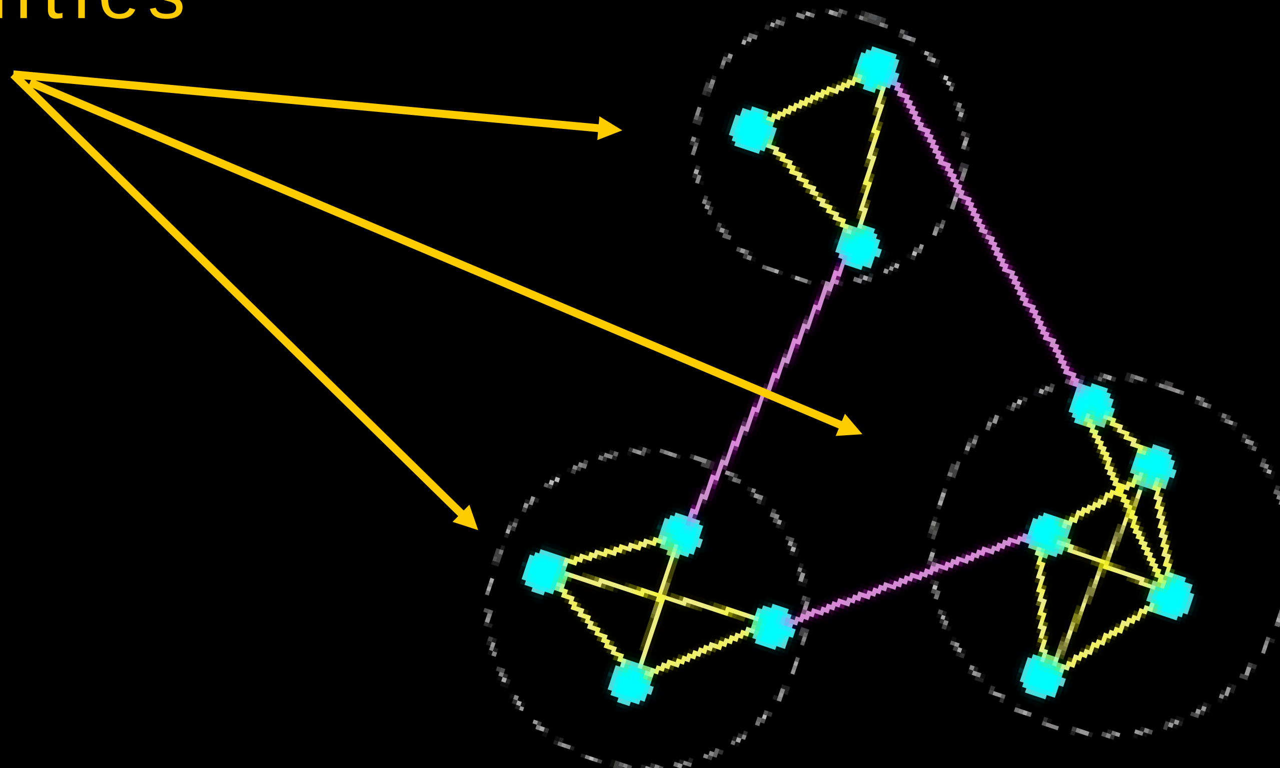
Community structure in graphs

Santo Fortunato

INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

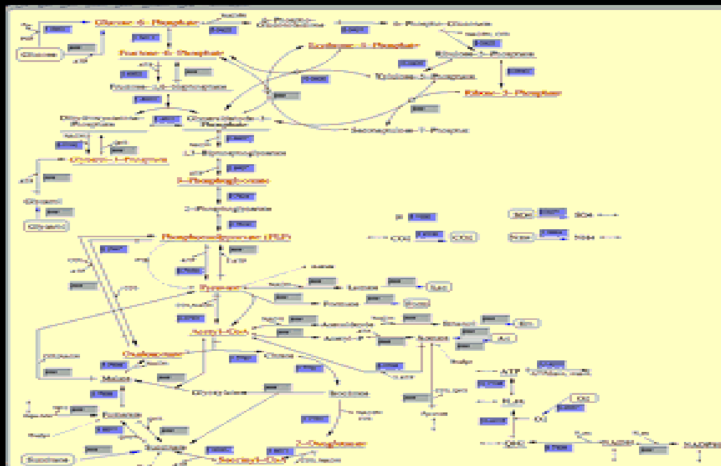
“Communities”



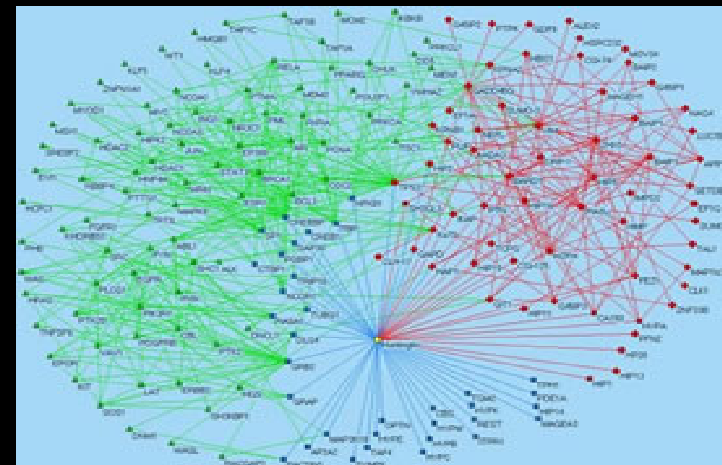
More links “inside” than “outside”

Graphs are “sparse”

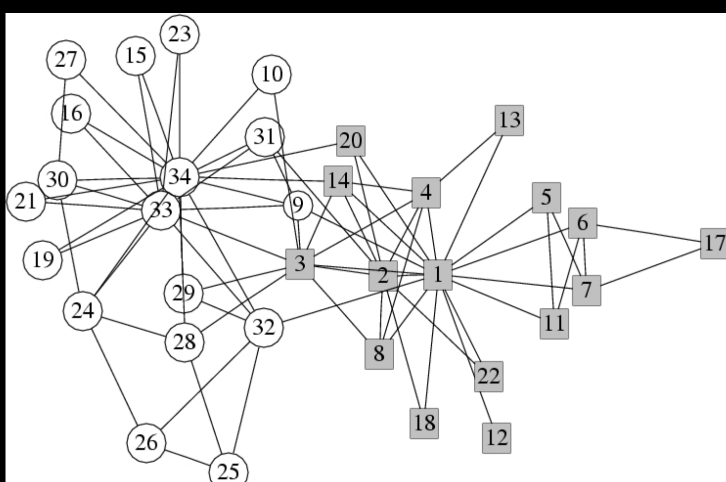
Metabolic



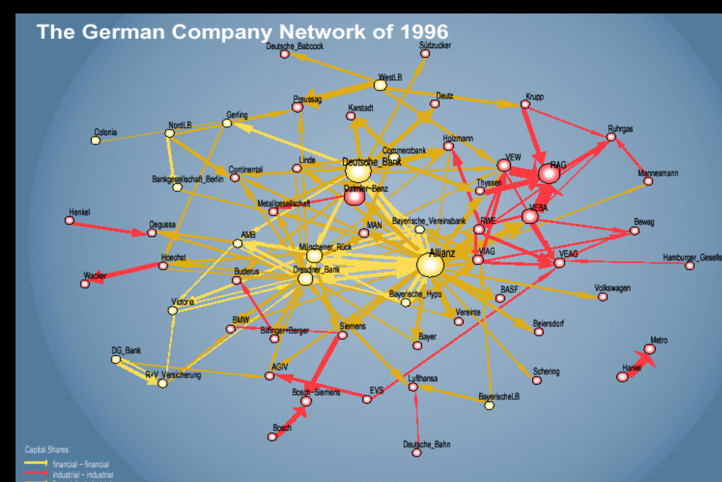
Protein-protein



Social



Economical



Outline

- Elements of community detection
- Graph partitioning
- Hierarchical clustering
- The Girvan-Newman algorithm
- New methods
- Testing algorithms
- Conclusions

Questions

- What is a community?
- What is a partition?
- What is a “good” partition?

