Community detection in networks

Santo Fortunato

Aalto University

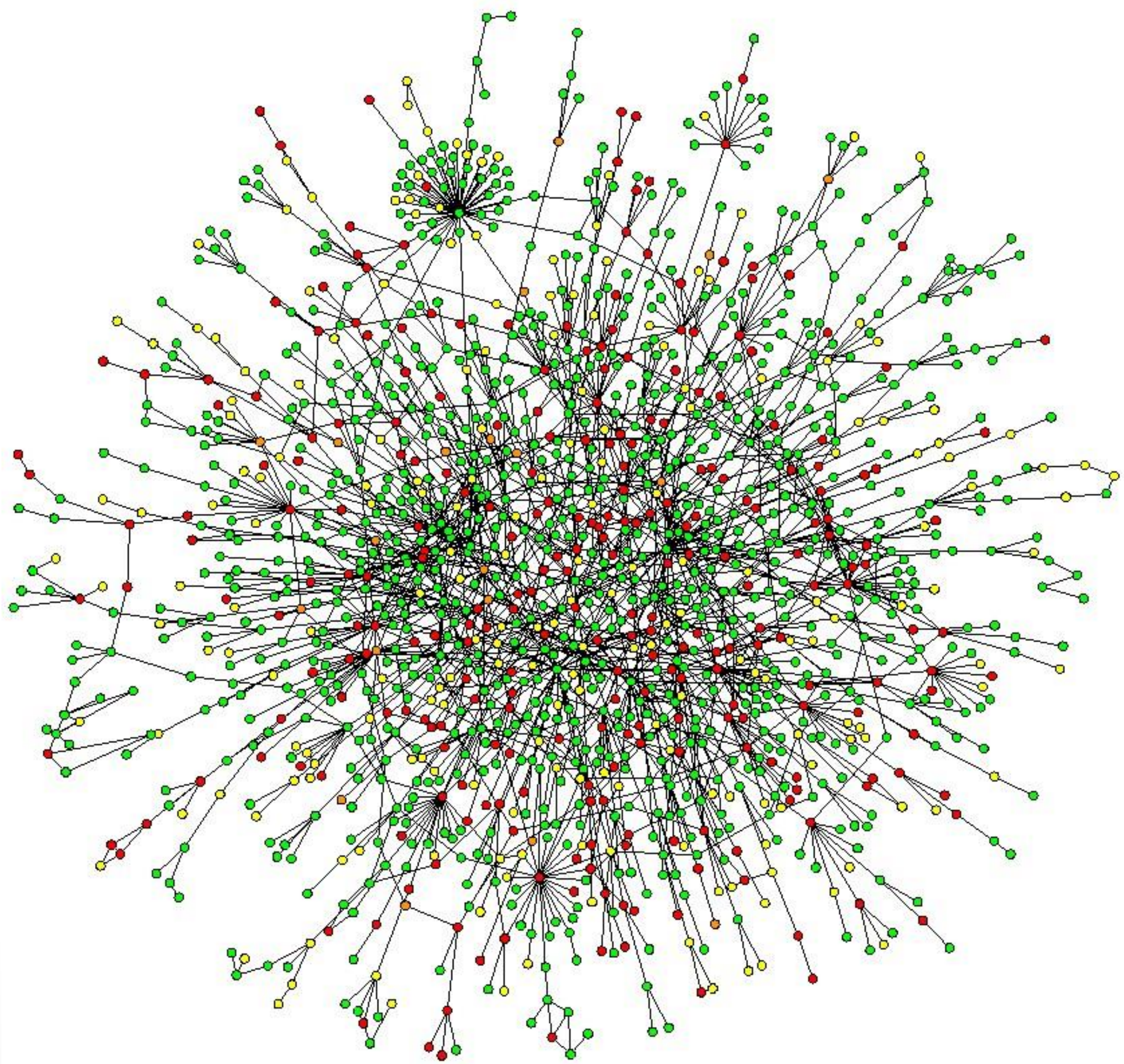
Outline

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

Networks

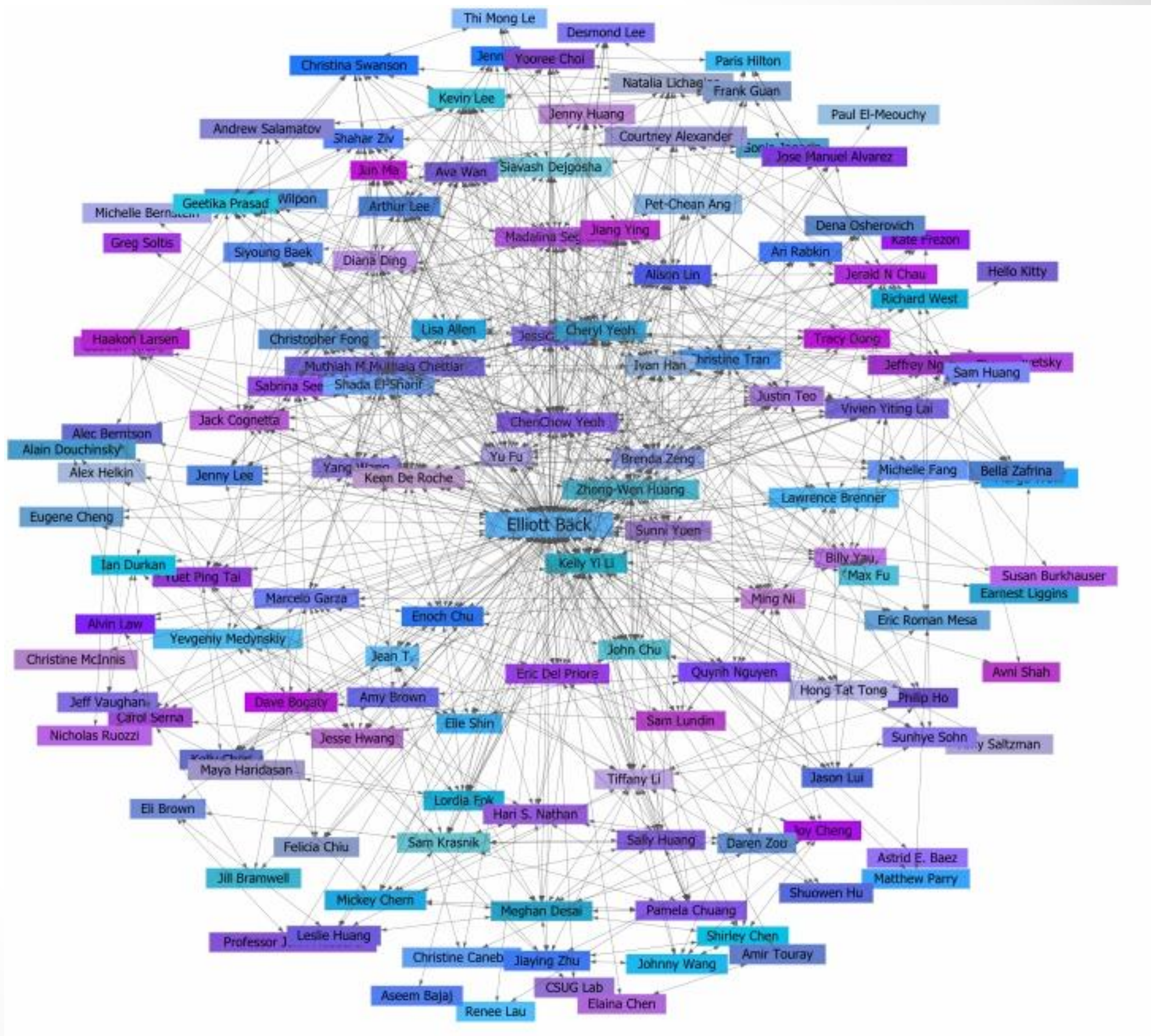
Protein-protein interaction networks

