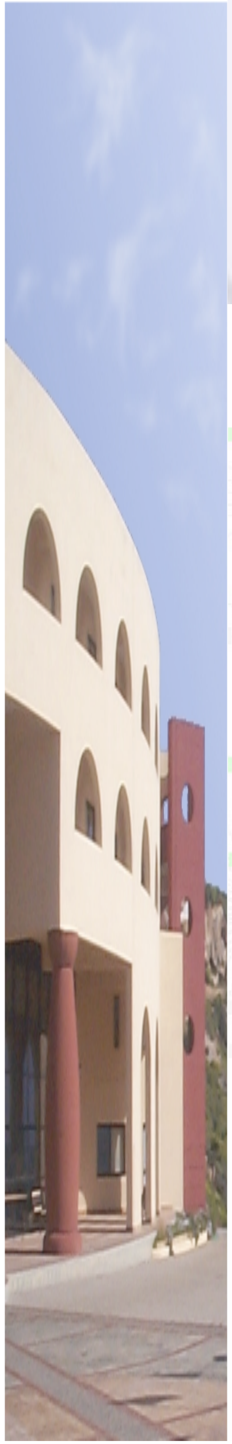


14-18 July 2008

Orthodox Academy of Crete

Kolympari - Chania - Greece

ΣΤΑΤΙΣΤΙΚΗ ΦΥΣΙΚΗ

STATISTICAL PHYSICS

2008

Module Recognition in Complex Networks by Dynamical Clustering

Alessandro Pluchino, Andrea Rapisarda* and Vito Latora*,*

*with Mikhail Ivanchenko** and Stefano Boccaletti****

CACTUS Chaos And Complexity Theoretical University Study Group Catania

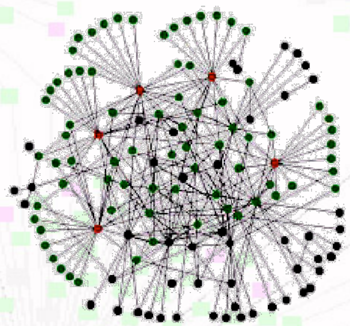
* Dipartimento di Fisica e Astronomia and INFN sezione di Catania University of Catania, Italy

** Moscow University

*** Nat.Ist. Applied Optycs - Florence

Outline

The problem:
Finding Community Structures in Complex networks



The approach:
Synchronization of Dynamical Oscillators in Weighted Networks

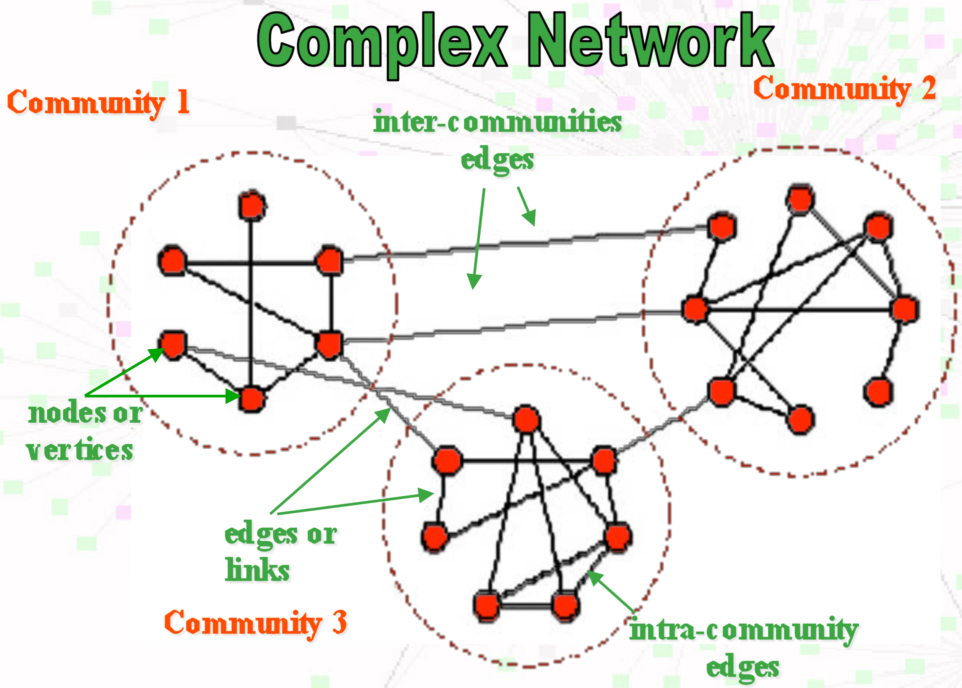
Our proposal:

Dynamical Clustering algorithm for the identification of Community Structures in Real and Trial Networks

Discussion and Numerical Results

The problem: Finding Community Structures in Complex Networks

An important open problem in complex networks analysis is the identification of modular structures:



Distinct modular structures, usually called Communities, can loosely be defined as subsets of nodes (vertices) which are more densely linked, when compared to the rest of the network.

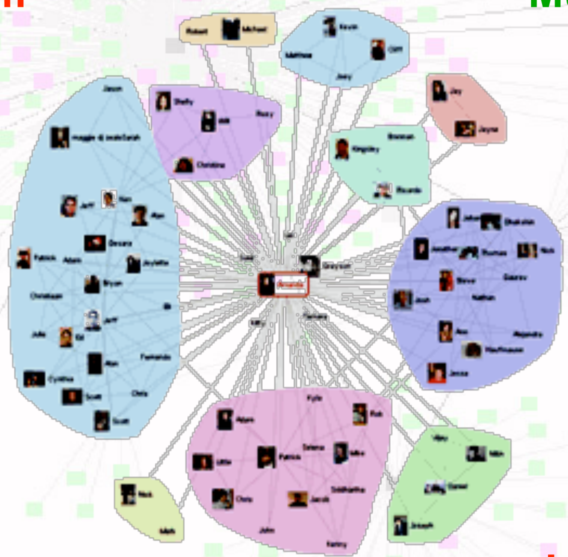
Communities, of course, are fundamental in social networks (parties, cultures, elites), but also in metabolic (biochemical patways) or neural networks (functional groups), in food webs and ecosystems (taxpnomic categories), in the world wide web (thematic pages), computer clusters and so on…

...thus many techniques has been developed in the years to deal with the problem of decting community structures in complex networks:

Graph Partitioning problem in computer science (NP complete)

Multi-Community Membership Methods

Spectral Analysis



Hierarchical Clustering Methods

Graph Equivalence through evolution of a physical analog

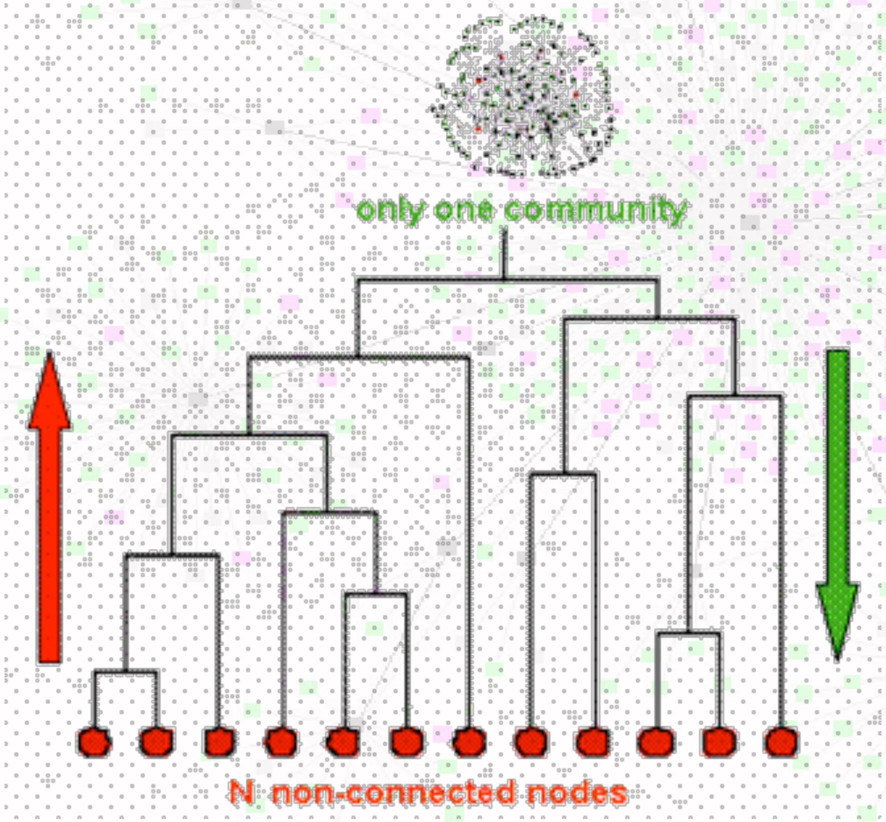
Simulated Annealing Techniques

Local Optimization of a Fitness Function

Dynamical Simplex Evolution

HIERARCHICAL CLUSTERING METHODS

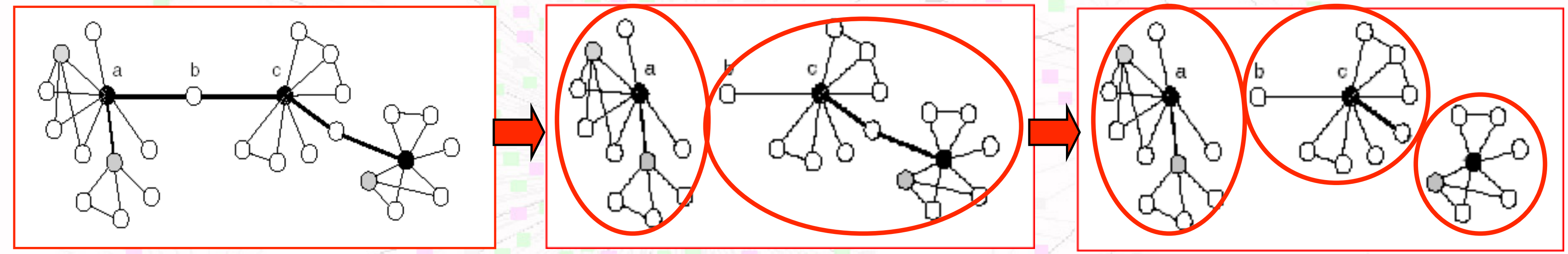
These techniques, firstly developed in social network analysis, are aimed at discovering natural divisions of networks into groups, based on various metric of similarity or strength of connection between vertices.

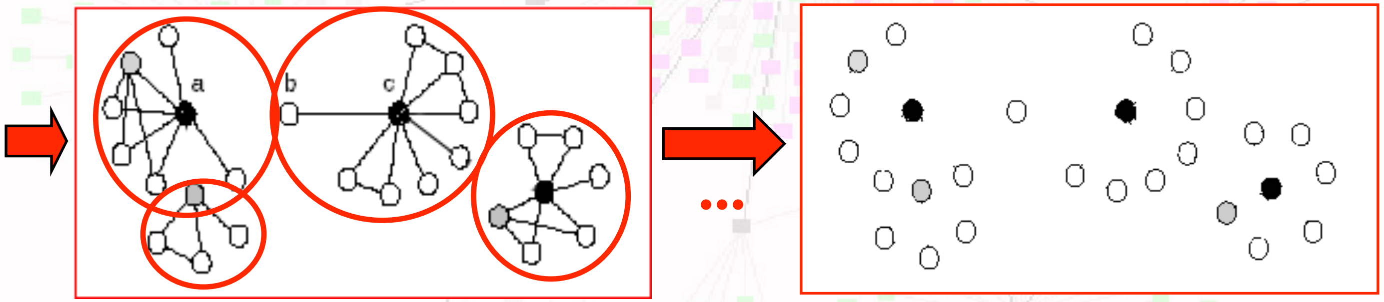


They fall into two broad classes: **agglomerative** and **divisive** methods, depending on whether they focus on the addition or the removal of edges to or from the network, and generating a dendrogram called hierarchical tree.

Divisive topological methods: progressively remove the edges of the network following their importance in connecting many pairs of nodes (expressed, for example, by the *edge betweenness**, i.e. the number of shortest paths which are making use of a given edge)

By doing this repeatedly, recalculating the betweenness at each step, the network breaks iteratively into smaller and smaller isolated clusters (communities or modules)…





…until it breaks into a collection of non-connected single nodes...

But which subdivision does give the best communities configuration for a given network?

