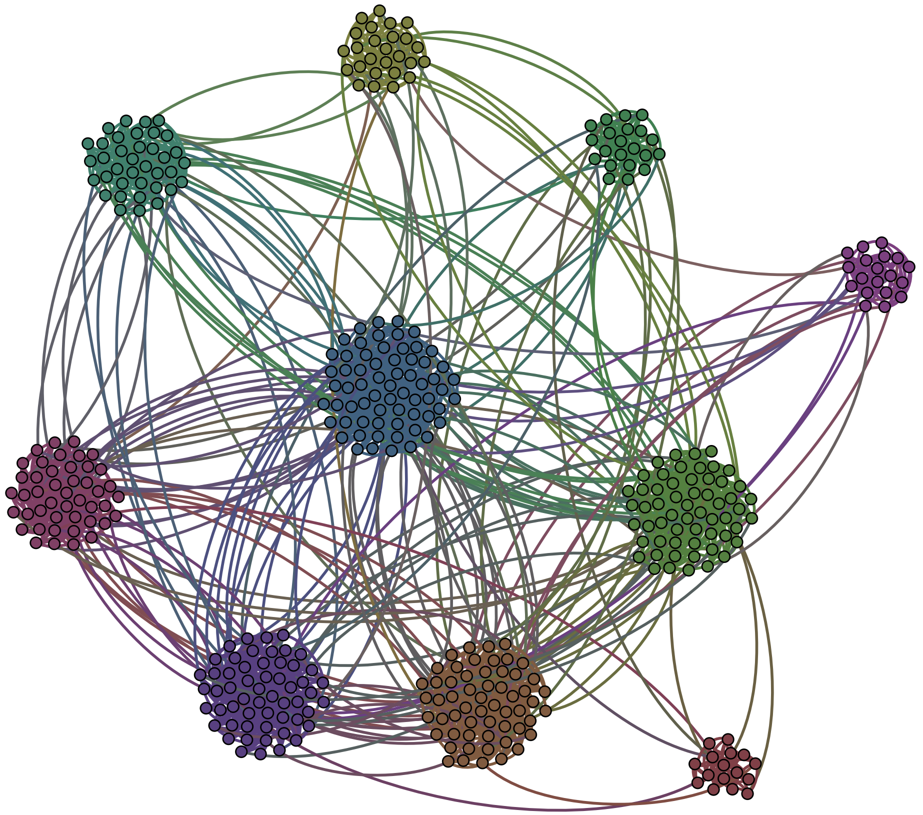


Principle: if p > q the groups are communities

Testing clustering algorithms

The LFR benchmark

Realistic feature: power law distributions of degree and community size



A. Lancichinetti, S. F., F. Radicchi, Phys. Rev. E 78, 046110 (2008)

https://sites.google.com/site/andrealancichinetti/files/

Testing clustering algorithms

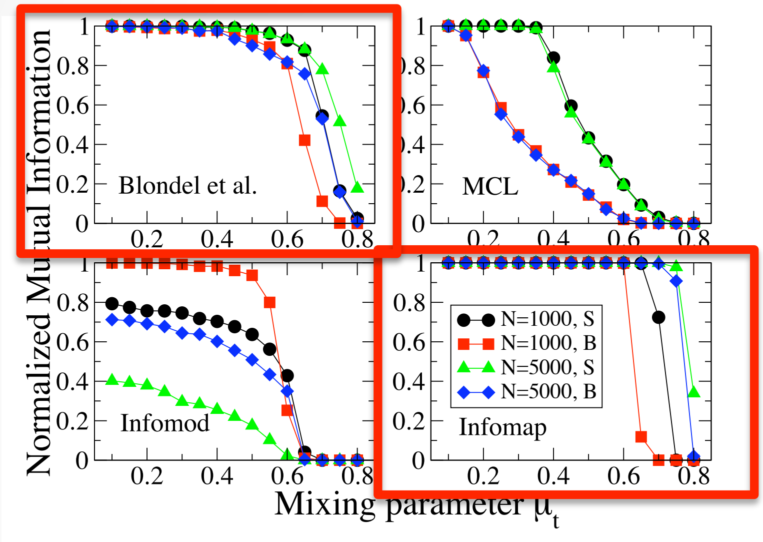
A comparative analysis

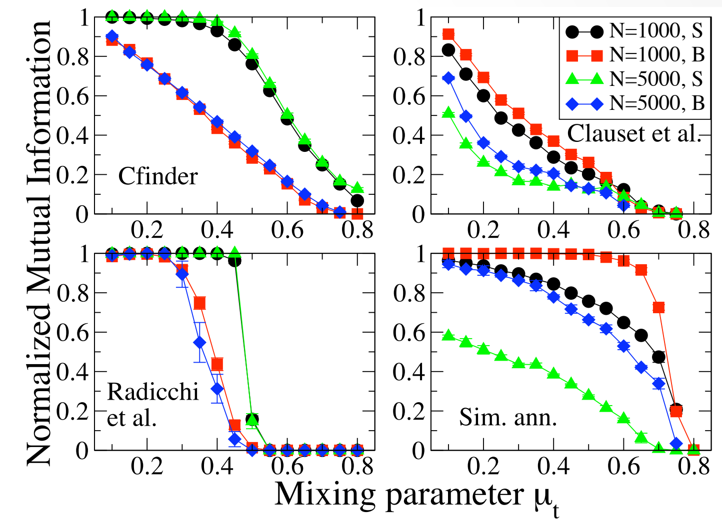
Author	Label	Order
Girvan & Newman	GN	$O(nm^2)$
Clauset et al.	Clauset et al.	$O(n \log^2 n)$
Blondel et al.	Blondel et al.	$O(m)$
Guimerà et al.	Sim. Ann.	parameter dependent
Radicchi et al.	Radicchi et al.	$O(m^4/n^2)$
Palla et al.	Cfinder	$O(\exp(n))$
Van Dongen	MCL	$O(nk^2)$, $k < n$ parameter
Rosvall & Bergstrom	Infomod	parameter dependent
Rosvall & Bergstrom	Infomap	$O(m)$
Donetti & Muñoz	DM	$O(n^3)$
Newman & Leicht	EM	parameter dependent
Ronhovde & Nussinov	RN	$O(n^\beta)$, $\beta \sim 1$