Network: simplest representation of a complex system



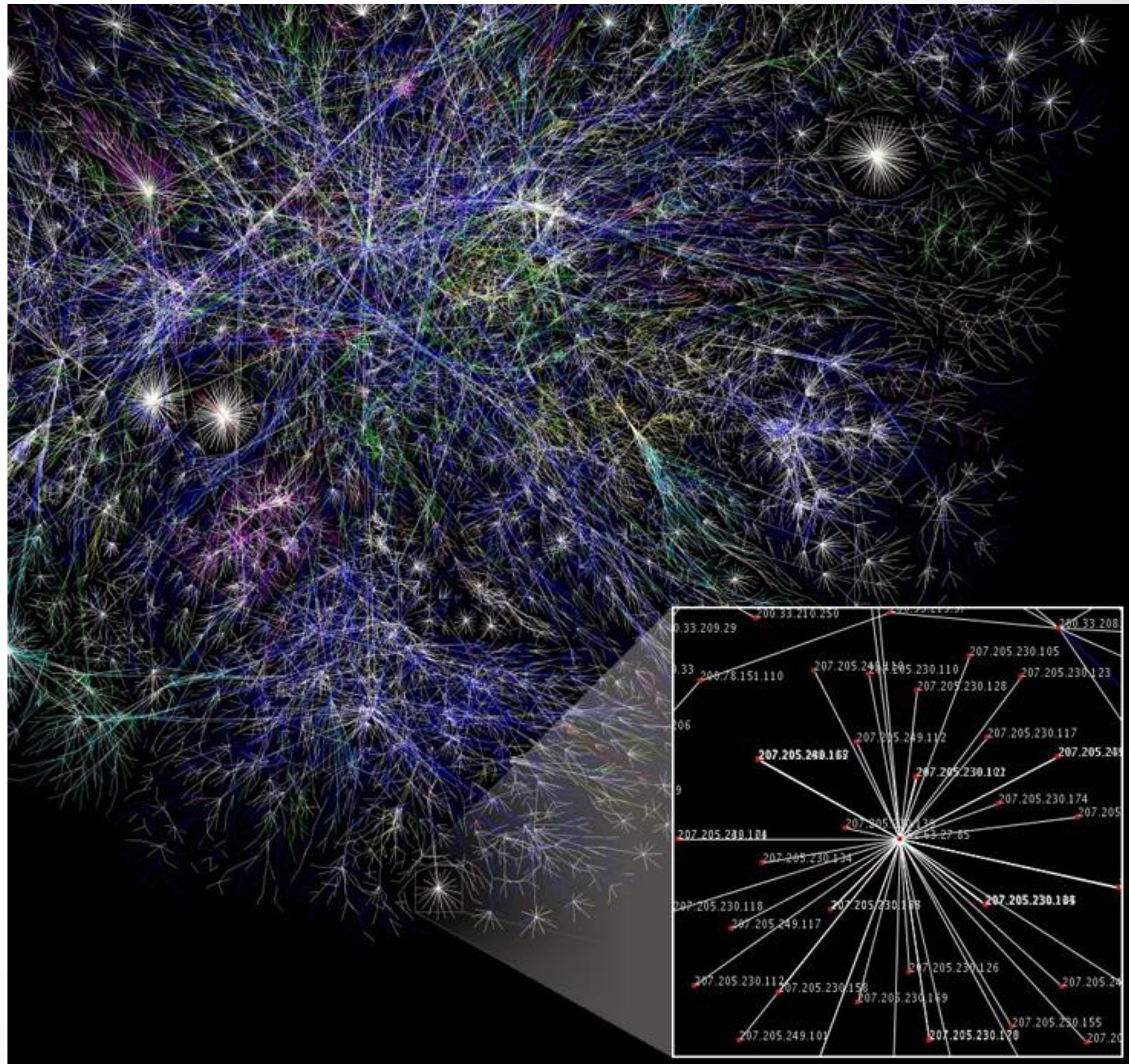
Networks

Social networks



Networks

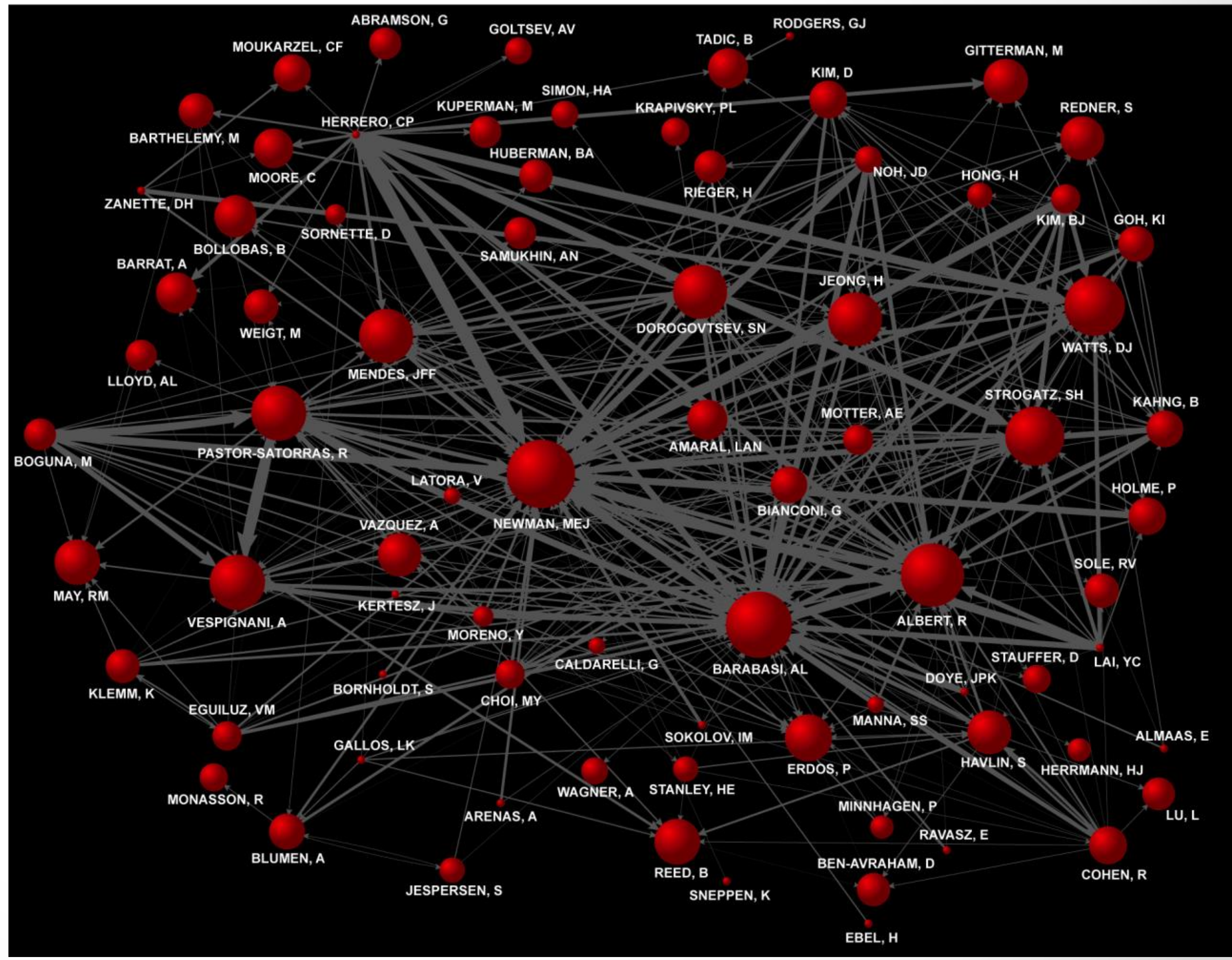
Internet



Networks

Citation networks

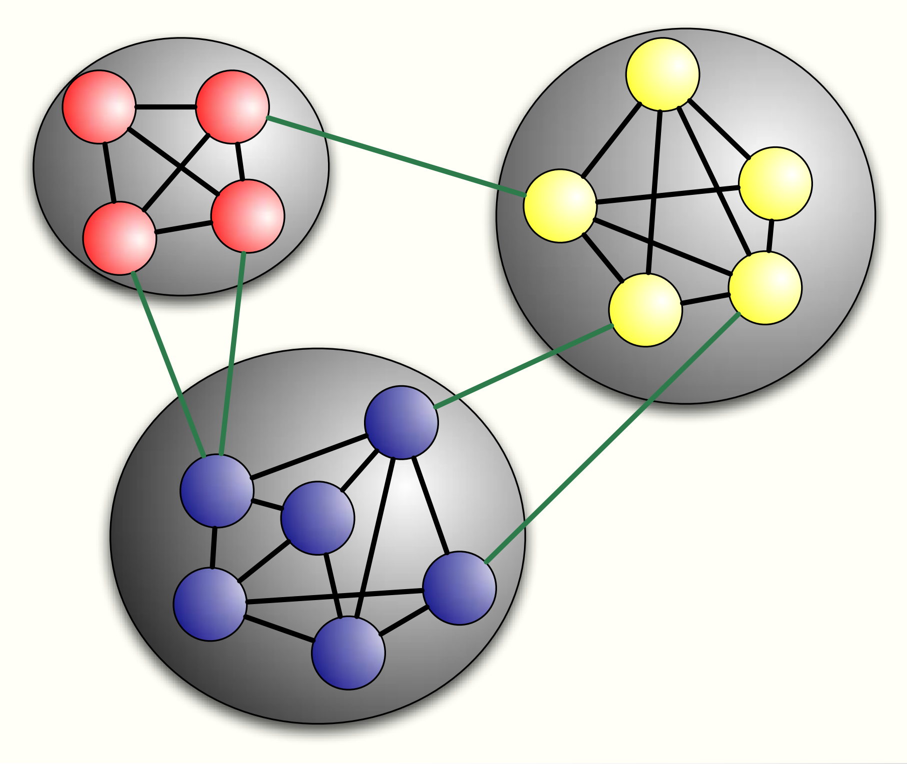
Important features of a system and its dynamics from purely structural information