*M.E.J.Newman and M.Girvan, 2004 *Phys. Rev. E* **69** 026113

In order to establish this, it is often used the "modularity" **Q** *, a quantity that, at each step, compares the fraction of intra-community edges with the expected value of the same quantity in an equivalent network with random connections (null model), and allows us to test which communities configuration found by the divisive algorithm is the best one:

modularity

$$Q = \sum_{i=1}^{n_C} (e_{ii} - b_i^{\,2})$$

fraction of edges that connect vertices in community i

fraction of edges that connect vertices in community i for a random network

Q=0 for only 1 com. or N isolated nodes

Tipically 0.3 < Q < 0.7

n_c is the number of communities

$||e||$ is a n_c x n_c matrix whose elements e_{ij} represent the fraction of total edges connecting a node in community i with a node in community j

$b_i = \sum_j e_{ij}$ represents the fraction of total edges connected to a node in community-i

*M.E.J.Newman and M.Girvan, 2004 *Phys. Rev. E* **69** 026113

S.Fortunato, M.Barthelemy, Proc.Natl.Acad.Sci. U.S.A. 104 (1), 36-41 (2007)

PNAS PNAS PNAS PNAS

Resolution limit in community detection

Santo Fortunato[†‡§] and Marc Barthélemy[†¶||]

[†]School of Informatics and Center for Biocomplexity, Indiana University, Bloomington, IN 47406; [‡]Fakultät für Physik, Universität Bielefeld, D-33501 Bielefeld, Germany; [§]Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, 10133 Torino, Italy; and [¶]Commissariat à l'Energie Atomique–Département de Physique Théorique et Appliquée, 91680 Bruyeres-Le-Chatel, France

Edited by David O. Siegmund, Stanford University, Stanford, CA, and approved November 6, 2006 (received for review July 17, 2006)

Detecting community structure is fundamental for uncovering the links between structure and function in complex networks and for practical applications in many disciplines such as biology and sociology. A popular method now widely used relies on the optimization of a quantity called modularity, which is a quality index for a partition of a network into communities. We find that modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined. This finding is confirmed through several examples, both in artificial and in real social, biological, and technological networks, where we show that modularity optimization indeed does not resolve a large number of modules. A check of the modules obtained through modularity optimization is thus necessary, and we provide here key elements for the assessment of the reliability of this community detection method.

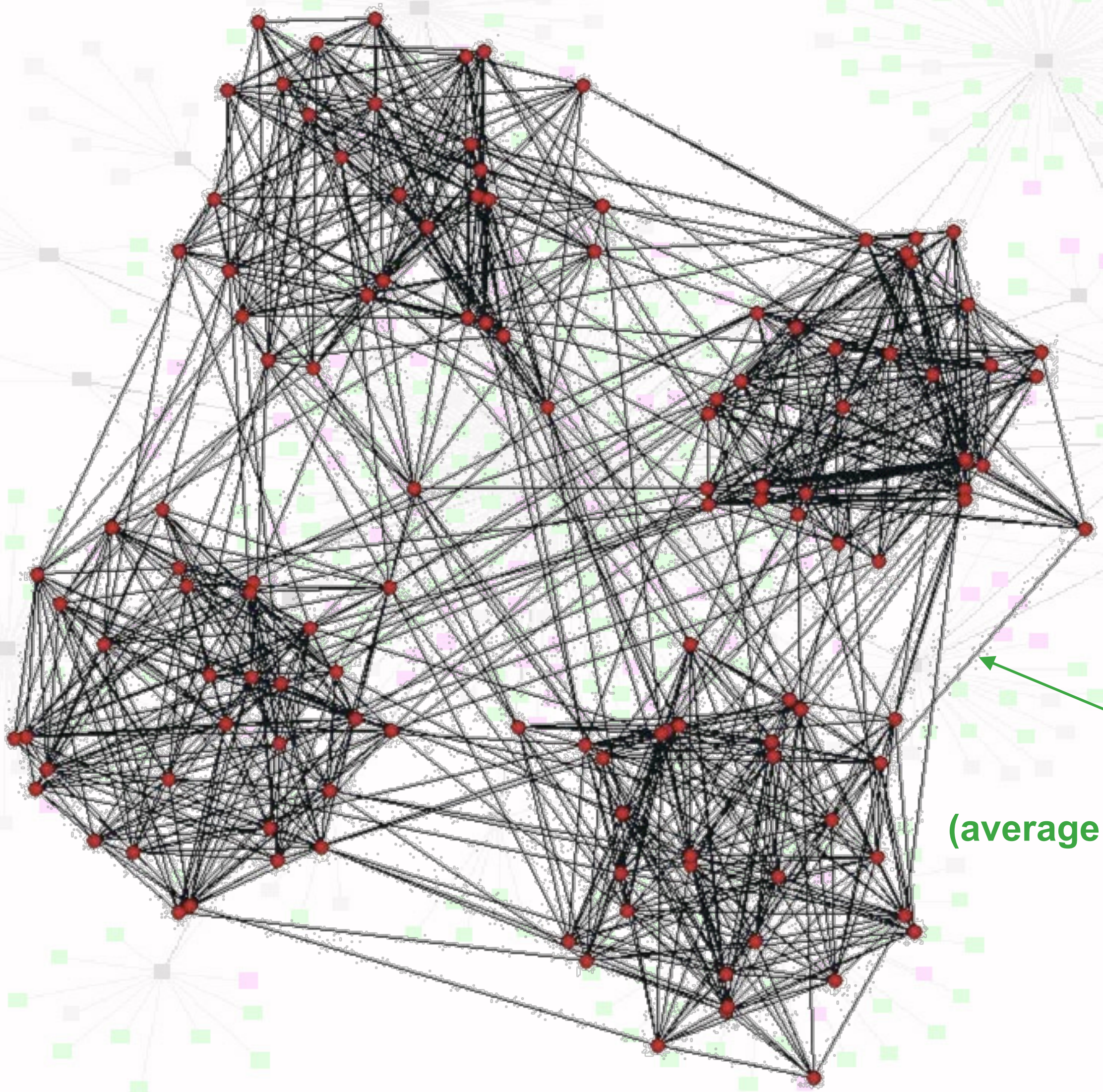
complex networks | modular structure | metabolic networks | social networks

annealing (27, 28), but this method is computationally very expensive.

Modularity optimization seems, therefore, to be a very effective method to detect communities, both in real and in artificially generated networks. However, modularity itself has not yet been thoroughly investigated, and only a few general properties are known. For example, it is known that the modularity value of a partition does not have a meaning by itself, but only when compared with the corresponding modularity expected for a random graph of the same size (29), as the latter may attain very high values due to fluctuations (27).

In this article, we present a critical analysis of modularity and of the applicability of modularity optimization to the problem of community detection. We show that modularity contains an intrinsic scale that depends on the total number of links in the network. Modules that are smaller than this scale may not be resolved, even in the extreme case where they are complete graphs connected by single bridges. The resolution limit of modularity actually depends on the degree of interconnectedness between pairs of communities and can reach values of the order of the size of the whole network. Tests performed on

Modularity in computer generated random trial networks
(N=128, <k>=16, 4 communities)

