Graph Wavelets and Multiscale Community Mining in networks

Pierre Borgnat, Nicolas Tremblay

CR1 CNRS – Laboratoire de Physique, ENS de Lyon, Université de Lyon

Équipe SISYPHE : Signaux, Systèmes et Physique

08/2014

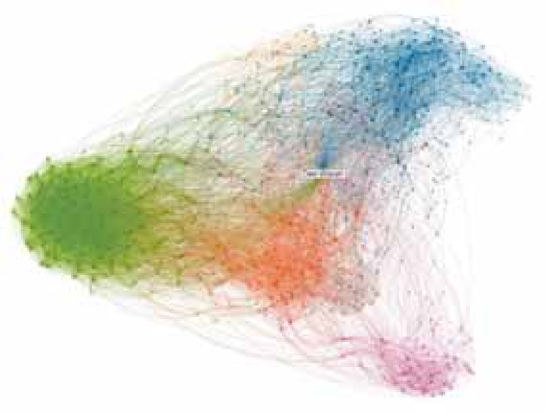
ENS DE LYON

Content of the talk

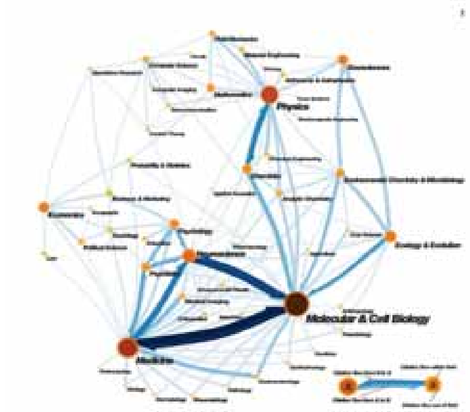
- General objective: revisit the classical question of finding communities in networks using multiscale processing methods on graphs.
- The things that will be discussed:

1. Recall the notion of community in networks
2. Recall spectral graph wavelets
3. Multiscale community mining with graph wavelets

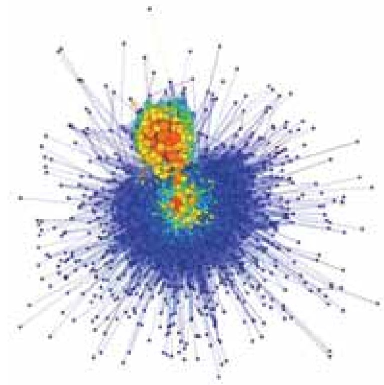
Examples of networks from our digital world



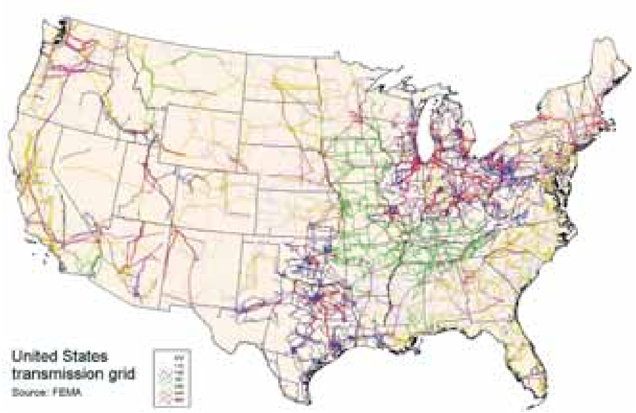
LinkedIn Network



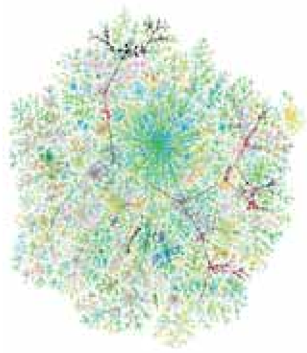
Citation Graph



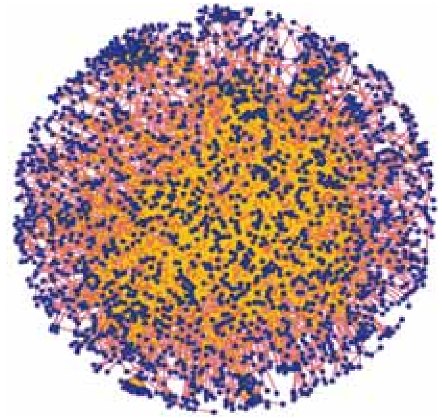
Vehicle Network



USA Power grid



Web Graph



Protein Network

Communities in networks

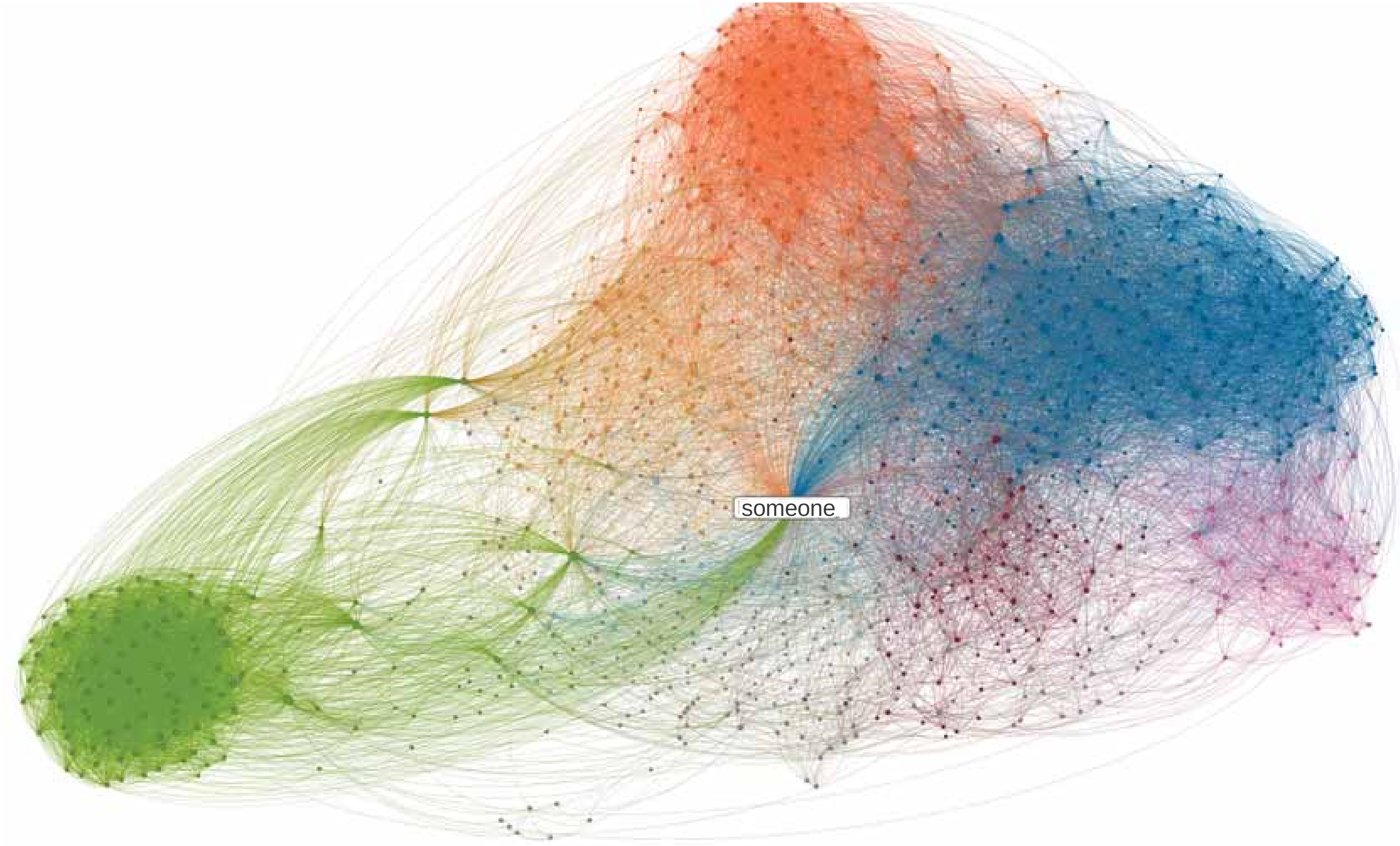
- Observed, real-world, networks are often inhomogeneous, made of communities (or modules): *groups of nodes having a larger proportion of links inside the group than with the outside*
- This is observed in various types of networks: social, technological, biological,...
- There exist several extensive surveys:

  [S. Fortunato, *Physic Reports*, 2010]

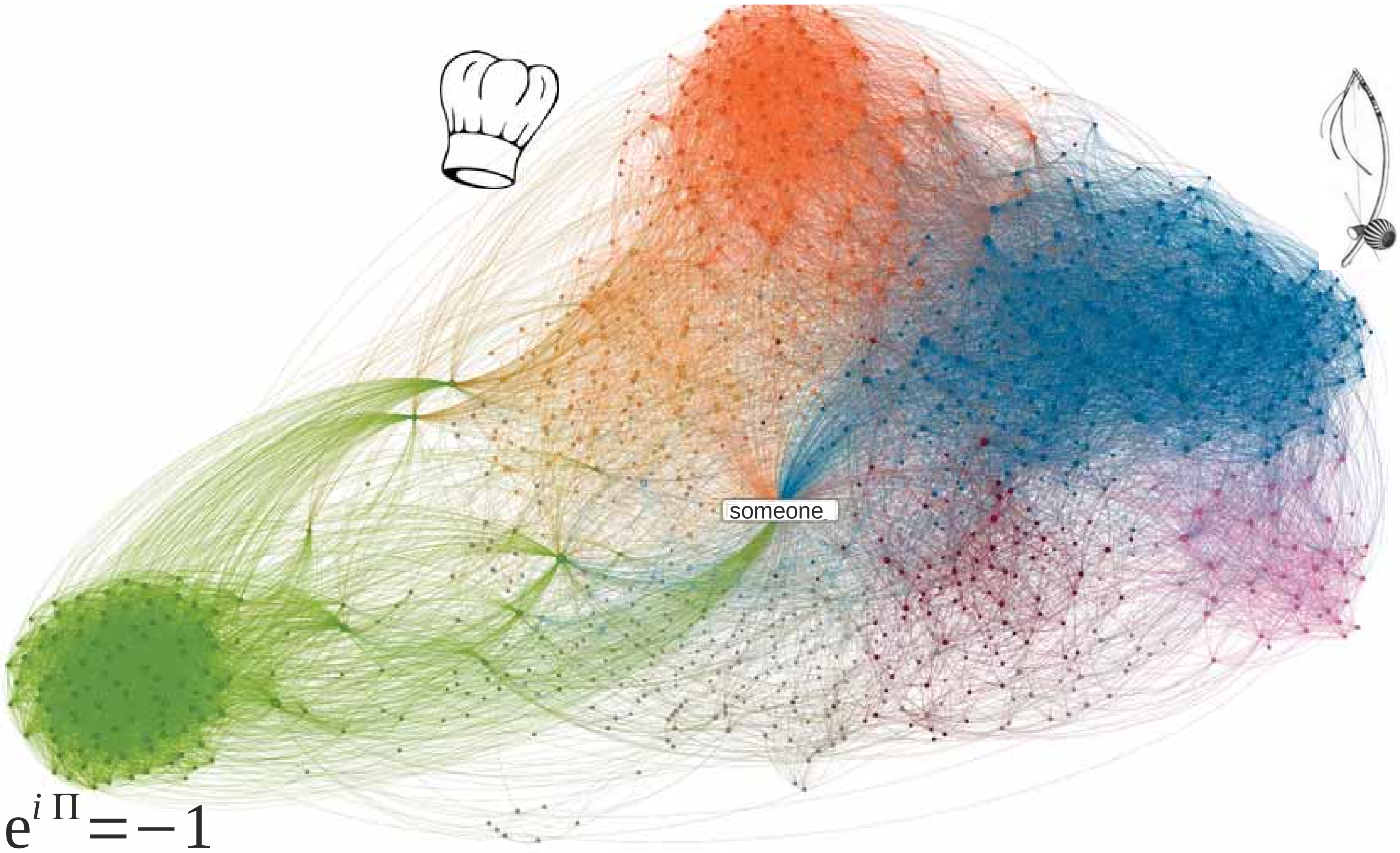
  [von Luxburg, *Statistics and Computating*, 2007]

  ...

Purpose of community detection?

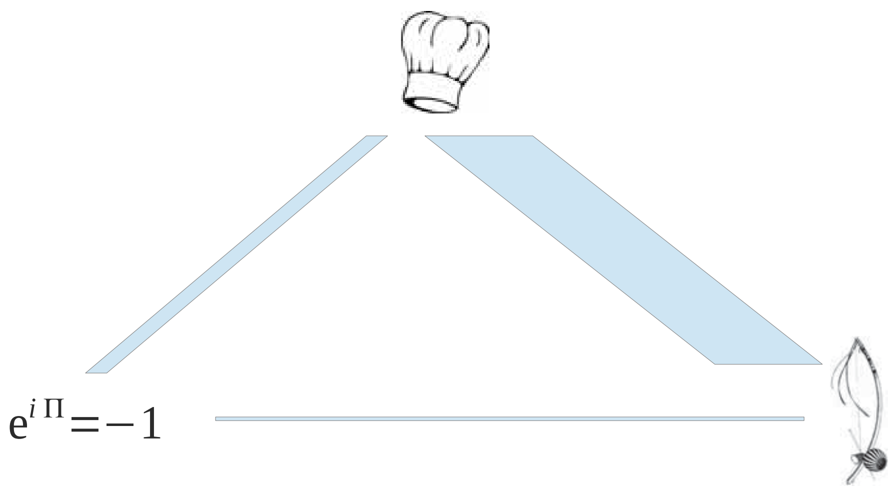


Purpose of community detection?



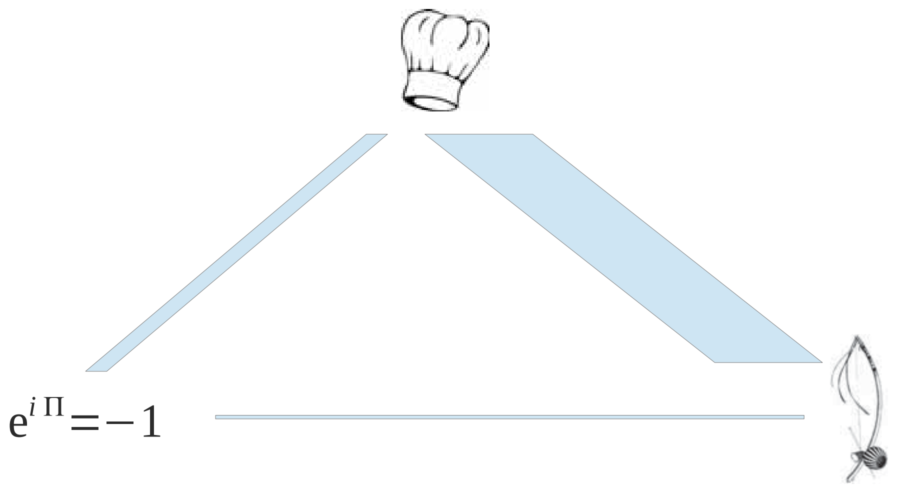
Purpose of community detection?

1) It gives us a sketch of the network:



Purpose of community detection?

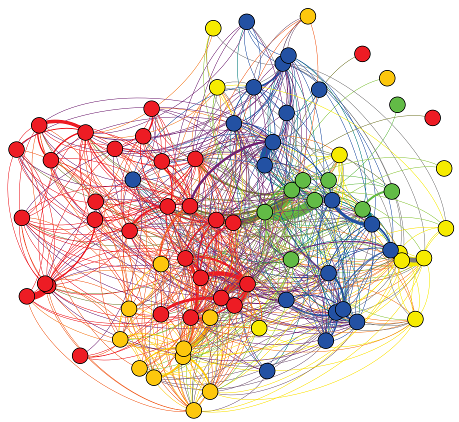
1) It gives us a sketch of the network:



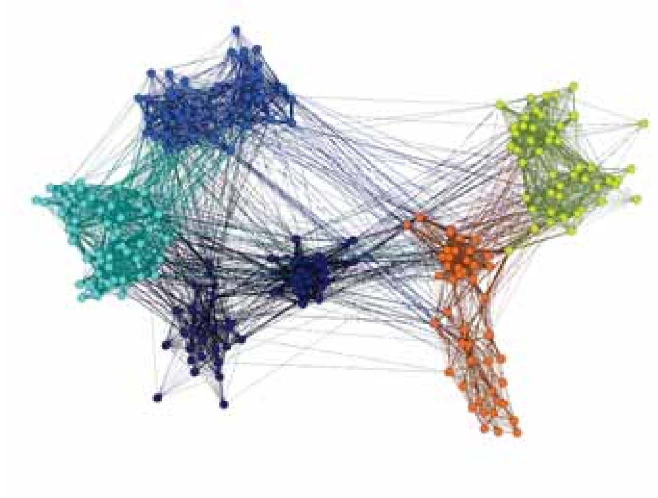
2) It gives us intuition about its components:

Some examples of networks with communities or modules

- Social face-to-face interaction networks [Sociopatterns; Barrat, Cattuto, et al.]

