A new fast algorithm to detect communities in networks

Santo Fortunato
Andrea Lancichinetti
Janos Kertész

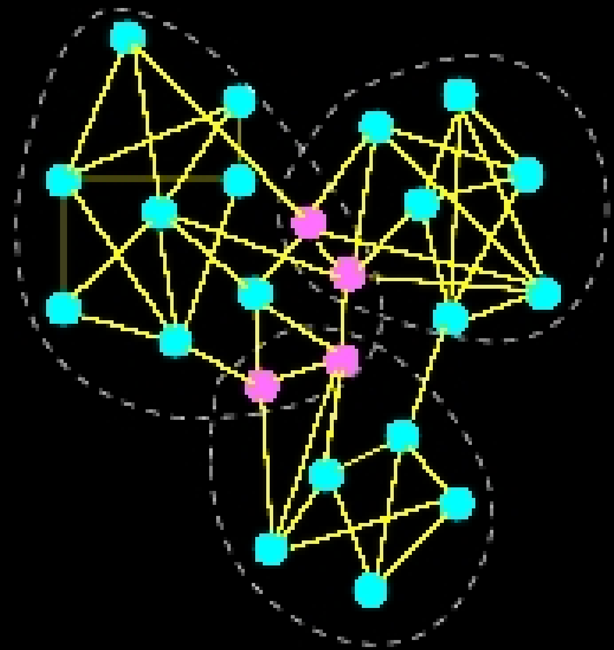
INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

Limits of current methods

- Overlapping communities
- Hierarchies
- Computer time

Overlapping communities

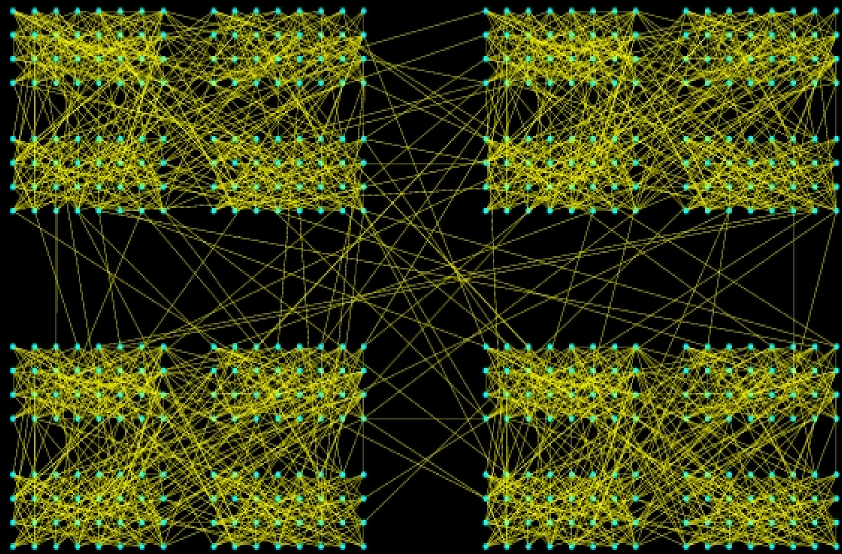
In real networks, vertices may belong to different modules



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, LNCS 4503, 1, 2007

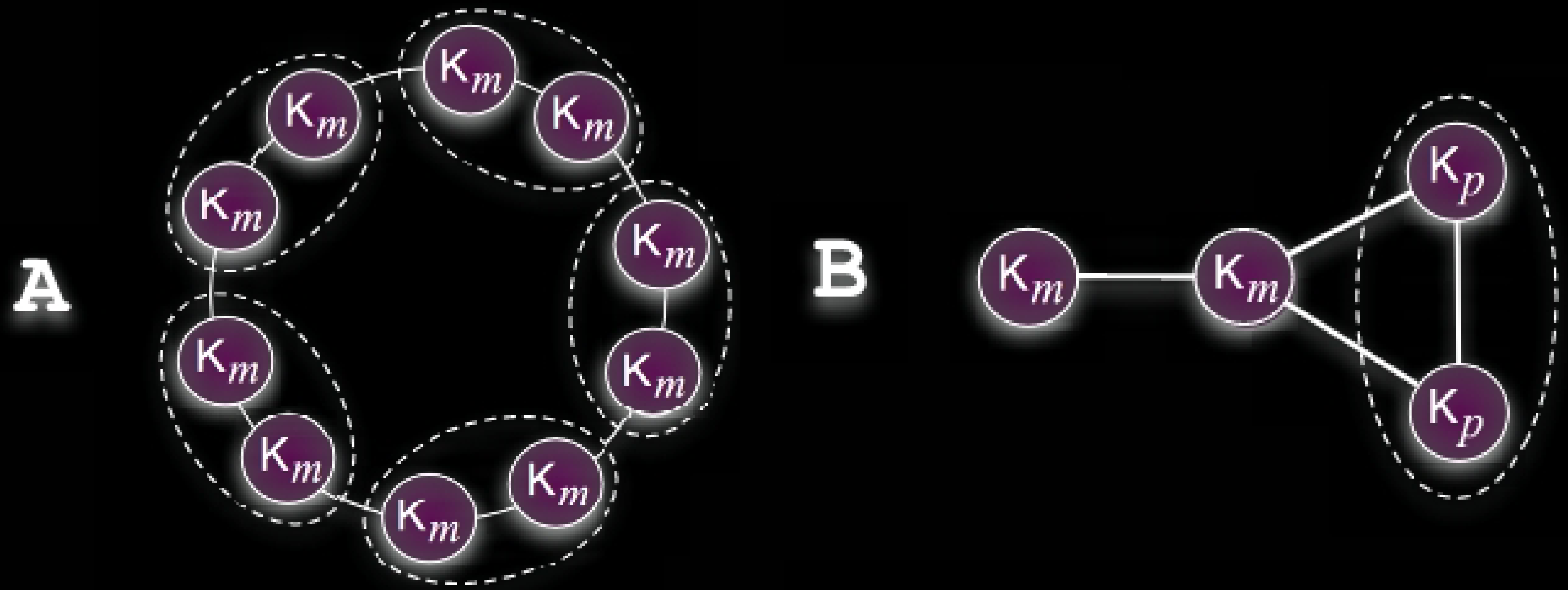
Computer time

Good algorithms run in a time $O(n^2)$

Some methods run in almost linear time!

- Greedy modularity optimization (Clauset, Newman, Moore, PRE 70, 066111, 2004)
- Wu-Huberman method (EPJB 38, 331, 2004)

The resolution limit of modularity optimization



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

Goal

Designing a FAST algorithm that accounts both for overlapping communities AND for hierarchies

Global or local?