Communities: definition

- Local criteria
- Global criteria
- Vertex similarity

In general, communities are indirectly defined by the particular algorithm used!

What is a partition?

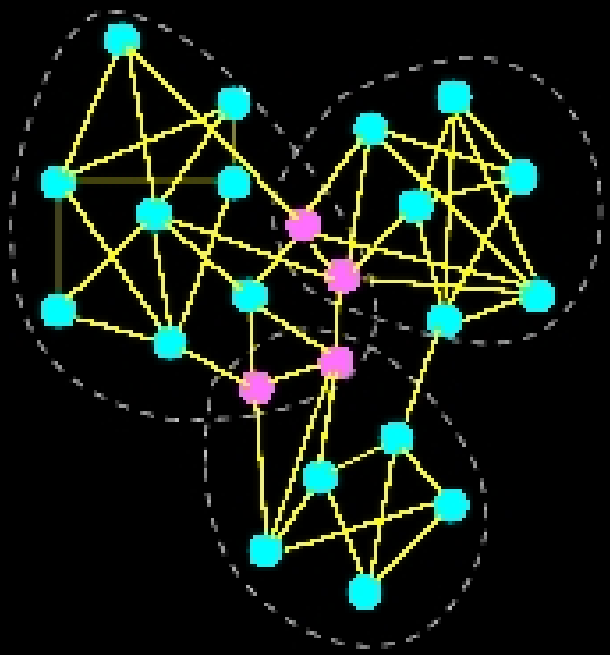
“A partition is a subdivision of a graph in groups of vertices, such that each vertex is assigned to one group”

Problems:

1) Overlapping communities
2) Hierarchical structure

Overlapping communities

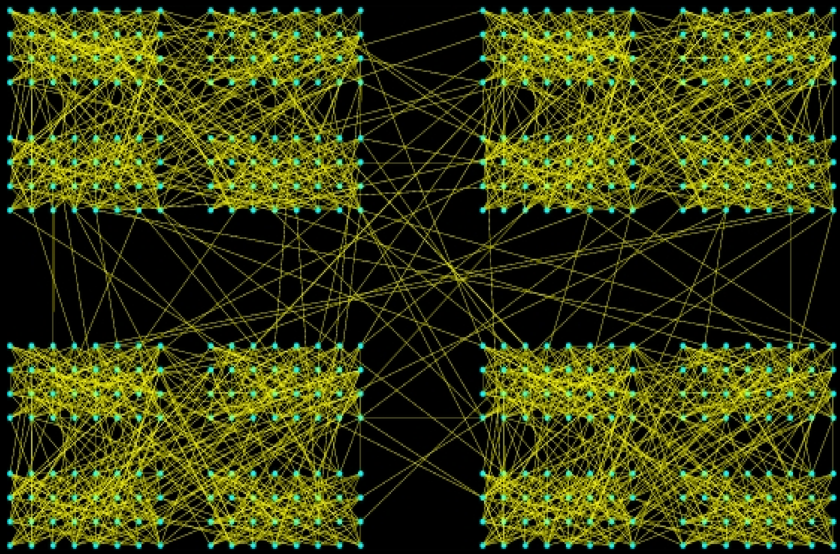
In real networks, vertices may belong to different modules