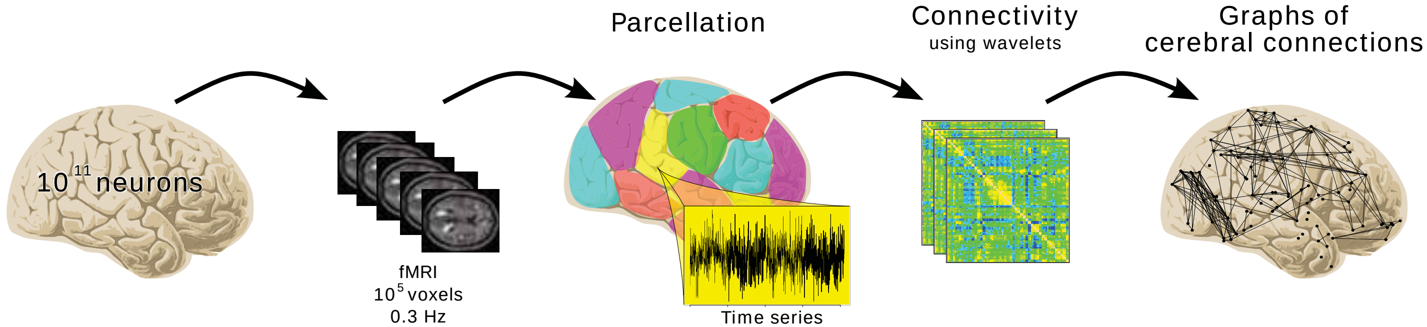


(Lab. physique, ENSL, 2013)



(école primaire; Sociopatterns, 2011)

- Brain networks [Bullmore, Achard, 2006]

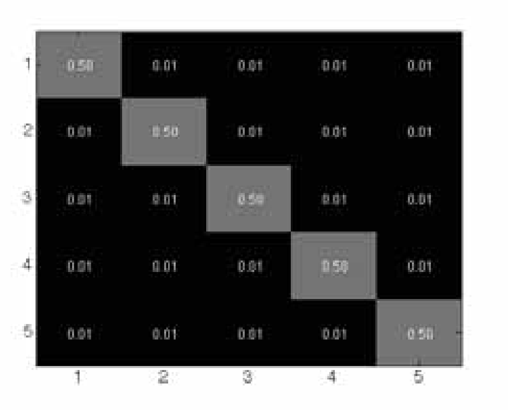


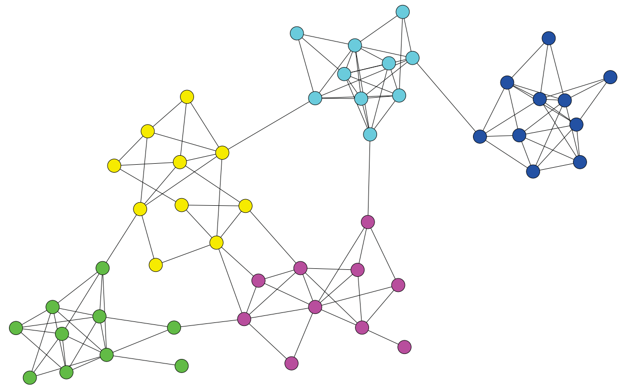
Classical methods to find communities in networks

- I will not pretend to make a full survey... Some important steps are:
- Cut algorithms (legacy from computer science)
- Spectral clustering (relaxed cut problem)
- Modularity optimization (physicists' contribution) [Newman, Girvan , 2004]
- Greedy modularity optimization a la Louvain (computer science strikes back) [Blondel et al., 2008]
- Using information compression [Rosvall, Bergstrom, 2008]
- Inference for stochastic-block models (e.g. with BP [Decelle et al., 2012]; with spectral approach [Lelarge, Massoulié,... 2012, 2014])

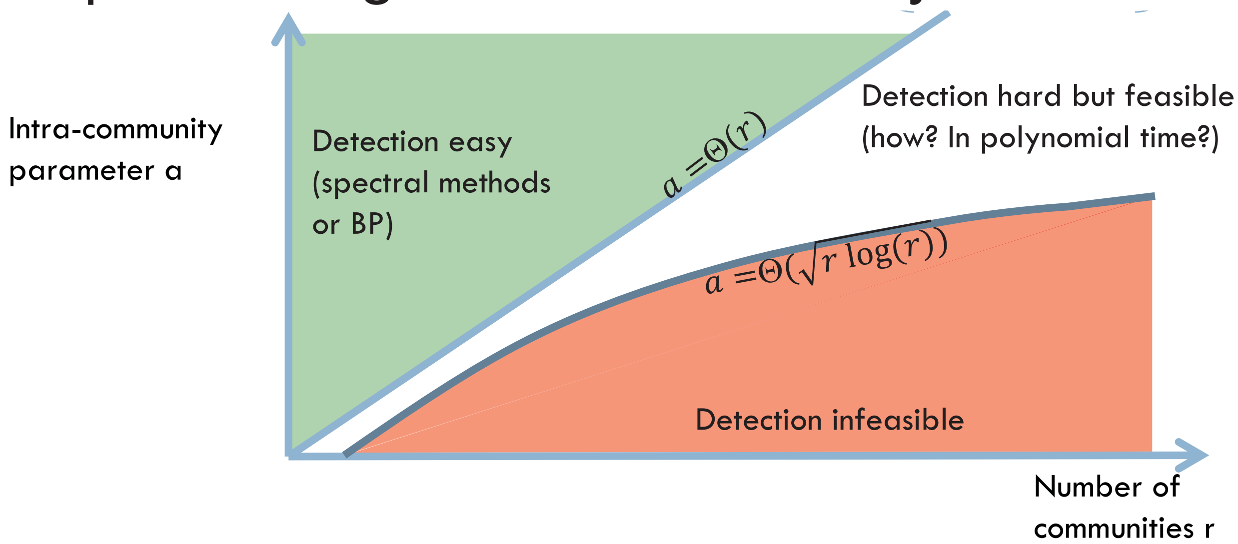
Parenthesis: Stochastic Block Model

- Representation: as a matrix, as a network





- Conjectured phase diagram of identifiability



[Decelle, Krzakala, Moore, Zdeborova, 2011]
[Lelarge, Massoulié, Xu, 2013]

Spectral analysis of networks

Spectral theory for network

This is the study of graphs through the **spectral analysis** (eigenvalues, eigenvectors) of matrices **related to the graph**: the adjacency matrix, the Laplacian matrices,....

Notations

$\mathcal{G} = (V, E, w)$	a weighted graph	
$N = \lvert V \rvert$	number of nodes	
A	adjacency matrix	$A_{ij} = w_{ij}$
d	vector of strengths	$d_i = \sum_{j \in V} w_{ij}$
D	matrix of strengths	$D = diag(d)$
f	signal (vector) defined on V	

Definition of the Laplacian matrix of graphs

Laplacian matrix

L	laplacian matrix	$L = D - A$
(λ_i)	L's eigenvalues	$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{N-1}$
(χ_i)	L's eigenvectors	$L\chi_i = \lambda_i \chi_i$

Note: $\chi_0 = \mathbf{1}$.

A simple example: the straight line

$$\longleftrightarrow \quad L = \begin{pmatrix} \cdots & & & & & & \\ \cdots & -1 & 0 & 0 & 0 & 0 & \\ \cdots & 2 & -1 & 0 & 0 & 0 & \\ & -1 & 2 & -1 & 0 & 0 & \\ & 0 & -1 & 2 & -1 & 0 & \\ & 0 & 0 & -1 & 2 & -1 & \\ & 0 & 0 & 0 & -1 & 2 & \cdots \\ & 0 & 0 & 0 & 0 & -1 & \cdots \\ & & & & & & \cdots \end{pmatrix}$$

For this regular line graph, L is the 1-D classical laplacian operator (i.e. double derivative operator).

Spectral clustering vs. Modularity

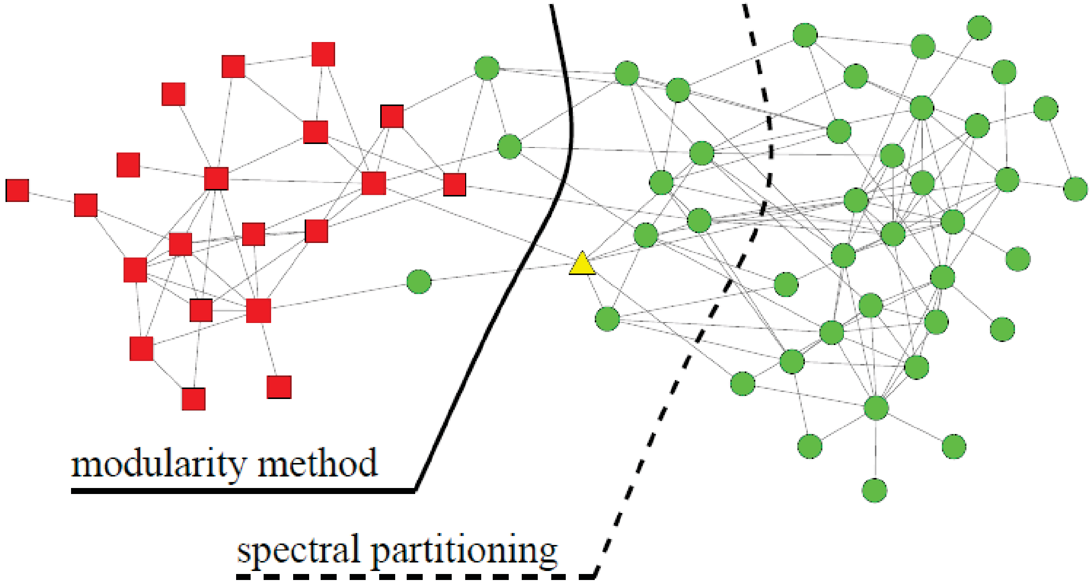
- **Spectral clustering**: relaxation of the optimization of the minimal cut. The cut size between groups of assignment $s_i = \pm 1$ is: $R = \frac{1}{2}\sum_{i,j} A_{ij}(1 - s_i s_j) = \frac{1}{4}\mathbf{s}^\top L\mathbf{s}$
- By spectral decomposition of L, $L_{ij} = \sum_{k=1}^{N-1} \lambda_k (\chi_k)_i (\chi_k)_j$, the minimum is for $s_i = (\chi_1)_i \ \rightarrow$ relaxed in $s_i = \mathrm{sign}((\chi_1)_\mathrm{i})$.
- Problems with spectral clustering:
  1) No assessment of the quality of the partitions
  2) No reference to comparison to some null hypothesis
- **Modularity** [Newman, 2003] (with $2m = \sum_i d_i$)

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{d_i d_j}{2m}\right]\delta(c_i, c_j)$$

- Null model: Bernoulli random graph with prob. $\frac{d_i d_j}{2m}$
- Q is between -1 and $+1$ ($\leq 1 - 1/n_c$ if n_c groups)

Spectral clustering vs. Modularity

- Comparison of optimization of cut and optimization of Q
- Modularity works well, better than spectral clustering



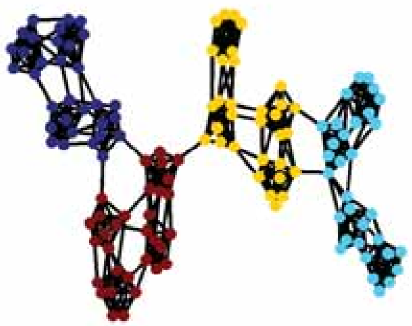
- More efficient algorithm: the greedy (ascending) Louvain approach (ok for millions of nodes !) [Blondel et al., 2008]

Existence of multiscale community structure in a graph

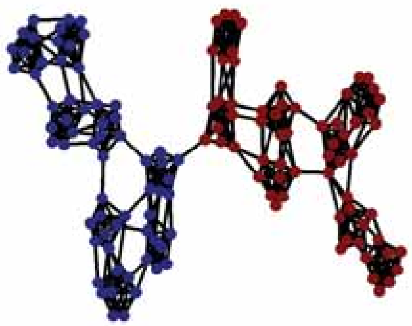
16 com. Q=0.80

8 com. Q=0.83

4 com. Q=0.74



2 com. Q=0.50



- All representations correct; modularity favours one
- Note: one could integrate a ad-hoc scale into modularity [Arenas et al., 2008; Reichardt and Bornholdt, 2006]

Relating the Laplacian of graphs to Signal Processing

Laplacian matrix

L or $\mathscr{L}$	laplacian matrix	$L = D - A$ or $\mathscr{L} = I - D^{-1/2} A D^{-1/2}$
(λ_i)	L's eigenvalues	$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{N-1}$
(χ_i)	L's eigenvectors	$L\,\chi_i = \lambda_i\,\chi_i$

A simple example: the straight line

$$\longleftrightarrow \quad L = \begin{pmatrix} \ddots & & & & & & \\ \ldots & -1 & 0 & 0 & 0 & 0 & \\ \ldots & 2 & -1 & 0 & 0 & 0 & \\ & -1 & 2 & -1 & 0 & 0 & \\ & 0 & -1 & 2 & -1 & 0 & \\ & 0 & 0 & -1 & 2 & -1 & \\ & 0 & 0 & 0 & -1 & 2 & \ldots \\ & 0 & 0 & 0 & 0 & -1 & \ldots \\ & & & & & & \ldots \end{pmatrix}$$

For this regular line graph, L is the 1-D classical laplacian operator (i.e. double derivative operator):

its eigenvectors are the Fourier vectors, and its eigenvalues the associated (squared) frequencies

Objective and Fundamental analogy

[Shuman, Vandergheynst et al., *IEEE SP Mag*, 2013]

Objective: Definition of a Fourier Transform adapted to graph signals

f : signal defined on V $\quad\longleftrightarrow\quad$ $\hat{f}$: Fourier transform of f

Fundamental analogy

On *any* graph, the eigenvectors χ_i of the Laplacian matrix L or $\mathscr{L}$ will be considered as the Fourier vectors, and its eigenvalues λ_i the associated (squared) frequencies.

- Works exactly for all regular graphs (+ Beltrami-Laplace)
- Conduct to natural generalizations of signal processing

The graph Fourier transform

- $\hat{f}$ is obtained from f's decomposition on the eigenvectors χ_i :

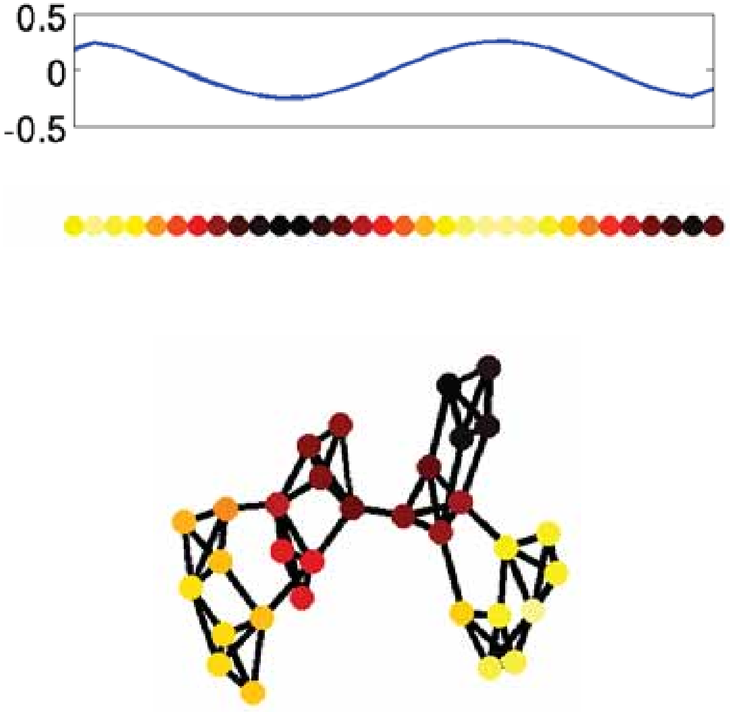
$$\hat{f} = \begin{pmatrix} < \chi_0, f > \\ < \chi_1, f > \\ < \chi_2, f > \\ \dots \\ < \chi_{N-1}, f > \end{pmatrix}$$

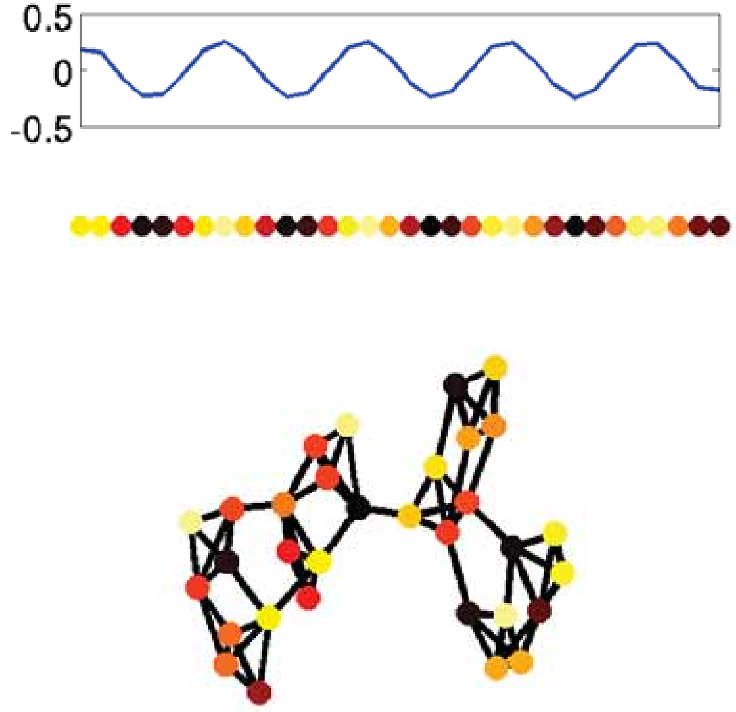
Define $\boldsymbol{\chi} = (\chi_0|\chi_1|\dots|\chi_{N-1})$: $\boxed{\hat{f} = \boldsymbol{\chi}^\top f}$

- Reciprocally, the inverse Fourier transform reads: $\boxed{f = \boldsymbol{\chi}\hat{f}}$
- Parseval theorem: $\forall(g, h) \quad < g, h > = < \hat{g}, \hat{h} >$
- Filtering: apply $g(\lambda_i)$ in the Fourier domain on the $\hat{f}(i)$.