Community structure

Communities: sets of tightly connected nodes

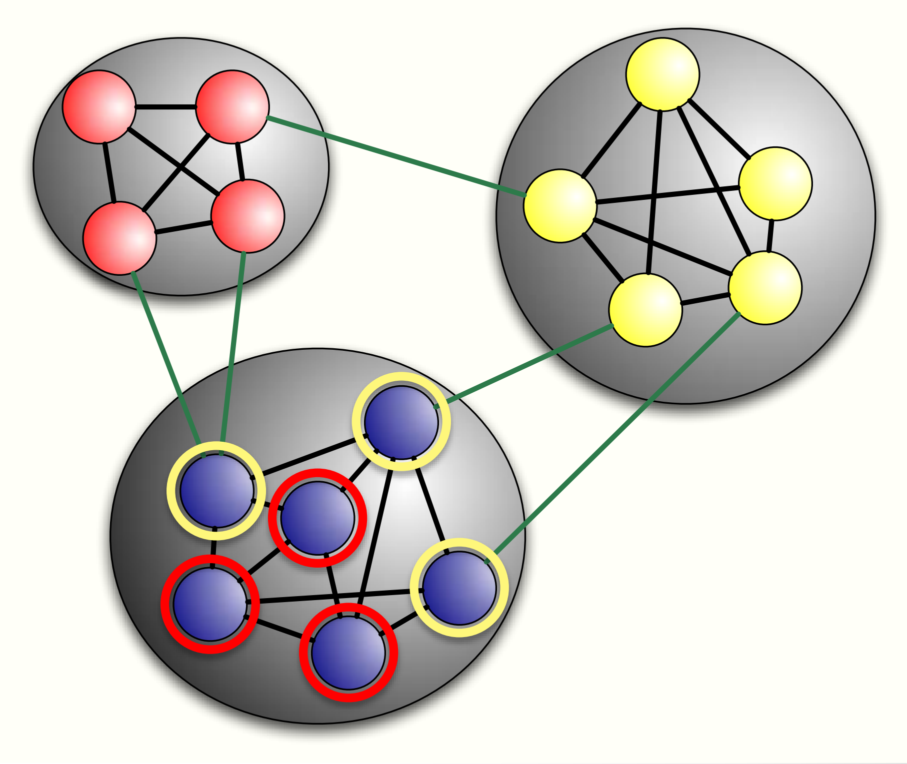
- People with common interests
- Scholars working on the same field
- Proteins with equal/similar functions
- Papers on the same/related topics
- …