G. Palla, I. Derényi, I. Farkas, T. Vicsek, Nature 435, 814, 2005

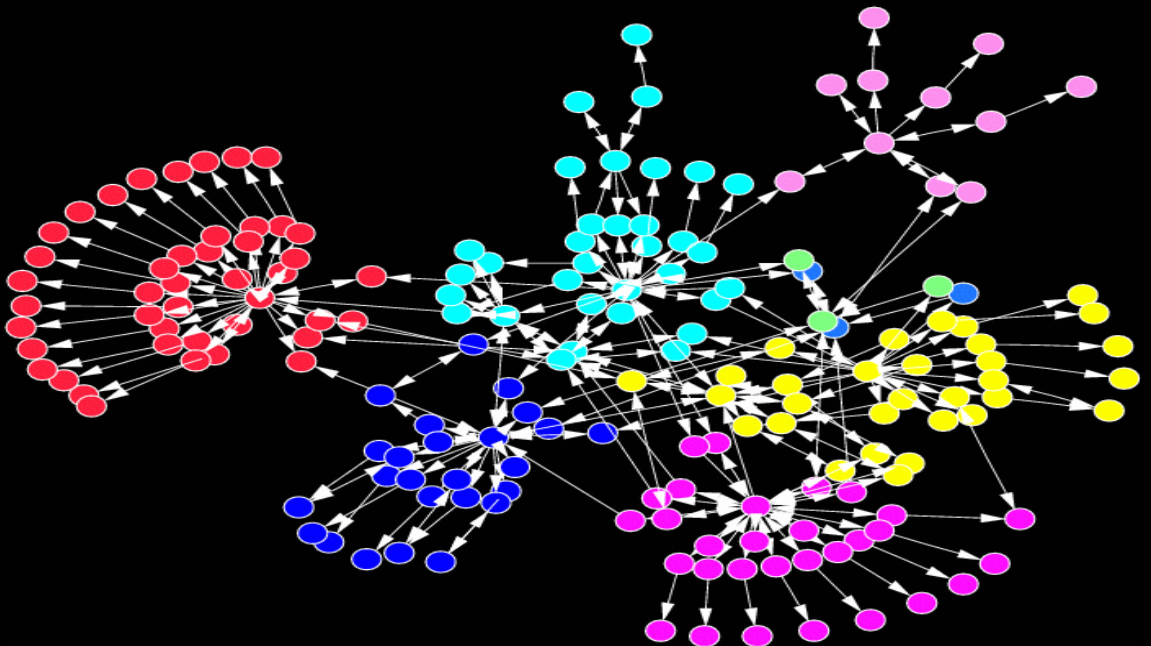
Hierarchies

Modules may embed smaller modules, yielding different organizational levels

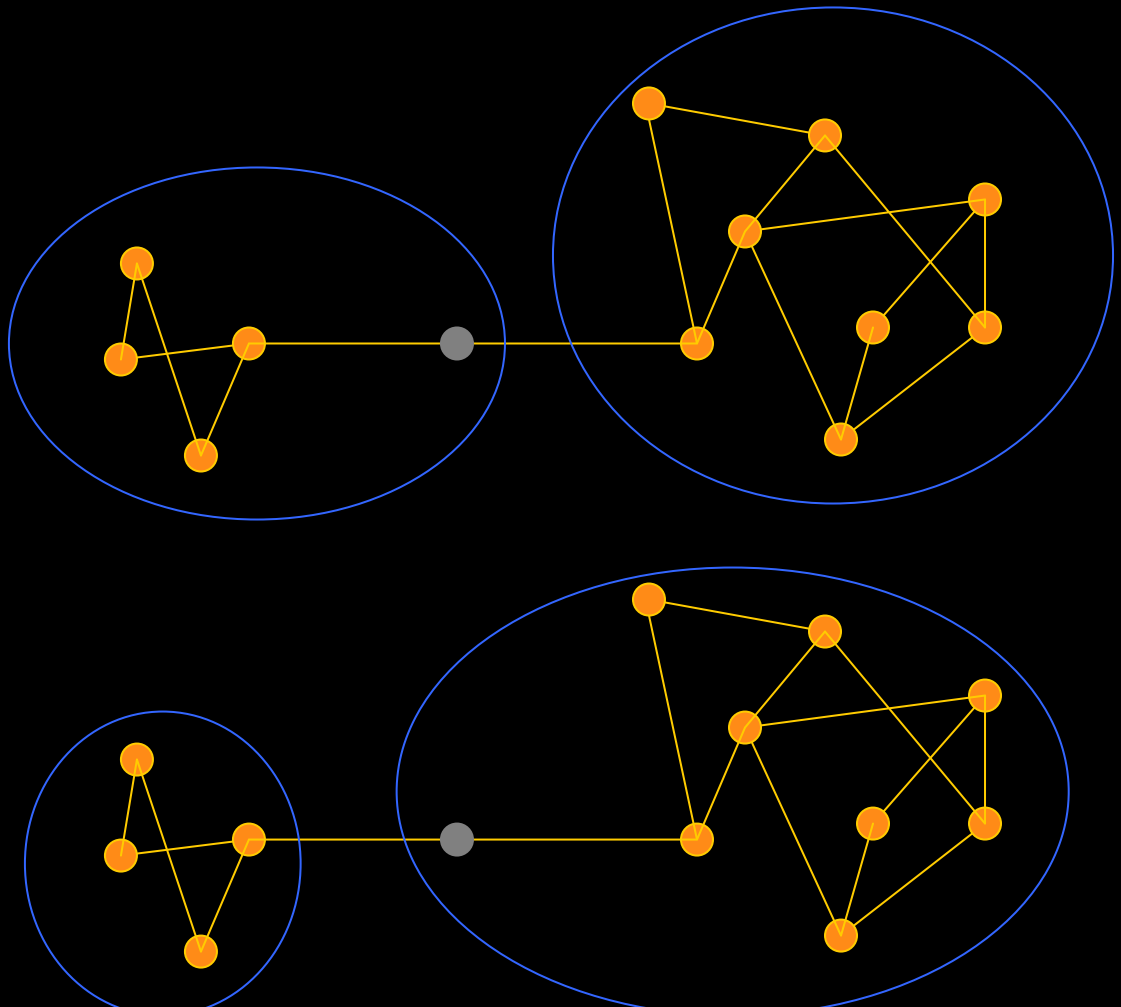


A. Clauset, C. Moore, M.E.J. Newman, *Nature* 453, 98, 2008

What is a “good” partition?



How can we compare different partitions?



?

Partition P_1 versus P_2: which one is better?

Quality function Q

Is $Q(P_1) > Q(P_2)$ or $Q(P_1) < Q(P_2)$?

Modularity

$$Q = \frac{1}{L}\sum_{i=1}^{n}\left(l_i - \frac{d_i^2}{4L}\right)$$

l_i = # links in module i

$\frac{d_i^2}{4L}$ = expected # of links in module i

History

- 1970s: Graph partitioning in computer science
- Hierarchical clustering in social sciences
- 2002: Girvan and Newman, PNAS 99, 7821-7826
- 2002-onward: methods of “new generation”, mostly by physicists

Graph partitioning

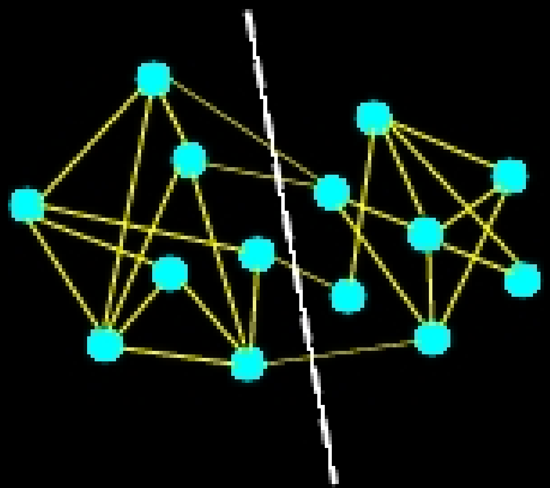
"Divide a graph in n parts, such that the number of links between them (*cut size*) is minimal"

Problems:

1. Number of clusters must be specified
2. Size of the clusters must be specified

If cluster sizes are not specified, the minimal cut size is zero, for a partition where all nodes stay in a single cluster and the other clusters are "empty"

Bipartition: divide a graph in two clusters of equal size and minimal cut size



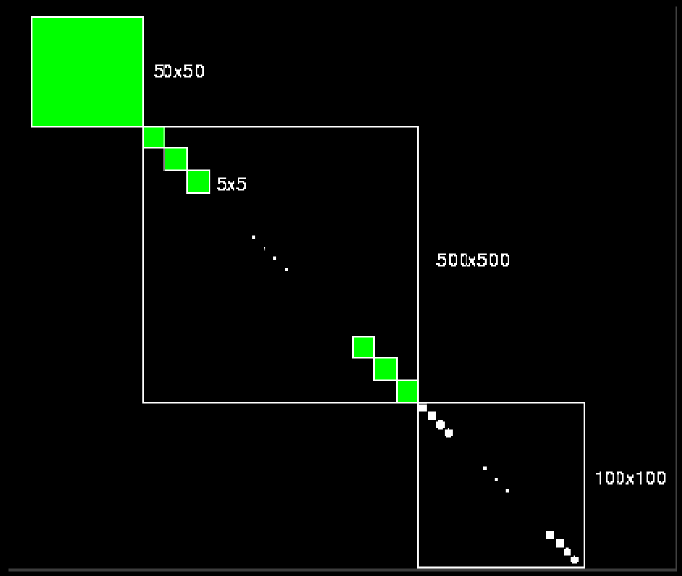
Spectral partitioning

Laplacian matrix L

$$L_{ij} = d_i \delta_{ij} - A_{ij}$$

Spectral properties of L:

- All eigenvalues are non-negative
- If the graph is divided in g components, there are g zero eigenvalues
- In this case L can be rewritten in a block-diagonal form



If the network is connected, but there are two groups of nodes weakly linked to each other, they can be identified from the eigenvector of the second smallest eigenvalue (*Fiedler vector*)

The Fiedler vector has both positive and negative components, their sum must be 0

If one wants a split into n_1 and $n_2=n-n_1$ nodes, one takes the n_1 largest (smallest) components of the Fiedler vector

Kernighan-Lin algorithm

Start: split in two groups

At each step, a pair of nodes of different groups are swapped so to decrease the cut size