Fourier modes: examples in 1D and in graphs

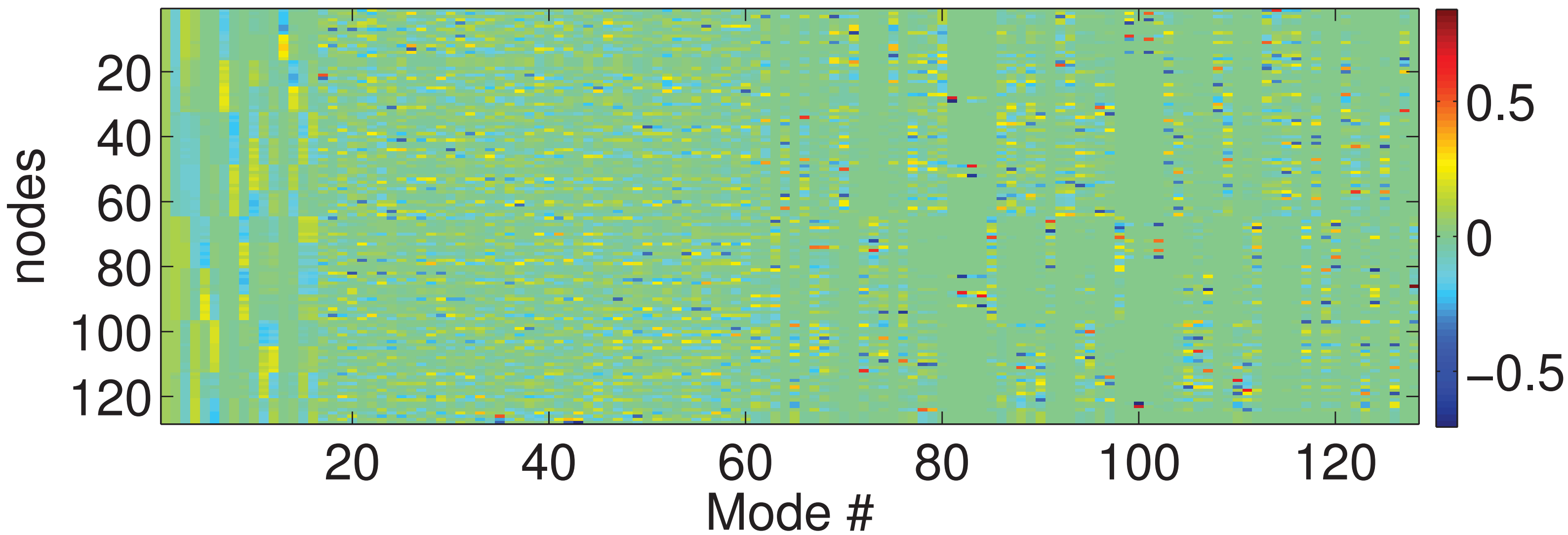
LOW FREQUENCY: HIGH FREQUENCY:

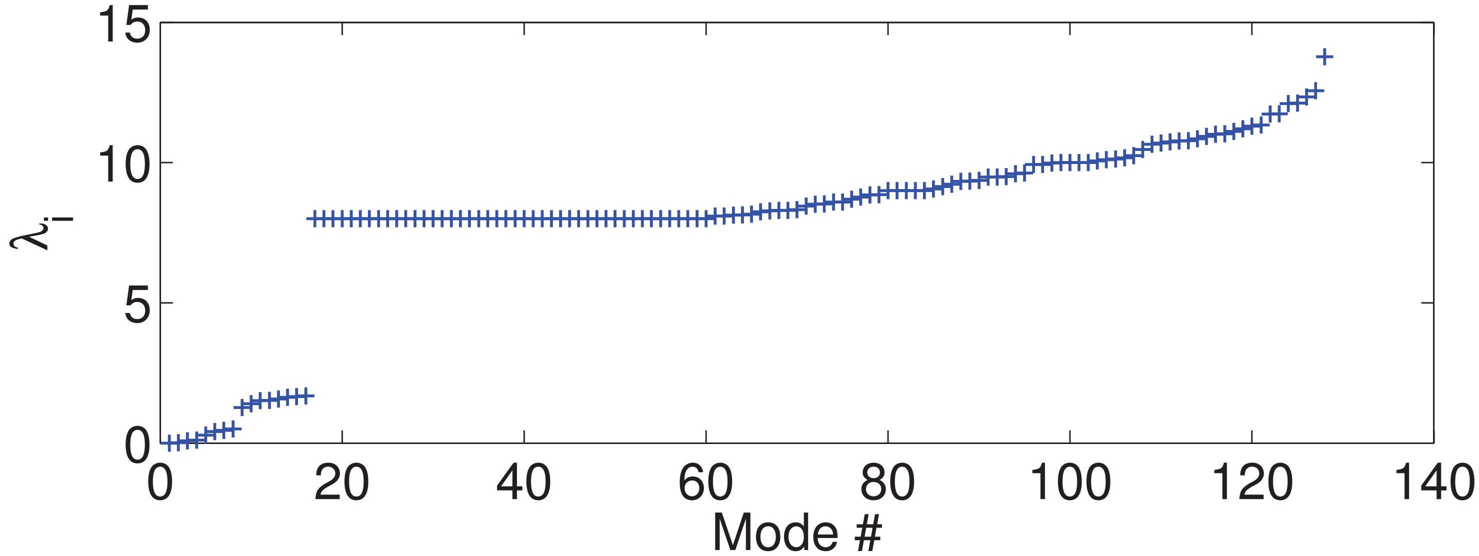




- Alternative Fourier transform: use the adjacency matrix A [Sandryhaila, Moura, *IEEE TSP*, 2013]

Spectral analysis: the χ_i and λ_i of a multiscale toy graph



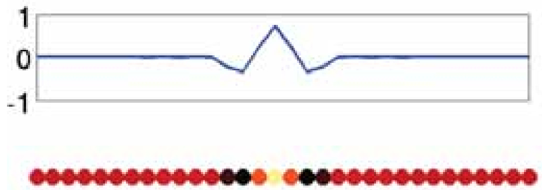


Spectral Graph Wavelets

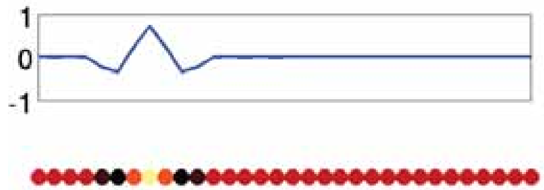
[Hammond et al., ACHA 2011]

- Fourier is a global analysis. Fourier modes (eigenvectors of the laplacian) are used in classical spectral clustering, but do not enable a jointly local and scale dependent analysis.
- For that classical signal processing (or harmonic analysis) teach us that we need **wavelets**.
- Wavelets : local functions that act as well as a filter around a chosen scale.

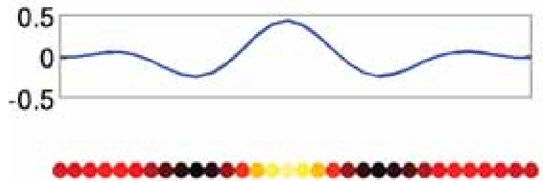
  A wavelet:



– Translated:



– Scaled



- Classical wavelets $\xrightarrow{\textit{by analogy}}$ Graph wavelets

Classical wavelets $\xrightarrow{\text{by analogy}}$ Graph wavelets

	Classical (continuous) world	Graph world
Real domain	x	node a
Fourier domain	ω	eigenvalues λ_i
Filter kernel	$\hat{\psi}(\omega)$	$g(\lambda_i) \Leftrightarrow \hat{\boldsymbol{G}}$
Filter bank	$\hat{\psi}(s\omega)$	$g(s\lambda_i) \Leftrightarrow \hat{\boldsymbol{G}}_{\boldsymbol{s}}$
Fourier modes	$\exp^{-i\omega x}$	eigenvectors χ_i
Fourier transf. of f	$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) \exp^{-i\omega x} dx$	$\hat{f} = \boldsymbol{\chi}^\top f$

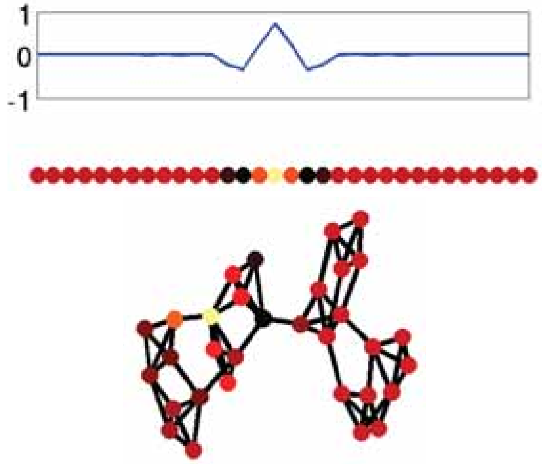
The wavelet at scale s centered around a is given by:

$$\psi_{s,a}(x) = \frac{1}{s}\psi\left(\frac{x-a}{s}\right) = \int_{-\infty}^{\infty} \hat{\delta}_a(\omega)\hat{\psi}(s\omega) \exp^{i\omega x} d\omega$$

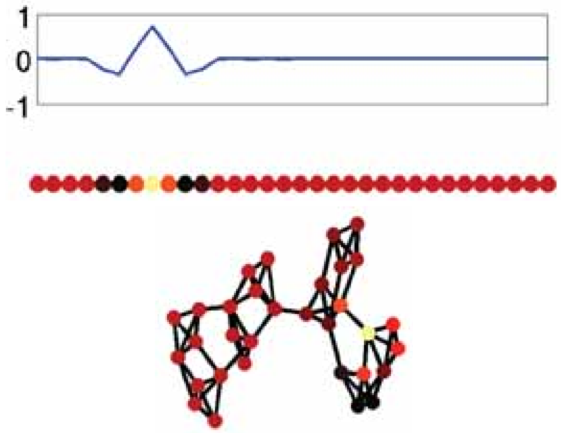
In the graph world: $\boxed{\psi_{s,a} = \boldsymbol{\chi}\, \hat{\boldsymbol{G}}_{\boldsymbol{s}} \hat{\delta}_a = \boldsymbol{\chi}\, \hat{\boldsymbol{G}}_{\boldsymbol{s}}\, \boldsymbol{\chi}^\top \delta_a}$

Examples of graph wavelets

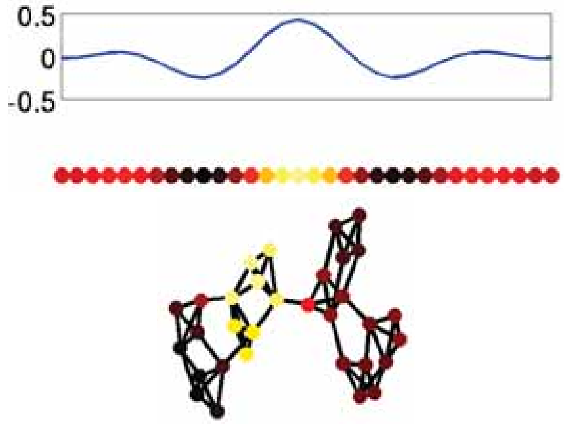
A WAVELET:



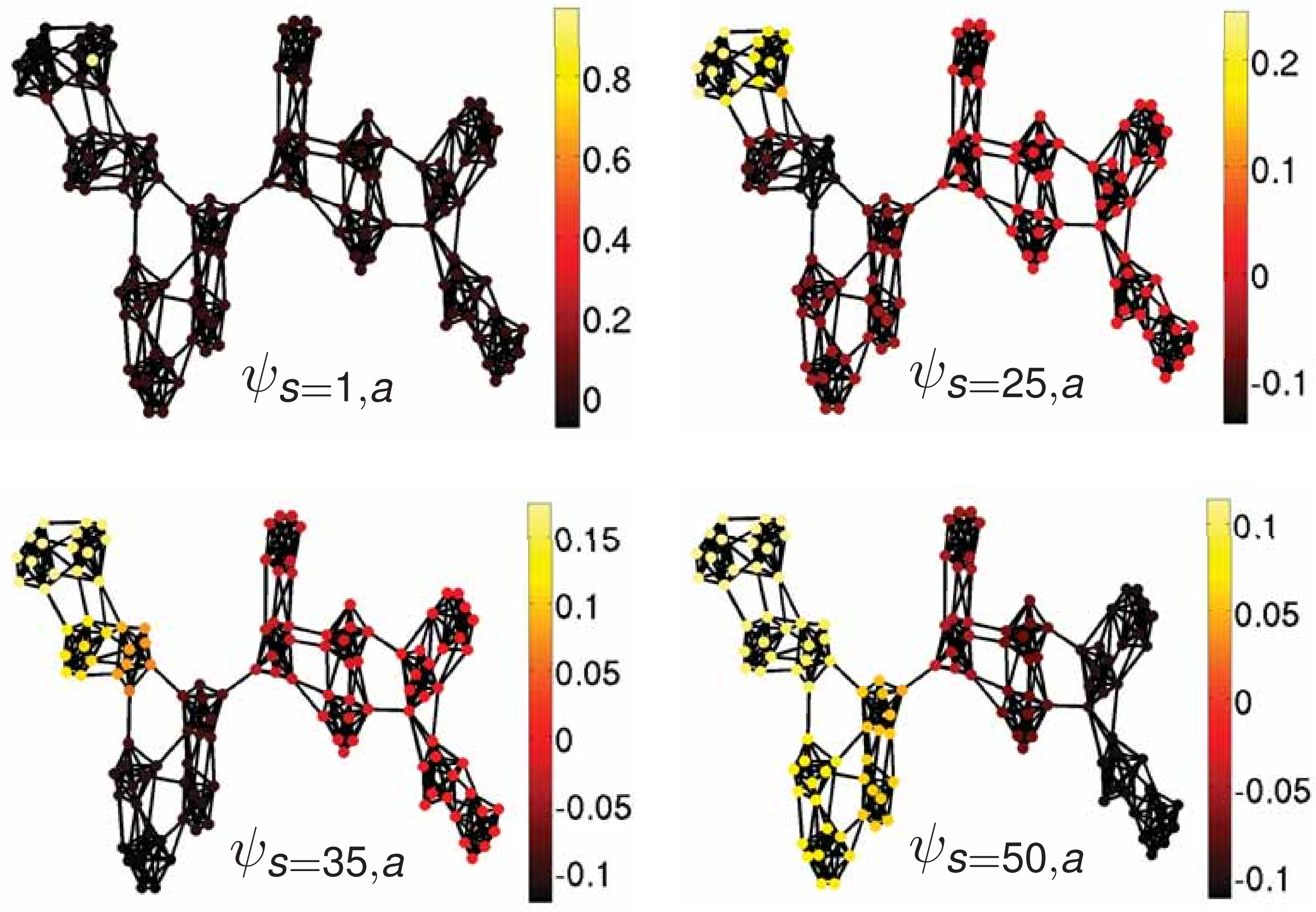
TRANSLATING:



SCALING:



Examples of wavelets: they encode the local topology



Example of wavelet filters

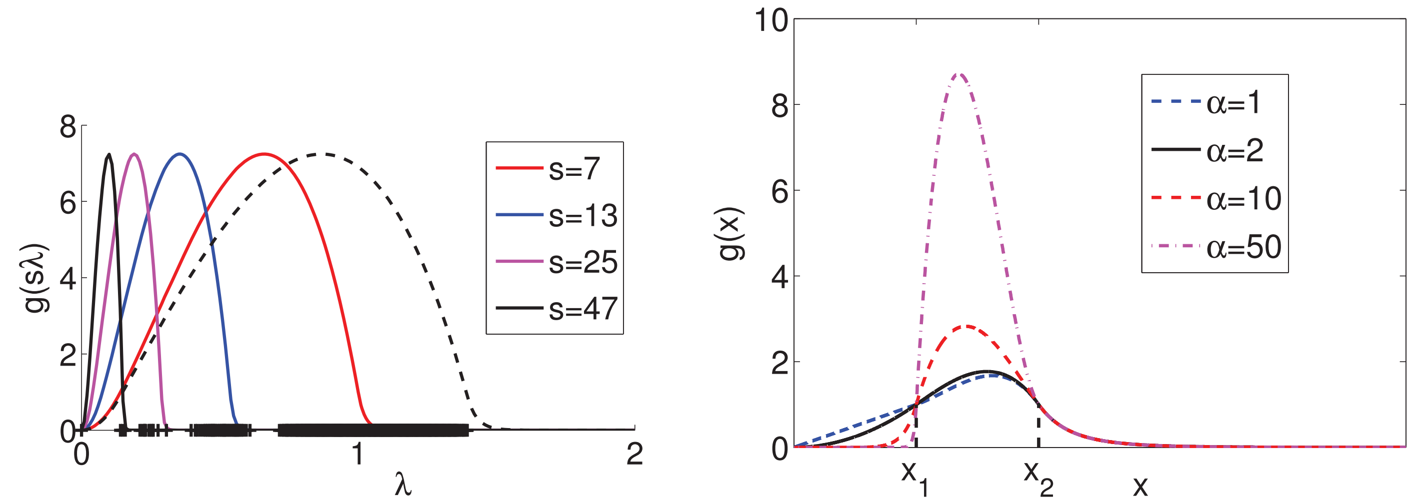
- More precisely, we will use the following kernel:

$$g(x;\alpha,\beta,x_1,x_2) = \begin{cases} x_1^{-\alpha} x^{\alpha} & \text{for} \quad x < x_1 \\ p(x) & \text{for} \quad x_1 \le x \le x_2 \\ x_2^{\beta} x^{-\beta} & \text{for} \quad x > x_2. \end{cases}$$

- To emphasize χ_1, the parameters are:

$$s_{min} = \frac{1}{\lambda_2}, \quad x_2 = \frac{1}{\lambda_2}, \quad s_{max} = \frac{1}{\lambda_2^2}, \quad x_1 = 1, \quad \beta = 1/\log_{10}\left(\frac{\lambda_3}{\lambda_2}\right)$$

- This leads to: (choice $\alpha = 2$)



A new method for multiscale community detection

[N. Tremblay, P. Borgnat, 2013]

General Ideas

- Take advantage of local topological information encoded in Graph Wavelets.
  Wavelet = ego-centered vision from a node
- Group together nodes whose local environments are similar at the description scale
- This will naturally offer a multiscale vision of communities

The method is based on:

1. wavelets (resp. scaling functions) as feature vectors
2. the correlation distance to compare them
3. the complete linkage clustering algorithm

1) Wavelets as features

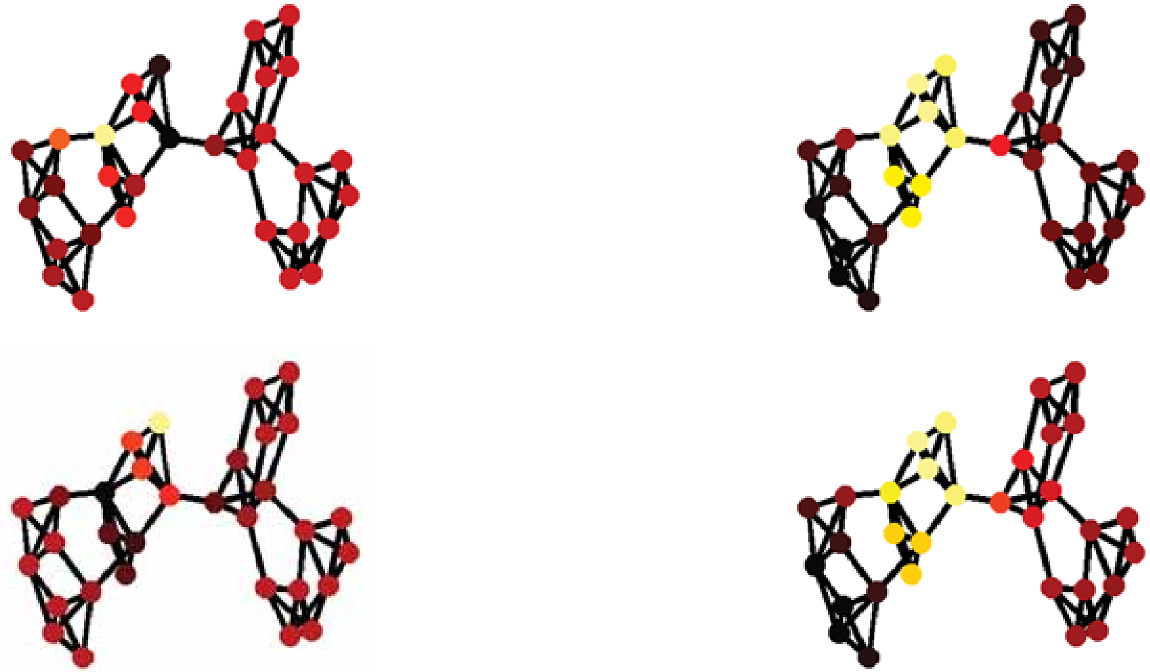
Each node a has feature vector $\psi_{s,a}$.
Globally, one will need $\boldsymbol{\Psi}_s$, all wavelets at a given scale s, i.e.

$$\boldsymbol{\Psi}_s = \left(\psi_{s,1}|\psi_{s,2}|\dots|\psi_{s,N}\right) = \chi \boldsymbol{G}_s \chi^\top.$$

AT SMALL SCALE: AT LARGE SCALE:

NODE A:

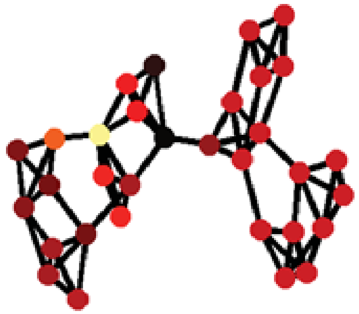
NODE B:

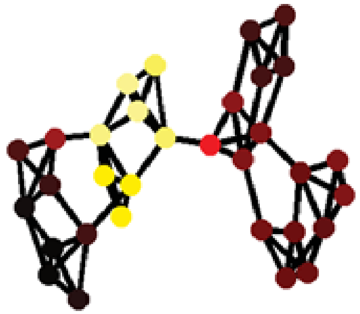


2) Correlation distances

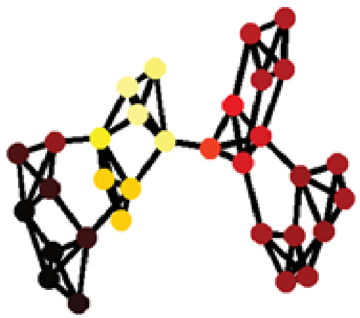
$$\boldsymbol{D}_s(a,b) = 1 - \frac{\psi_{s,a}^{\top}\psi_{s,b}}{||\psi_{s,a}||_2||\psi_{s,b}||_2}.$$

NODE A:





NODE B:



CORR. COEF.: -0.50 0.97

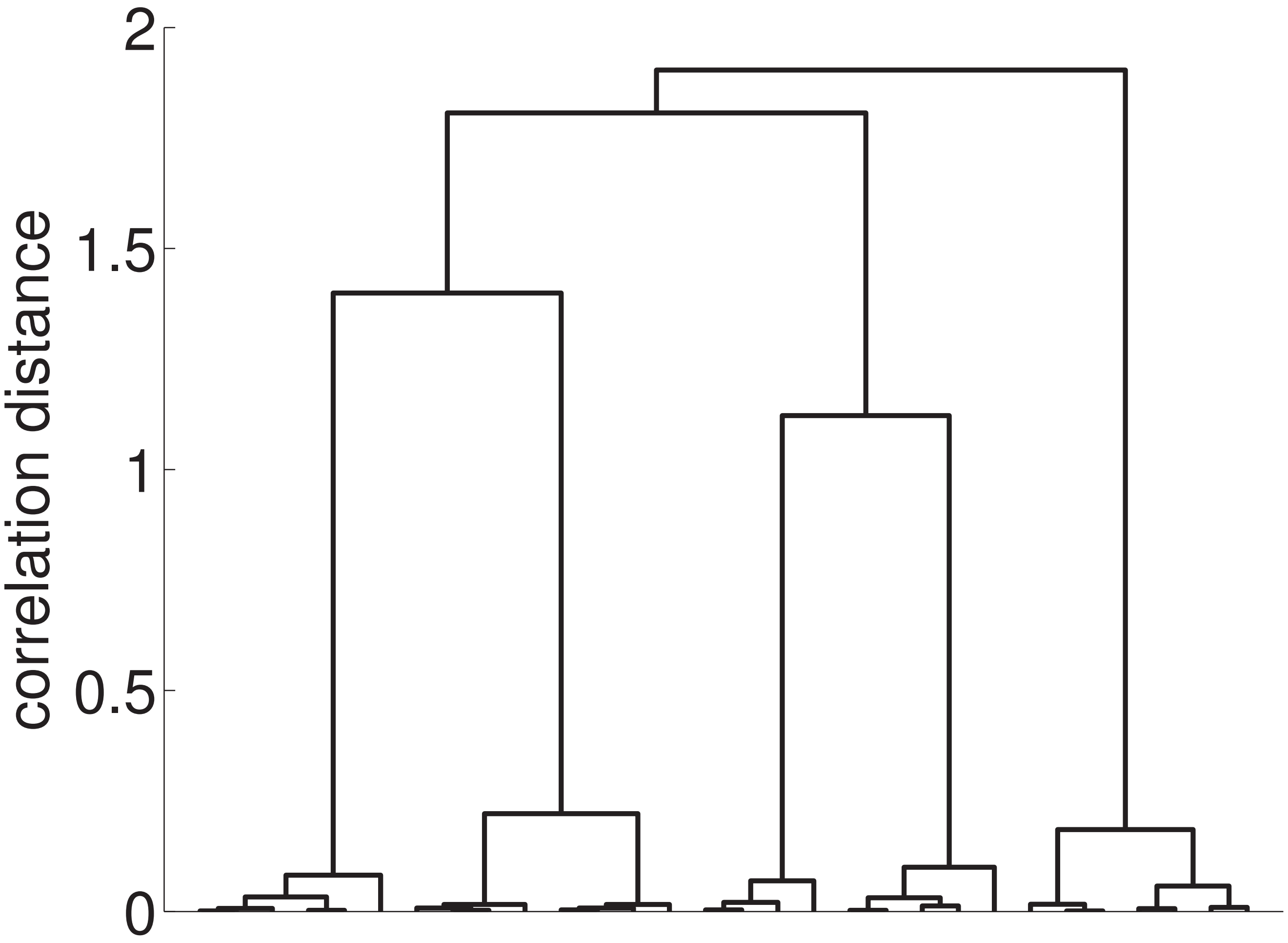
3) Complete linkage clustering and dendrogram

- Bottom to top hierarchical algorithm:
  start with as many clusters as nodes and work the way up to fewer clusters (by linking subclusters together) until reaching one global cluster.

- Computation of the distance between two subclusters:
  the maximum distance between all pairs of nodes, taking one from each cluster

- Output: a dendrogram

Example of a dendrogram **at a given scale s**

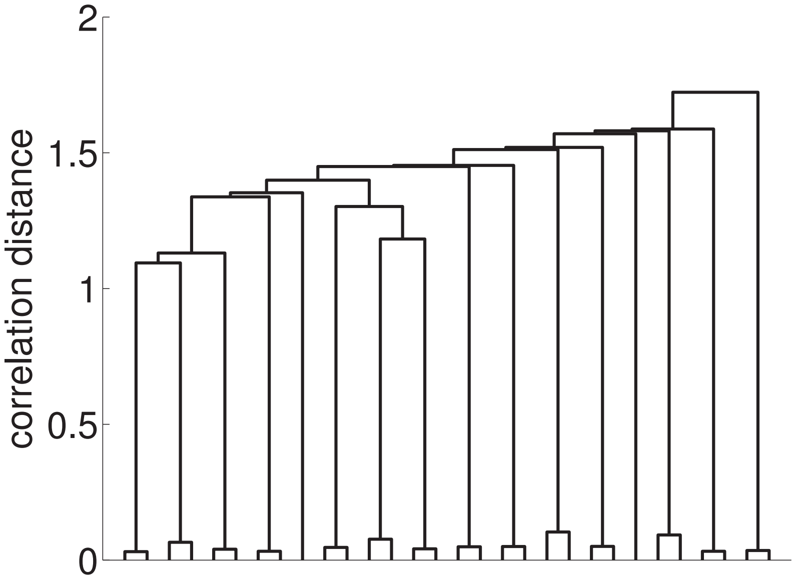


The big question: where should we cut the dendrogram?

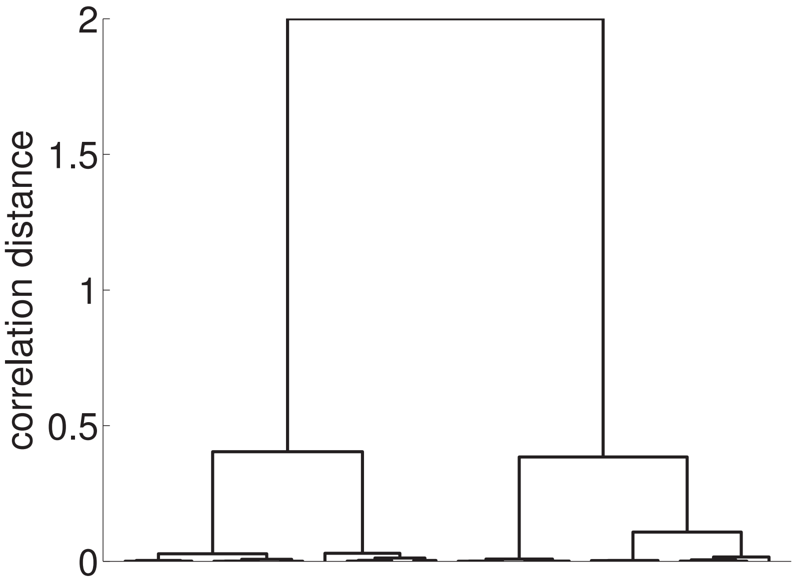
Dendrogram cut at maximal gap

Simplest method: cut the dendrogram at its maximal gap.

At small scale:



At large scale:

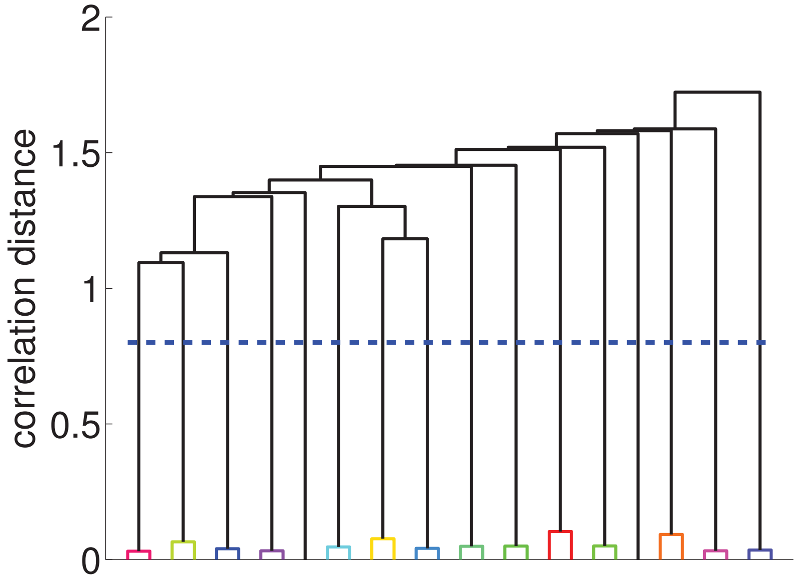


Note: we use the toy graph

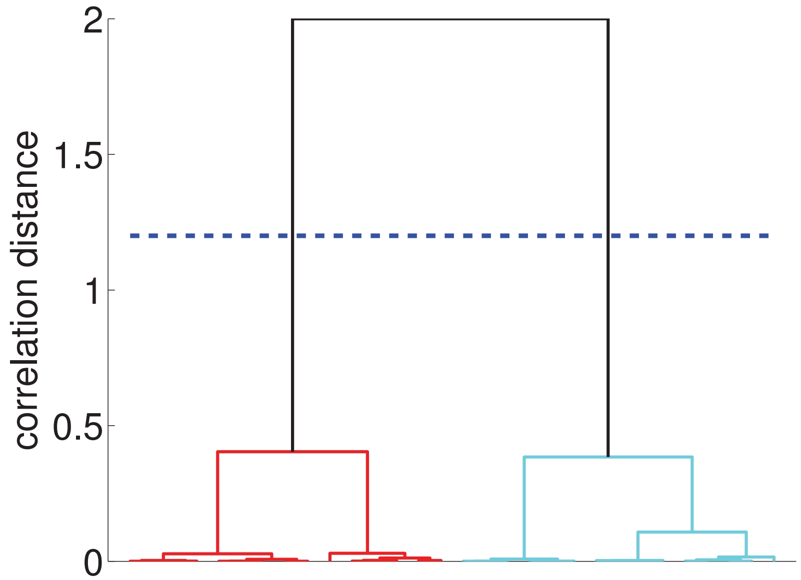
Dendrogram cut at maximal gap

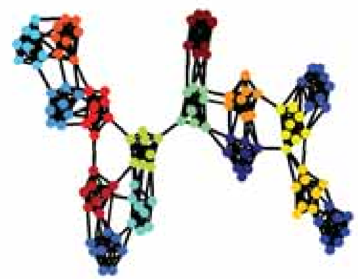
Simplest method: cut the dendrogram at its maximal gap.