“Global” community: a cluster of nodes with some property relative to the whole network

“Local” community: a cluster of nodes with a property relative to the nodes themselves and (possibly) their neighbors

The method

Basic rule: finding local communities about individual nodes

A local community is built by maximizing a *fitness function*

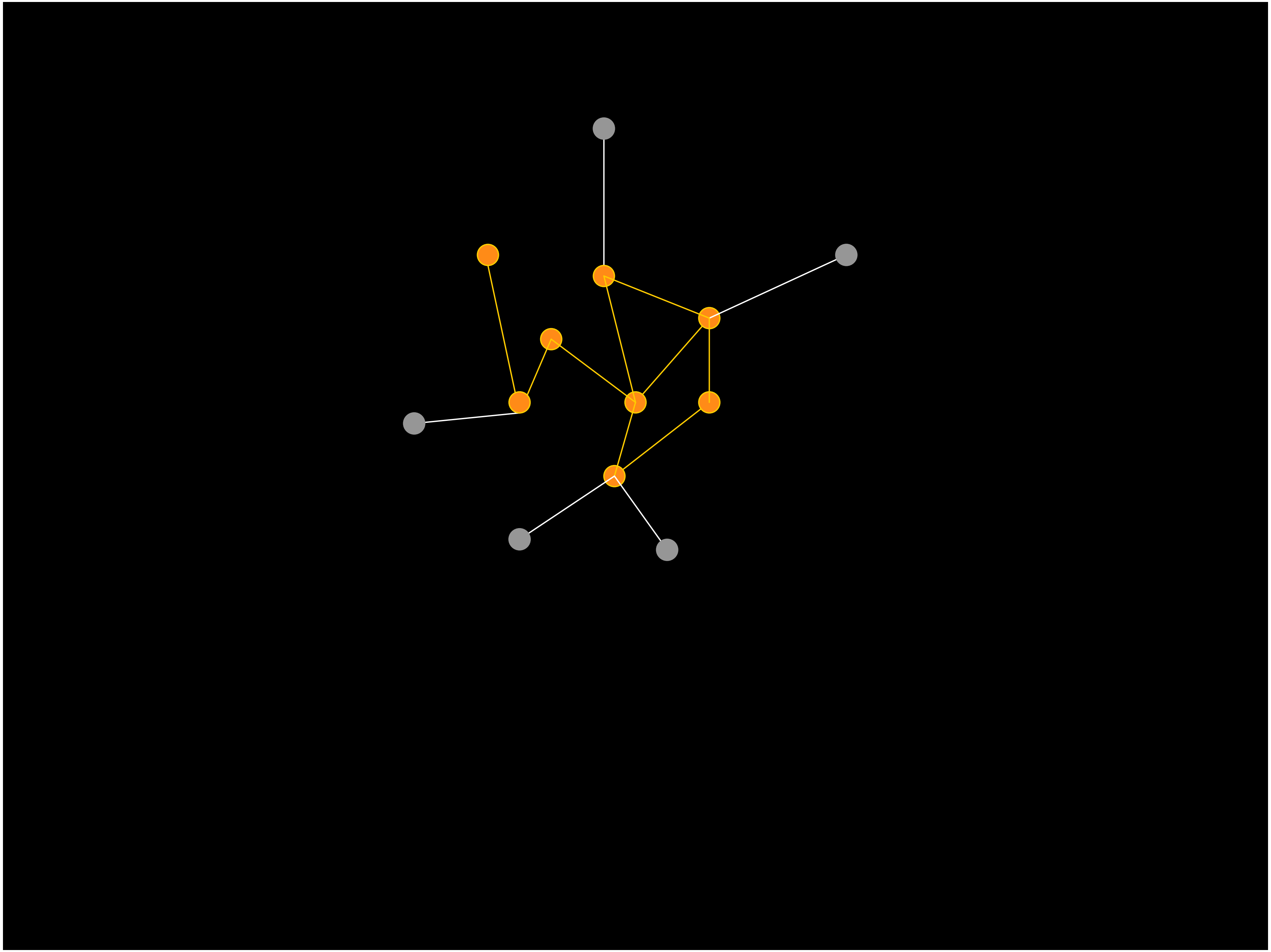
The fitness function depends on a parameter that tunes the size of the communities

The fitness function

Several options

$$f_i = \frac{k_{in}^i}{(k_{in}^i + k_{out}^i)^\alpha}$$

Resolution parameter **α>0**



Node fitness

Node A, cluster i

$$f_i^A = f_{i \cup A} - f_{i - A}$$

Positive fitness if the fitness of cluster i increases due to the addition of node A