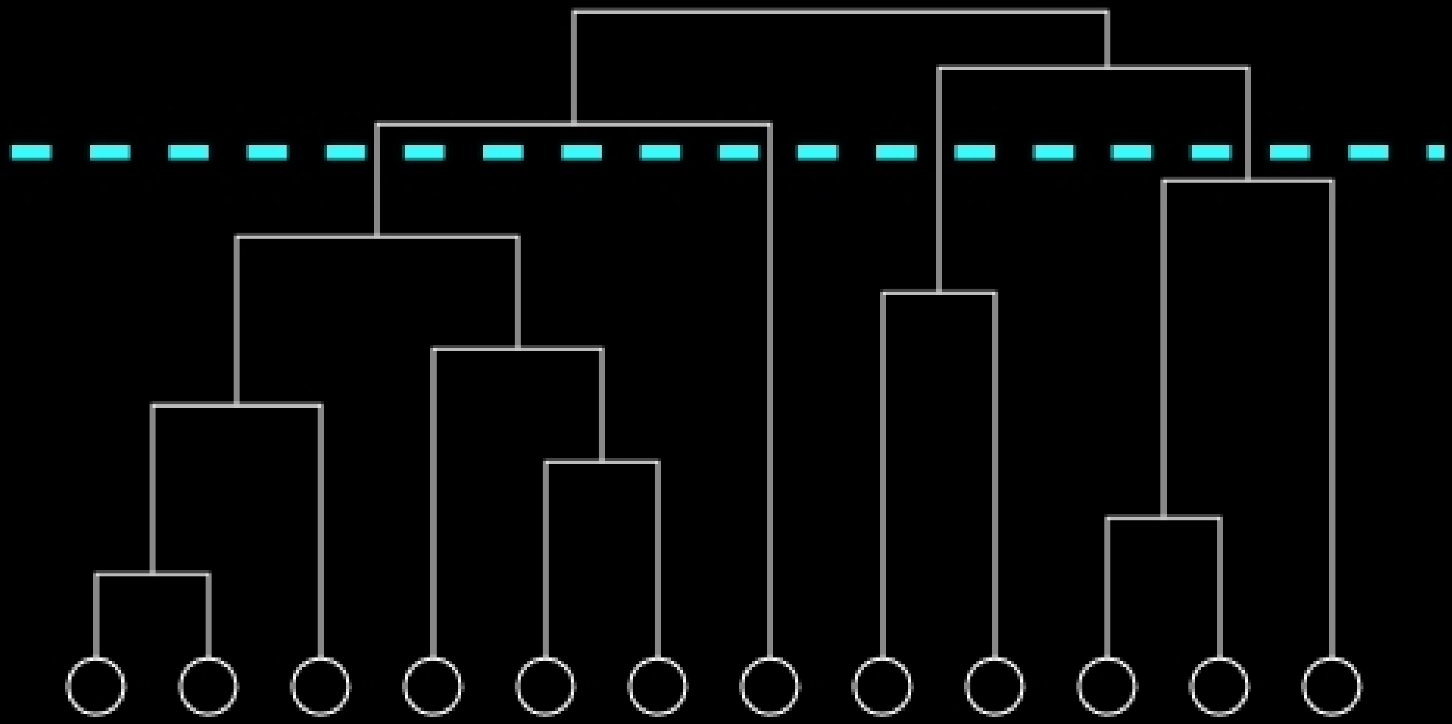
Sometimes swaps are allowed that increase the cut size, to avoid local minima

Hierarchical clustering

Very common in social network analysis

1. A criterion is introduced to compare nodes based on their similarity
2. A similarity matrix X is constructed: the similarity of nodes i and j is X_{ij}
3. Starting from the individual nodes, larger groups are built by joining groups of nodes based on their similarity

Final result: a hierarchy of partitions (dendrogram)



Problems of traditional methods

- Graph partitioning: one needs to specify the number and the size of the clusters
- Hierarchical clustering: many partitions recovered, which one is the best?

One would like a method that can predict the number and the size of the partition and indicate a subset of “good” partitions

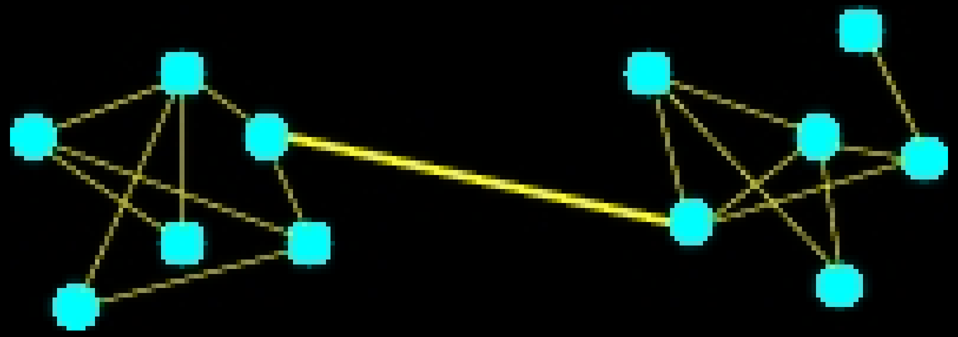
Girvan-Newman algorithm

M. Girvan & M.E.J Newman,
PNAS 99, 7821-7826 (2002)

Divisive method: one removes the links that connect the clusters, until the latter are isolated

How to identify intercommunity links?
Betweenness

Link-betweenness: number of shortest paths crossing a link



Steps

1. Calculate the betweenness of all links
2. Remove the one with highest betweenness
3. Recalculate the betweenness of the remaining edges
4. Repeat from 2

The process delivers a hierarchy of partitions: which one is the best?

The best partition is the one corresponding to the highest modularity Q

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

The algorithm runs in a time $O(n^3)$ on a sparse graph (i.e. when $m \sim n$)

New methods

- Divisive algorithms
- Modularity optimization
- Spectral methods
- Dynamics methods
- Clique percolation
- Statistical inference

Modularity optimization

$$Q = \frac{1}{L}\sum_{i=1}^{n}\left(l_i - \frac{d_i^2}{4L}\right)$$

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

Problem: NP-complete!