At small scale:



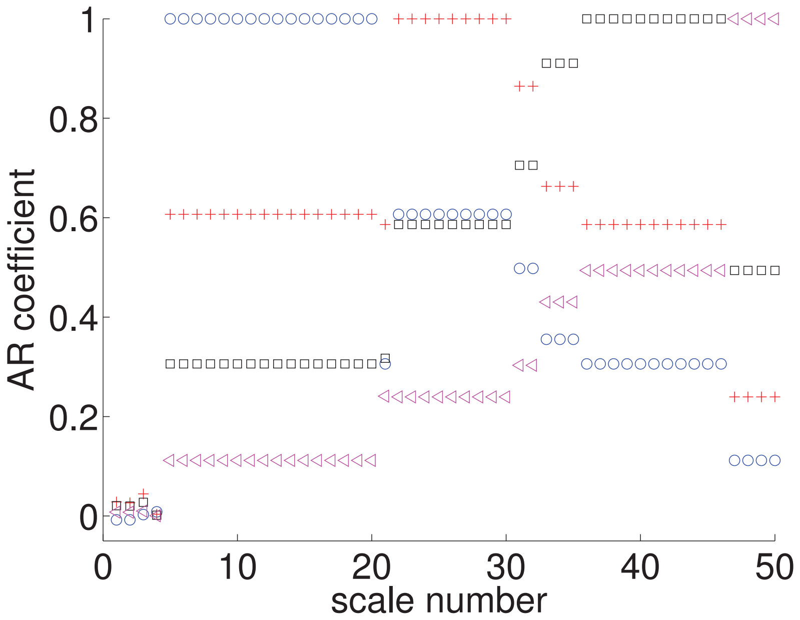
At large scale:

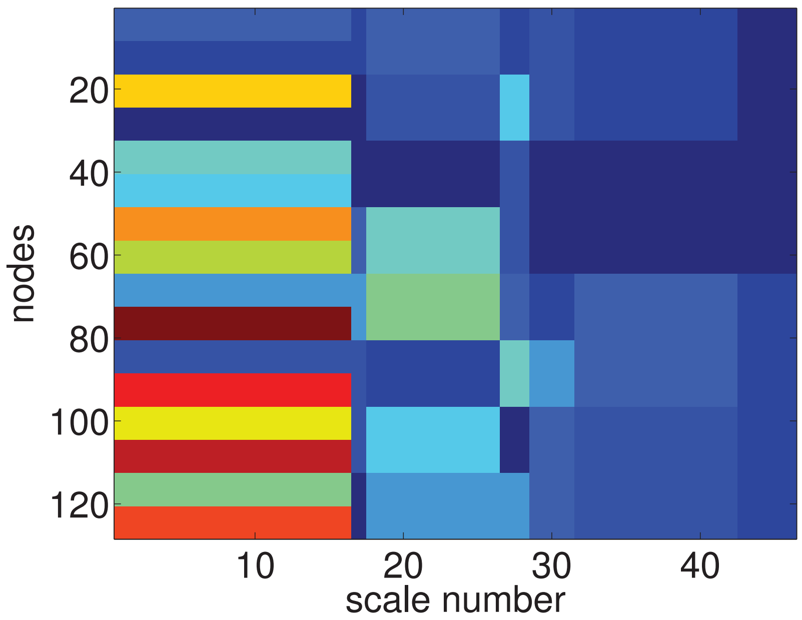




Note: we use the toy graph

Dendrogram cut at maximal gap

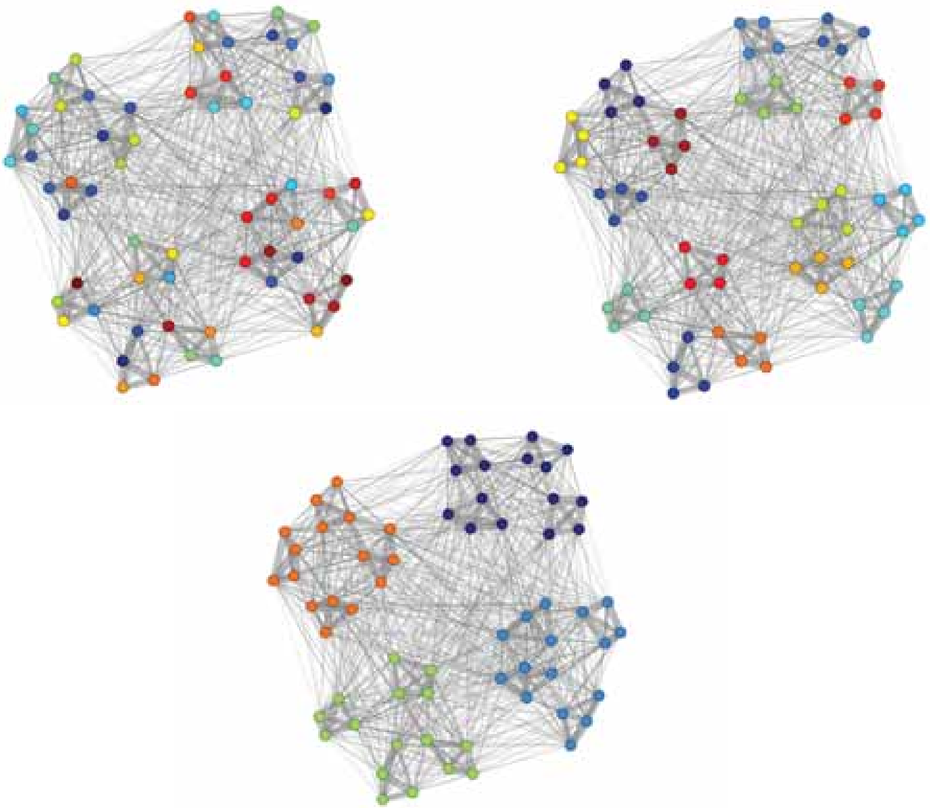




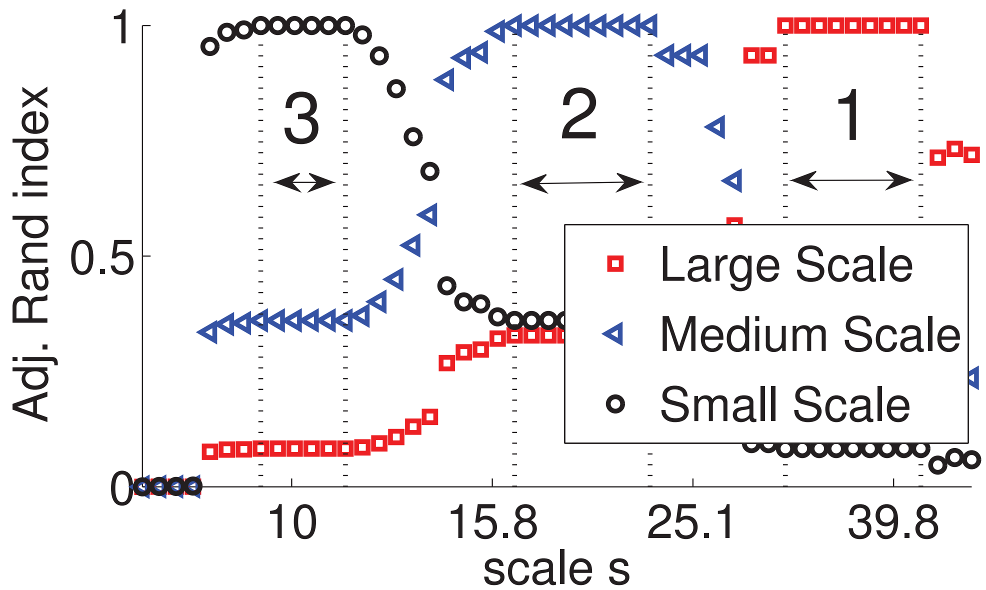
- Improvement: cut at average maximal gap

The Sales-Pardo benchmark

- Three community structures nested in one another
- Parameters:
    - sizes of the communities ($N = 640$)
    - ρ tunes how well separated the different scales are
    - $\bar{k}$ is the average degree; the sparser is the graph, the harder it is to recover the communities.



Results on the Sales-Pardo benchmark



The case of larger networks

- Limit of the method: computation of the $N \times N$ matrix of the wavelets $\mathbf{\Psi}_s$.
- Improvement: use of random features.
- Let $\boldsymbol{r} \in \mathbb{R}^N$ be a random vector on the nodes of the graph, composed of N independent normal random variables of zero mean and finite variance σ^2.
- Define the feature $f_{s,a} \in \mathbb{R}$ at scale s associated to node a as

$$f_{s,a} = \boldsymbol{\psi}_{s,a}^{\top} \boldsymbol{r} = \sum_{k=1}^{N} \psi_{s,a}(k) r(k).$$

The case of larger networks

- Let us define the correlation between features

$$\mathrm{Cor}(f_{s,a}, f_{s,b}) = \frac{\mathbb{E}((f_{s,a} - \mathbb{E}(f_{s,a}))(f_{s,b} - \mathbb{E}(f_{s,b})))}{\sqrt{\mathrm{Var}(f_{s,a})\mathrm{Var}(f_{s,b})}}.$$

- It is easy to show that:

$$\mathrm{Cor}(f_{s,a}, f_{s,b}) = \frac{\psi_{s,a}^{\top}\psi_{s,b}}{||\psi_{s,a}||_2||\psi_{s,b}||_2}.$$

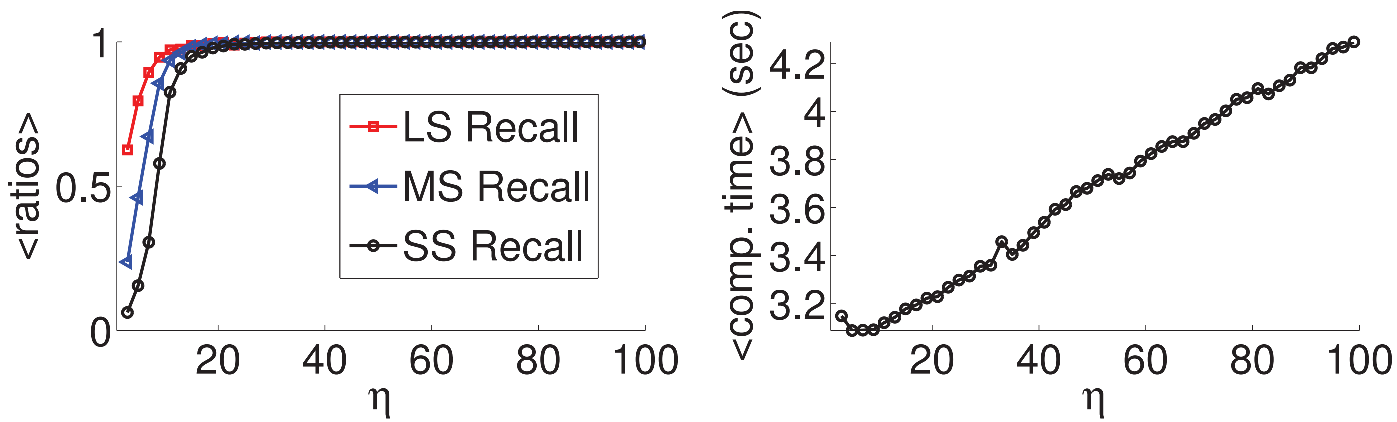
- Therefore, the sample correlation estimator $\hat{C}_{ab,\eta}$ satisfies:

$$\lim_{\eta \to +\infty} \hat{C}_{ab,\eta} = \frac{\psi_{s,a}^{\top}\psi_{s,b}}{||\psi_{s,a}||_2||\psi_{s,b}||_2} = 1 - \boldsymbol{D}_s(a, b).$$

- This leads to a faster algorithm.

Results on the Sales-Pardo benchmark

- As a function of η, the number of random vectors used



Stability of the communities

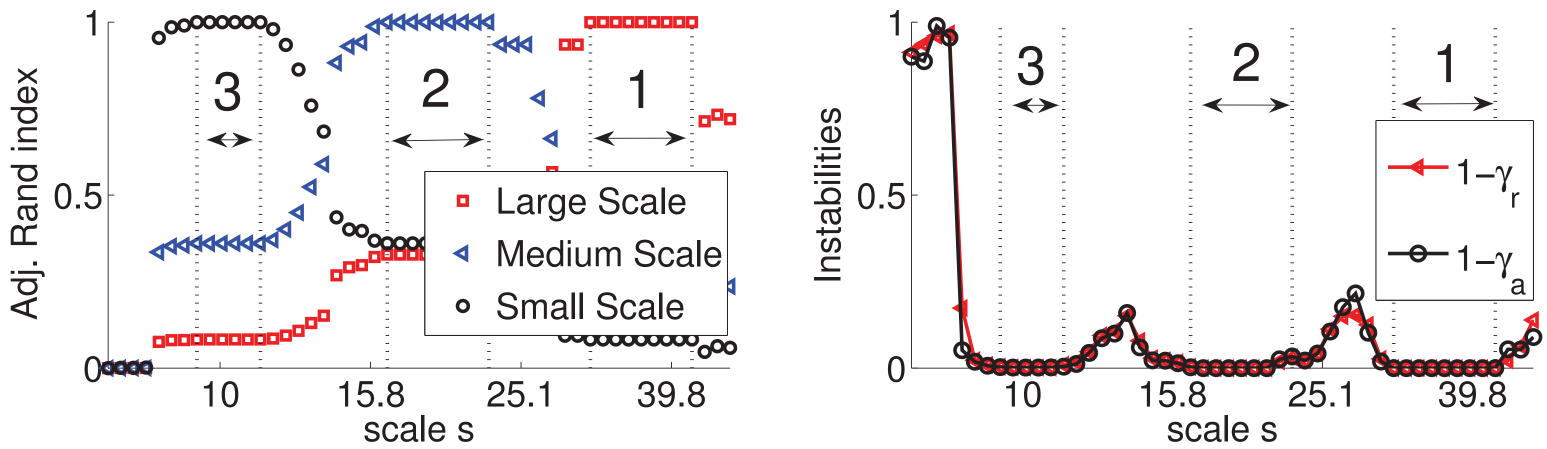
- Not all partitions are relevant: only those stable enough convey information about the network
- Lambiotte's approach to stability:
  Create B resampled graphs by randomly adding $\pm p\%$ (typically $p = 10$) to the weight of each link and computing the corresponding B sets of partitions $\{P_s^b\}_{b\in[1,B],s\in\mathcal{S}}$.
  Then, stability:

$$\gamma_r(s) = \frac{2}{B(B-1)} \sum_{(b,c)\in[1,B]^2, b\neq c} \mathtt{ari}(P_s^b, P_s^c), \qquad (1)$$

- New approach: we have a stochastic algorithm.
  Consider J sets of η random signals and compute the associated sets of partitions $\{P_s^j\}_{j\in[1,J],s\in\mathcal{S}}$. Let stability be:

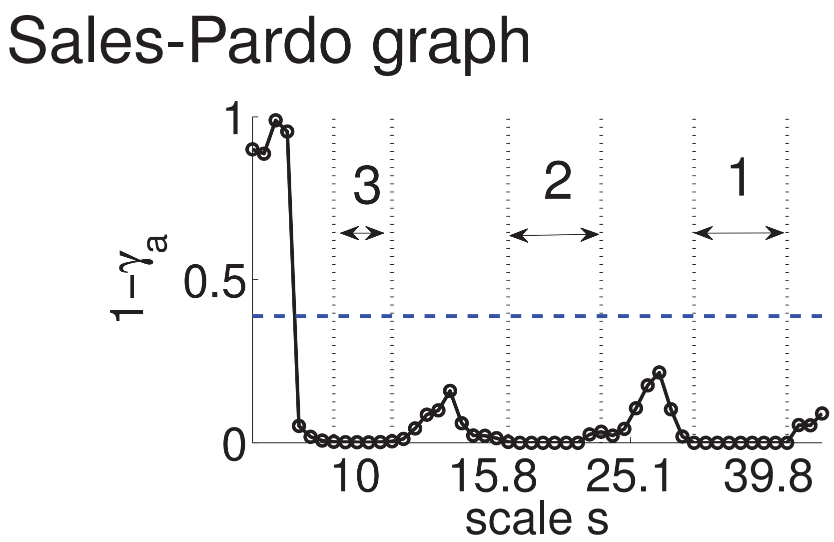
$$\gamma_a(s) = \frac{2}{J(J-1)} \sum_{(i,j)\in[1,J]^2, i\neq j} \mathtt{ari}(P_s^i, P_s^j). \qquad (2)$$

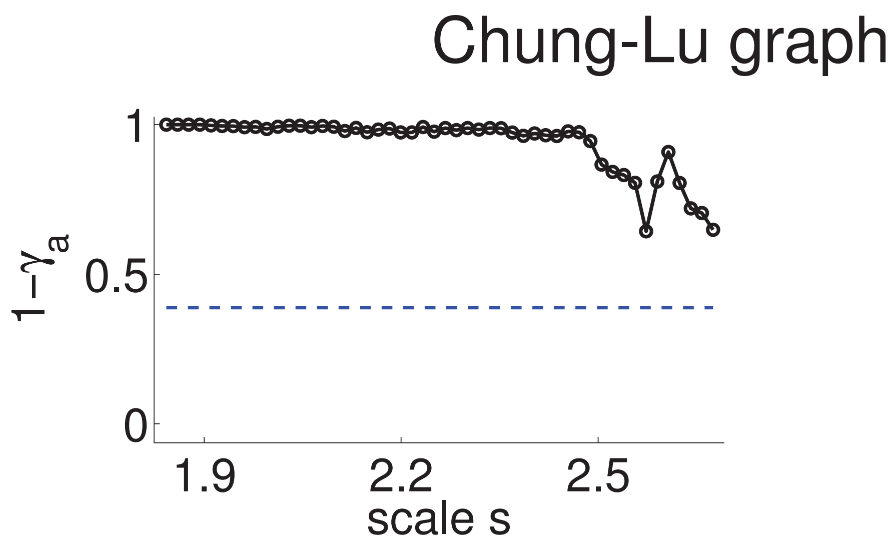
Results with stabilities on the Sales-Pardo benchmark



In addition: statistical test of relevance of the communities

- It is possible to design a data-driven test on γ_a (computation of a numerical threshold for the configuration (or Chung-Lu) model).
- Result: threshold for $1 - \gamma_a$ above which the partition in communities is irrelevant.