Community detection

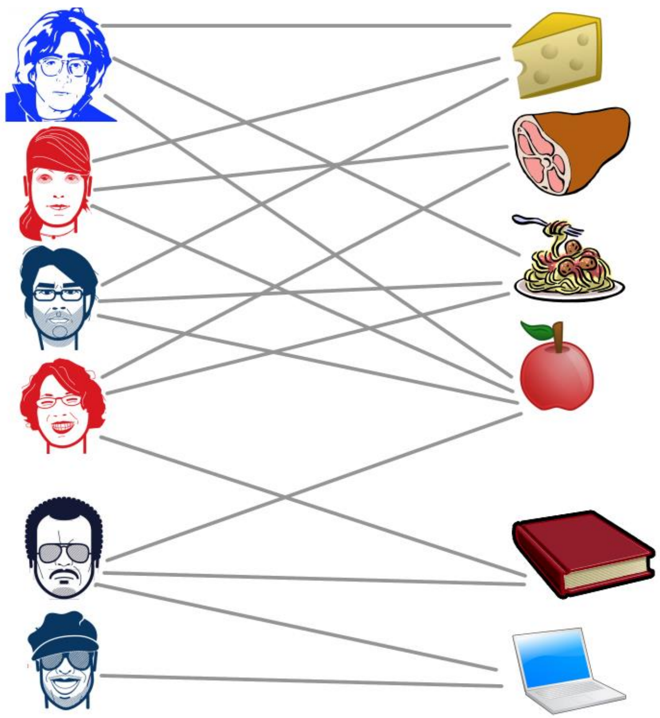
Theoretical reasons

- Organization
- Node features
- Node classification
- Missing links



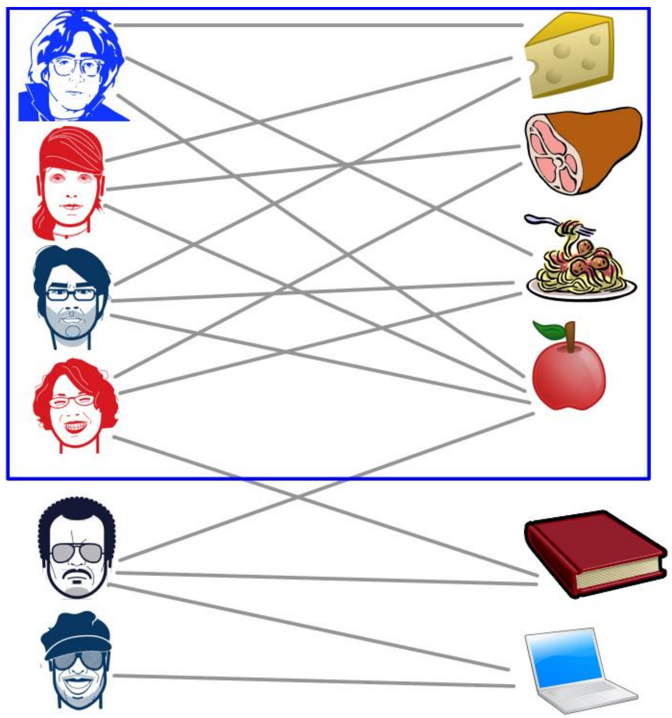
Community detection

Practical reasons: recommendation systems



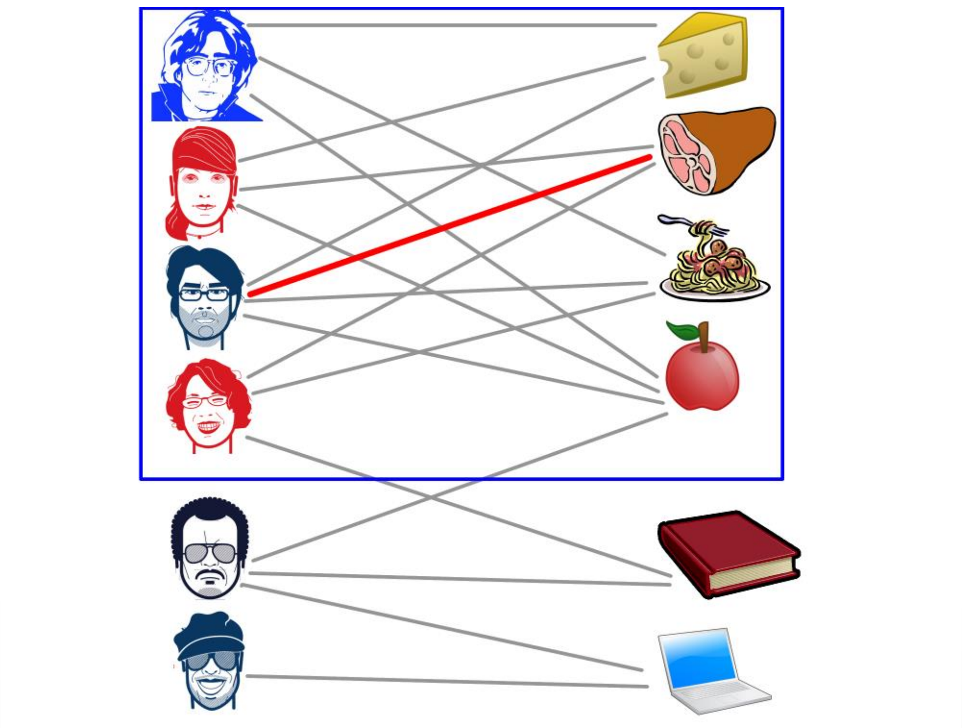
Community detection

Practical reasons: recommendation systems



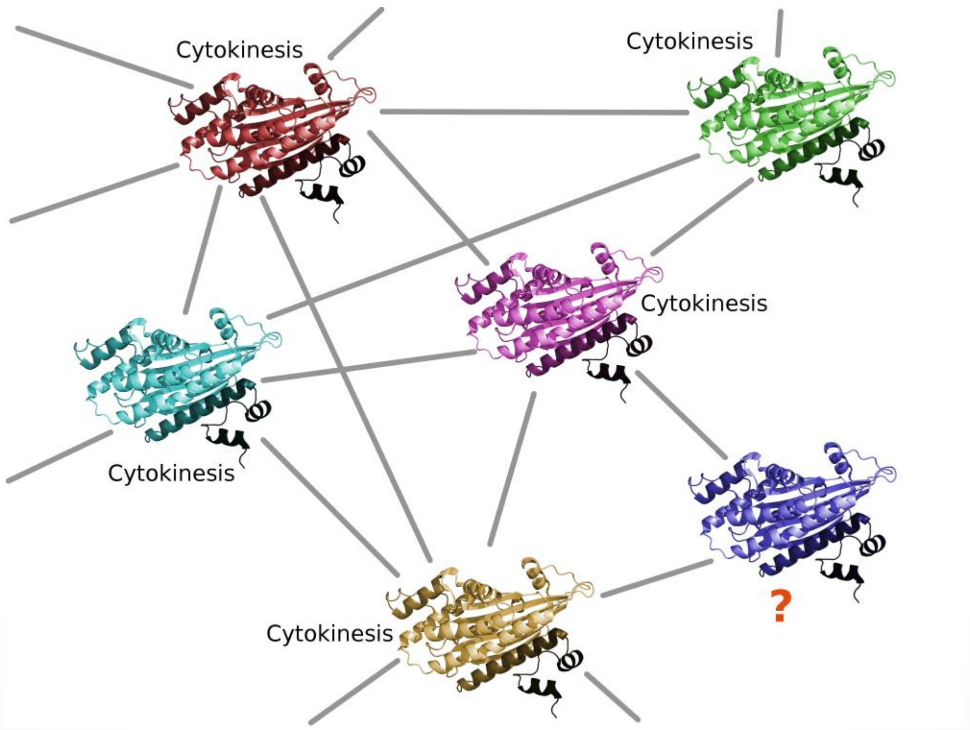
Community detection

Practical reasons: recommendation systems



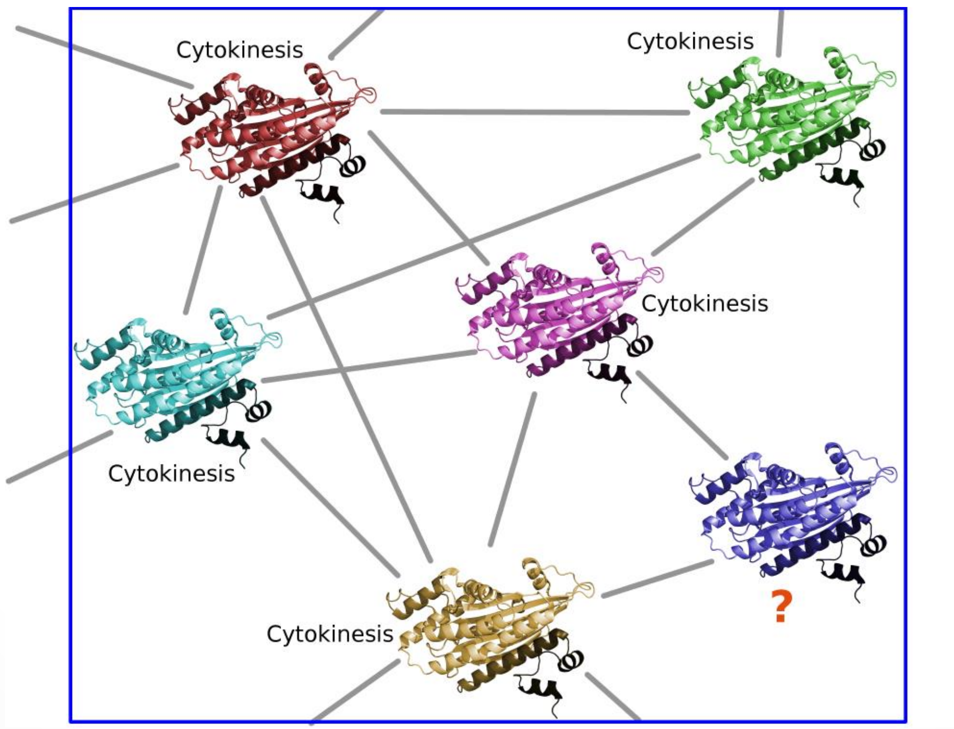
Community detection