1) Greedy algorithms
2) Simulated annealing
3) Extremal optimization
4) Spectral optimization

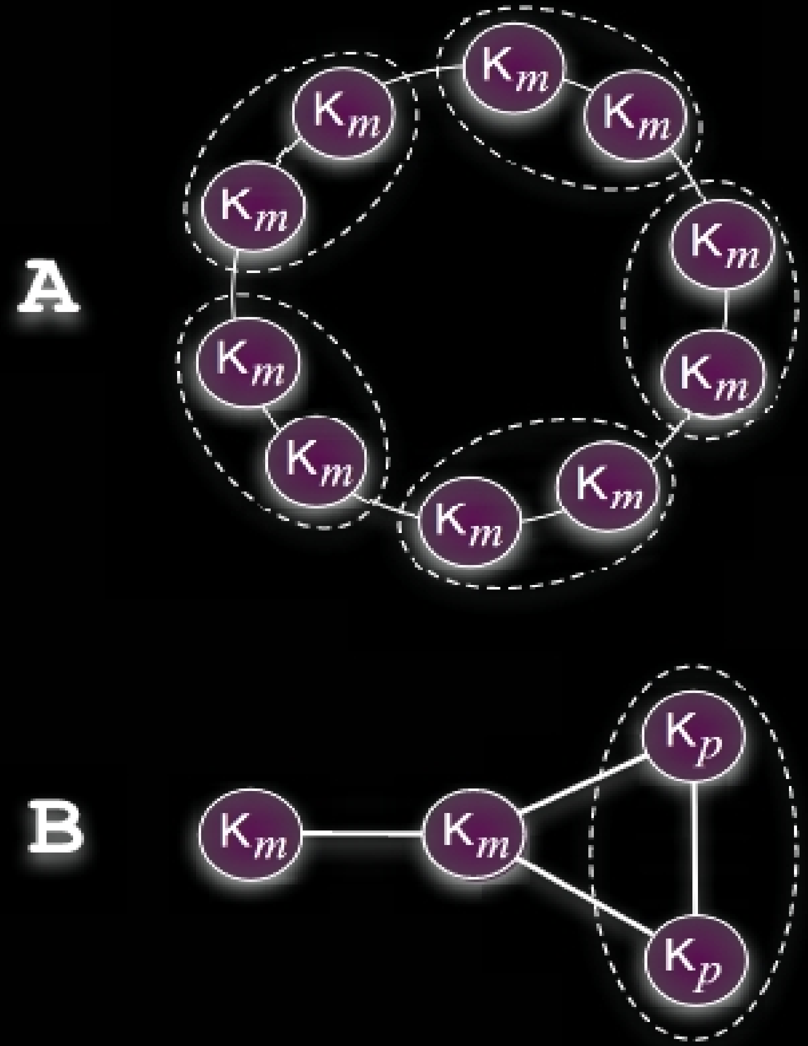
Greedy algorithm

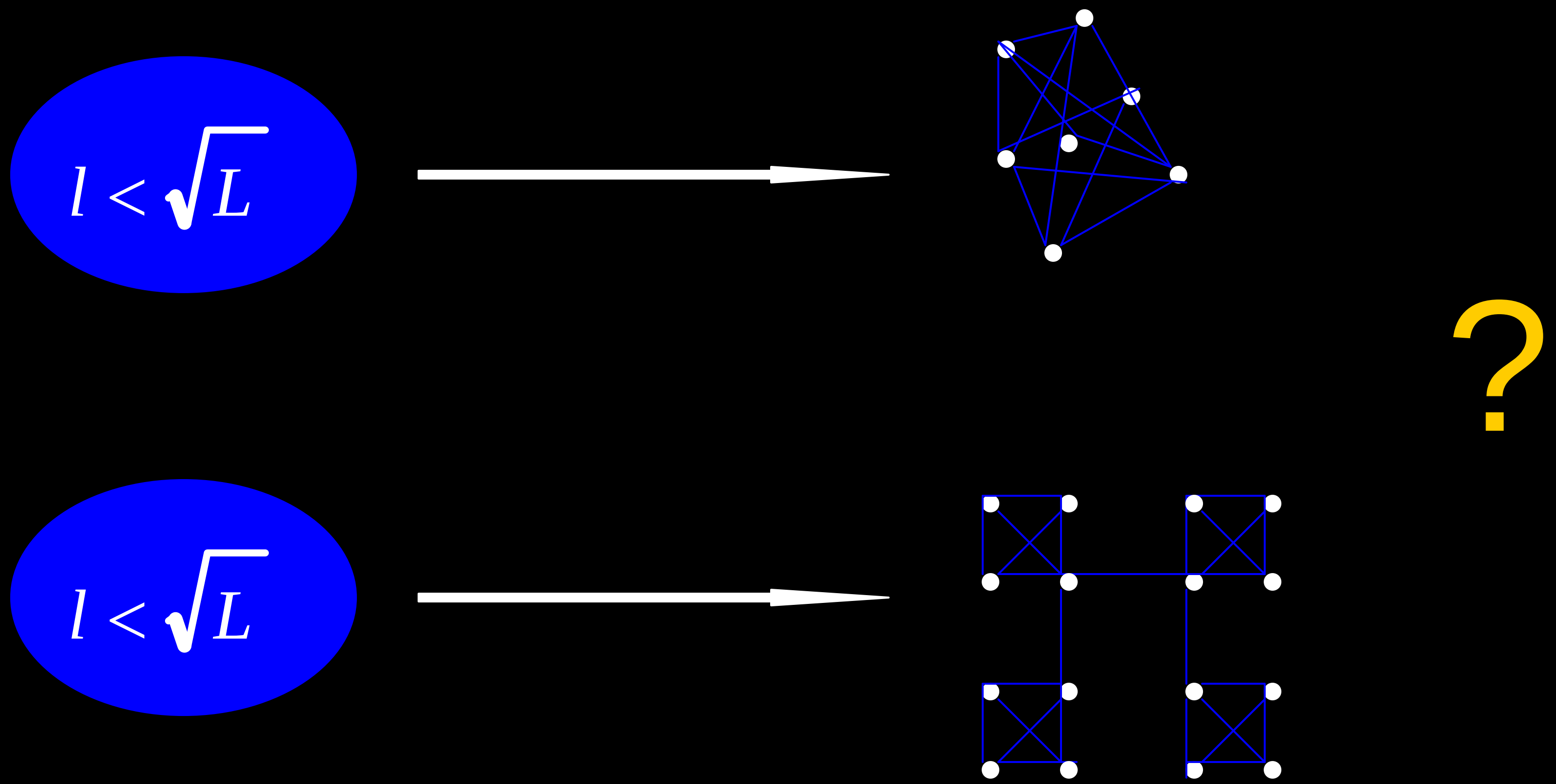
M.E.J. Newman,
Phys. Rev. E 69, 066133, 2004

- Start: partition with one node in each community
- Merge groups of nodes so to obtain the highest increase of Q
- Continue until all nodes are in the same community
- Pick the partition with largest modularity

CPU time $O(n^2)$

Resolution limit of modularity





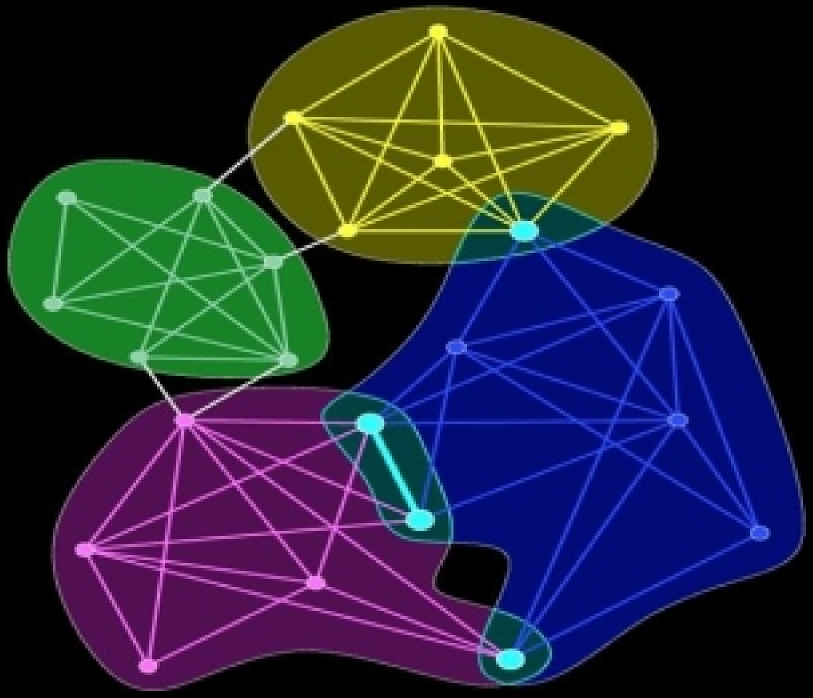
S.F. & M. Barthélemy, PNAS 104, 36 (2007)