Steps of the algorithm

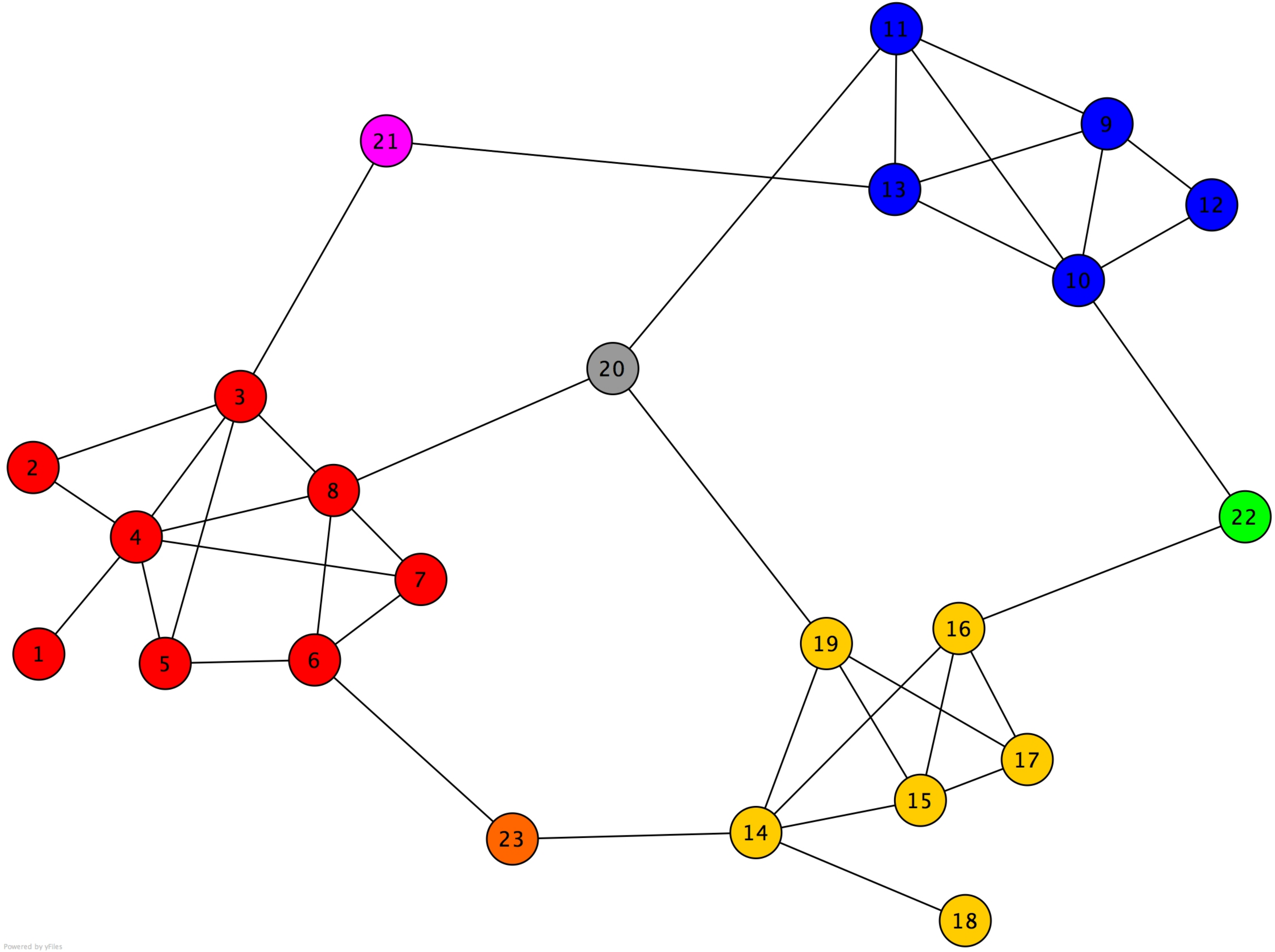
α is fixed

1. Take a node A at random
2. Look for community of A
3. Pick a node B at random not yet assigned to a community; the community of node B *may overlap with the others*
4. Repeat from 2

Building a node's community

Cluster with s nodes

- The neighboring node with the largest (positive) fitness is added to the group
- If a node is added, the fitness of all nodes of the group is recalculated
- Nodes with negative fitness are removed
- The process is repeated until all neighboring nodes have negative fitness (maximal cluster)


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

Computer time

The time to “close” a community with s nodes goes (about) as $O(s^2)$

The average CPU time is of the order of $O(ns_{Max})$

The worst-case time scales as $O(n^2)$

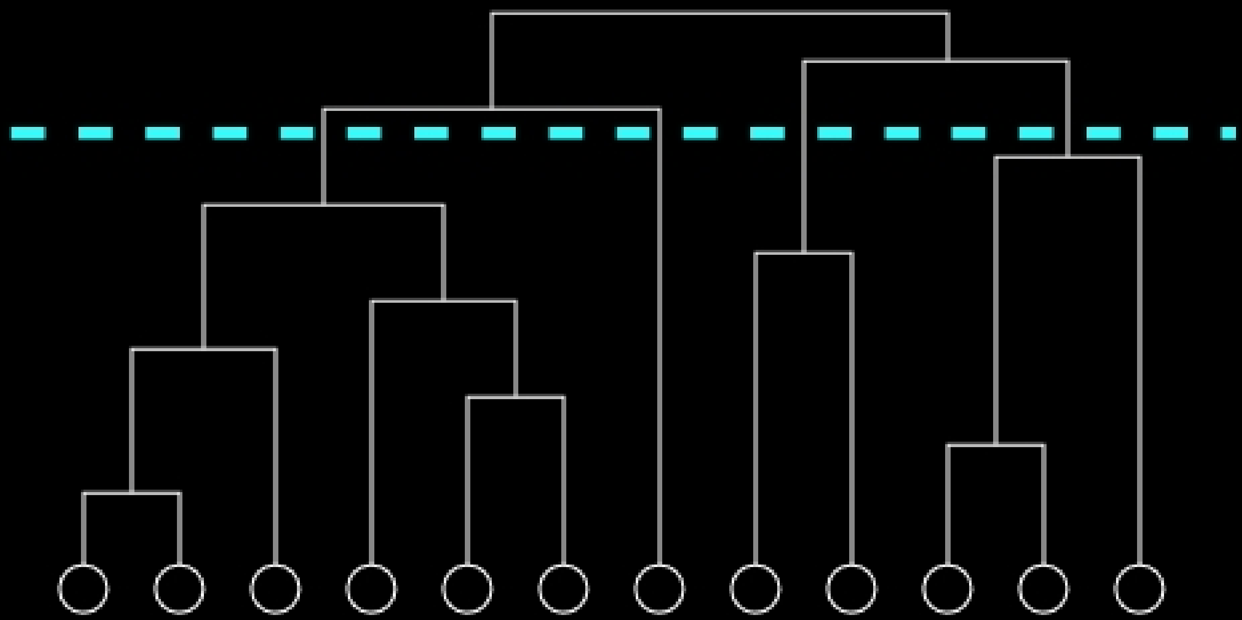
Resolution & hierarchies

Different values of the resolution parameter α yield partitions with different cluster sizes

α small → large communities

α large → small communities

By varying α hierarchical structure can be recovered



For hierarchical networks, the depth of the dendrogram varies as log n → the number of α-values is of the order of log n

Quality of partitions

The method delivers many partitions: which one(s) is the best?