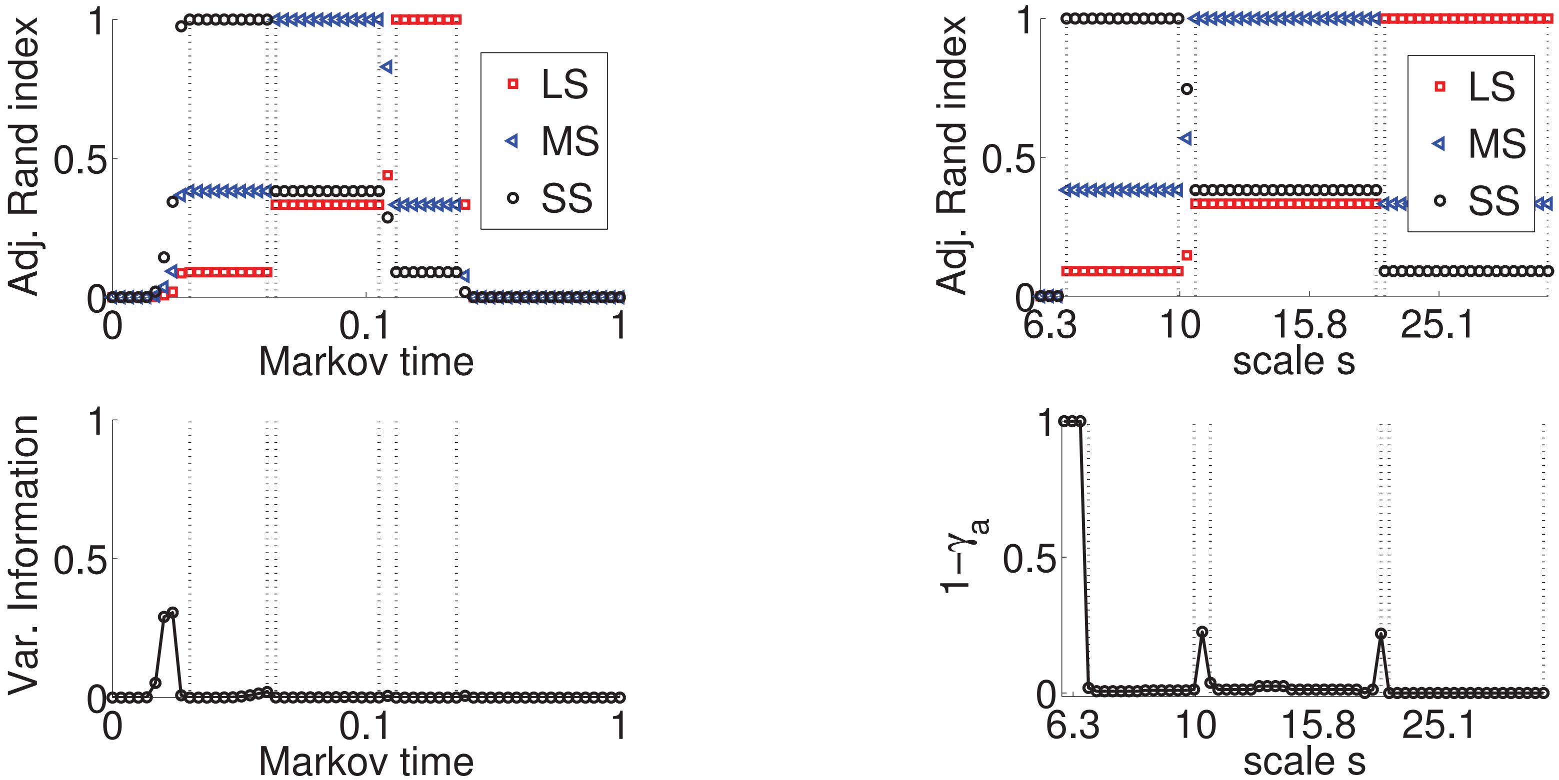


Comparison on larger Sales-Pardo graphs

$N = 6400$ nodes

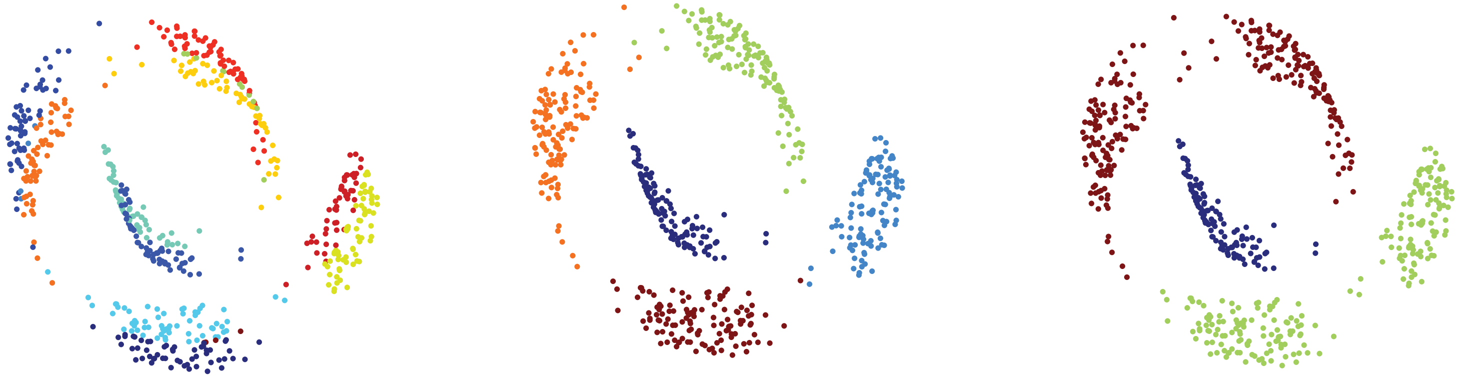
Schaub-Delvenne

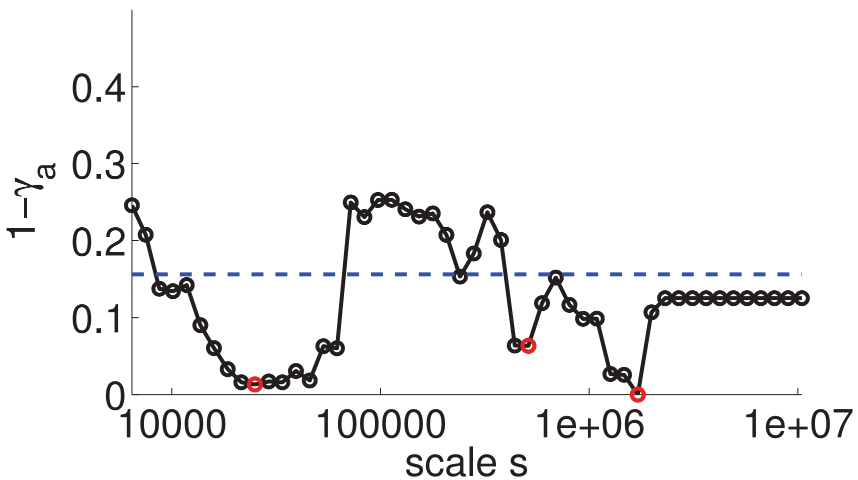
Wavelets



Sensor network on the swiss roll manifold

- Three scale ranges of relevant community structure



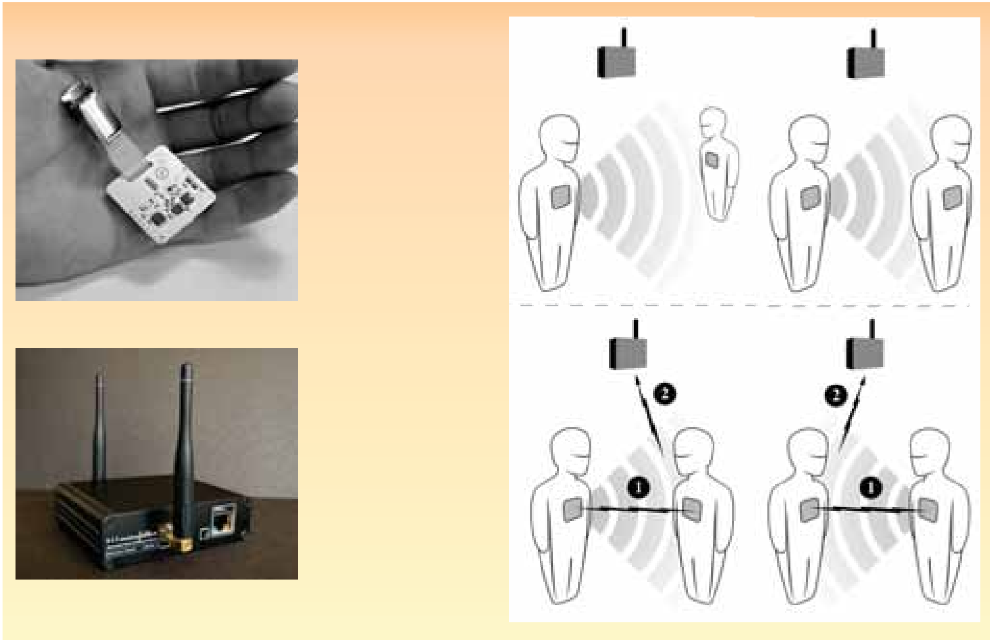


The dynamic social network of a primary school

Collaboration with A. Barrat (CPT Marseille), C. Cattuto (ISI, Turin)

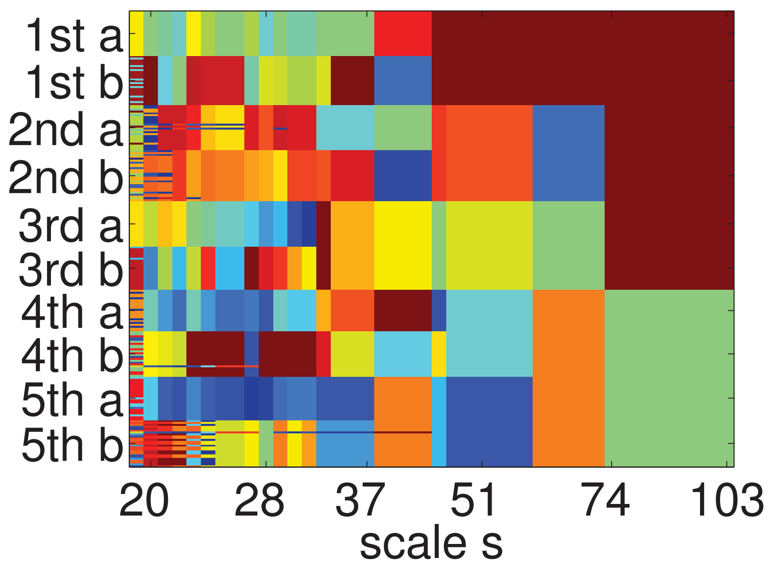
Sociopatterns project

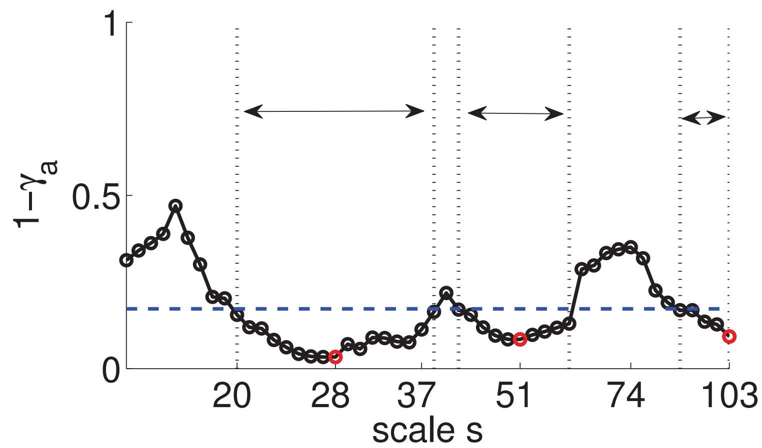
- Acquisition of face-to-face human contacts (resolved in time) using active RFID tags and + fixed antenna



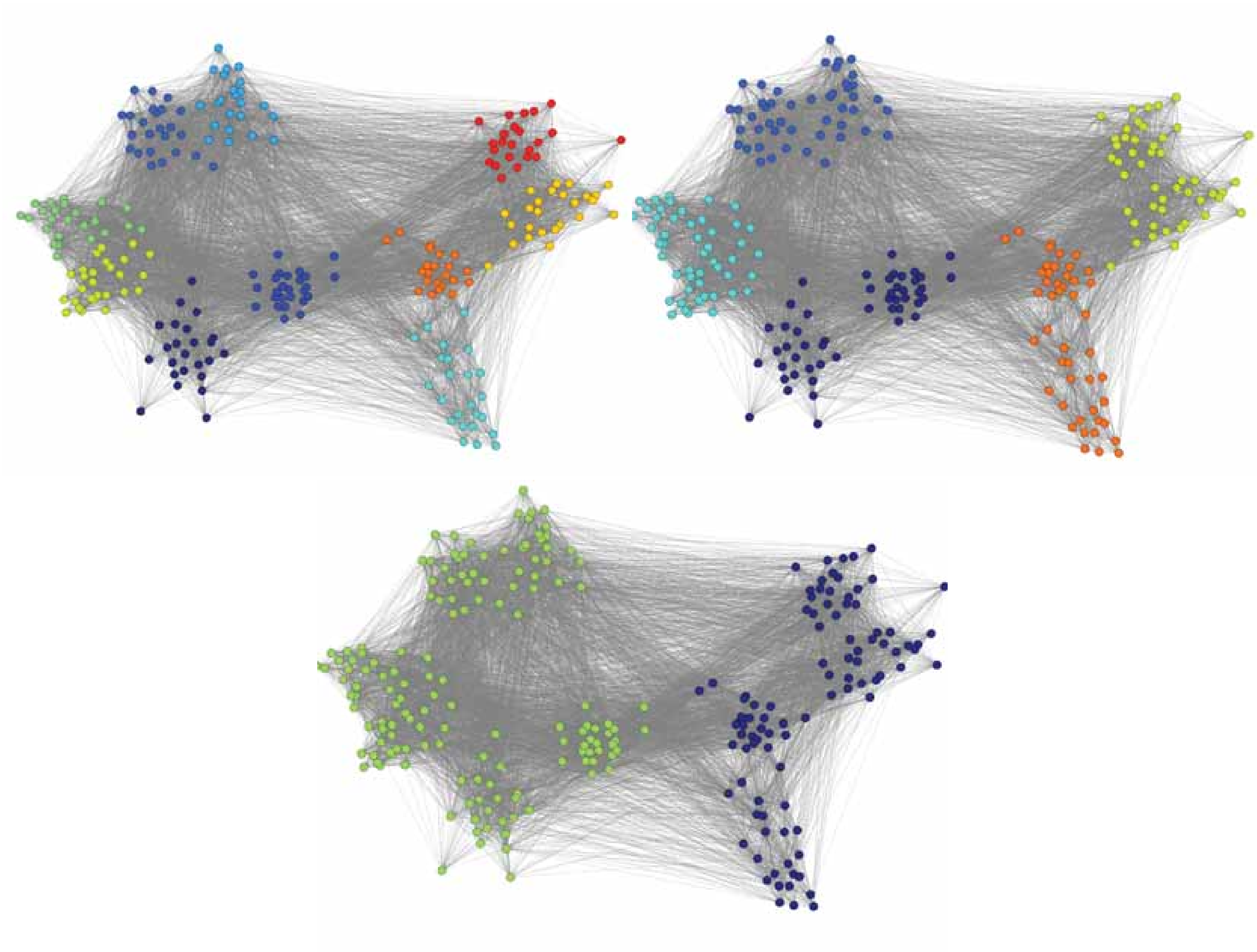
- Interest: social studies, spreading processes (of information, of epidemic,...), contact dynamics,...
- Time for a movie!