Dynamic algorithms

- Potts model
- Synchronization
- Random walks

Clique percolation

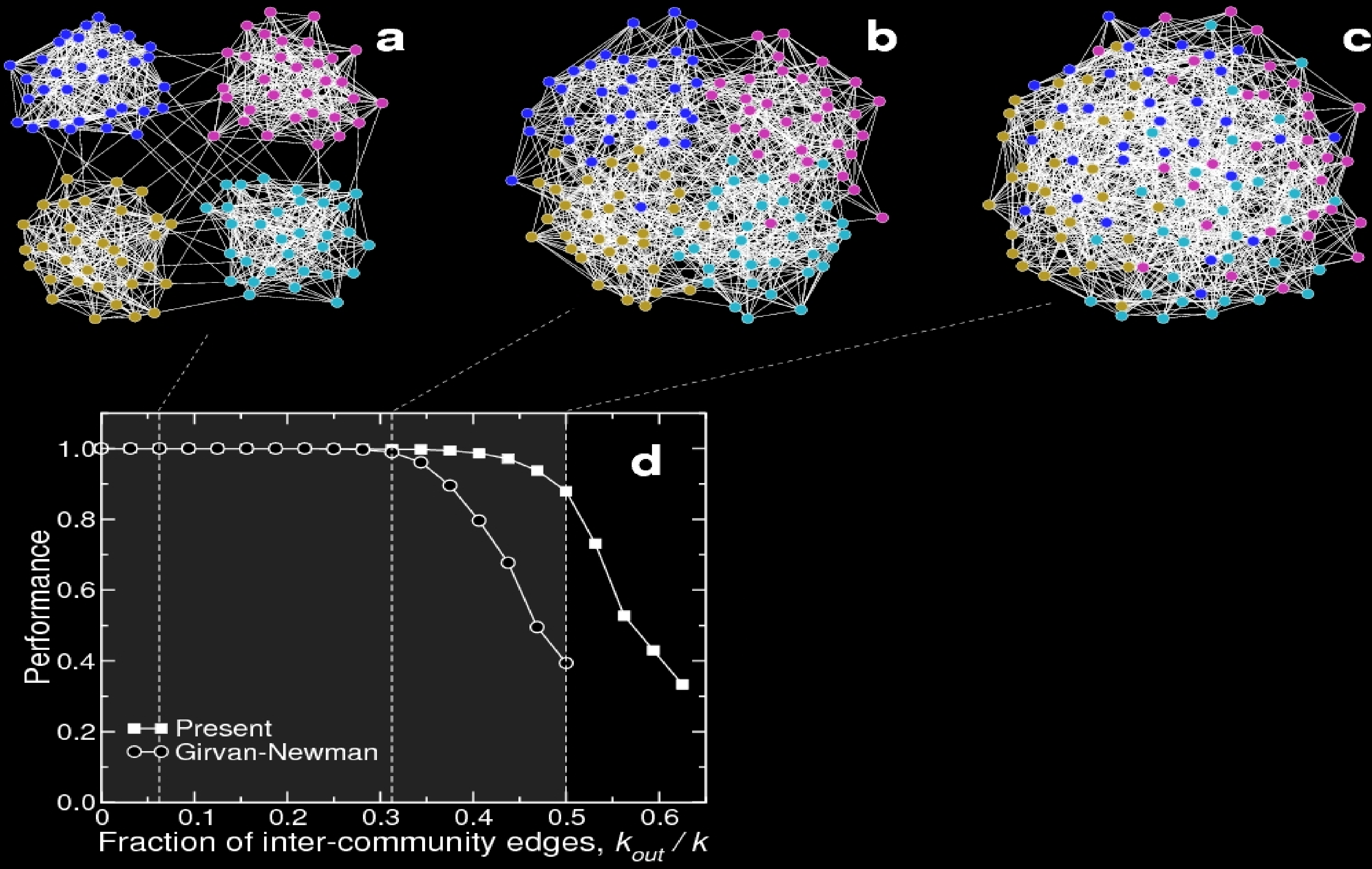
G. Palla, I. Derényi, I. Farkas, T. Vicsek,
Nature 435, 814, 2005