Answer: the best partition is the *most stable* **in the range of α**

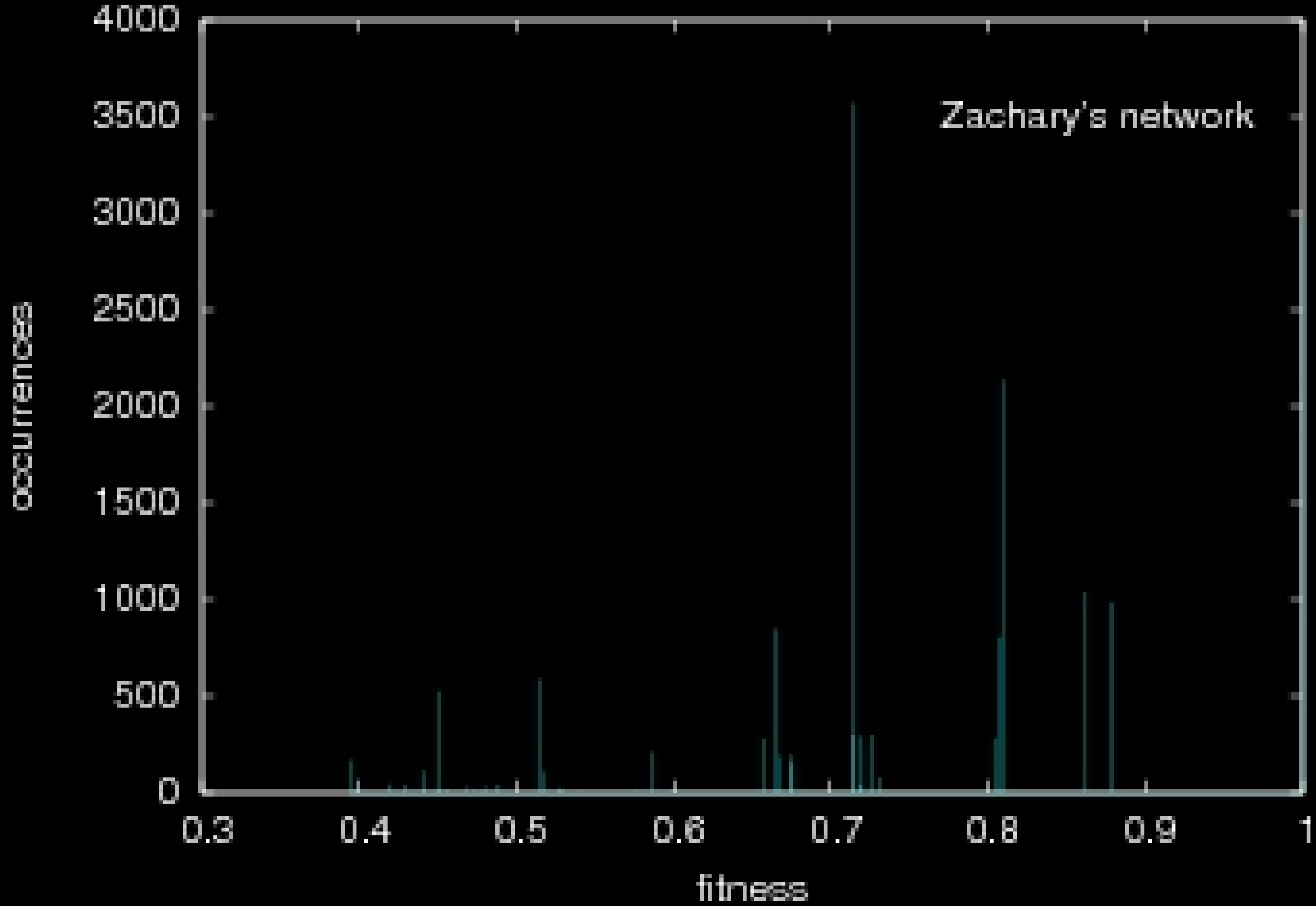
$$F(\alpha = 1) = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{k_{in}^i}{k_{in}^i + k_{out}^i}$$

Stable partitions are recovered in large ranges of α

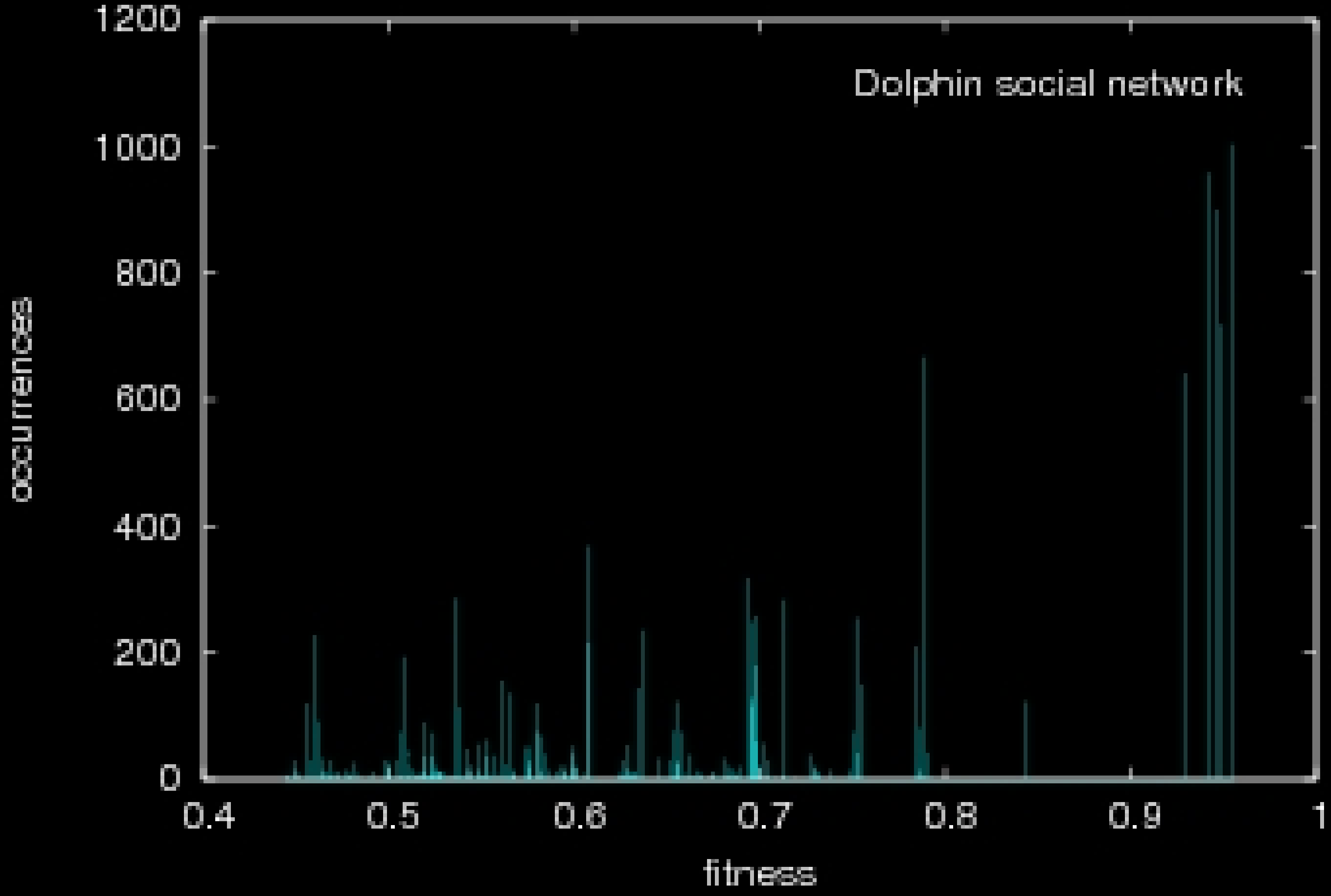
Recipe

- The “best” partition corresponds to the one found for the largest number of α-values!
- This can be revealed by plotting the histogram of the fitness values of all partitions found: peaks of the histogram correspond to stable partitions (“community spectroscopy”)