Multi-scale Communities in Primary School





Multi-scale Communities in Primary School



Conclusion

- Wavelet $\psi_{s,a}$ gives an ”egocentered view“ of the network seen from node *a* at scale *s*
- Correlation between these different views gives us a distance between nodes at scale *s*
- This enables multi-scale clustering of nodes in communities
- Associated to a notion of stability and of statistical detection of relevance

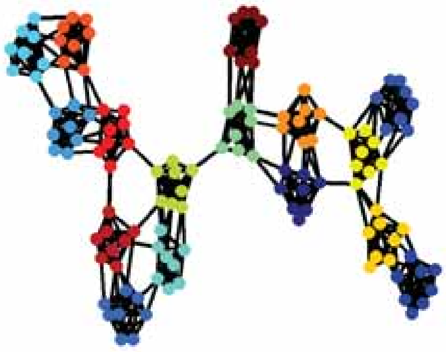
- I hope also that you were interested in
the emerging field of graph signal processing for networks.

http://perso.ens-lyon.fr/pierre.borgnat

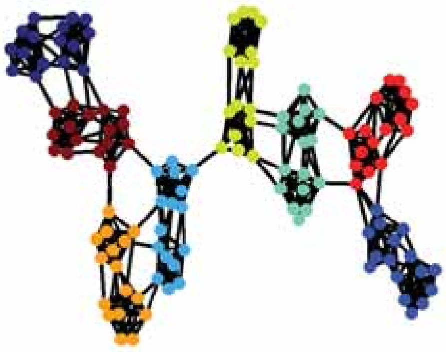
Acknowledgements: thanks to Nicolas Tremblay for borrowing many of his figures or slides.

A toy graph for introducing the method

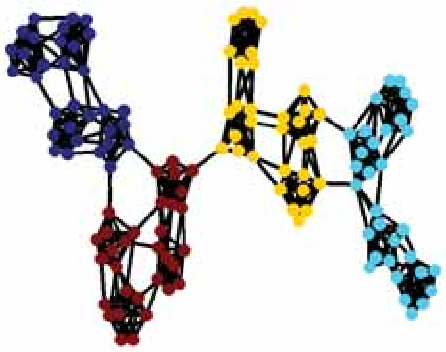
smallest scale (16 com.):



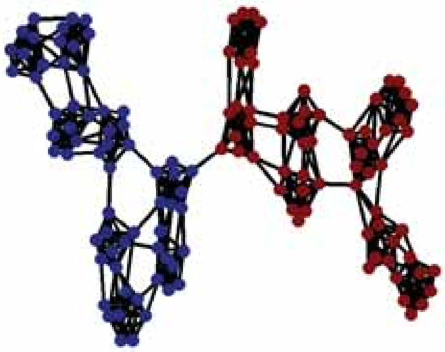
small scale (8 com.):



medium scale (4 com.):

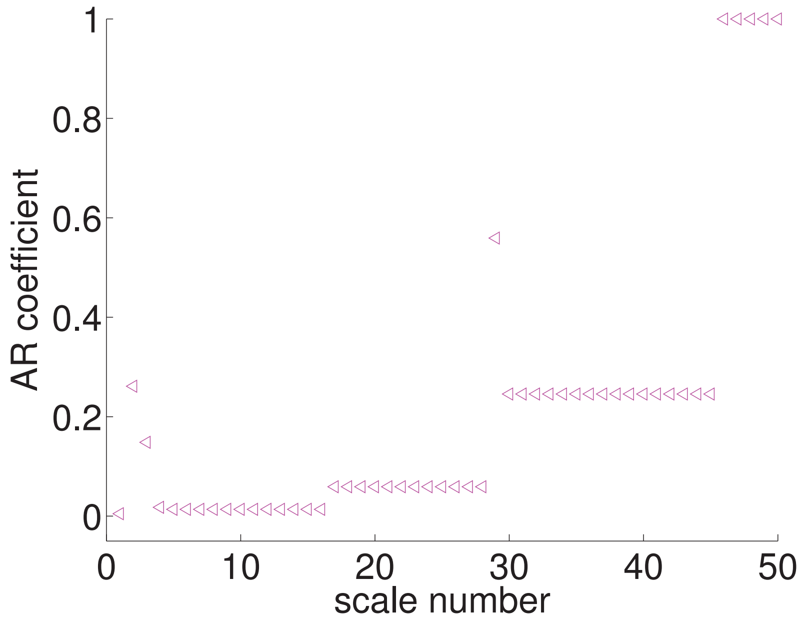


large scale (2 com.):



Dendrogram cut with prior knowledge

Let us cheat by using prior knowledge on the number of communities we are looking for.
If we cut each dendrogram in two clusters

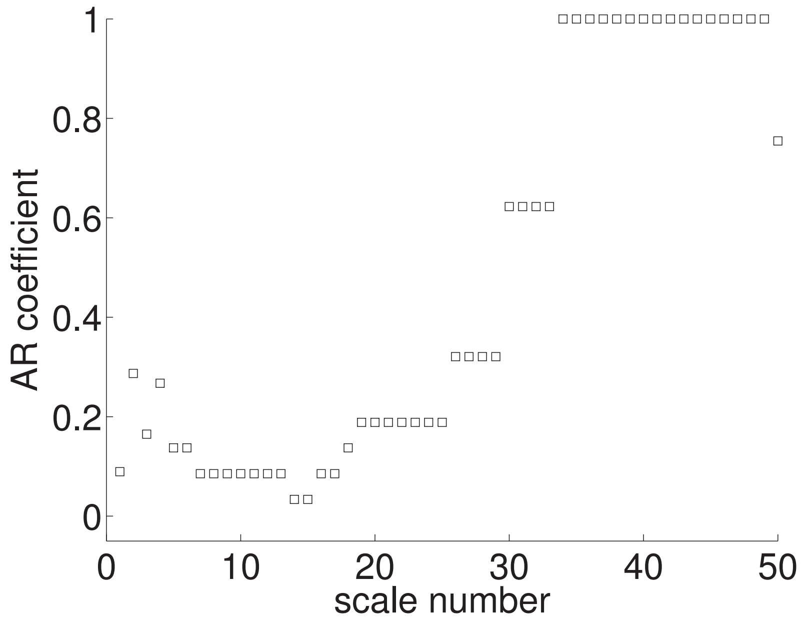


Using wavelets as features

Conclusion: the dendrograms at different scales contain the community structure at various scales.

Dendrogram cut with prior knowledge

Let us cheat by using prior knowledge on the number of communities we are looking for.
If we cut each dendrogram in four clusters

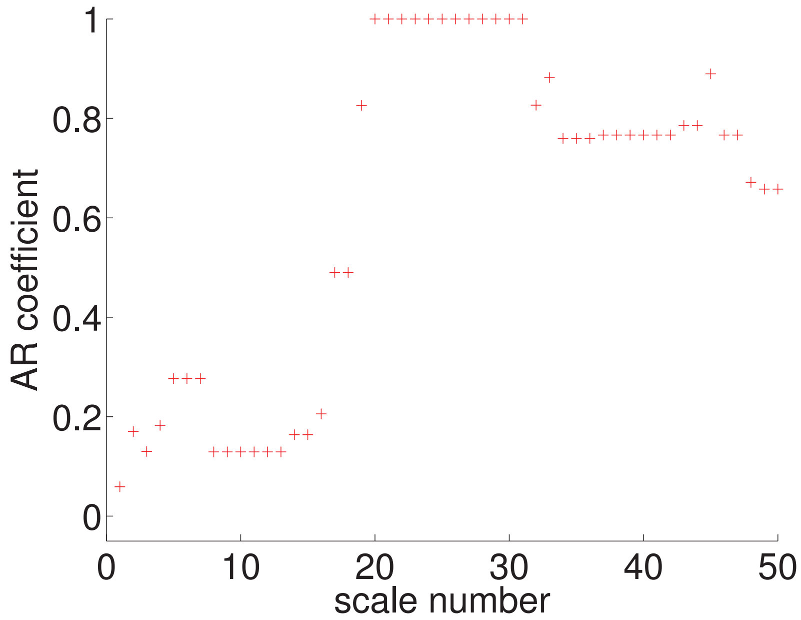


Using wavelets as features

Conclusion: the dendrograms at different scales contain the community structure at various scales.

Dendrogram cut with prior knowledge

Let us cheat by using prior knowledge on the number of communities we are looking for.
If we cut each dendrogram in eight clusters

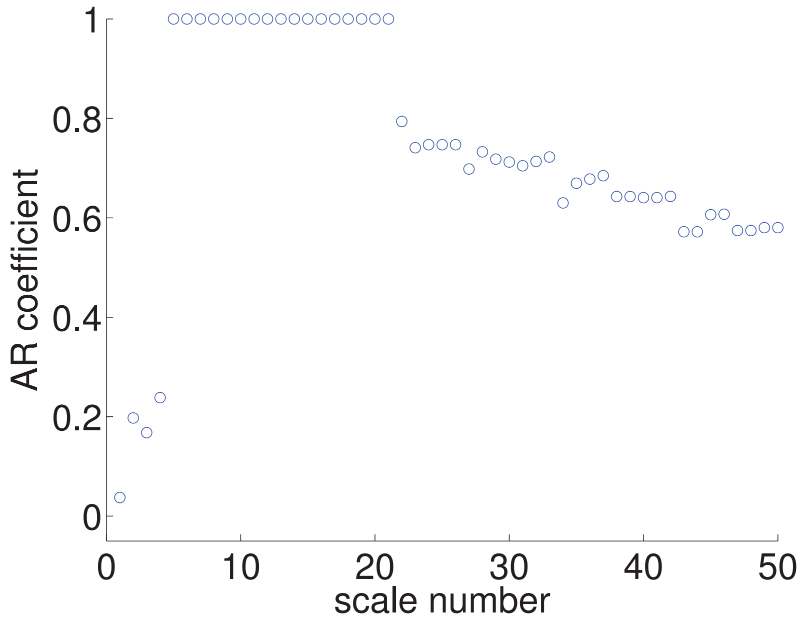


Using wavelets as features

Conclusion: the dendrograms at different scales contain the community structure at various scales.

Dendrogram cut with prior knowledge

Let us cheat by using prior knowledge on the number of communities we are looking for.
If we cut each dendrogram in sixteen clusters

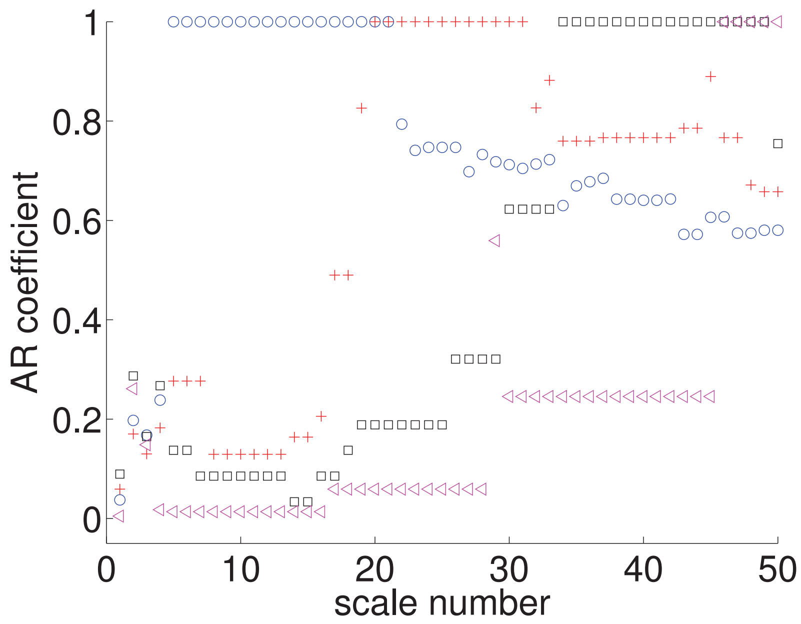


Using wavelets as features

Conclusion: the dendrograms at different scales contain the community structure at various scales.

Dendrogram cut with prior knowledge

Let us cheat by using prior knowledge on the number of communities we are looking for.
The four levels of communities.



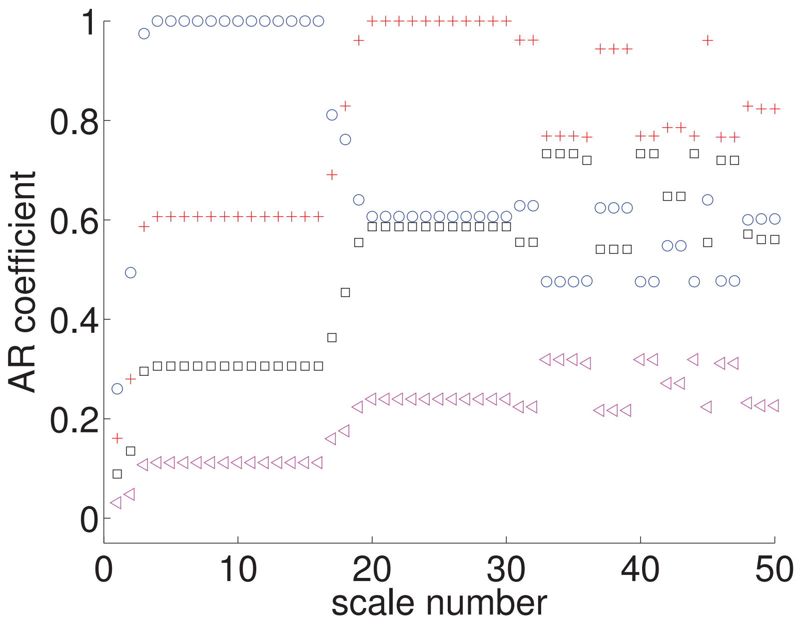
Using wavelets as features

Conclusion: the dendrograms at different scales contain the community structure at various scales.

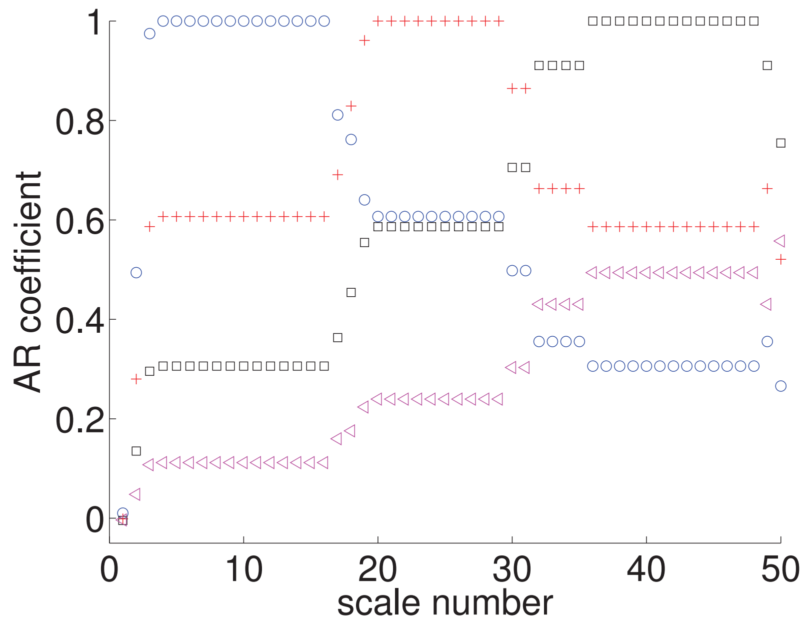
Dendrogram cut with modularity

- By max. of with classical modularity Q
- or by max. of a filtered modularity [Arenas, Delvenne,...]

Classical Modu Opt.