Practical reasons: unknown protein functions



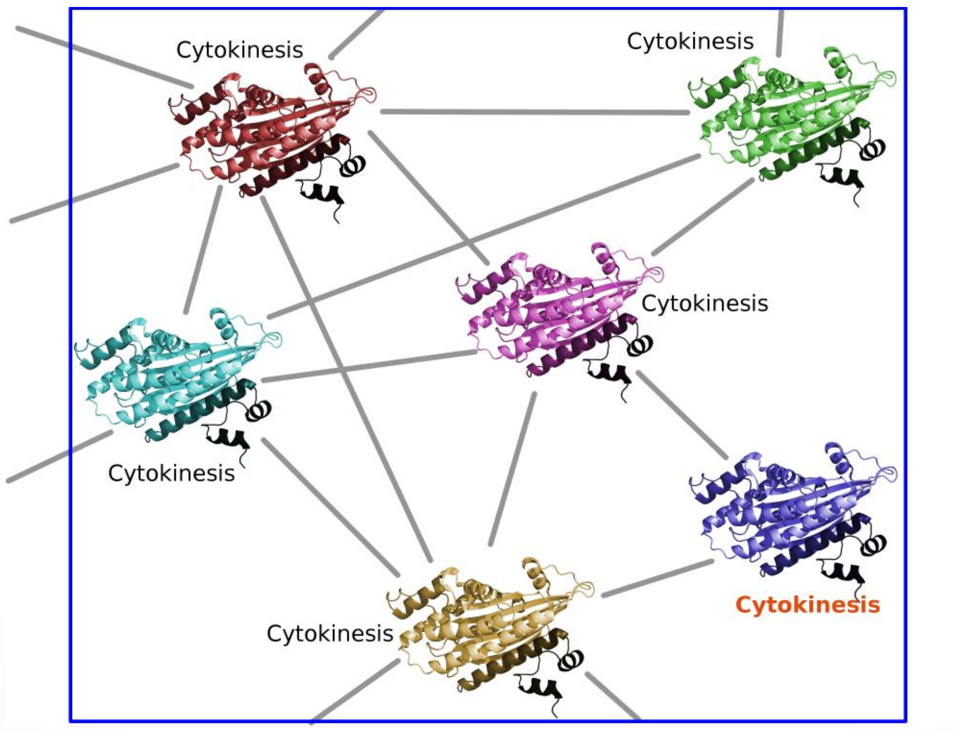
Community detection

Practical reasons: unknown protein functions

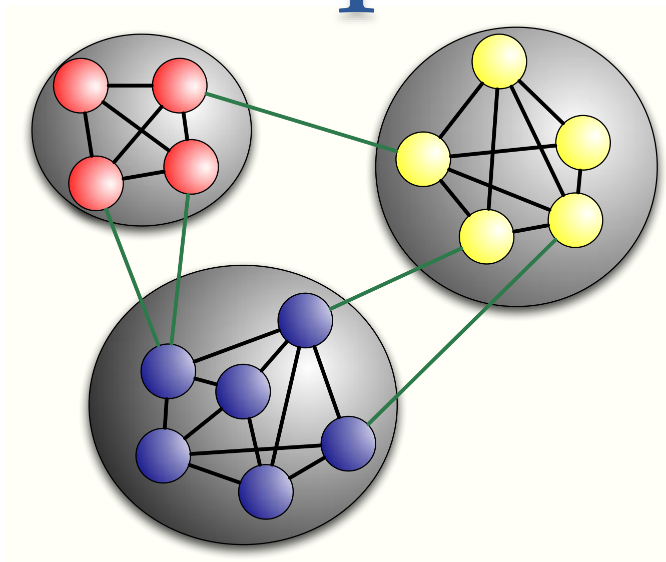


Community detection

Practical reasons: unknown protein functions



Difficult problem!

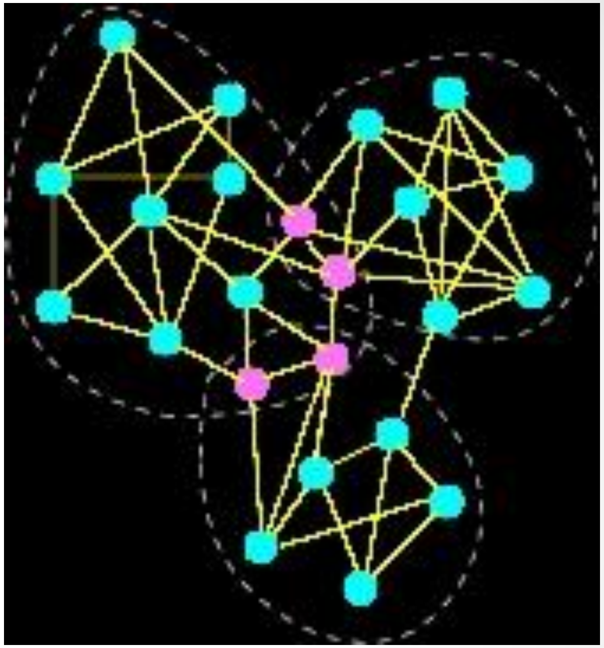


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1								1		1	1			1	1
2						1		1						1	
3	1											1			1
4				1						1		1			1
5							1	1					1	1	
6		1						1							1
7			1						1			1	1		
8		1				1	1							1	