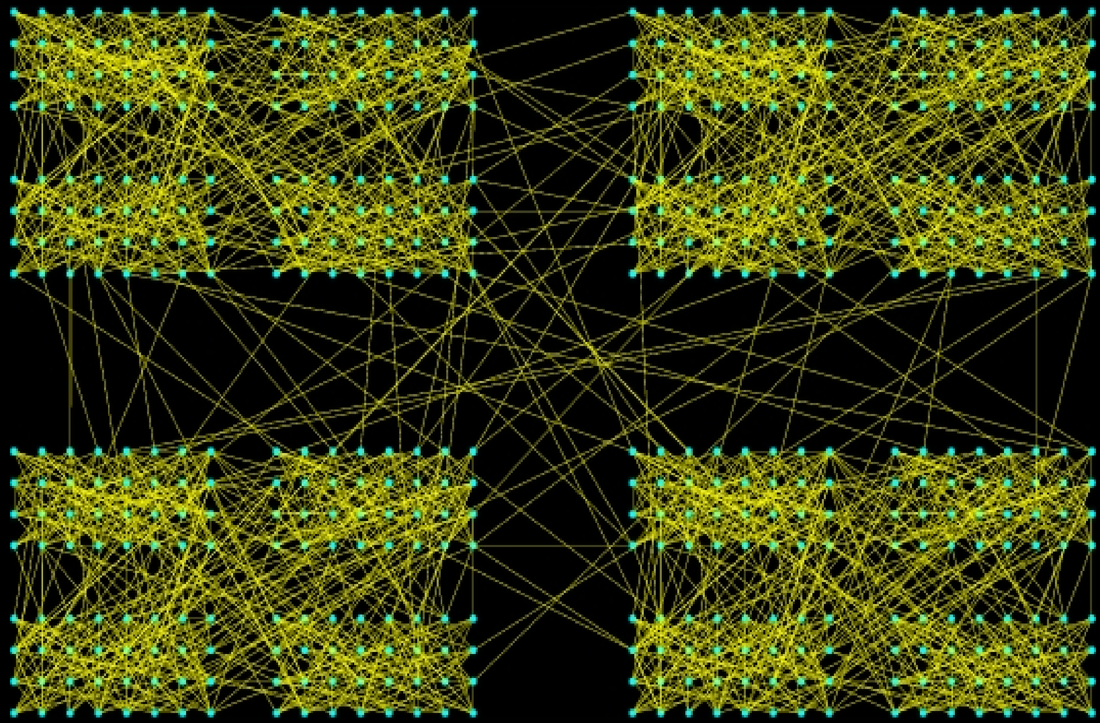
Zachary's karate club



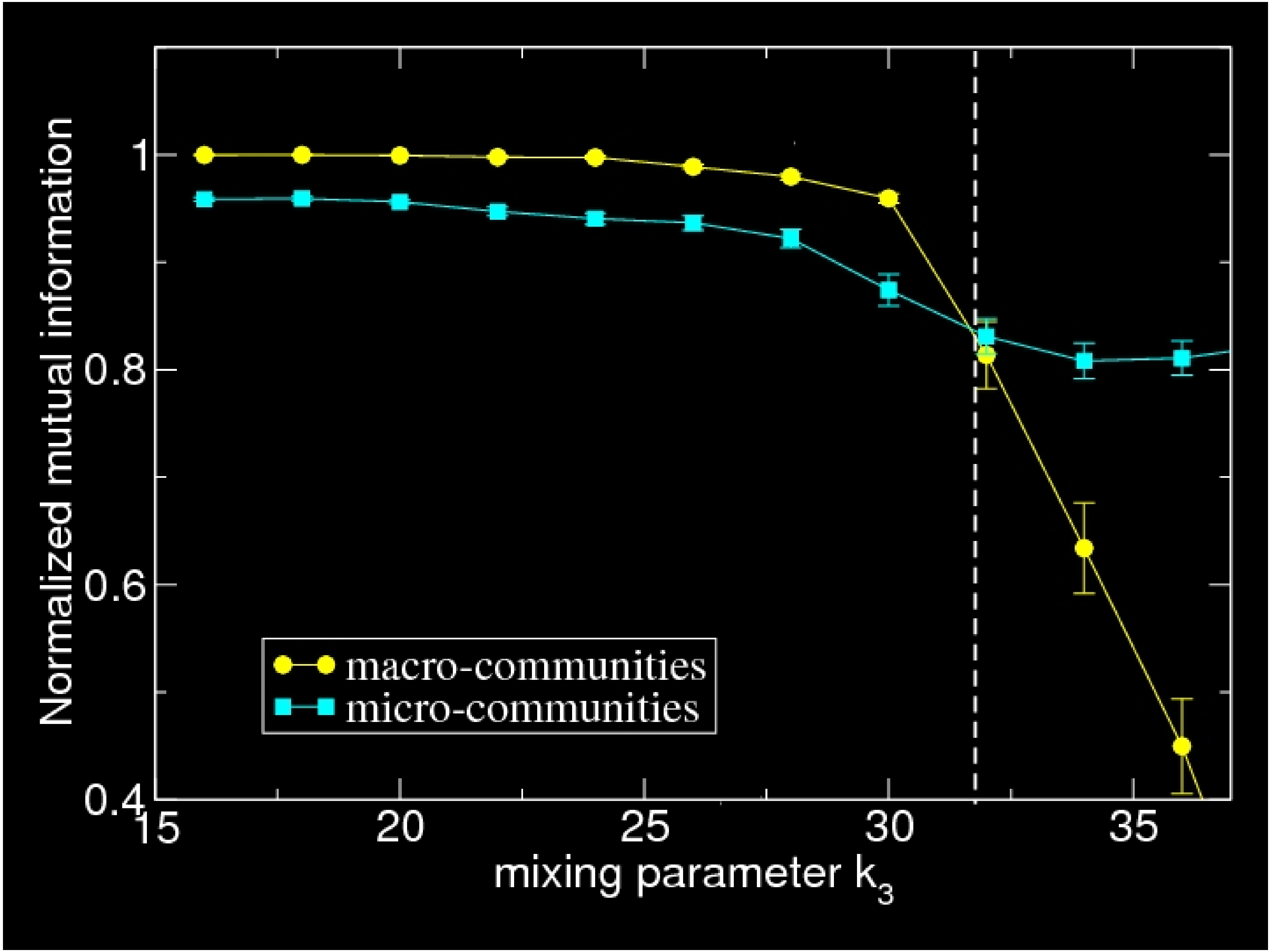
Dolphins’ network



Hierarchical benchmark

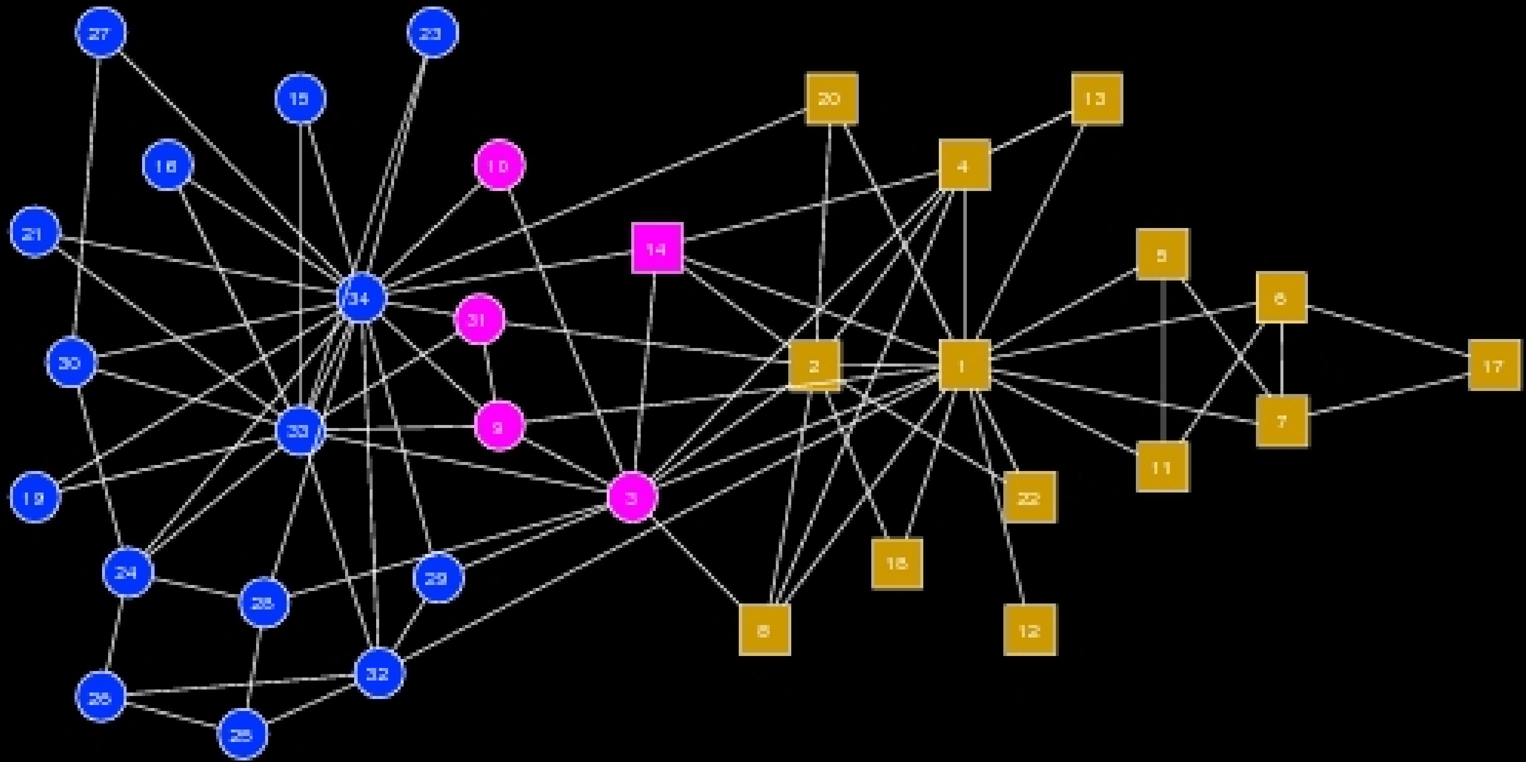