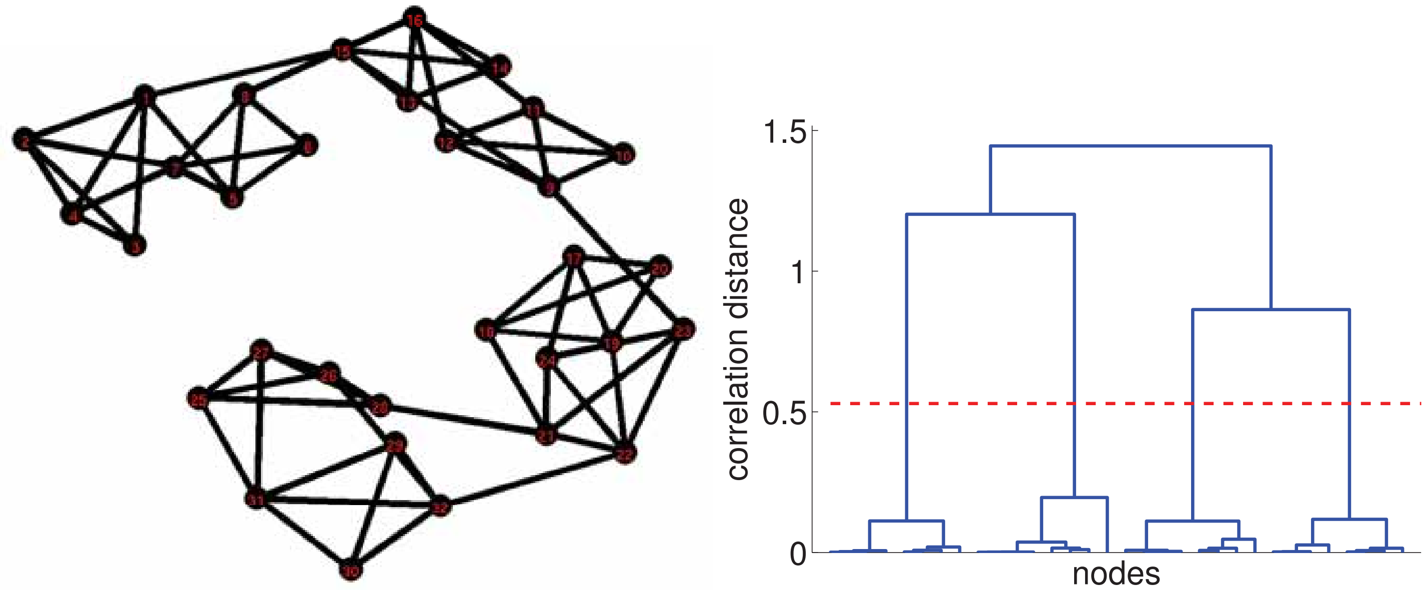


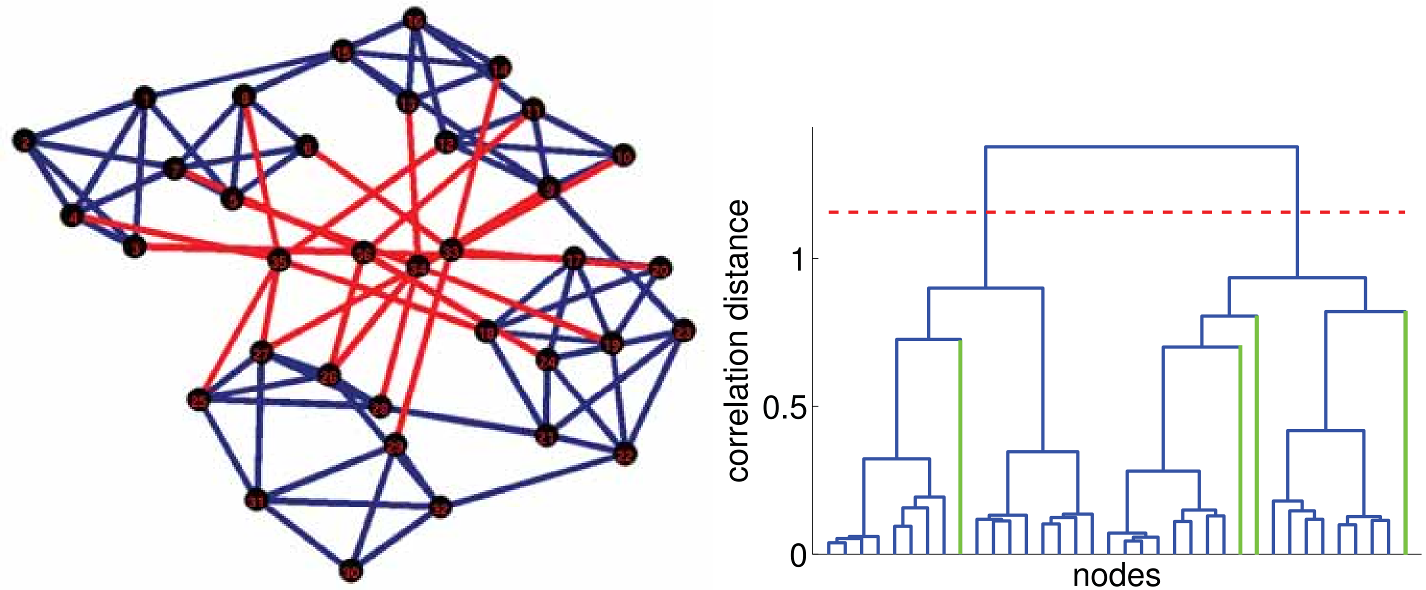
Filtered Modu Opt.



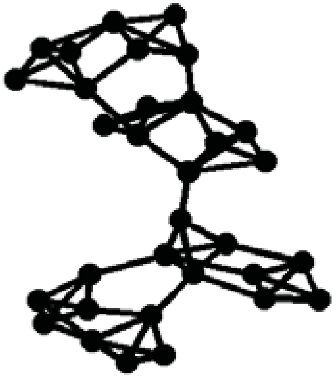
- The solutions are not really good at all scale.

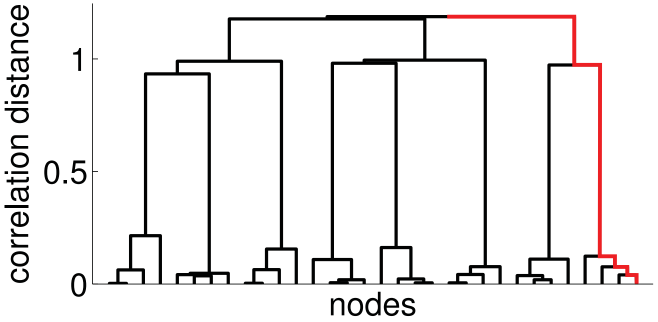
Dendrogram cut at maximal gap: non robust to outliers

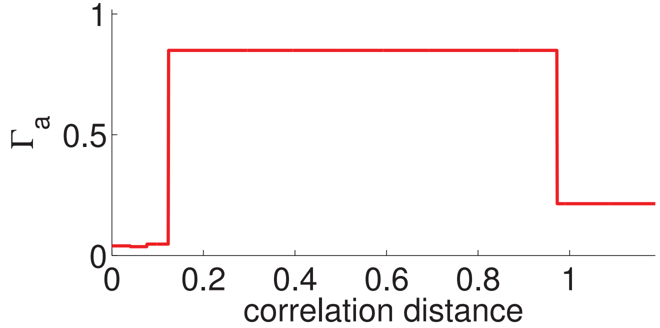




Dendrogram cut at maximal average gap

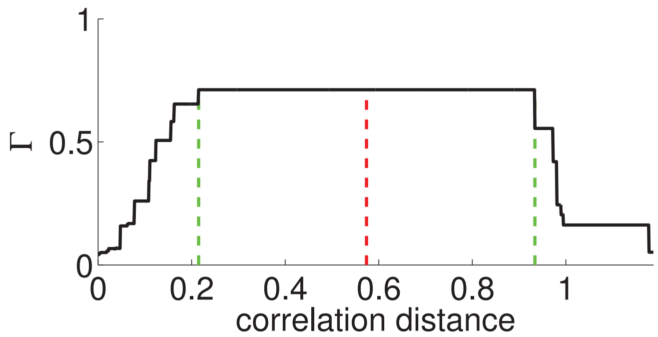


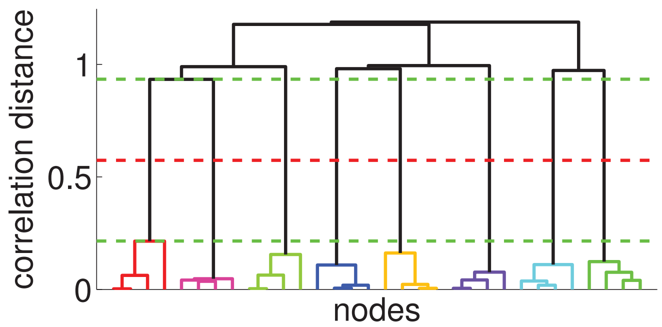




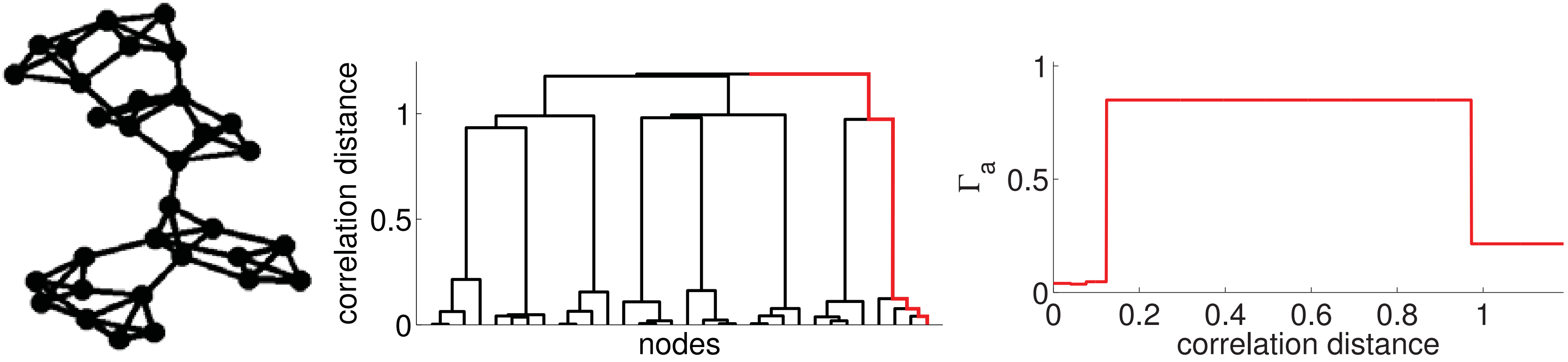
$$\Gamma = \frac{1}{N \max(\text{corr. dist.})} \sum_{a \in \mathcal{V}} \Gamma_a$$

At small scale



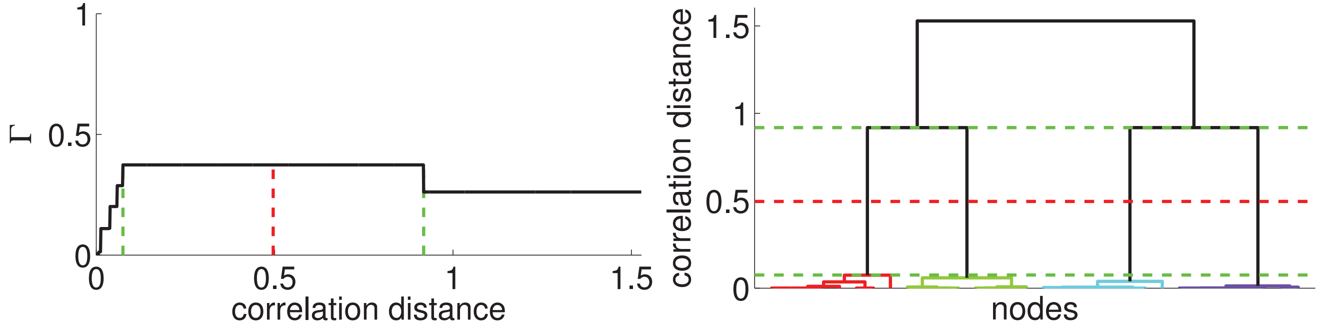


Dendrogram cut at maximal average gap

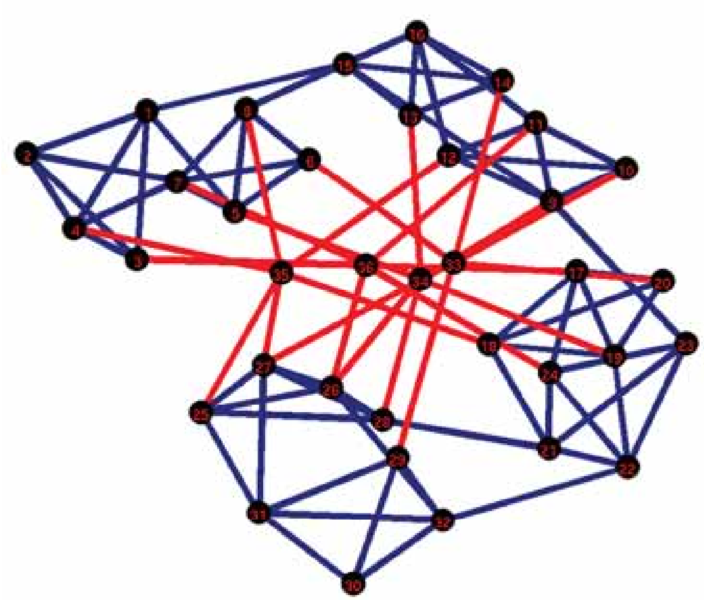


$$\Gamma = \frac{1}{N \max(\text{corr. dist.})} \sum_{a \in \mathcal{V}} \Gamma_a$$

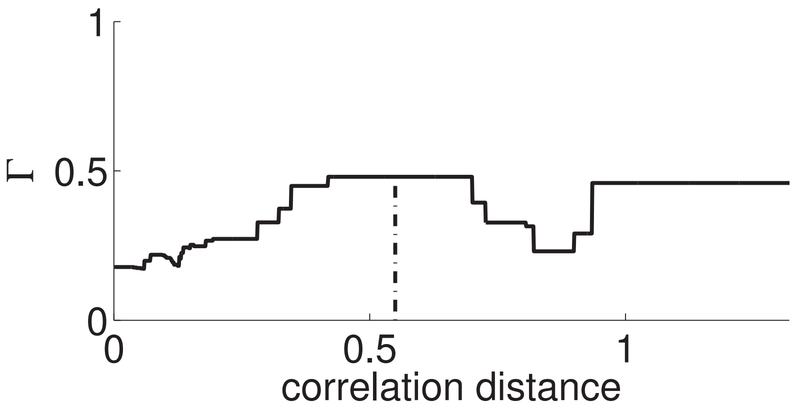
At larger scale

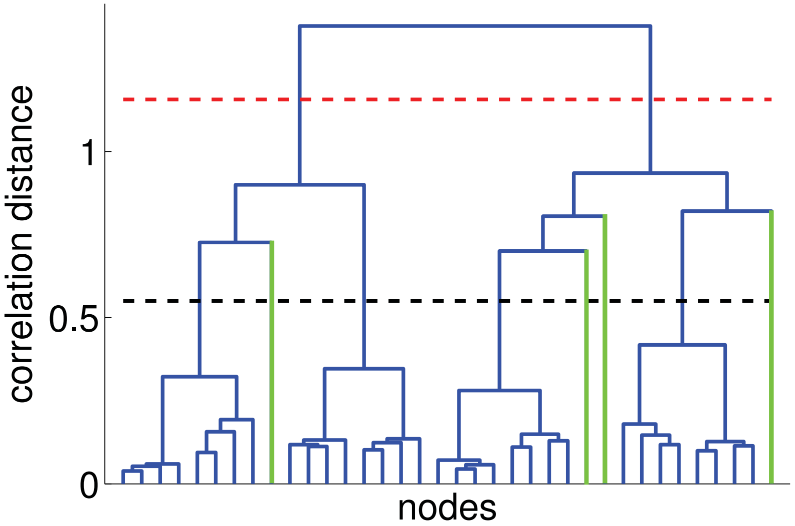


Dendrogram cut at maximal average gap



For the previous graph:





Recall: The Adjusted Rand Index

Let:

- $\mathcal{C}$ and $\mathcal{C}'$ be two partitions we want to compare.
- a be the $\#$ of pairs of nodes that are in the same community in $\mathcal{C}$ *and* in the same community in $\mathcal{C}'$
- b be the $\#$ of pairs of nodes that are in different communities in $\mathcal{C}$ *and* in different communities in $\mathcal{C}'$
- c be the $\#$ of pairs of nodes that are in the same community in $\mathcal{C}$ *and* in different communities in $\mathcal{C}'$
- d be the $\#$ of pairs of nodes that are in different communities in $\mathcal{C}$ *and* in the same community in $\mathcal{C}'$

$a + b$ is the number of “agreements“ between $\mathcal{C}$ and $\mathcal{C}'$.
$c + d$ is the number of “disagreements“ between $\mathcal{C}$ and $\mathcal{C}'$.

The Adjusted Rand Index

The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

The Adjusted Rand index AR is the corrected-for-chance version of the Rand index:

$$AR = \frac{R - ExpectedIndex}{MaxIndex - ExpectedIndex}$$