9			1				1					1	1		
10		1						1						1	1
11							1			1		1	1		
12			1				1		1						1
13				1			1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

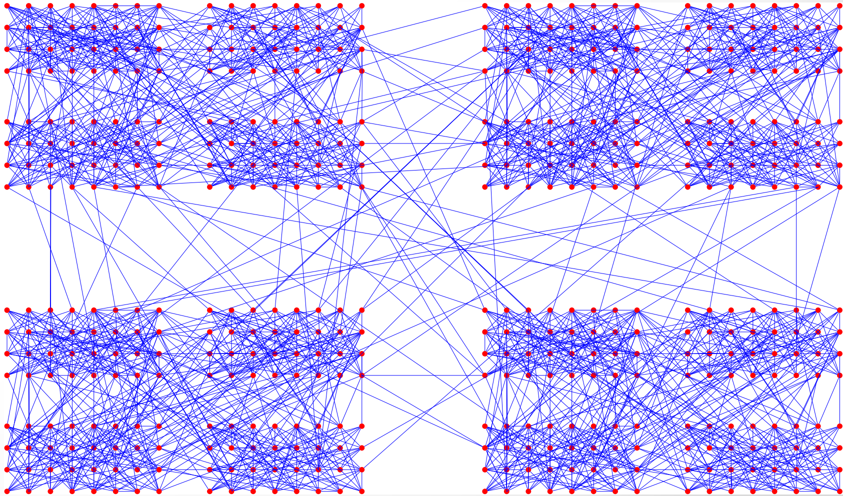
	6	2	8	14	5	7	13	11	9	12	10	4	1	3	15
6		1	1	1											
2	1		1	1							1				
8	1	1		1		1									
14	1	1	1			1									
5			1	1		1	1			1					
7					1		1	1	1						
13					1	1		1		1					
11						1	1		1	1					
9						1		1		1			1		
12					1		1	1	1		1				
10		1										1	1	1	1
4									1		1		1		1
1				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