Two levels: 4 communities of 128 nodes, each including 4 communities of 32

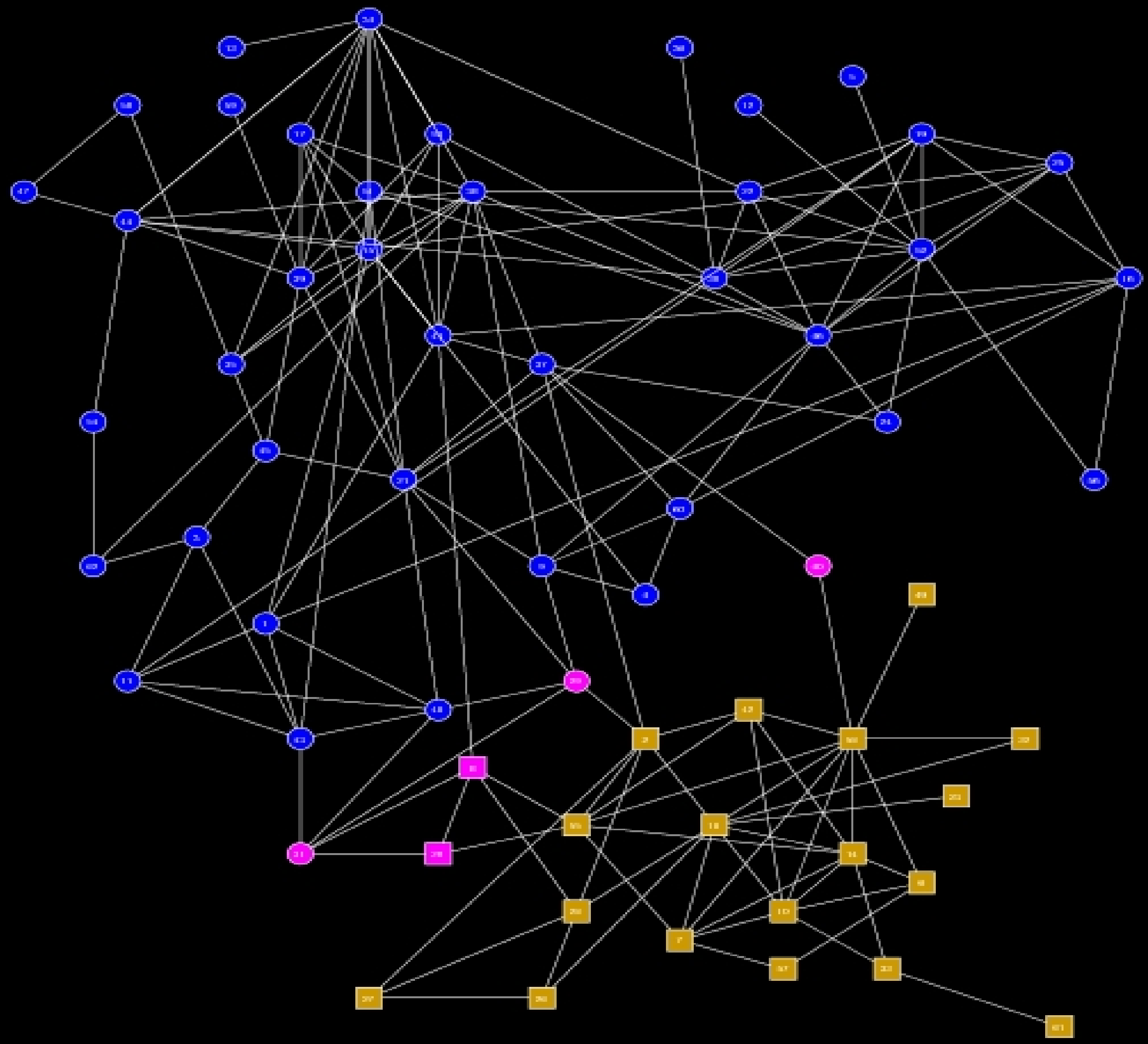


Normalized mutual information
1
0.8
0.6
0.4
15
20
25
30
35
mixing parameter k_3
macro-communities
micro-communities