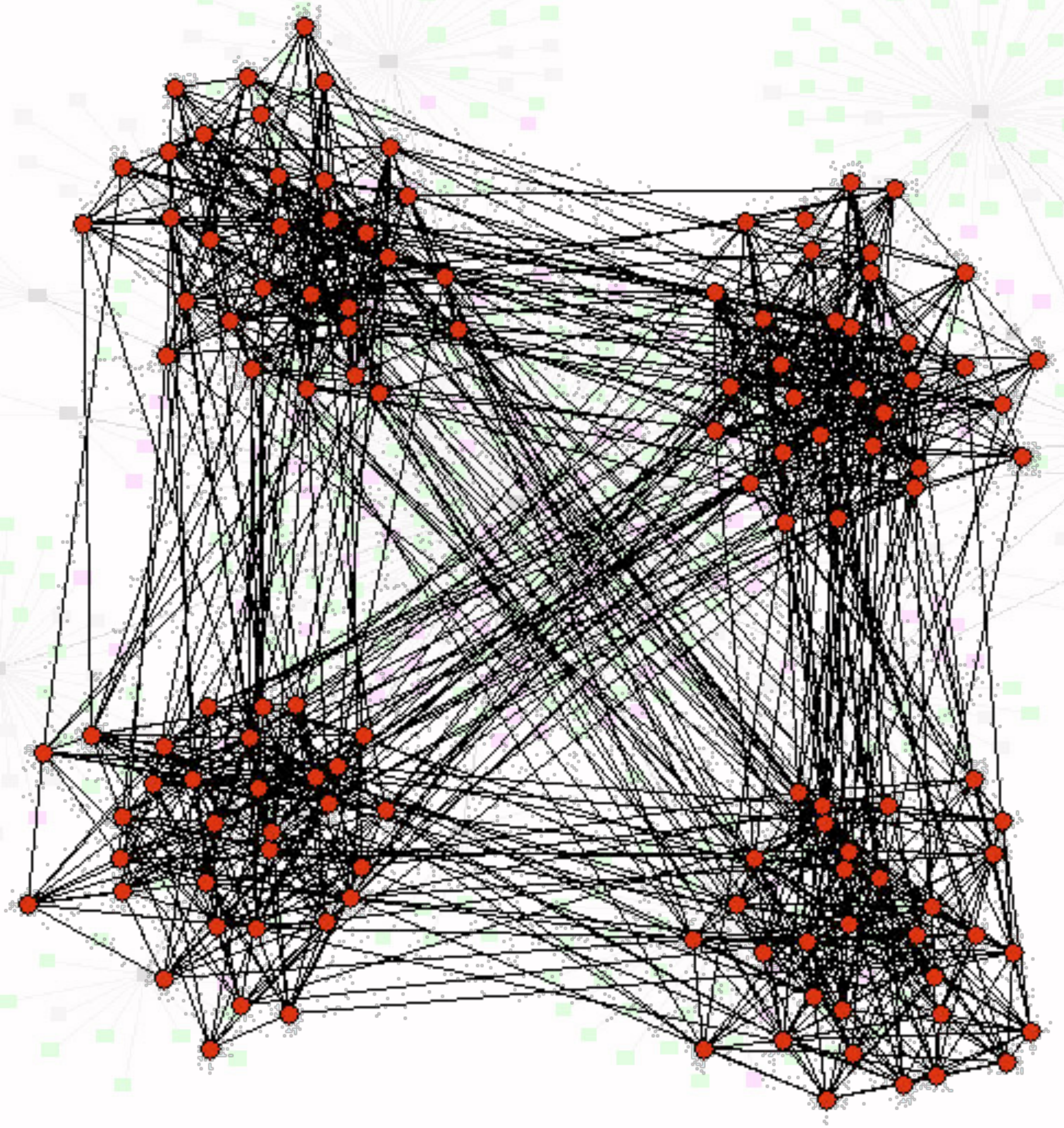


zout = 2
Q ~ 0.7

zout

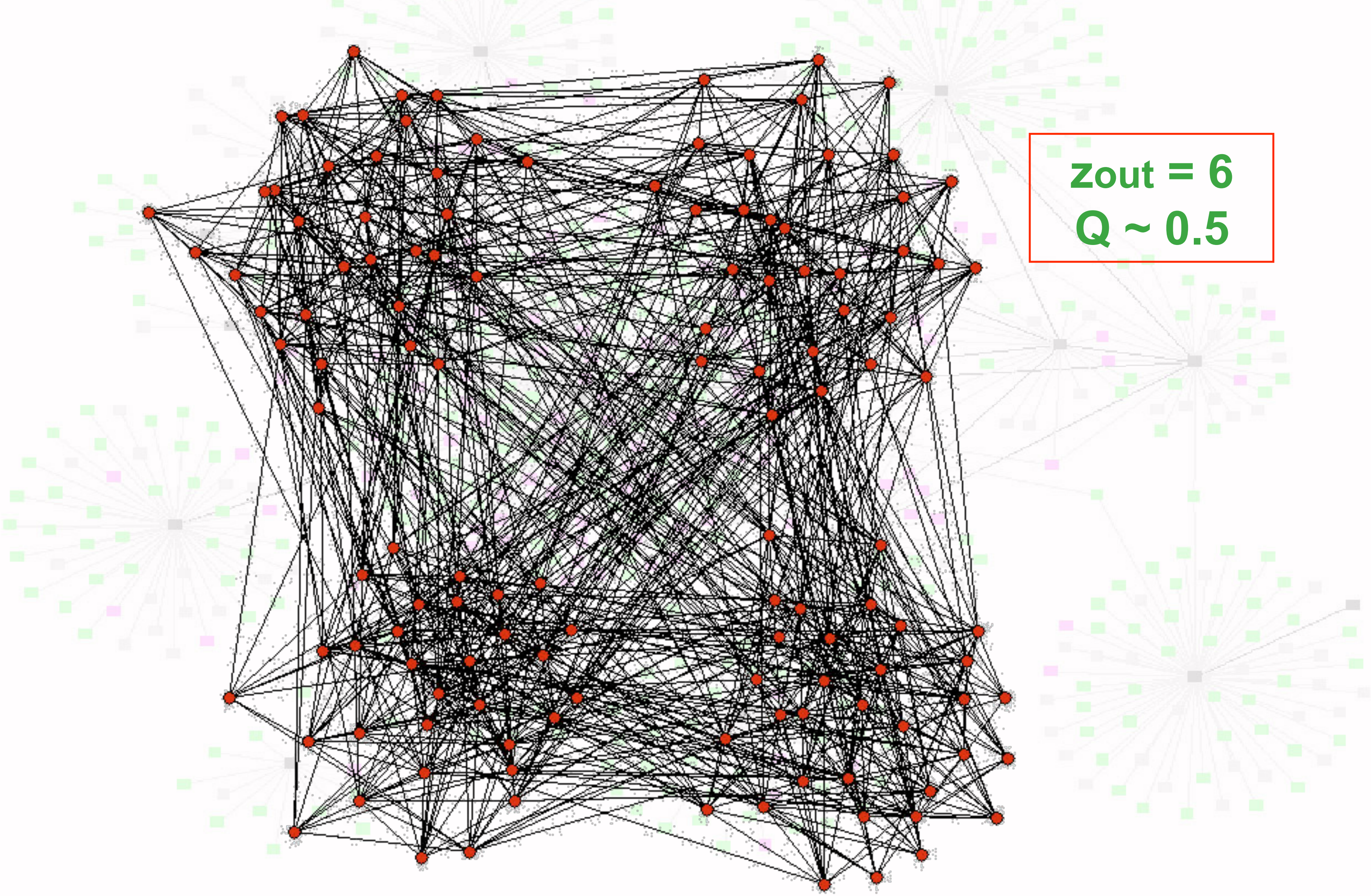
(average number of inter-community edges per node)

Modularity in computer generated random trial networks (N=128, <k>=16, 4 communities)



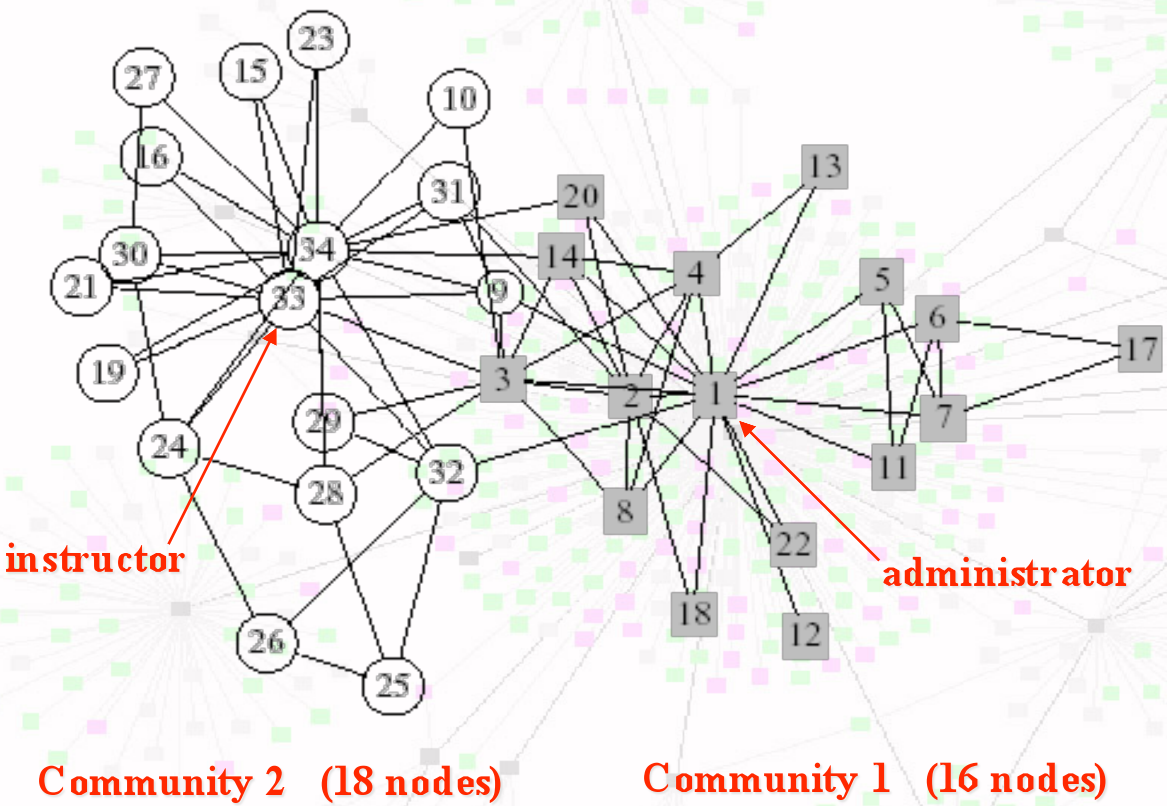
zout = 4
Q ~ 0.6

Modularity in computer generated random trial networks (N=128, <k>=16, 4 communities)



zout = 6
Q ~ 0.5

Zachary’s Karate Club friendships network



Girvan Newman
Shortest-path edge-betweenness divisive method

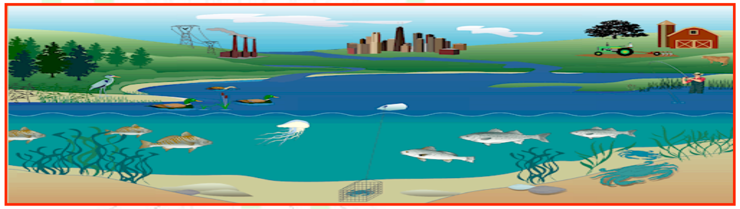
M.E.J.Newman and M.Girvan, 2004 *Phys. Rev. E* **69** 026113

W.Zachary (1977) *J.Anthropol.Res.* **33** 452-473

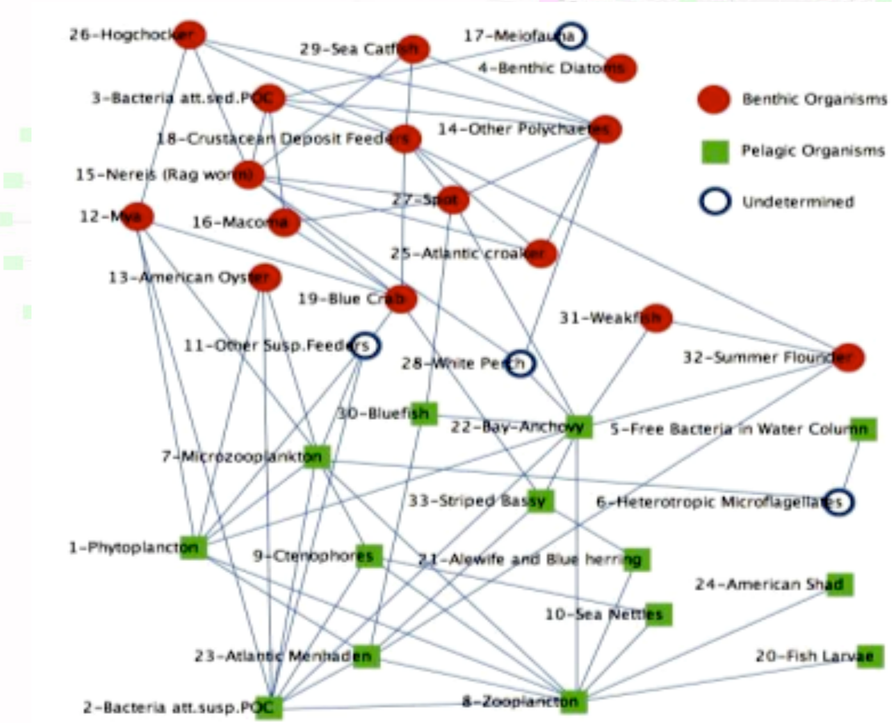
Chesapeake Bay food web (USA)

The Chesapeake Bay -- the largest estuary in the U.S. -- is a complex ecosystem that includes important habitats and food webs. The Bay itself, its rivers, wetlands, trees and land all provide homes, protection or food for complex groups of species, with impressive combinations of relationships.

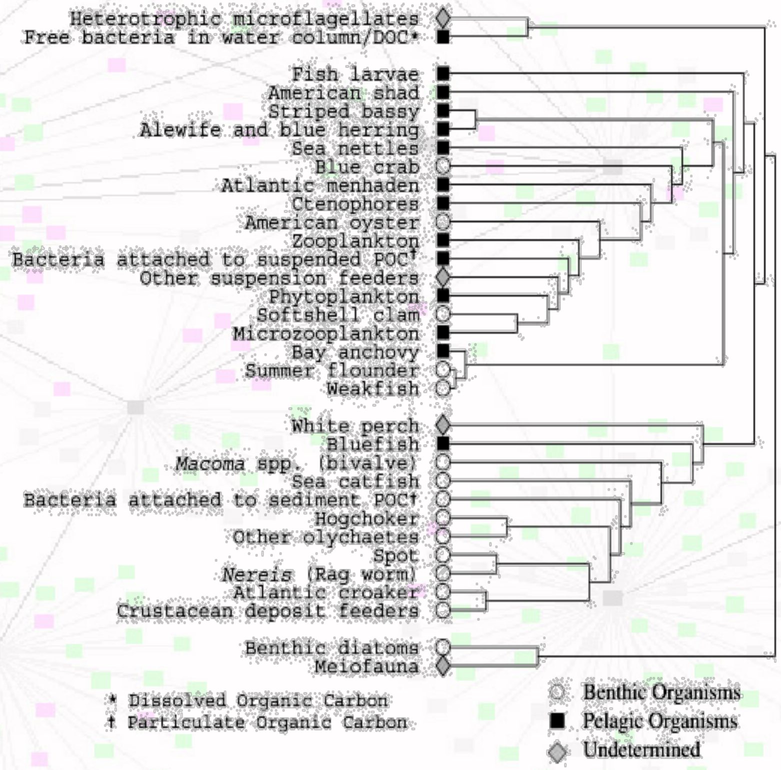
D.Baird & R.Ulanowicz (1989) *Ecol.Monogr.* **59** 329-364



Predatory Relationships NETWORK among the 33 most important taxa



Hierarchical tree obtained with GN algorithm



Girvan M, Newman M E J. Community structure in social and biological networks. In *Proc. the National Academy of Science* , USA, 2002, 99(12): 7821-7826.

Topological Divisive Algorithms like GN show in general a good sensitivity but have the problem of recalculating betweennesses at each step

Since a single-step calculus of all the edge-betweennesses takes $O(N^2)$ **operations, and the whole process takes K~N steps, these algorithms are quite slow –** $O(N^3)$ on a spars graph–

Other algorithms are quicker but in many cases their sensitivity in the identification of an unknown number of communities is not very good...