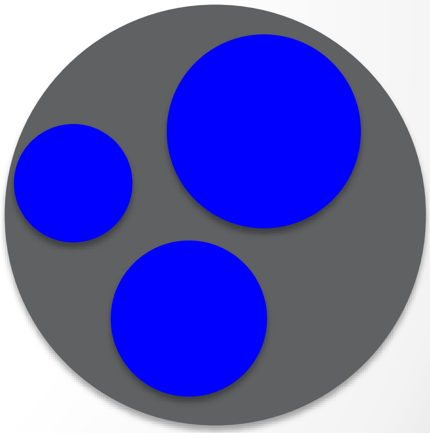
Global optimization

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



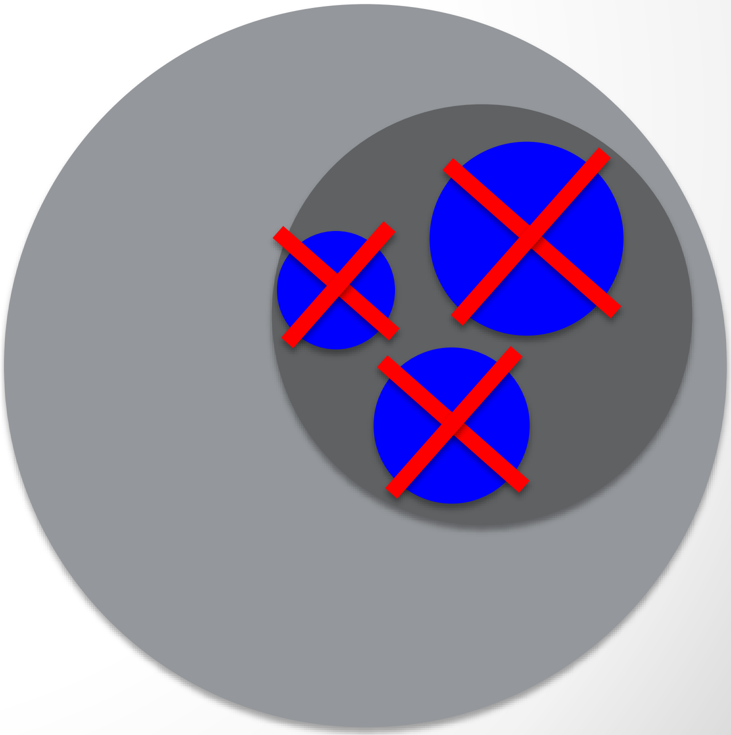
Global optimization

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

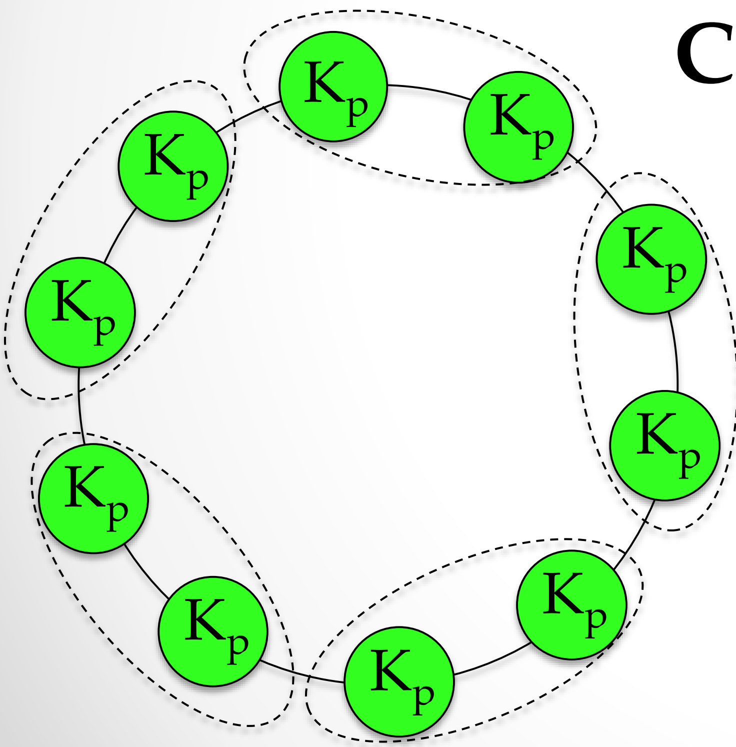
Problem: NP-complete (Brandes et al., 2007)!

Resolution limit

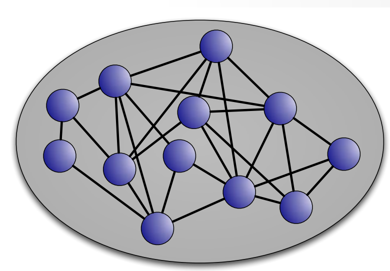
$$Q = \frac{1}{m}\sum_{c=1}^{n_c}\left[l_c - \frac{1}{4}\left(\frac{d_c}{\sqrt{m}}\right)^2\right]$$

modularity's scale

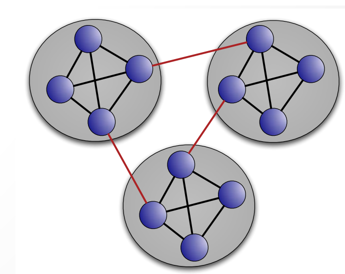
Consequences



$d_c < \sqrt{m}$ ⟶



$d_c < \sqrt{m}$ ⟶



?

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

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

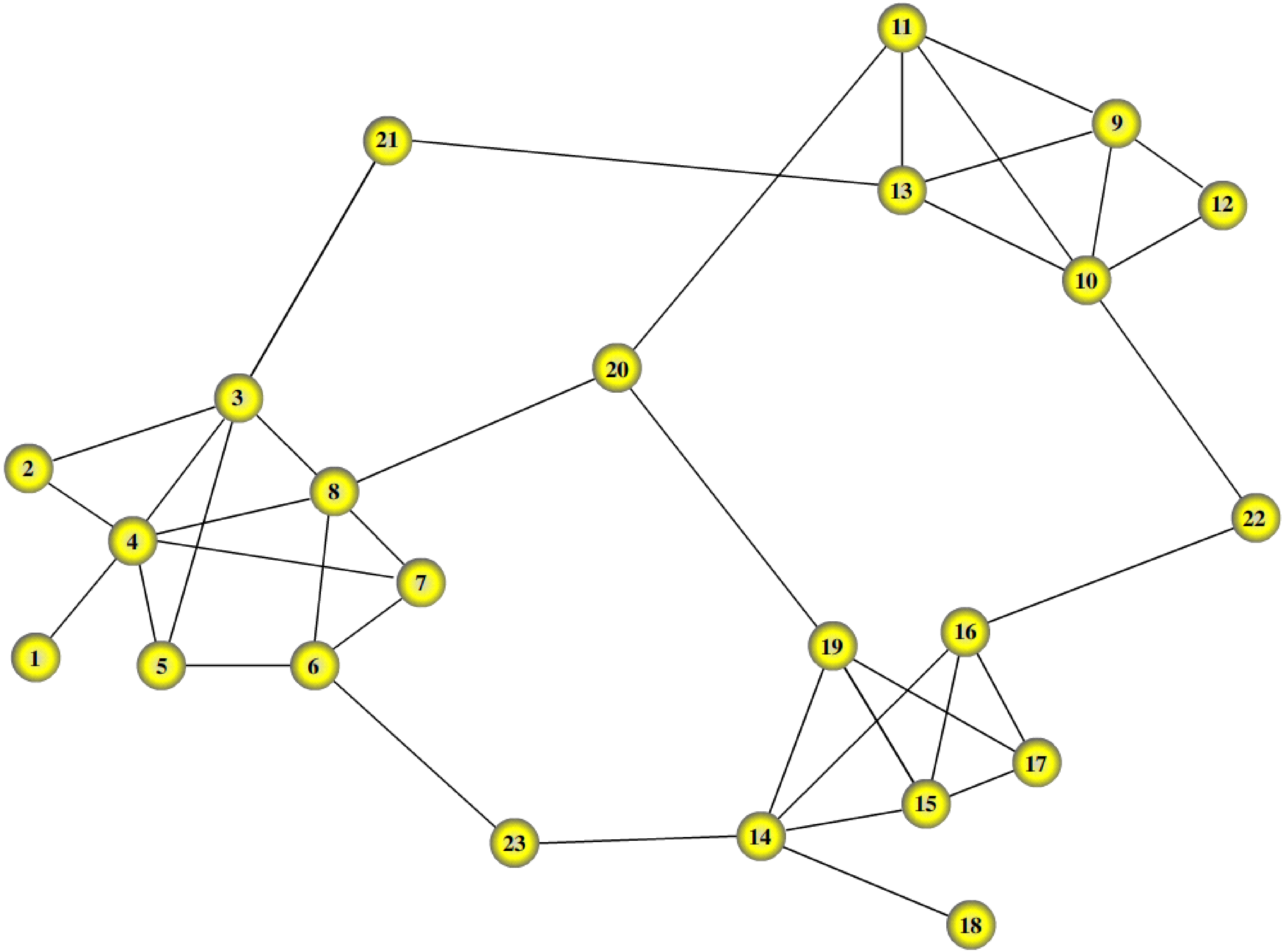
Implementation:

- 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)

Fitness of cluster *C* : $$f_{\mathcal{C}} = \frac{k_{in}^{\mathcal{C}}}{(k_{in}^{\mathcal{C}} + k_{out}^{\mathcal{C}})^{\alpha}} = \frac{2l_{\mathcal{C}}}{d_{\mathcal{C}}^{\alpha}}$$

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



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

Local optimization: OSLOM

Basics:

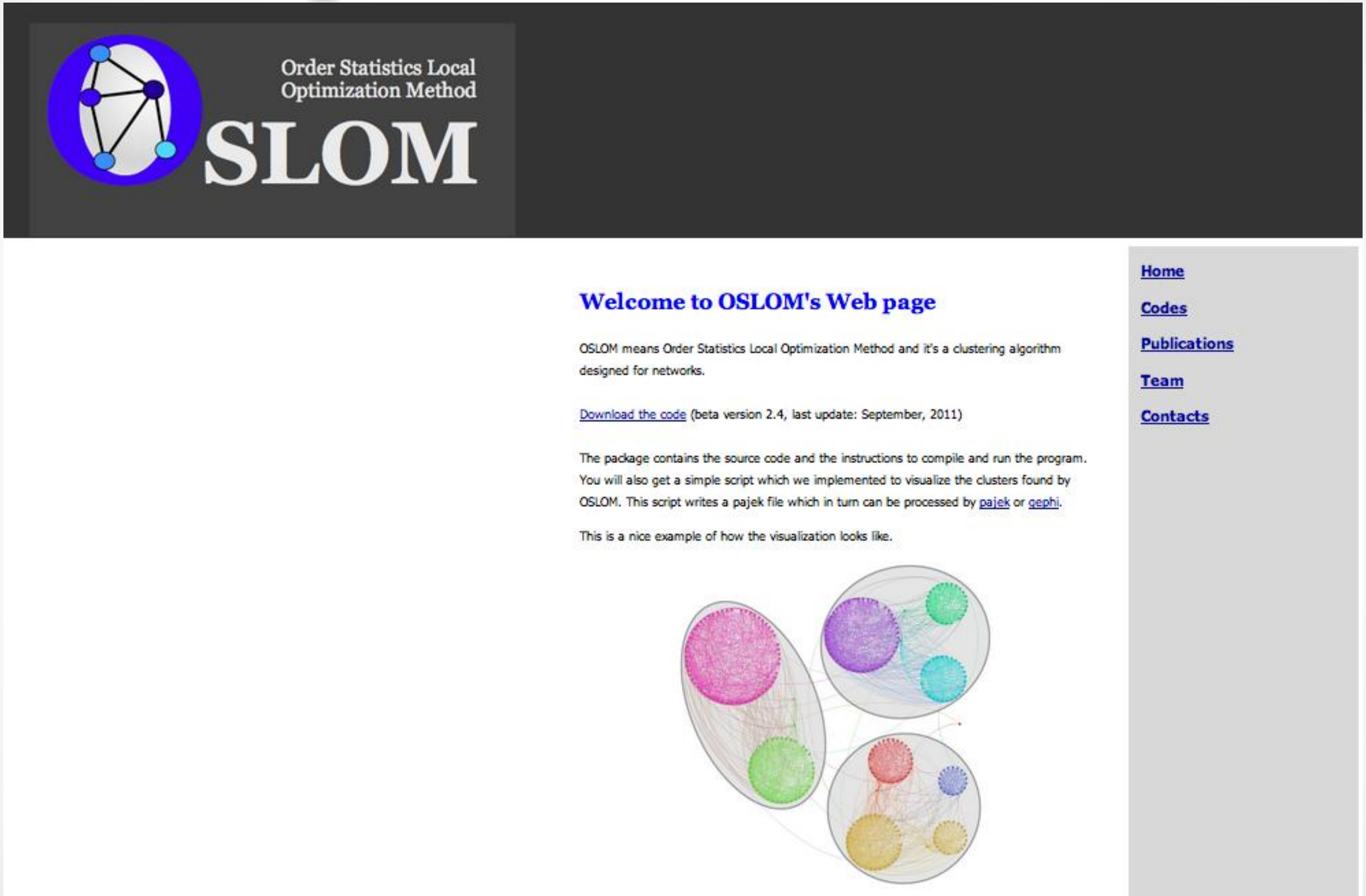
- LFM with fitness expressing the statistical significance of a cluster with respect to random fluctuations
- Statistical significance evaluated with Order Statistics

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/

NetCom Analyzer

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

Summary

1) What is a community? **No unique answer**! **Definition is system- and problem-dependent**
2) Magic method? **No such thing**! **Domain dependent methods?**
3) **Global optimization** methods have important limits: **local optimization** looks more natural and promising
4) Attention on **validation**

Total citations Cited by 1874

Citations per year

Scholar articles

Community detection in graphs
S Fortunato - Physics Reports, 2010
Cited by 1874 - Related articles - All 24 versions

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

Top 25 Hottest Articles

Physics and Astronomy

Most Cited Physics Reports Articles

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

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.

Acknowledgements

Alex Arenas

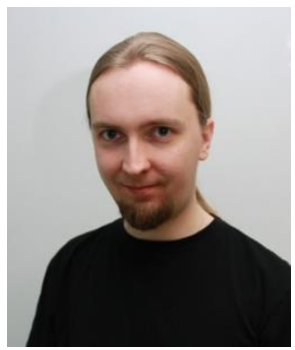
Marc Barthelémy

Alberto Fernández

Sergio Gómez

Janos Kertész

Mikko Kivelä

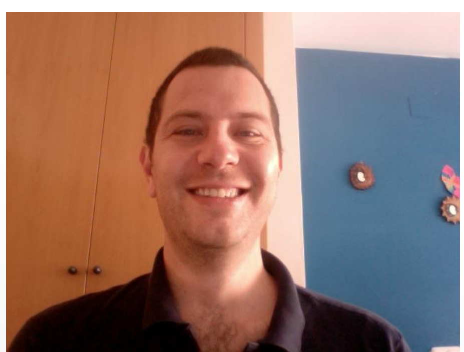


Andrea Lancichinetti

Vito Latora

Massimo Marchiori

Filippo Radicchi



José J. Ramasco

Jari Saramäki