Testing algorithm

- Artificial networks
- Real networks with known community structure

Benchmark of Girvan & Newman



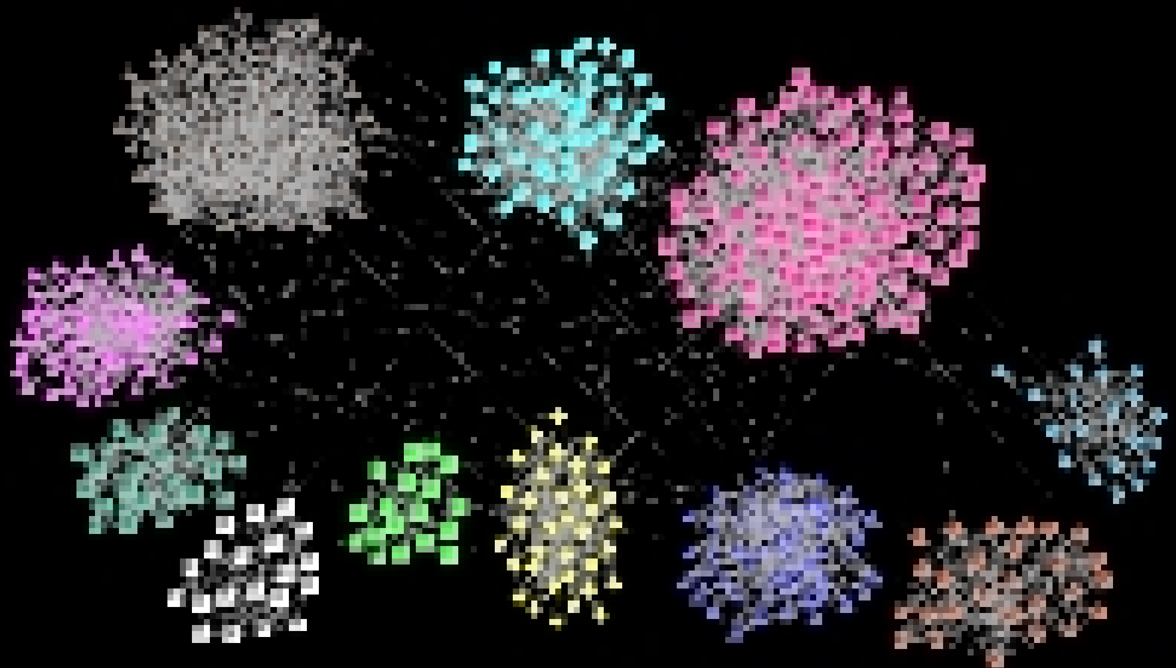
Problems

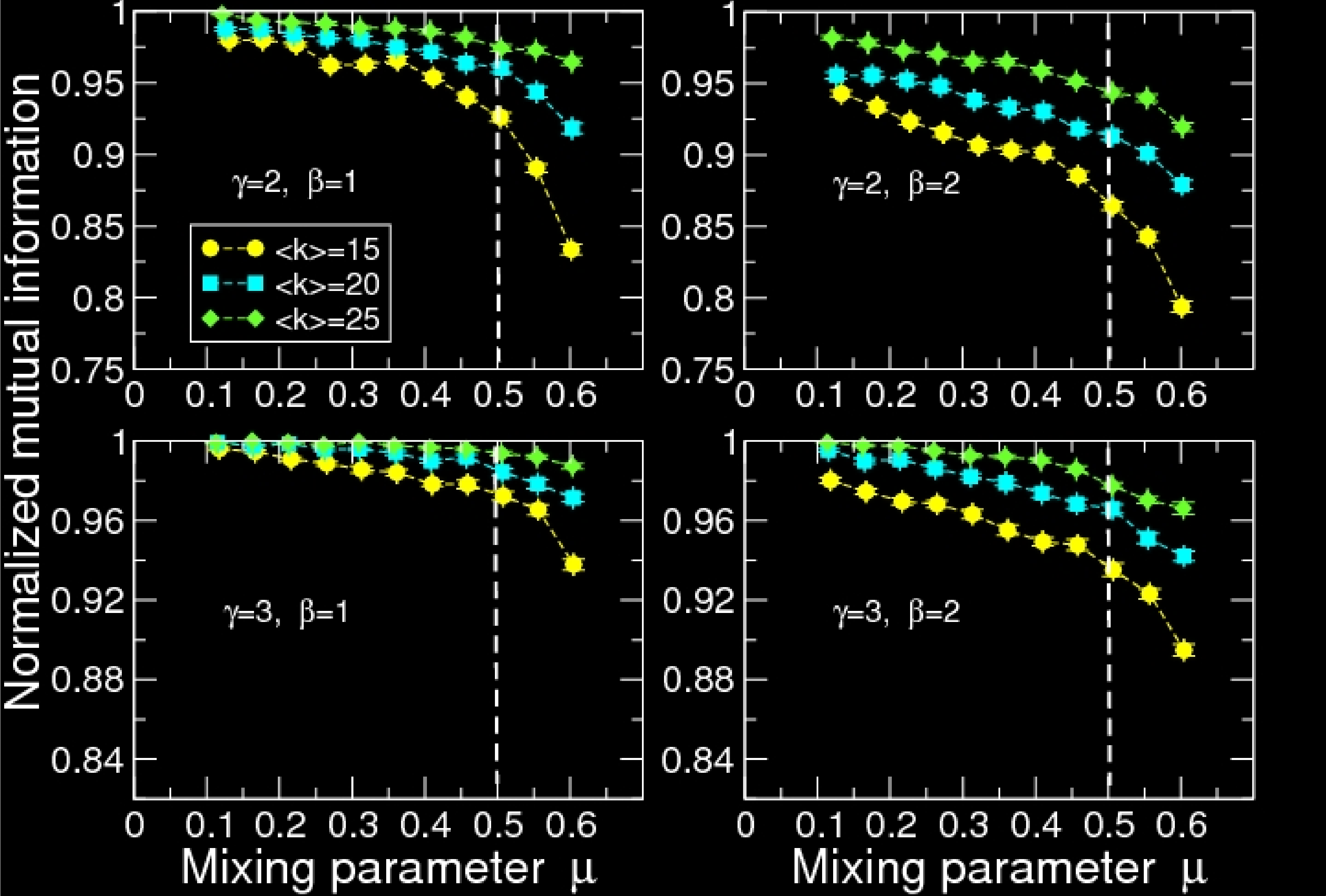
- All nodes have the same degree
- All communities have equal size

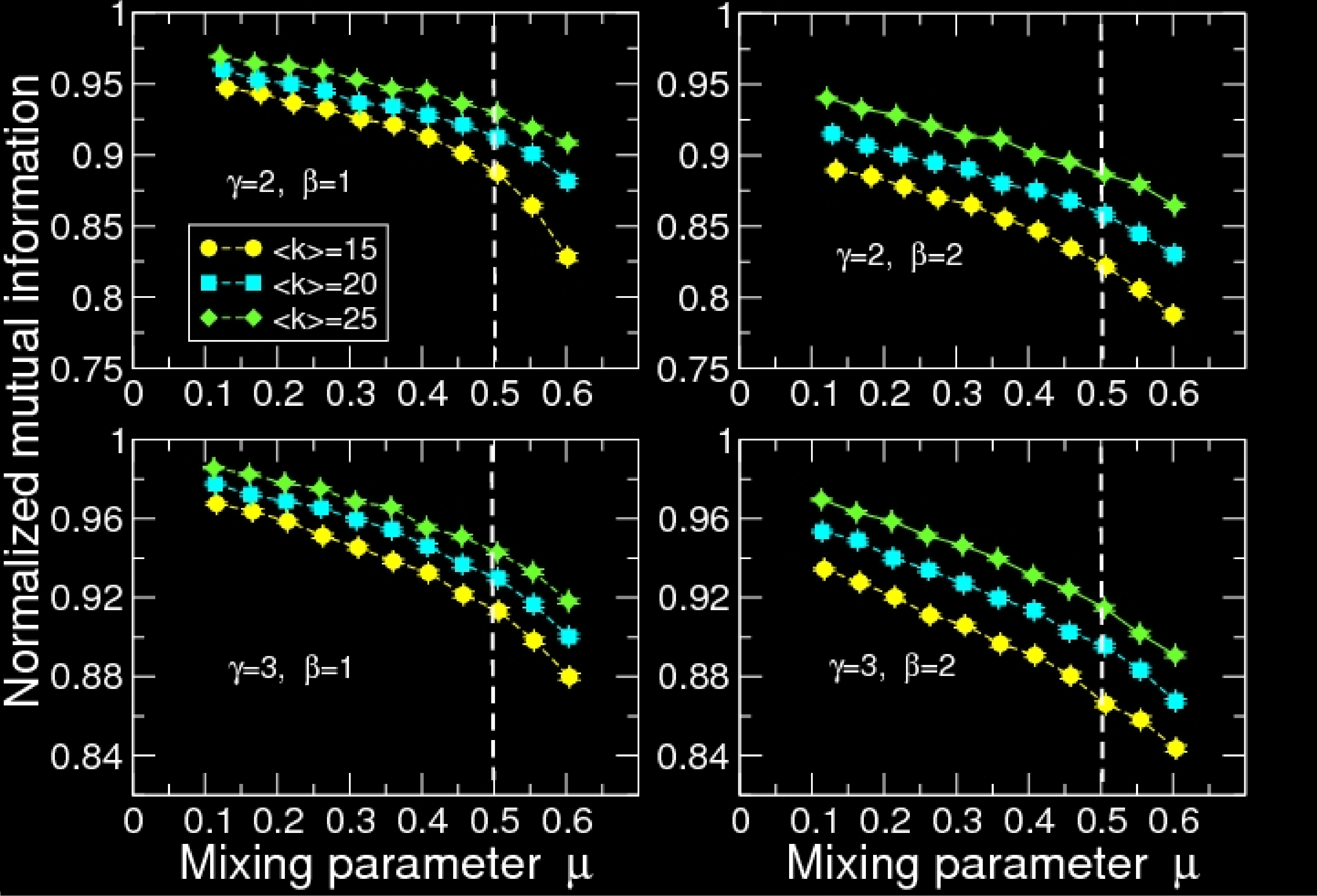
In real networks the distributions of degree and community size are highly heterogeneous!