We propose a
DIFFERENT DIVISIVE
HIERARCHICAL APPROACH
based on
Synchronization of Dynamical
Oscillators in Weighted Networks
(Dynamical Clustering)

An example for Synchronization

The Kuramoto model* is the simplest models for synchronization available on the market and consists of N fully coupled phase oscillators with intrinsic natural frequencies ω_i and coupling parameter K:

coupling strenght

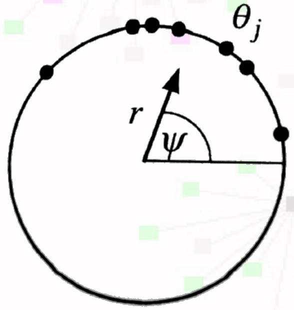
$$\frac{d\vartheta_i(t)}{dt} = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\vartheta_j - \vartheta_i), \qquad i = 1,\dots,N$$

natural (fixed) frequencies

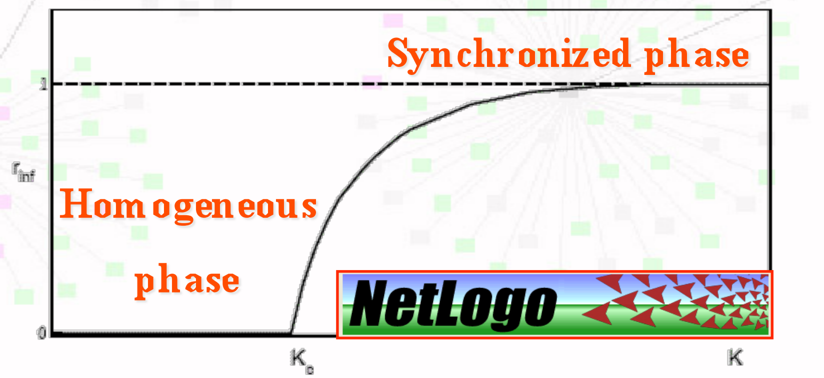
phases of oscillators

$$\vartheta_i(t) \in [0, 2\pi)$$

The coherence of the system is measured by the mean field order parameter r (0<r(t)<1):



$$re^{i\Psi} = \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j}$$



Asymptotic order parameter r_∞ as a function of the coupling in the Kuramoto model

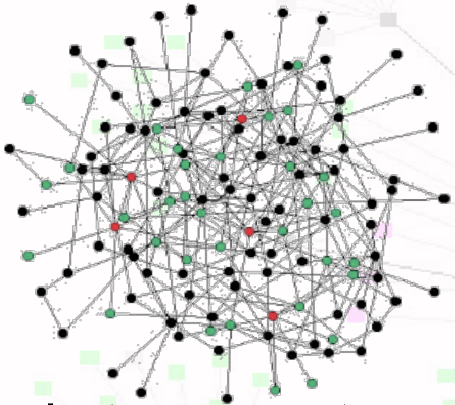
*proposed by Y.Kuramoto in 1975

Weighting procedure of a Complex Network

*M.Chavez, D.U.Hwang, A.Amann, H.G.E.Hentschel and S.Boccaletti, *Phys. Rev. Lett.* **94** 218701 (2005)

Suppose to have a (unweighted, undirected) network of N coupled identical oscillators*. The equation of motion reads:

Network with N nodes



coupling strenght

$$\underbrace{\dot{\vec{x}}_i = \vec{F}(\vec{x}_i)}_{\text{dynamical system defined over each node of the network}} - \sigma \sum_{j=1}^{N} G_{ij} \vec{H}(\vec{x}_i - \vec{x}_j), \qquad i = 1, \dots, N$$

coupling matrix

coupling function

Let us now to perform an opportune choice of the coupling matrix G_{ij} in the network equation, by means of a weighting procedure that assignes to each edge a **load l_{ij}** equal to its betweenness (i.e. the number of shortest paths that are making use of that edge):

coupling matrix G=G(α)

$$\dot{\vec{x}}_i = \vec{F}(\vec{x}_i) - \sigma \sum_{j \in K_i}^{N} \frac{l_{ij}^{\alpha(t)}}{\sum_{j \in K_i} l_{ij}^{\alpha(t)}} \vec{H}(\vec{x}_i - \vec{x}_j), \qquad i = 1, \dots, N$$

where **α(t)** is a real tunable parameter and K_i is the set of neighbors of the i^{th} node.

Tuning the synchronization of a network of oscillators for finding community structures

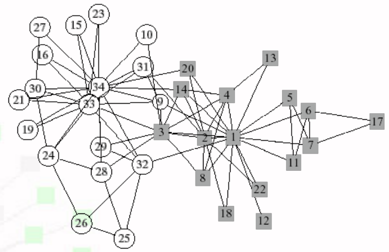
DYNAMICAL CLUSTERING ALGORITHM

$$\dot{\vec{x}}_i = \vec{F}(\vec{x}_i) - \sigma \sum_{j \in K_i}^{N} \frac{l_{ij}^{\alpha(t)}}{\sum_{j \in K_i} l_{ij}^{\alpha(t)}} \vec{H}(\vec{x}_i - \vec{x}_j), \qquad i = 1,\ldots,N$$

1. **At variance with the topological methods we calculate the edge betweennesses (i.e. the edge's loads l_{ij}) only one time for a given network;**
2. $t = 0: \ \alpha(0) \sim 0$ **We fix the coupling strenght σ so that the system starts from a state which rapidly synchronizes in frequency;**
3. $t > 0: \ \alpha(t) \to -\infty$ **Decreasing α at each time-step, the edges with a great betweenness will be weighted less and less and the oscillators progressively desynchronize;**
4. **We look at clusters of nodes (communities) oscillating with a common phase or frequency and we select the clusters configuration with the highest modularity Q.**

S.Boccaletti, M.Ivanchenko, V.Latora, A.P. and A.Rapisarda - Physical Review E **75** (2007) 045102(R)

First tests on the Karate Club Network:

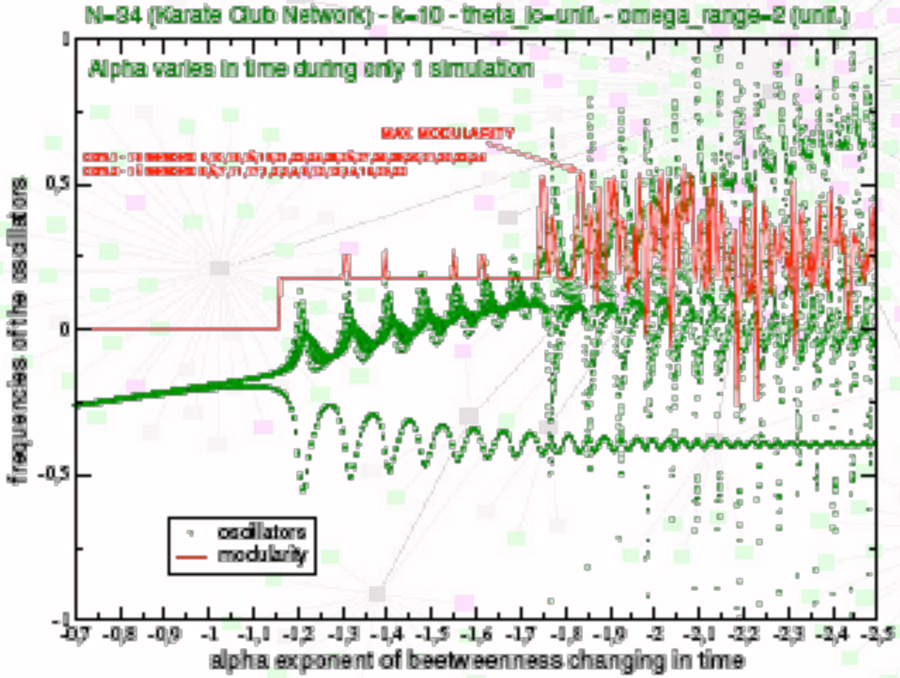


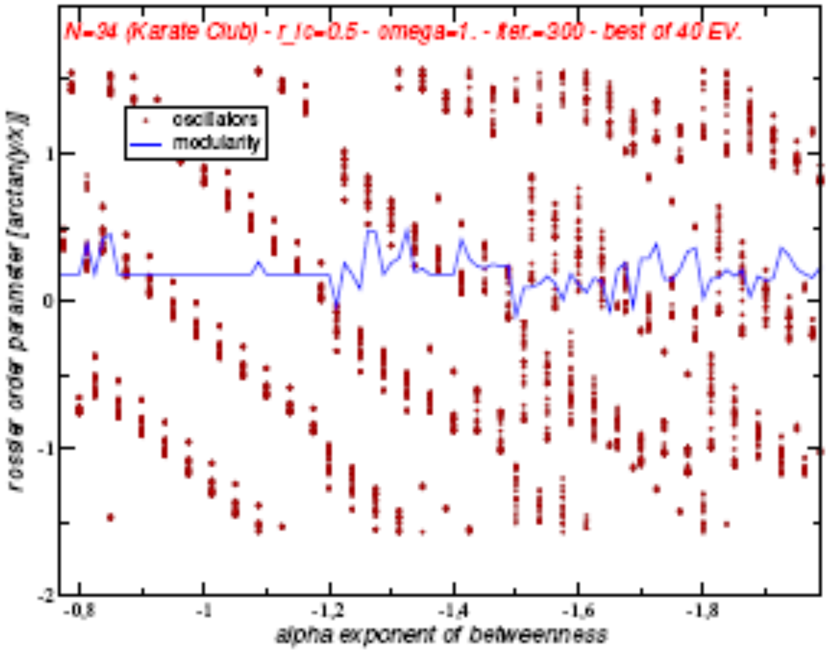
Kuramoto's non identical 1D oscillators

$$\dot{\vartheta}_i = \omega_i + \frac{\sigma}{\sum_{j\in K_i} l_{ij}^{\alpha(t)}} \sum_{j\in K_i} l_{ij}^{\alpha(t)} \sin(\theta_j - \theta_i) \qquad i = 1,\ldots,N$$

Chaotic Rössler identical 3D oscillators

$$\begin{cases} \dot{x}_i = -\omega y_i - z_i - \dfrac{\sigma}{\sum_{j\in K_i} l_{ij}^{\alpha(t)}} \sum_{j\in K_i} l_{ij}^{\alpha(t)} (x_i - x_j) \\ \dot{y}_i = \omega x_i + 0.165 y_i \qquad\qquad i = 1,\ldots,N \\ \dot{z}_i = 0.2 + z_i (x_i - 10) \end{cases}$$





The Opinion Changing Rate (OCR) model

It is a modification of the Kuramoto model and consists of the following rate equations describing the opinions evolution of N fully interacting agents:

istantaneous frequencies

coupling strenght

$$\dot{x}_i = \omega_i + \frac{\sigma}{N}\sum_{j=1}^{N} \beta \sin(x_j - x_i)\, e^{-\beta|x_j - x_i|}, \qquad i = 1,\ldots,N$$

intrinsic frequencies

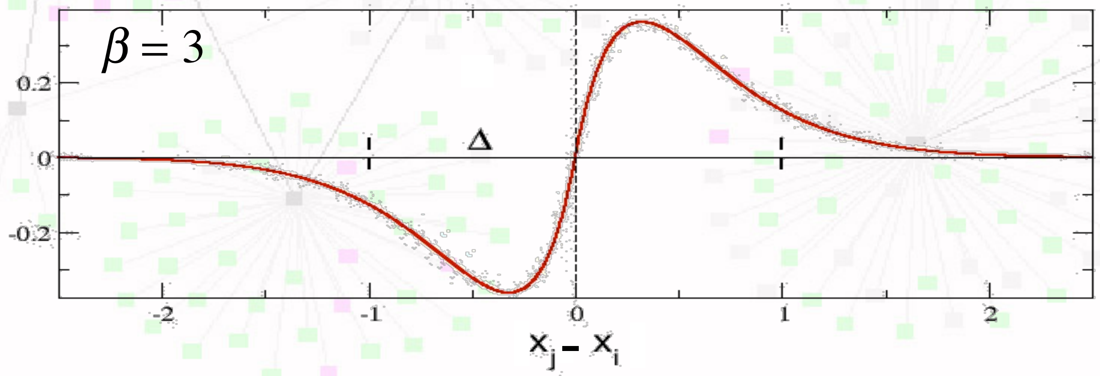
opinions

$x_i(t) \in \,]-\infty,+\infty[$

$x_i(0) \in \,]-1,+1[$

$\omega_i \in [0,1]$

The interaction potential decreases for distant opinions:



A.P., V.Latora, A.Rapisarda, *Int.Journ.of Mod.Phys. C* **16** 515 (2005)

OCR-HK on weighted networks: Dynamical Clustering (DC) Algorithm

In order to apply the DC algorithm to the OCR system we further modified the standard OCR model forcing the oscillators natural frequencies to follow the so called Heigselmann-Krause dynamics, a process which improves the performance of the algorithm and minimizes the dependence on the initial distribution of natural frequencies:

loads (betweennesses)

tuning parameter ($\delta\alpha \approx 10^{-3}$)