A. Lancichinetti, S. F., Phys. Rev. E 80, 056117 (2009)

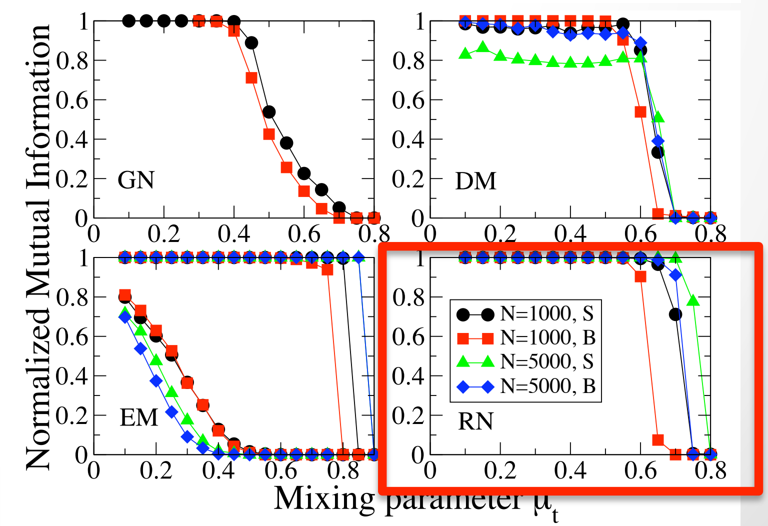
Testing clustering algorithms

A comparative analysis



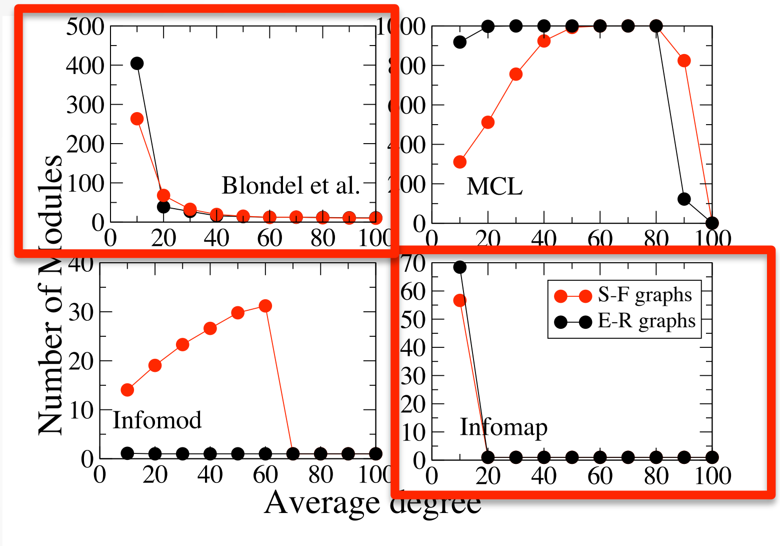


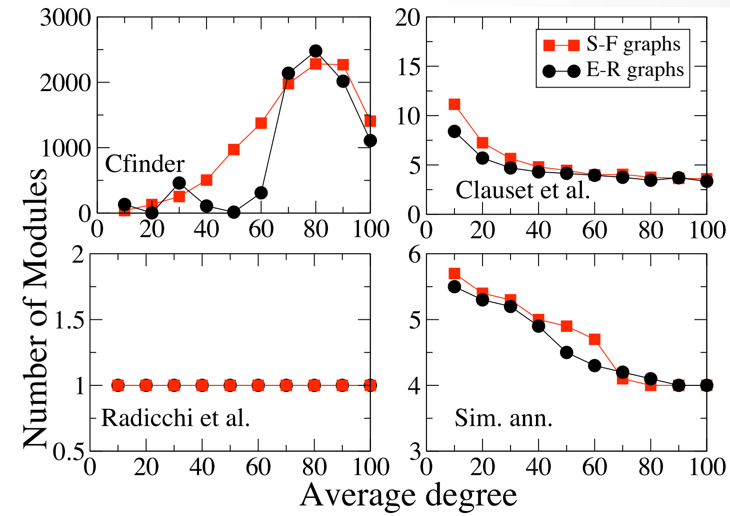
LFR benchmark graphs

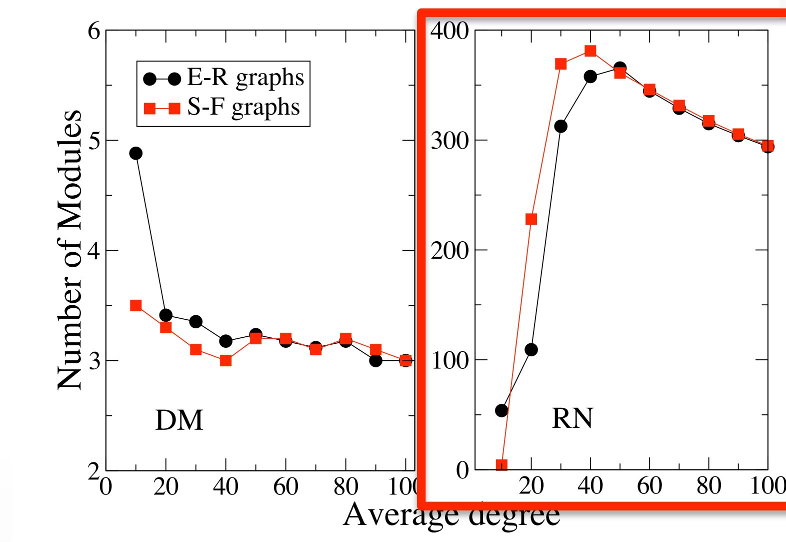


Testing clustering algorithms

A comparative analysis







Random graphs: no clusters!

Testing clustering algorithms

Limits of artificial benchmarks:

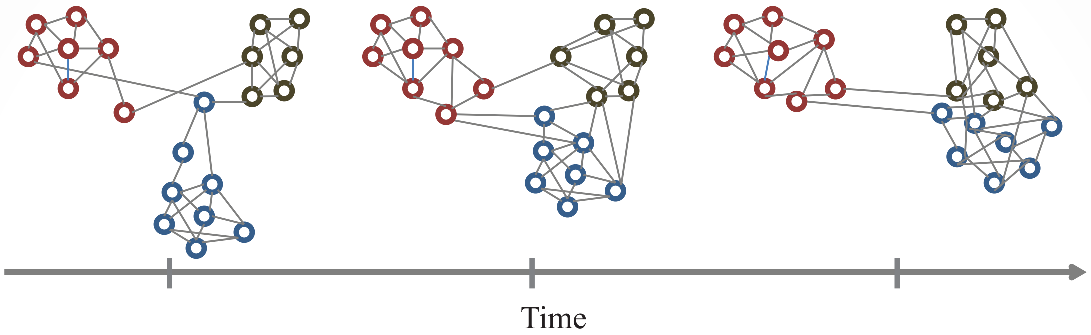
- Relationships with real community structure unclear
- Risk of creating algorithms performing well on the benchmarks and not so well on real networks

Solution: real networks with ground truth classification?

Yang & Leskovec (2012): *arXiv: 1205.6233, arXiv: 1205.6228*

Warning: classification must be reliable!

Dynamic clustering



Typical approach:

- Find the clusters for each snapshot with a static method
- Associate clusters of different snapshots

Limit: partitions independent of the history of the system

Alternative: exploiting the structural information of different snapshots -> *cluster stability*