New benchmark (A. Lancichinetti, S. F., F. Radicchi, arXiv:0805.4770)

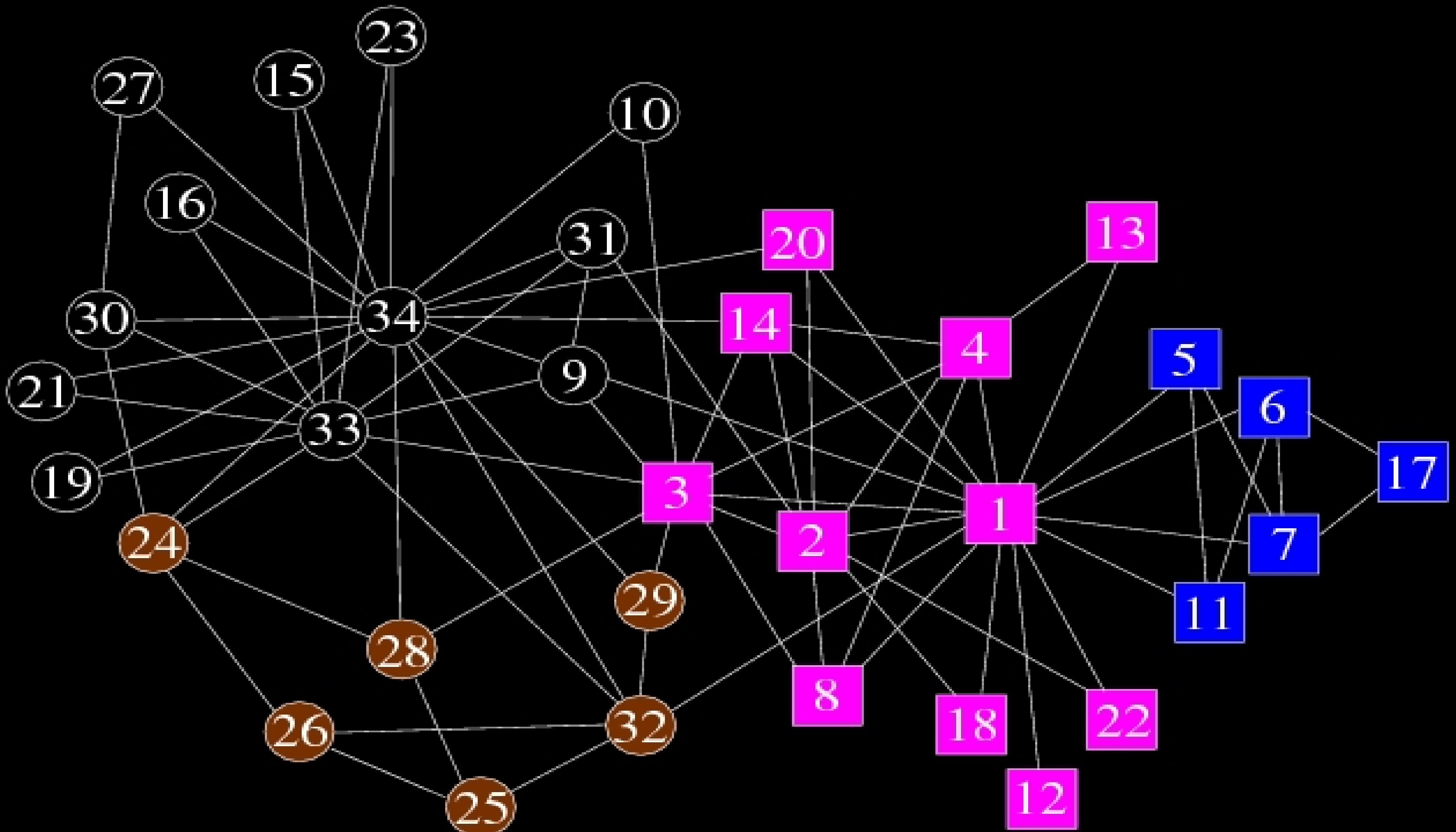
- Power law distribution of degree
- Power law distribution of community size
- A mixing parameter μ sets the ratio between the external and the total degree of each node




Normalized mutual information
Mixing parameter μ
Mixing parameter μ
γ=2, β=1
γ=2, β=2
γ=3, β=1
γ=3, β=2
<k>=15
<k>=20
<k>=25
1
0.95
0.9
0.85
0.8
0.75
0.96
0.92
0.88
0.84
0
0.1
0.2
0.3
0.4
0.5
0.6


Normalized mutual information
Mixing parameter μ
Mixing parameter μ
γ=2, β=1
γ=2, β=2
γ=3, β=1
γ=3, β=2
<k>=15
<k>=20
<k>=25
1
0.95
0.9
0.85
0.8
0.75
0.96
0.92
0.88
0.84
0
0.1
0.2
0.3
0.4
0.5
0.6

Real networks



Open problems

- Overlapping communities
- Hierarchies
- Directed graphs
- Weighted graphs
- Computational complexity
- Testing?

Community Structure in Graphs

Santo Fortunato[a], Claudio Castellano[b]

[a] Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, Torino, Italy
[b] SMC, INFM-CNR and Dipartimento di Fisica, "Sapienza" Università di Roma, P. le A. Moro 2, 00185 Roma, Italy

Abstract

Graph vertices are often organized into groups that seem to live fairly independently of the rest of the graph, with which they share but a few edges, whereas the relationships between group members are stronger, as shown by the large number of mutual connections. Such groups of vertices, or communities, can be considered as independent compartments of a graph. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. The task is very hard, though, both conceptually, due to the ambiguity in the definition of community and in the discrimination of different partitions and practically, because algorithms must find "good" partitions among an exponentially large number of them. Other complications are represented by the possible occurrence of hierarchies, i.e. communities which are nested inside larger communities, and by the existence of overlaps between communities, due to the presence of nodes belonging to more groups. All these aspects are dealt with in some detail and many methods are described, from traditional approaches used in computer science and sociology to recent techniques developed mostly within statistical physics.

1 Introduction

The origin of graph theory dates back to Euler's solution [1] of the puzzle of Königsberg's bridges in 1736. Since then a lot has been learned about graphs and their mathematical properties [2]. In the 20th century they have also become extremely useful as representation of a wide variety of systems in different areas. Biological, social, technological, and information networks can be studied as graphs, and graph analysis has become crucial to understand the features of these systems. For instance, social network analysis started in the 1930's and has become one of the most important topics in sociology [3, 4]. In recent times, the computer revolution has provided scholars with a huge amount of data and computational resources to process and analyse these data. The size of real networks one can potentially handle has also grown considerably, reaching

S. F., C. Castellano, arXiv:0712.2716