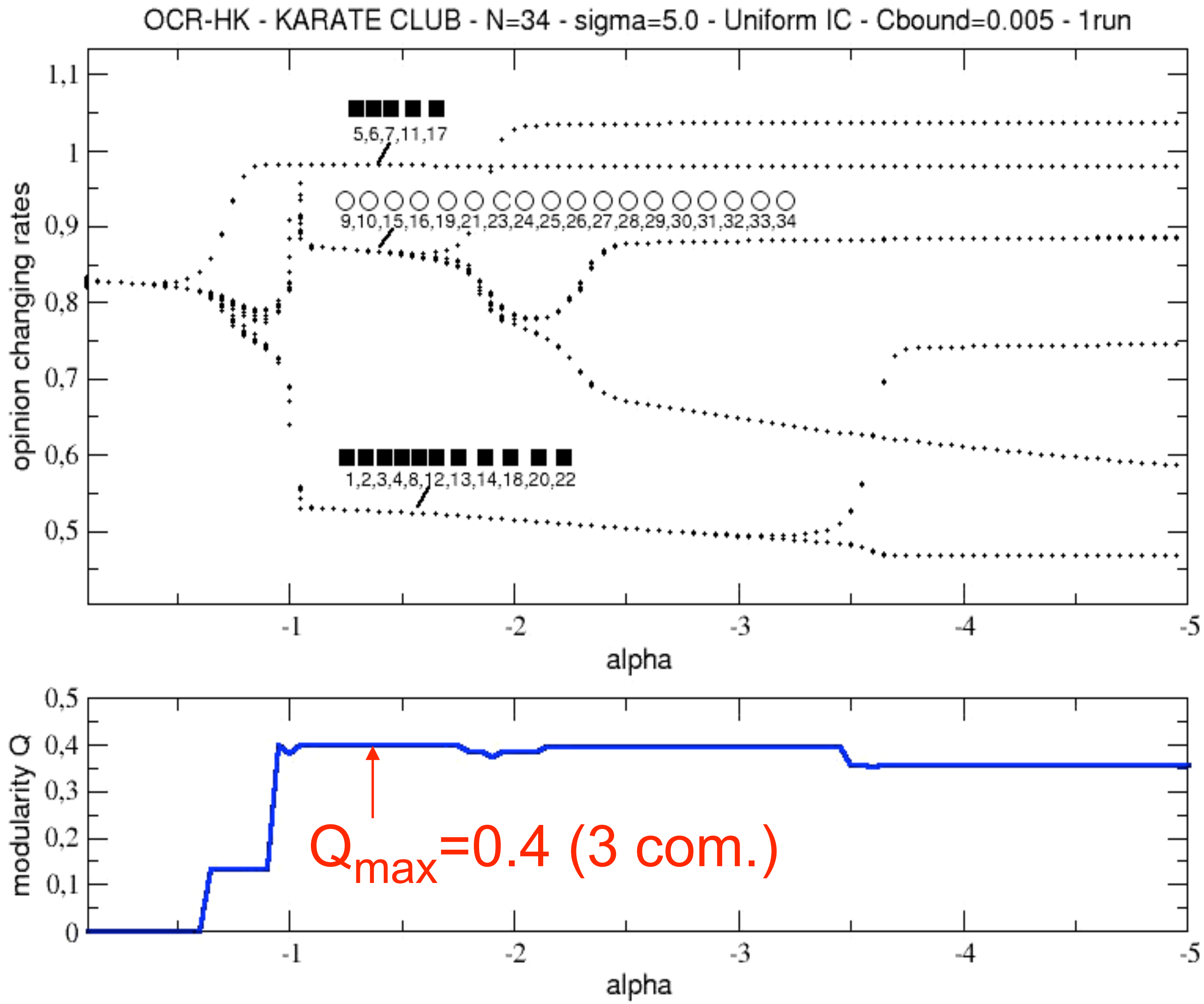
$$\dot{x}_i(t) = \omega_i(t) + \frac{\sigma}{\sum_{j \in K_i} l_{ij}^{\alpha(t)}} \sum_{j \in K_i}^{N} \beta\, l_{ij}^{\alpha(t)} \sin(x_j - x_i)\, e^{-\beta |x_j - x_i|}, \qquad i = 1, \dots, N$$

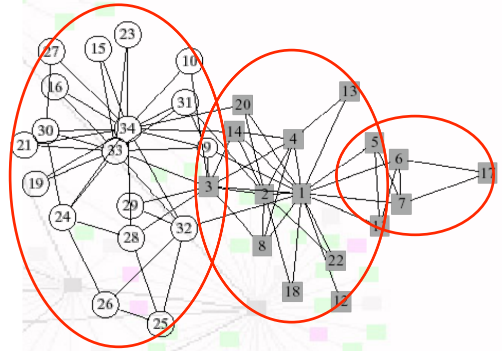
neighbours of node -i in the selected netwrok

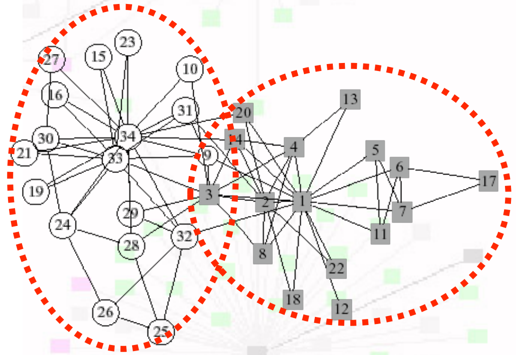
intrinsic frequencies, updated in time with HK dynamics (based on the concept of "confidence bound")

see S.Boccaletti, M.Ivanchenko, V.Latora, A.P. and A.Rapisarda Physical Review E 75 (2007) 045102(R) for further details

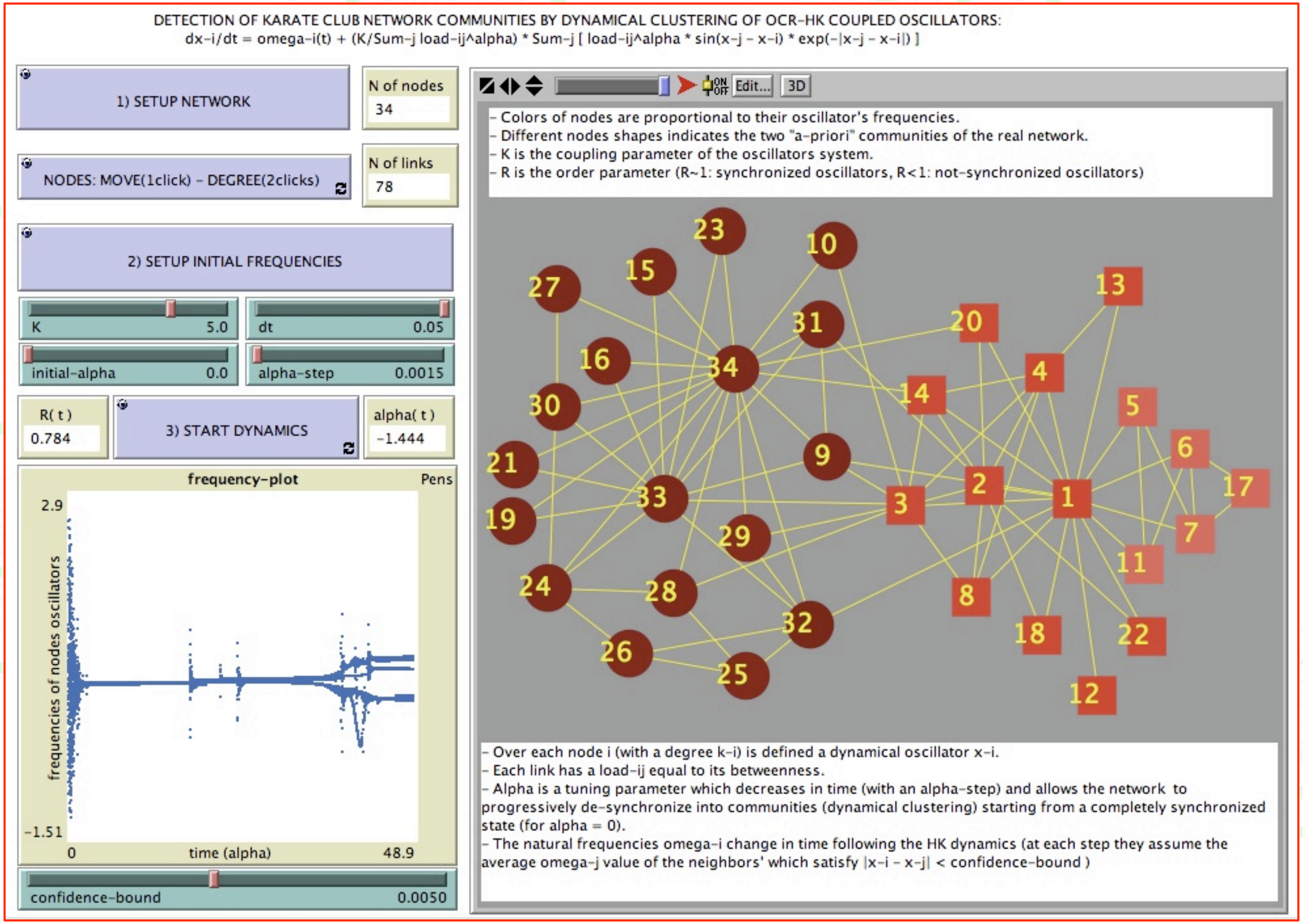
OCR-HK Tests on real networks: Karate Club