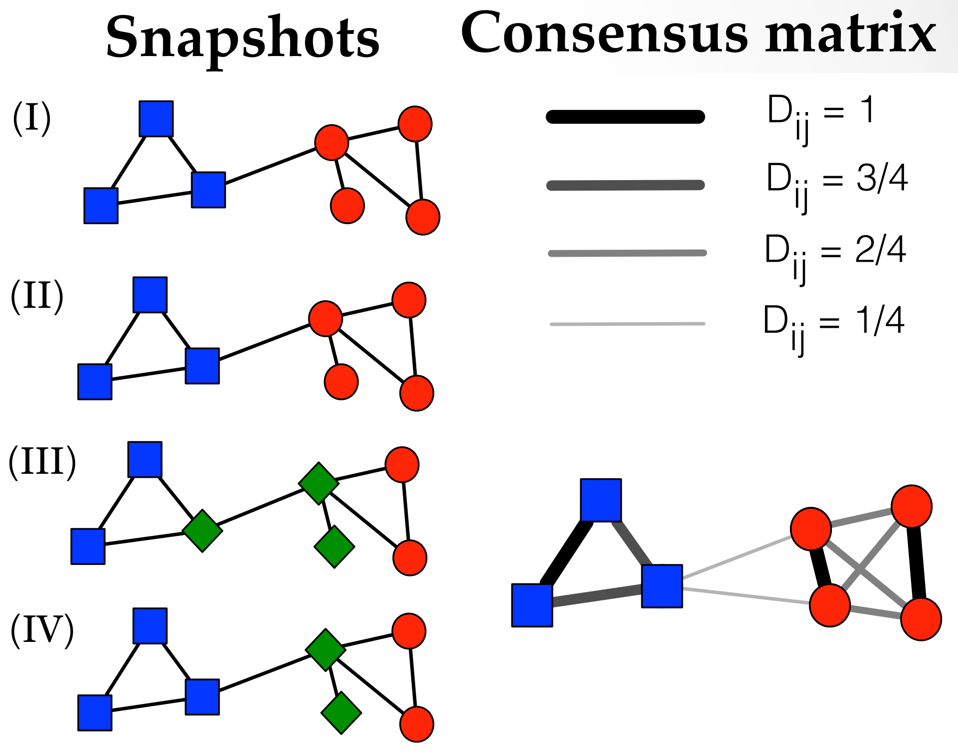
Dynamic clustering

Consensus clustering

Goal: finding *median* partition of network

Steps:

1) Compute partition of each snapshot with static algorithm
2) Compute consensus partition for sequences of k consecutive snapshots



Resulting partitions more accurate and stable!

A. Lancichinetti, S. F., Sci. Rep. 2, 336 (2012)

NetCom Analyzer

http://www.netcom-analyzer.org/

Summary of the talk

1) **Global optimization** methods have important limits: **local optimization** looks more natural and promising
2) **Validation**:

   a) artificial benchmarks useful, not 100% reliable

   b) real networks with ground truth information

Summary of the field

1) What is a community? **No unique answer! Definition is system- and problem-dependent**
2) Magic method? **No such thing! Domain dependent methods?**
3) **Low complexity** techniques (down to linear!)
4) **Versatile** methods: directed networks, weighted networks, overlapping communities, hierarchy
5) Attention on **validation**
6) **Constraints**: a (new) method should

   a) not split cliques

   b) not merge cliques, if well-separated

   c) not find communities in random graphs

Total citations Cited by 1258

Citations per year

Scholar articles **Community detection in graphs**
S Fortunato - Physics Reports, 2010

Article history:
Accepted 5 November 2009
Available online 4 December 2009
editor: I. Procaccia

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly

S. F., Phys. Rep. 486, 75-174 (2010)

Most Cited Physics Reports Articles

The most cited articles published since 2008, extracted from SciVerse Scopus.

Top 25 Hottest Articles
Physics and Astronomy

Community detection in graphs

Volume 486, Issues 3-5, February 2010, Pages 75-174

Fortunato, S.

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks. © 2009 Elsevier B.V.

Acknowledgements

Alex Arenas

Marc Barthelémy

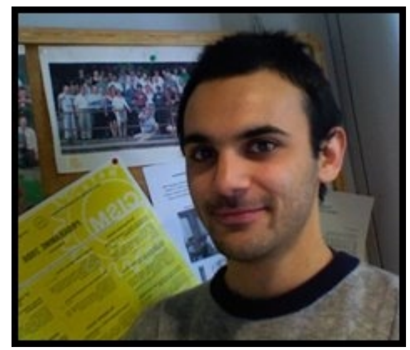
Alberto Fernández

Sergio Gómez

Janos Kertész

Mikko Kivelä

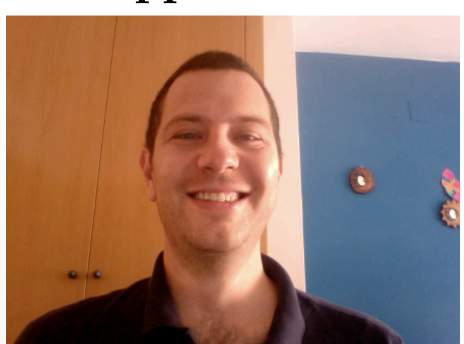
Andrea Lancichinetti



Vito Latora

Massimo Marchiori

Filippo Radicchi



José J. Ramasco

Jari Saramäki