Zachary's karate club

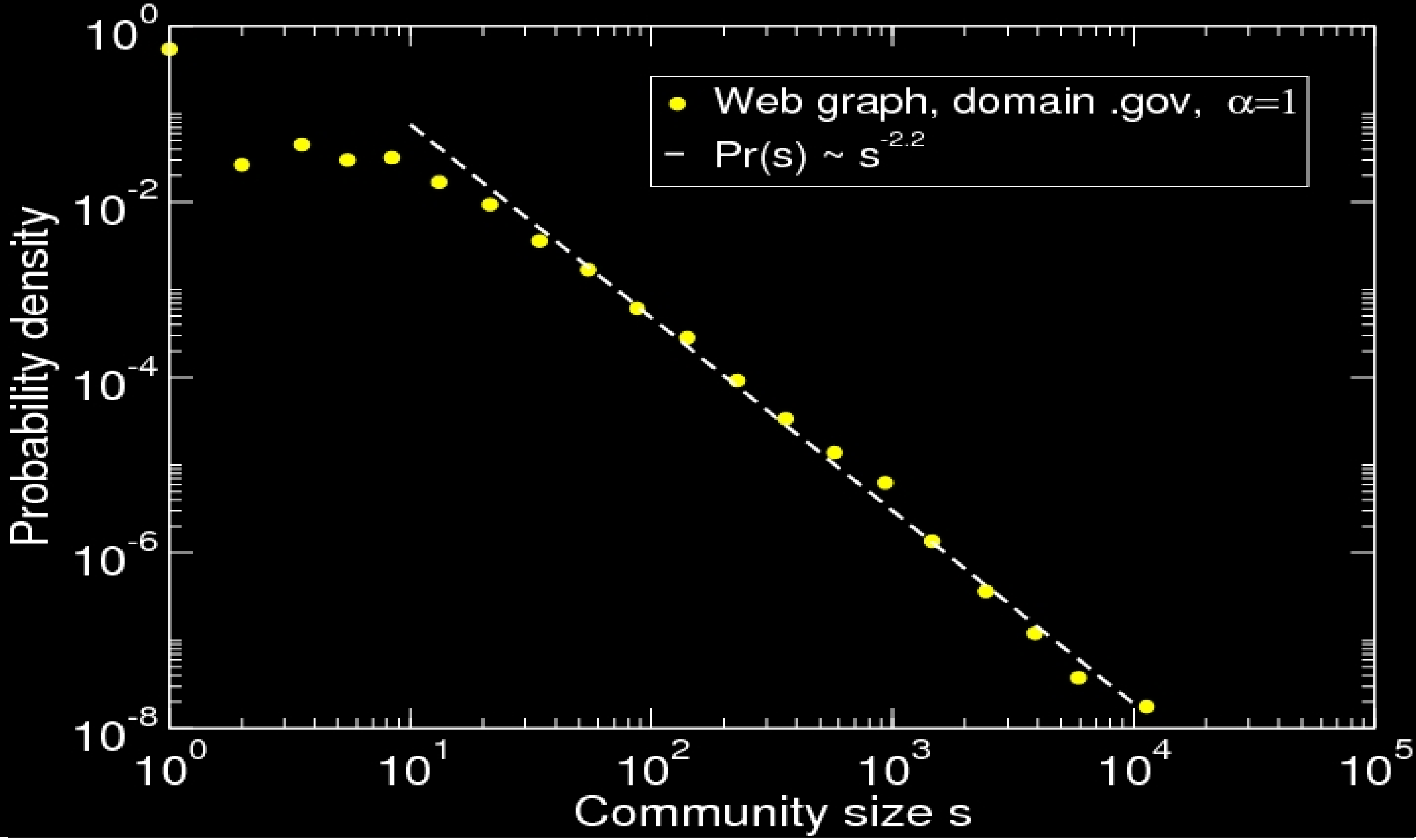


Lusseau’s dolphins network



Web graph: domain .gov

774908 URLs, 4711340 links



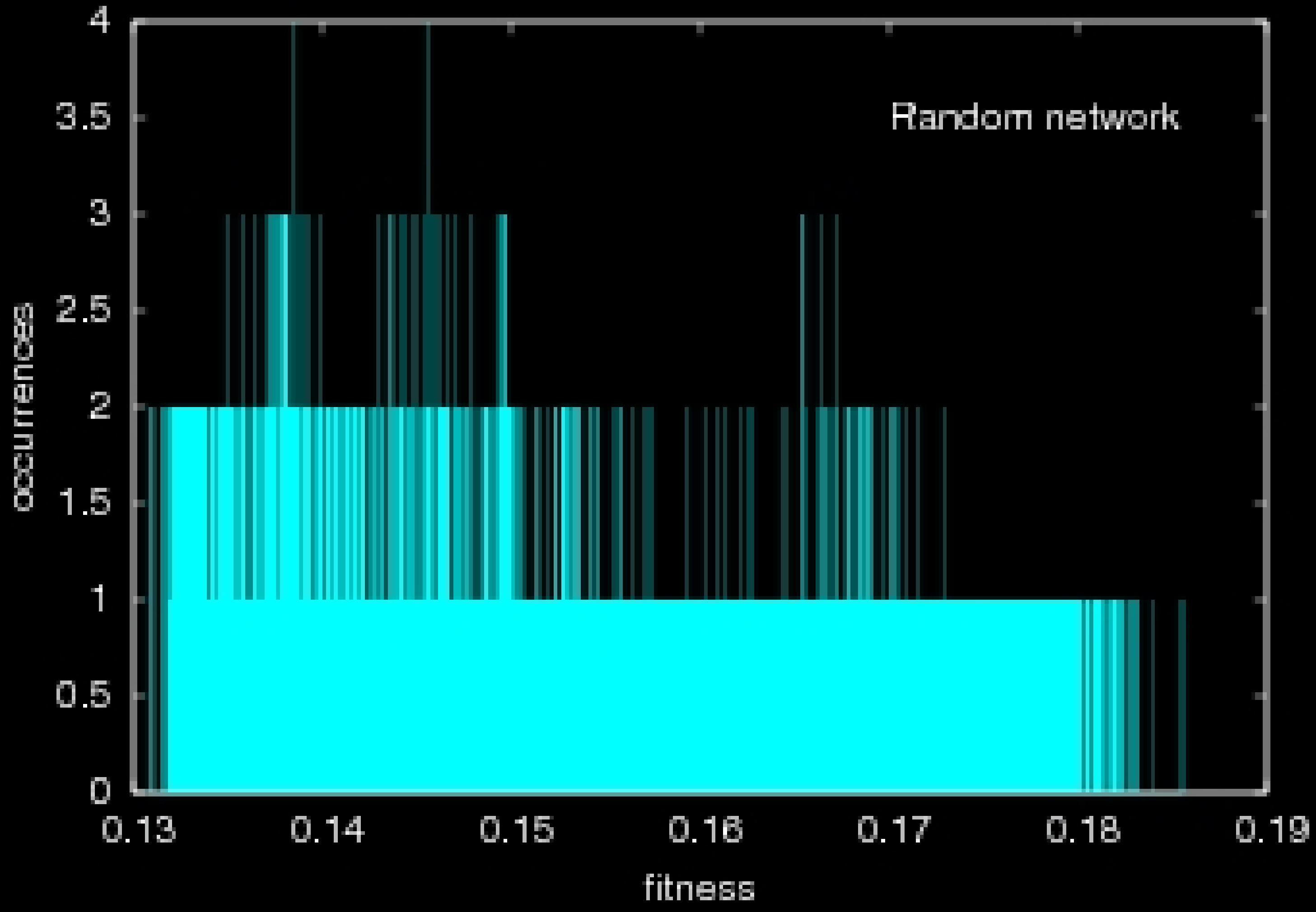
Significance of community structure

Not all networks have community structure!

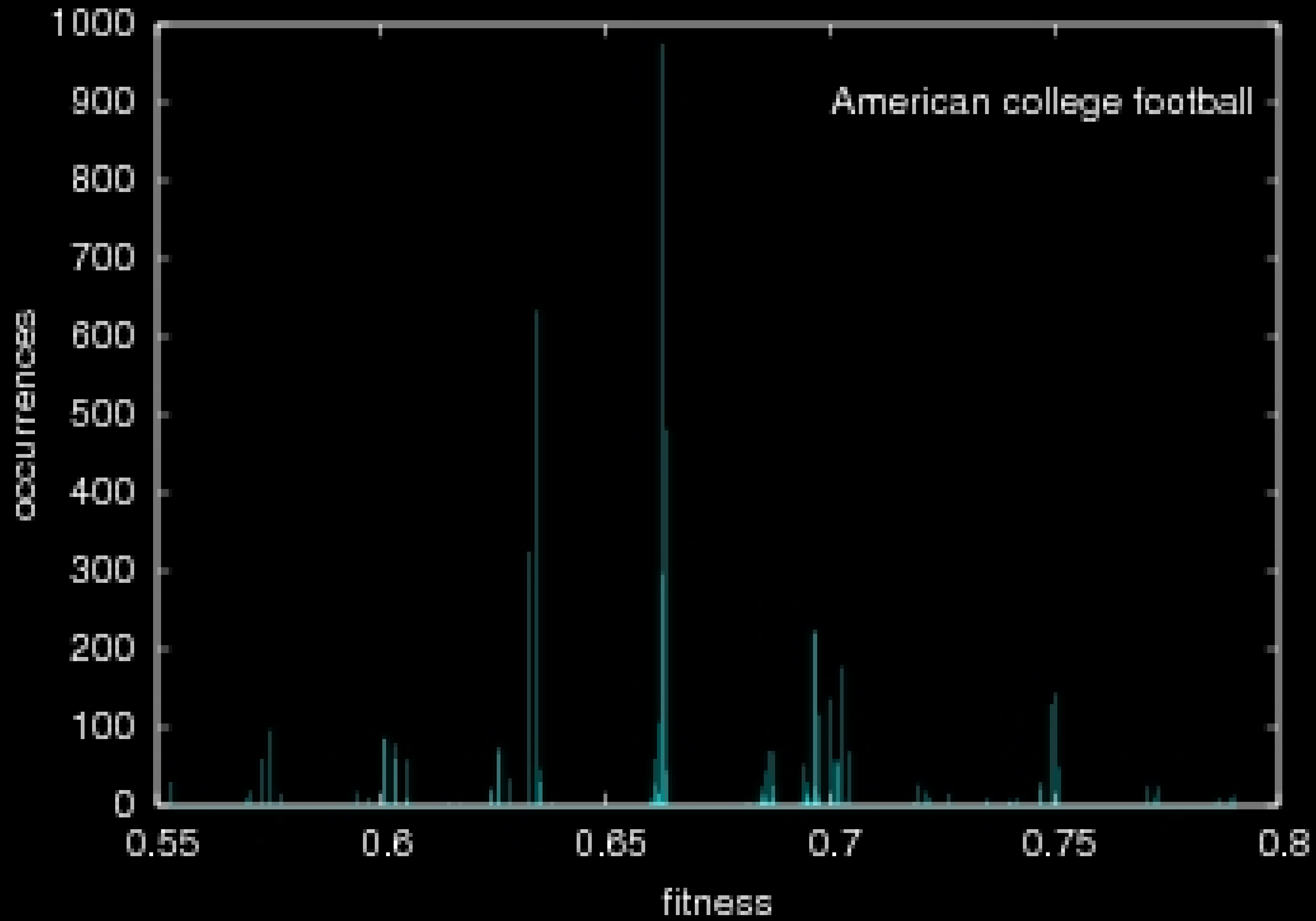
Ex. Random graphs

A good algorithm should indicate both the presence and the absence of community structure

Random graph



College football



Summary

Our method is:

- Fast
- Easy to implement
- It finds overlapping nodes
- It finds hierarchies
- Tests on artificial and real networks give excellent results

So use it!

http://arxiv.org/abs/0802.1218

International Workshop

Sociophysics: Status and Perspectives

ISI Foundation, Torino, Italy, 26-29 May 2008

- Home
- Program
- Invited Speakers
- Participants
- How to apply
- General Info
- Local Info
- Application Form

INSTITUTE FOR SCIENTIFIC INTERCHANGE FOUNDATION

Home

Sociophysics aims at a statistical physics modeling of large scale social phenomena, like opinion formation, cultural dissemination, the origin and evolution of language, crowd behavior, social contagion. The last years have witnessed the attempt to study collective phenomena emerging from the interactions of individuals as elementary units in social structures. A lot of work has been carried on, especially in the design of microscopic models, whereas comparatively little attention has been paid to a quantitative description of social phenomena and to the promotion of an effective cooperation between physicists and social scientists. The workshop is an occasion to review the state of the art of the field and to identify open problems and future research directions.

Program Committee

Claudio Castellano	SMC, INFM-CNR and University of Rome "La Sapienza", Italy
Santo Fortunato	ISI Foundation, Italy
Sidney Redner	Boston University, USA
Felix Reed-Tsochas	University of Oxford, UK

http://www.isi.it/progetti/sociophysics08/