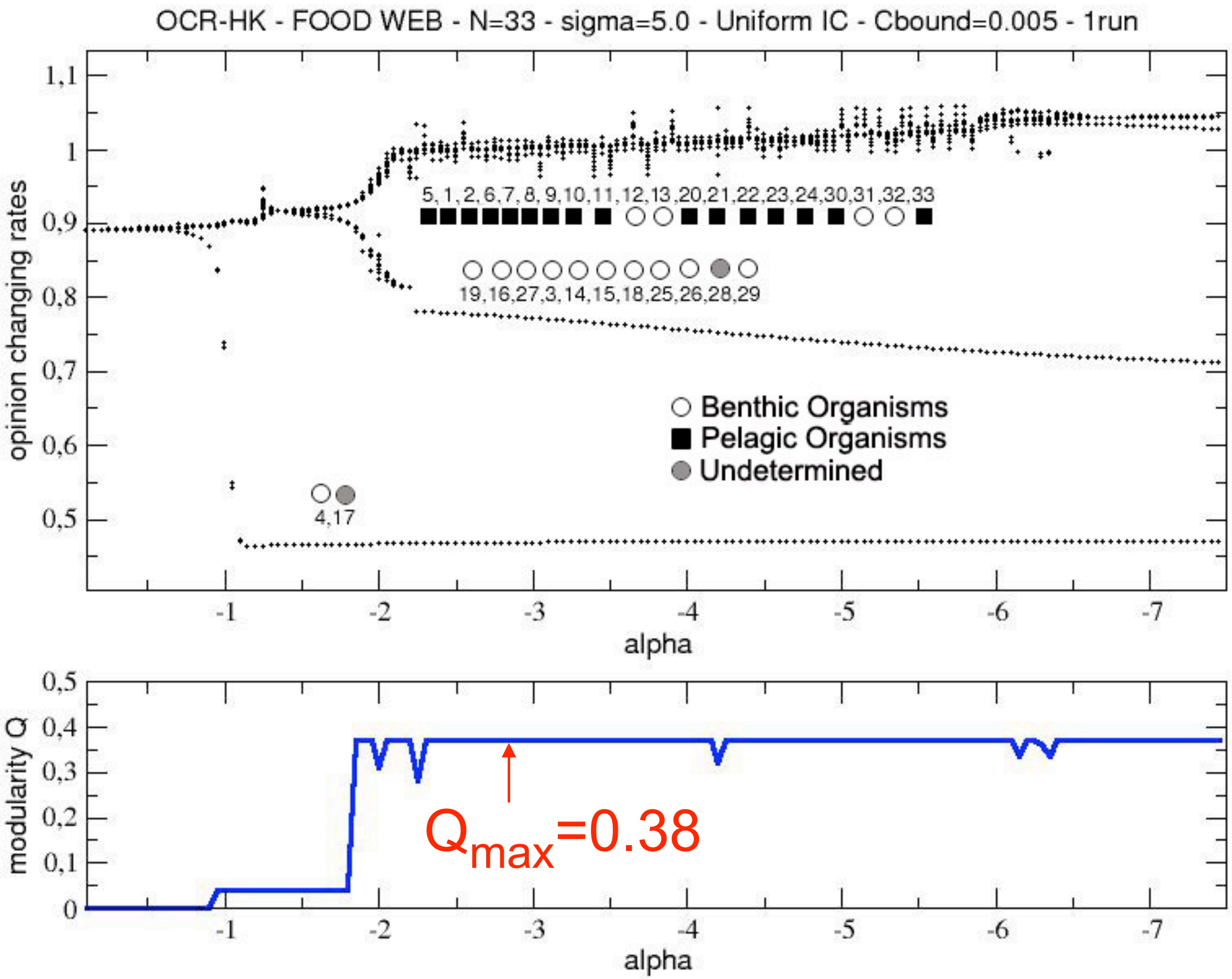




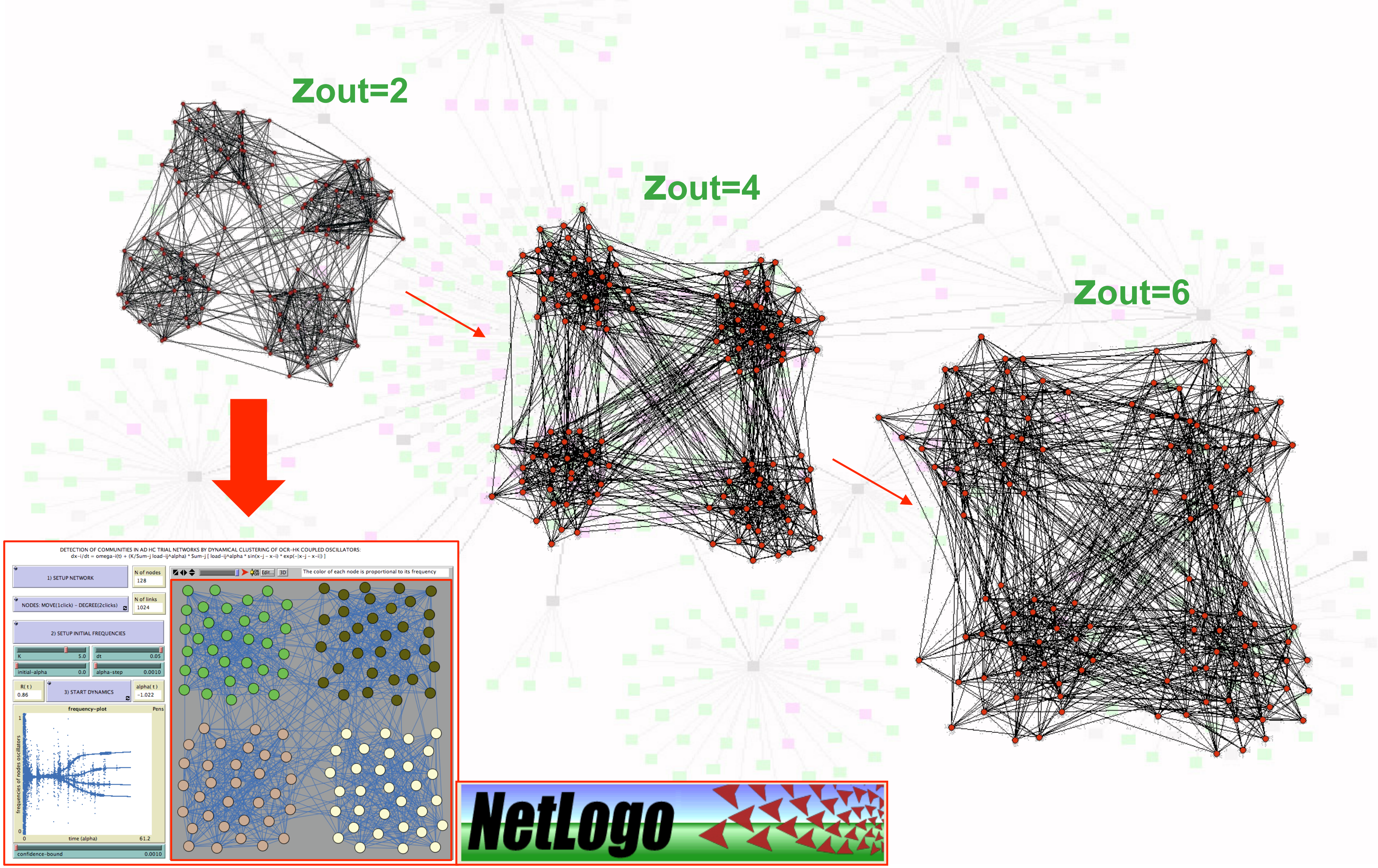
NetLogo

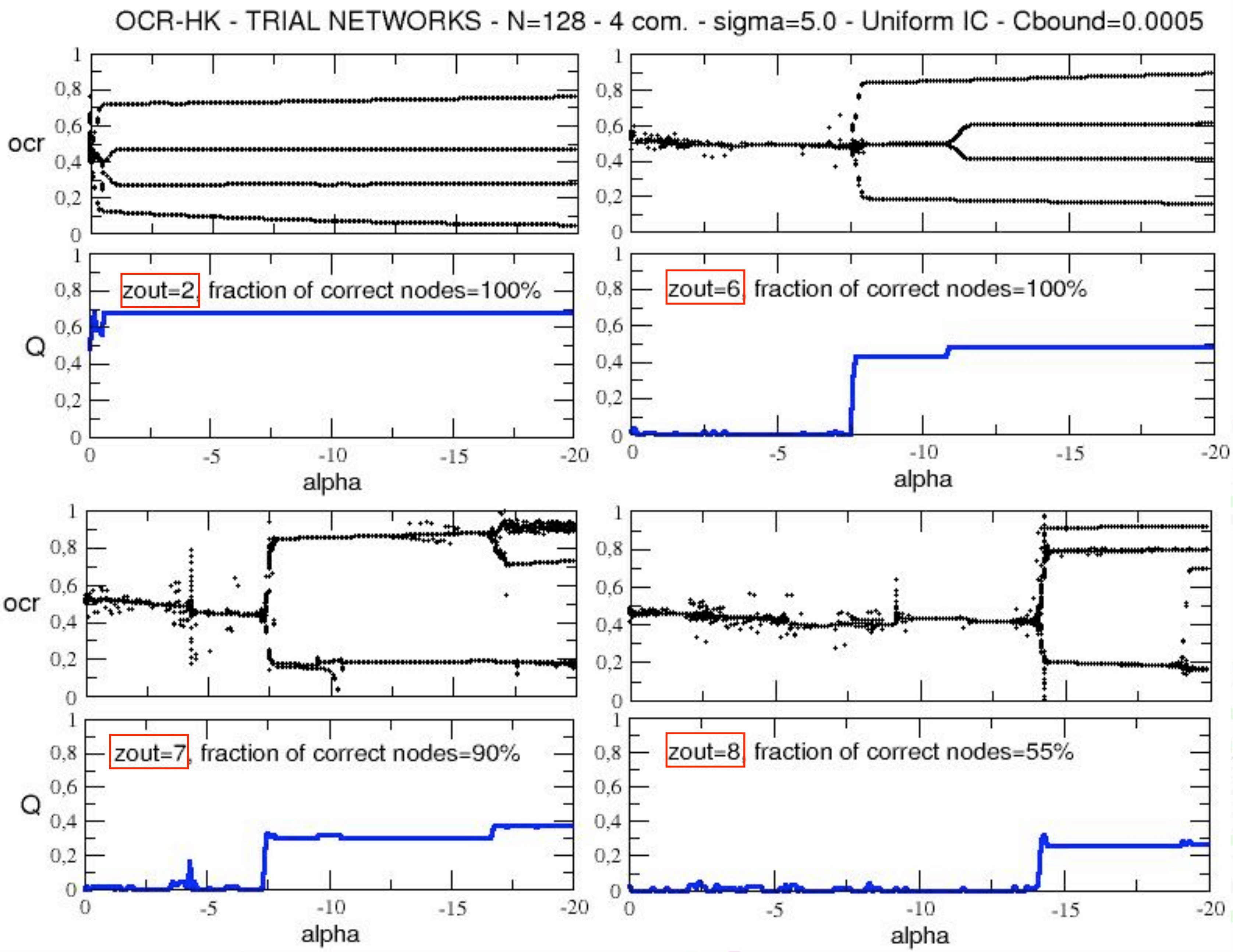

DETECTION OF KARATE CLUB NETWORK COMMUNITIES BY DYNAMICAL CLUSTERING OF OCR-HK COUPLED OSCILLATORS:
dx-i/dt = omega-i(t) + (K/Sum-j load-ij^alpha) * Sum-j [load-ij^alpha * sin(x-j - x-i) * exp(-|x-j - x-i|)]
1) SETUP NETWORK
N of nodes
34
NODES: MOVE(1click) - DEGREE(2clicks)
N of links
78
2) SETUP INITIAL FREQUENCIES
K
5.0
dt
0.05
initial-alpha
0.0
alpha-step
0.0015
R(t)
0.784
3) START DYNAMICS
alpha(t)
-1.444
frequency-plot
Pens
2.9
-1.51
frequencies of nodes oscillators
0
time (alpha)
48.9
confidence-bound
0.0050
Edit...
3D
- Colors of nodes are proportional to their oscillator's frequencies.
- Different nodes shapes indicates the two "a-priori" communities of the real network.
- K is the coupling parameter of the oscillators system.
- R is the order parameter (R~1: synchronized oscillators, R<1: not-synchronized oscillators)
- Over each node i (with a degree k-i) is defined a dynamical oscillator x-i.
- Each link has a load-ij equal to its betweenness.
- Alpha is a tuning parameter which decreases in time (with an alpha-step) and allows the network to progressively de-synchronize into communities (dynamical clustering) starting from a completely synchronized state (for alpha = 0).
- The natural frequencies omega-i change in time following the HK dynamics (at each step they assume the average omega-j value of the neighbors' which satisfy |x-i - x-j| < confidence-bound)

OCR-HK Tests on real networks: Chesapeake Bay food web



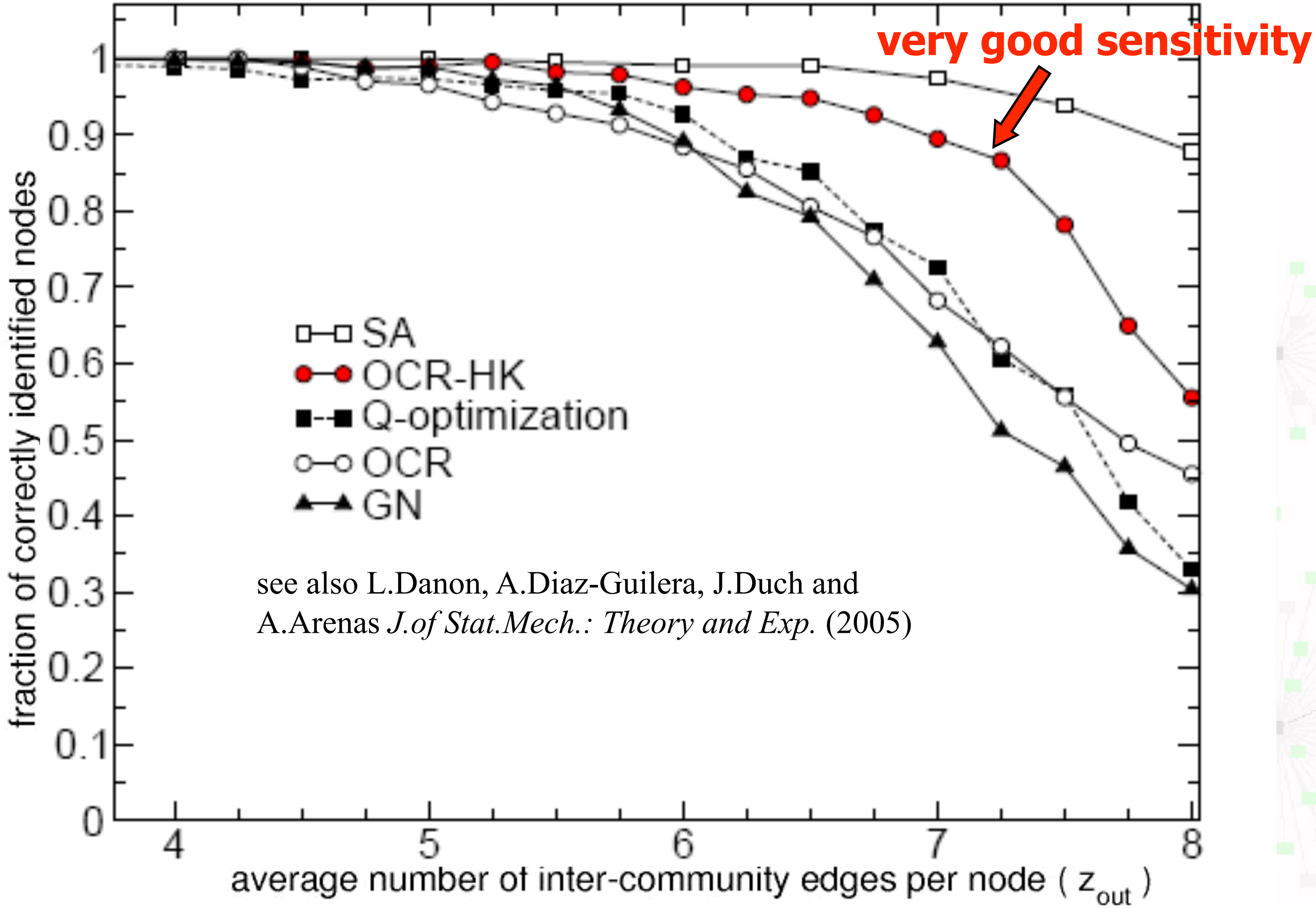
OCR-HK Tests on computer generated random trial networks with increasing *zout* (N=128, <k>=16, 4 communities)



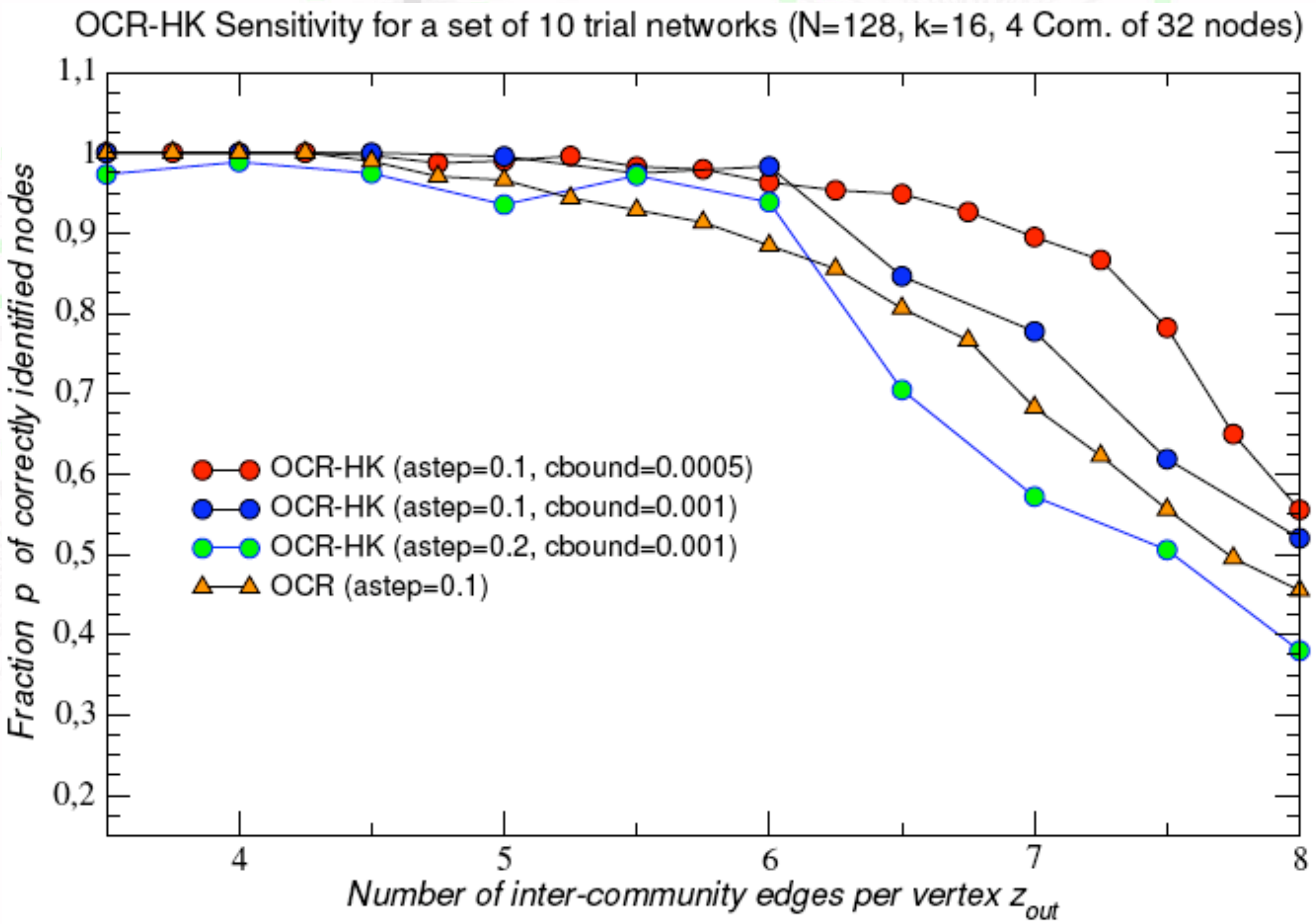

OCR-HK - TRIAL NETWORKS - N=128 - 4 com. - sigma=5.0 - Uniform IC - Cbound=0.0005
ocr
Q
alpha
zout=2, fraction of correct nodes=100%
zout=6, fraction of correct nodes=100%
zout=7, fraction of correct nodes=90%
zout=8, fraction of correct nodes=55%

Dynamical Clustering on random trial networks

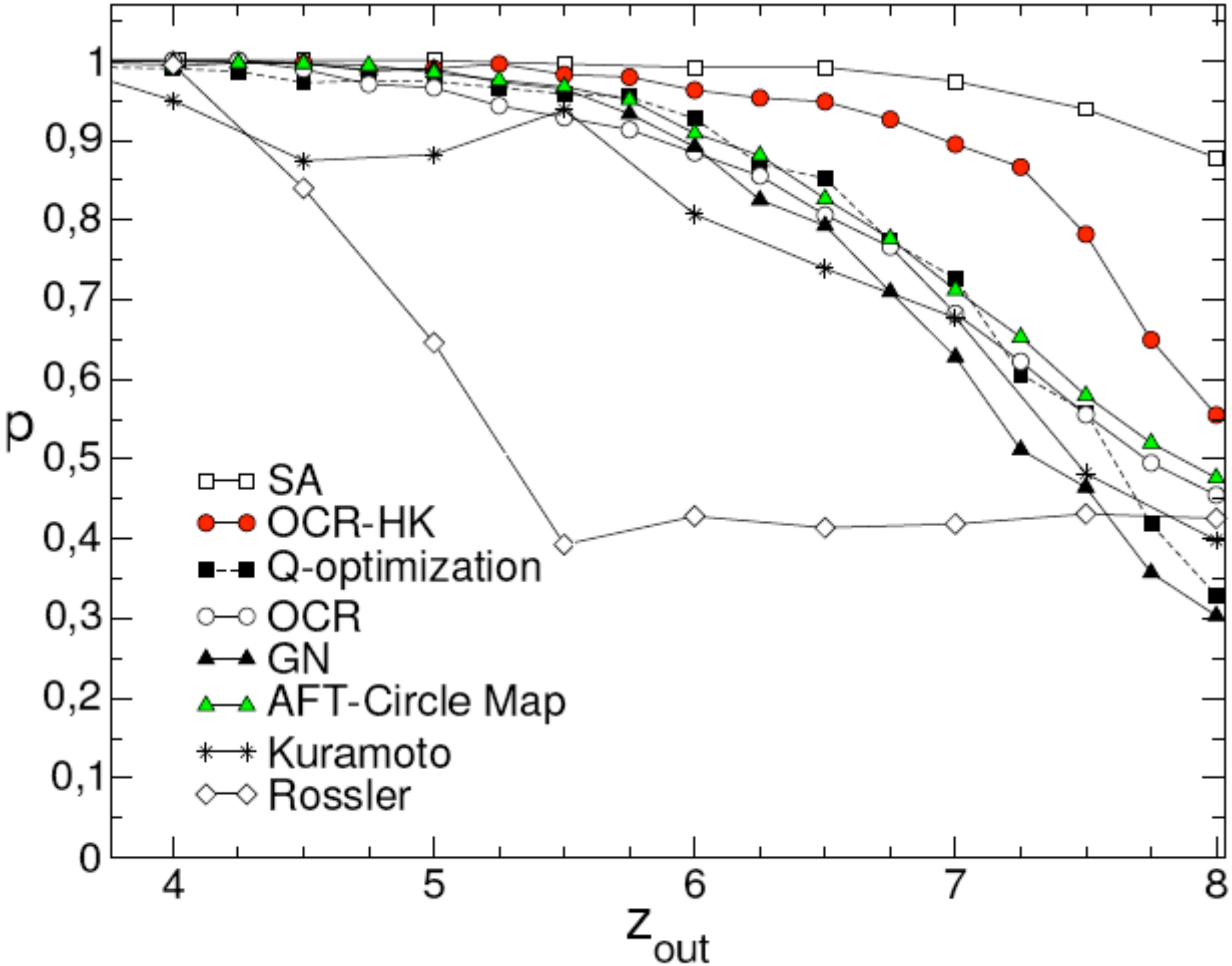
Sensitivity test



Sensitivity for different values of α-step and confidence bound



Sensitivity tests with other dynamical systems
(Kuramoto, Rössler, Circle-Map)



Dynamical Clustering on random trial networks

Computational cost

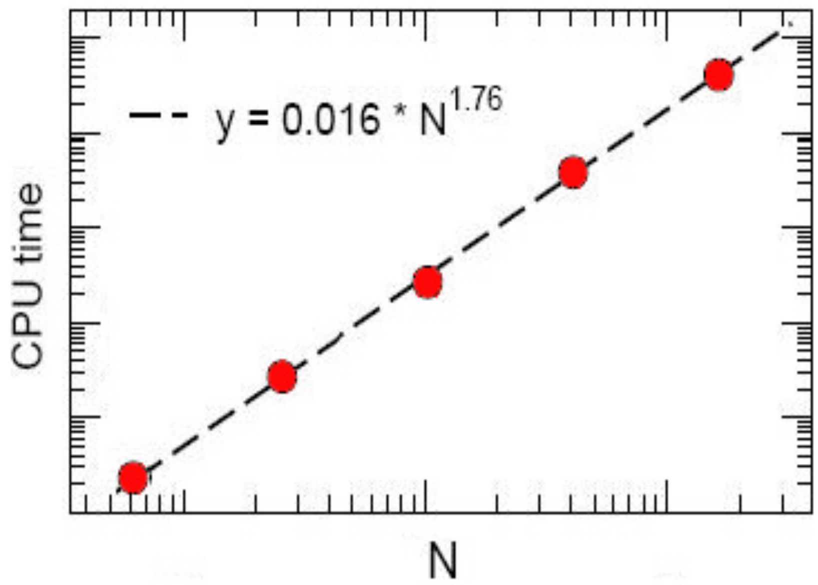
1.initial betwenness calculation: $O(N^2)$

+

2.dynamical clustering evolution time: $O(N^{1.76})$

very low global computational cost

$O(\sim N^2)$



L.Danon et al., *J.of Stat.Mech.: Theory and Exp.* (2005)

Author	Ref.	Label	Order
Eckmann & Moses	[13]	EM	$O(m\langle k^2\rangle)$
Zhou & Lipowsky	[14]	ZL	$O(n^3)$
Latapy & Pons	[15]	LP	$O(n^3)$
Newman	[24]	NF	$O(n\log^2 n)$
Newman & Girvan	[25]	NG	$O(m^2 n)$
Girvan & Newman	[32]	GN	$O(n^2 m)$
Guimerà et al.	[27, 43]	SA	parameter dependent
Duch & Arenas	[31]	DA	$O(n^2\log n)$
Fortunato et al.	[33]	FLM	$O(n^4)$
Radicchi et al.	[34]	RCCLP	$O(n^2)$
Donetti & Muñoz	[35, 36]	DM/DMN	$O(n^3)$
Bagrow & Bollt	[37]	BB	$O(n^3)$
Capocci et al.	[38]	CSCC	$O(n^2)$
Wu & Huberman	[39]	WH	$O(n+m)$
Palla et al.	[40]	PK	$O(\exp(n))$
Reichardt & Bornholdt	[41]	RB	parameter dependent

The best identification methods scales with the network size as $O(N\log^2 N)$

Summary

- The problem of finding the best modular subdivision of a network is fundamental but it is also a formidable task

- Divisive topological methods have a good sensitivity but have also an high computational cost

- We developed a new algorithm based on a dynamical clustering tecnique that shows a very high sensitivity both for real and trial networks, and at the same time is very fast

- It makes an interesting bridge between researches in complex network and those in synchronization of dynamical systems

- Further investigations are in progress and regard the application of our algorithm to larger real networks (also weighted and/or directed ones) and to networks with overlapping or nested communities

Thank you for the attention!

CACTUS
Chaos And Complexity Theoretical University Study
Group
Catania

http://www.ct.infn.it/cactus

Netlogo Simulations Lab: http://www.ct.infn.it/cactus/simulab.html

Some references

A.Pluchino, A.Rapisarda and V.Latora, arXiv: 0806.4276v2, Eur. Phys. J. B in press

S.Boccaletti, M.Ivanchenko, V.Latora, A.Pluchino and A.Rapisarda - Physical Review E 75 (2007) 045102(R)

A.Pluchino and A.Rapisarda, Proocedings of American Institute of Physics 965 (2007) p.323 (arXiv:0711.1726)

A.Pluchino, V.Latora, A.Rapisarda, Eur. Phys. J. B 50 (2006) 169 (arXiv:physics/0510141)

A.Pluchino, V.Latora, A.Rapisarda, *Int.Journ.of Mod.Phys. C* 16 (2005) 515 (arXiv:cond-mat/0410217)

S.Fortunato, V.Latora, A.Pluchino, A.Rapisarda, *Int.Journ.of Mod.Phys. C 16 (2005) 1535* (arXiv:physics/0504017)

Further Tests with other dynamical systems (in progress...)

AVT algorithm: _ **variable in time**

AFT algorithm: _ **fixed in time**

Kuramoto's non identical 1D oscillators

$$\dot{\vartheta}_i = \omega_i + \frac{\sigma}{\sum_{j\in K_i} l_{ij}^{\alpha(t)}} \sum_{j\in K_i} l_{ij}^{\alpha(t)} \sin(\theta_j - \theta_i) \qquad i = 1,\dots,N$$

Chaotic Rössler identical 3D oscillators

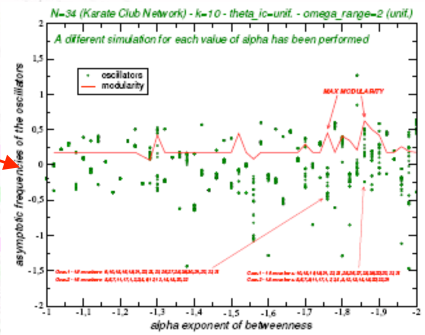
$$\begin{cases} \dot{x}_i = -\omega y_i - z_i - \dfrac{\sigma}{\sum_{j\in K_i} l_{ij}^{\alpha(t)}} \sum_{j\in K_i} l_{ij}^{\alpha(t)} (x_i - x_j) \\ \dot{y}_i = \omega x_i + 0.165 y_i \qquad\qquad i = 1,\dots,N \\ \dot{z}_i = 0.2 + z_i (x_i - 10) \end{cases}$$

Sine-Circle Map: non identical 1D oscillators

$$x_i(n+1) = x_i(n) + \omega_i + \frac{\sigma}{\sum_{j\in K_i} l_{ij}^{\alpha(t)}} \sum_{j\in K_i} l_{ij}^{\alpha(t)} \sin(x_j - x_i) \qquad i = 1,\dots,N$$

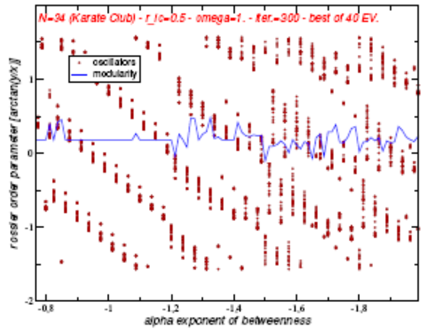
Karate Club

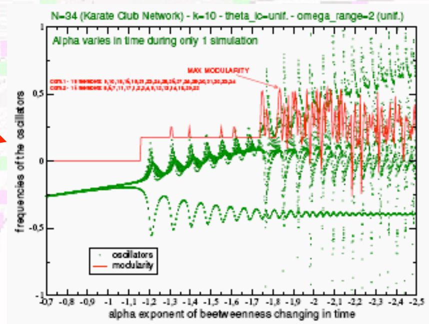
AFT



Karate Club

AVT



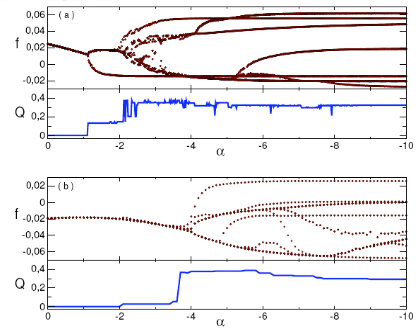


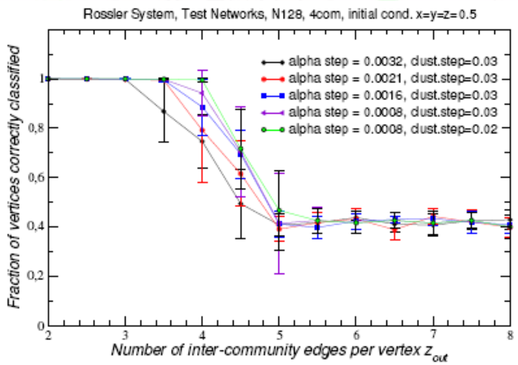
Karate Club

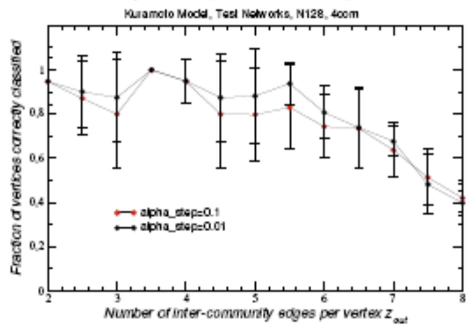
AFT

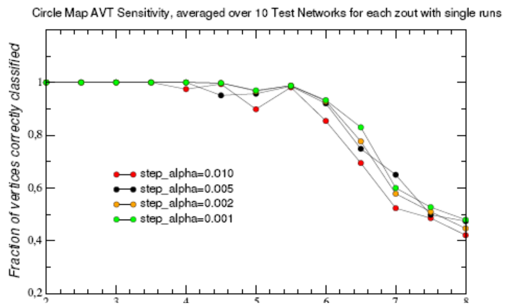
Chesapeake Bay

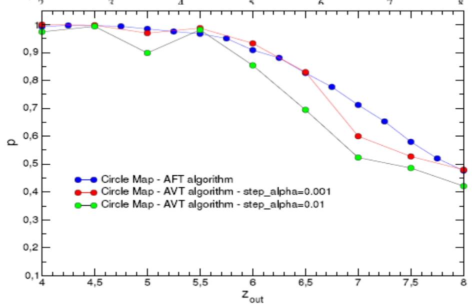
AFT









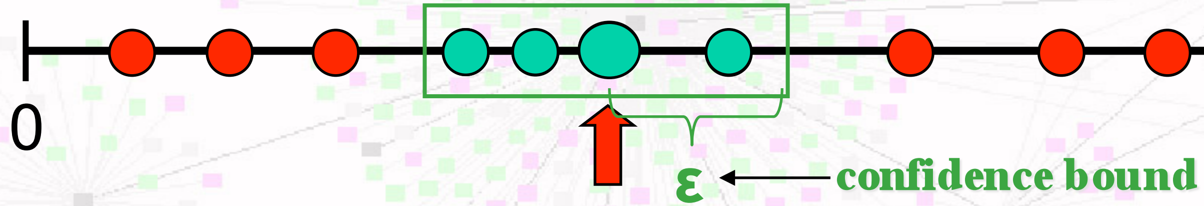


S.Boccaletti, M.Ivanchenko, V.Latora, A.P. and A.Rapisarda – in preparation

Heigselmann-Krause Dynamics: the OCR-HK model

The Hegselmann-Krause (HK) opinion dynamics* is based on the presence of a parameter **e**, called "confidence bound", which expresses the range of compatibility of the opinions.

The 1-D opinion space is represented by the points of a [0,1] line, where the opinions are uniformly distributed:



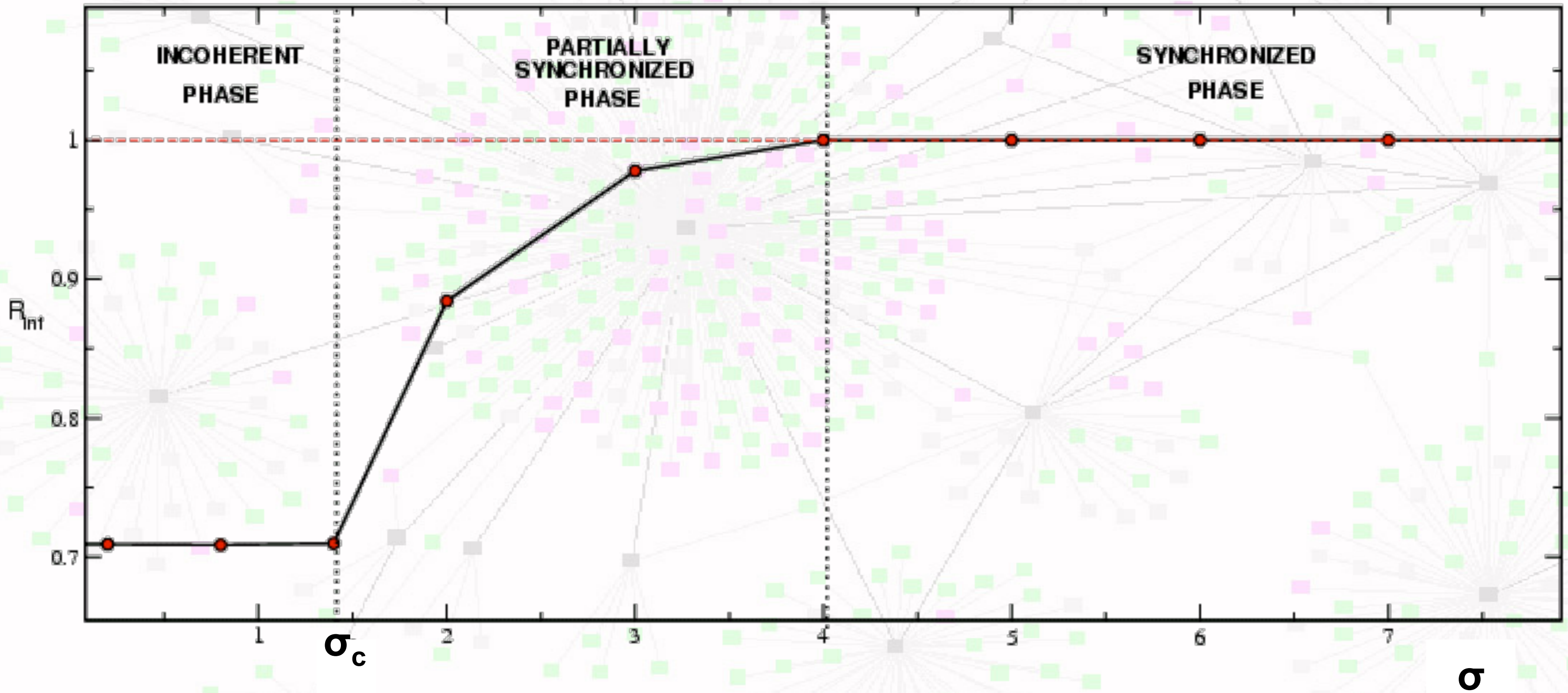
At each step, one chooses at random one opinion and checks how many opinions are compatible with him, i.e. are inside the confidence bound…

…at the next step, the agent takes the <u>average opinion</u> of its compatible neighbours...

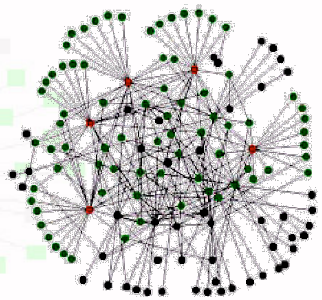
*R. Hegselmann and U. Krause, *Journal of Articial Societies and Social Simulation 5*, issue 3, paper 2 (jasss.soc.surrey.ac.uk) (2002);

Phase transition for the asymptotic order parameter R_∞ at σ_C~1.4

$$R(t) = 1 - VAR(x_i)$$



OCR-HK: Dynamical Clustering Algorithm



istantaneous frequencies (opinion changing rates)

loads (betweennesses)

tuning parameter

$$\dot{x}_i(t) = \omega_i(t) + \frac{\sigma}{\sum_{j \in K_i} l_{ij}^{\alpha(t)}} \sum_{j \in K_i}^{N} \beta \, l_{ij}^{\alpha(t)} \sin(x_j - x_i) \, e^{-\beta |x_j - x_i|}, \qquad i = 1, \dots, N$$

neighbours of node -i in the selected netwrok

intrinsic frequencies, updated with HK dynamics

1.We start at α=0 from a state with uniformly distributed frequencies which rapidly synchronize (since we set $\sigma > \sigma_c$);

2.We let α to decrease in time during a single run and we look desynchronizes and we look for clusters in frequency;

3.We repeat the procedure for several runs, with different initial frequency distributions, then we select the configuration with the highest score of modularity Q