Nonlinear time series analysis

Multivariate analysis

Cristina Masoller

Universitat Politecnica de Catalunya, Terrassa, Barcelona, Spain

Cristina.masoller@upc.edu

www.fisica.edu.uy/~cris

http://www.fisica.edu.uy/~cris/teaching/curitiba_masoller_multi.pdf

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

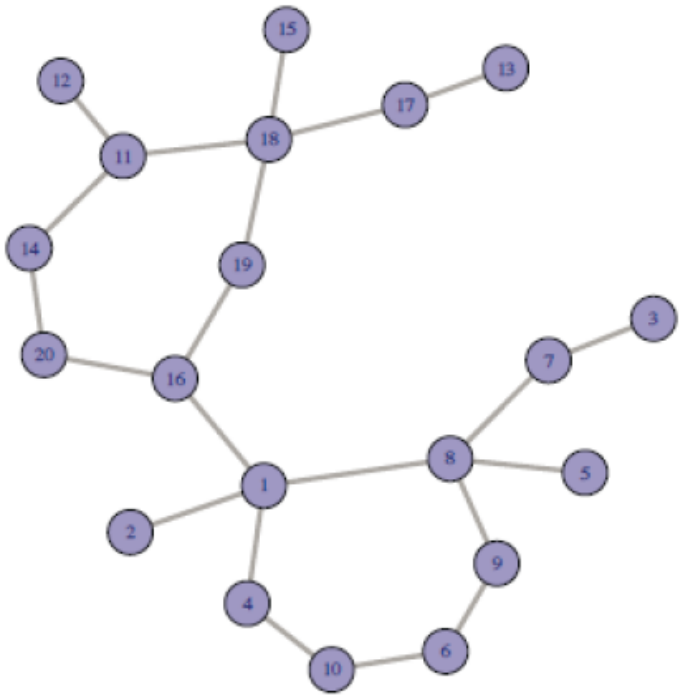
Campus d'Excel·lència Internacional

Outline

- **Introduction**
  - Historical developments: from dynamical systems to complex systems
- **Univariate analysis**
  - Methods to extract information from a time series.
  - Applications.
- **Bivariate analysis**
  - Extracting information from two time series.
  - Correlation, directionality and causality.
  - Applications.
- **Multivariate analysis**
  - Many time series: complex networks.
  - Network characterization and analysis.
  - Climate networks.

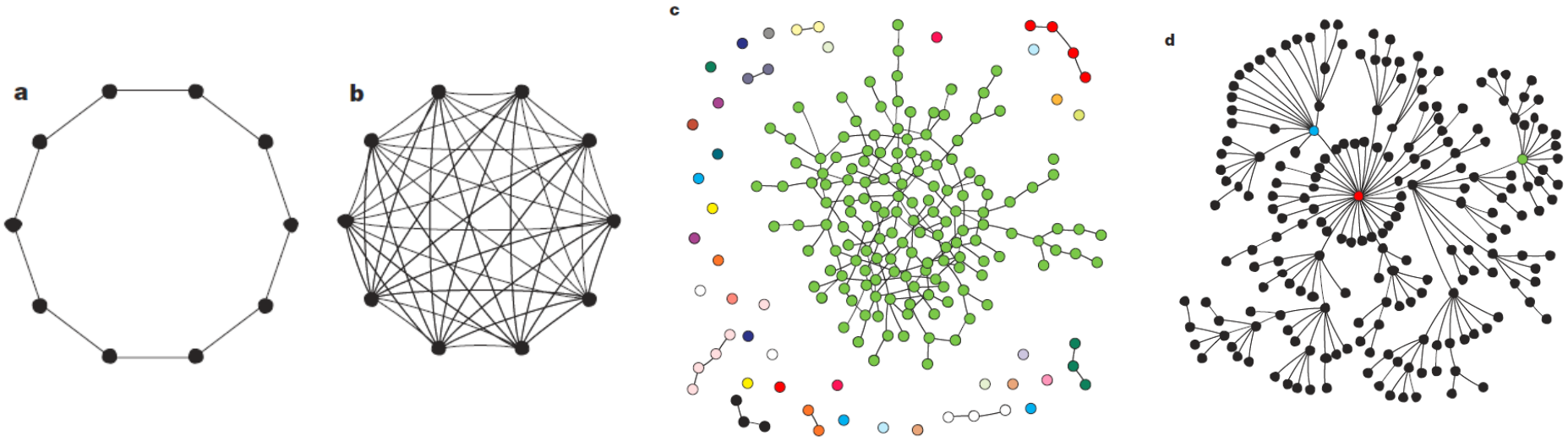
What is a network?

- A graph: a set of “nodes” connected by a set of “links”.
- Nodes and links can be weighted or unweighted.
- Links can be directed or undirected.



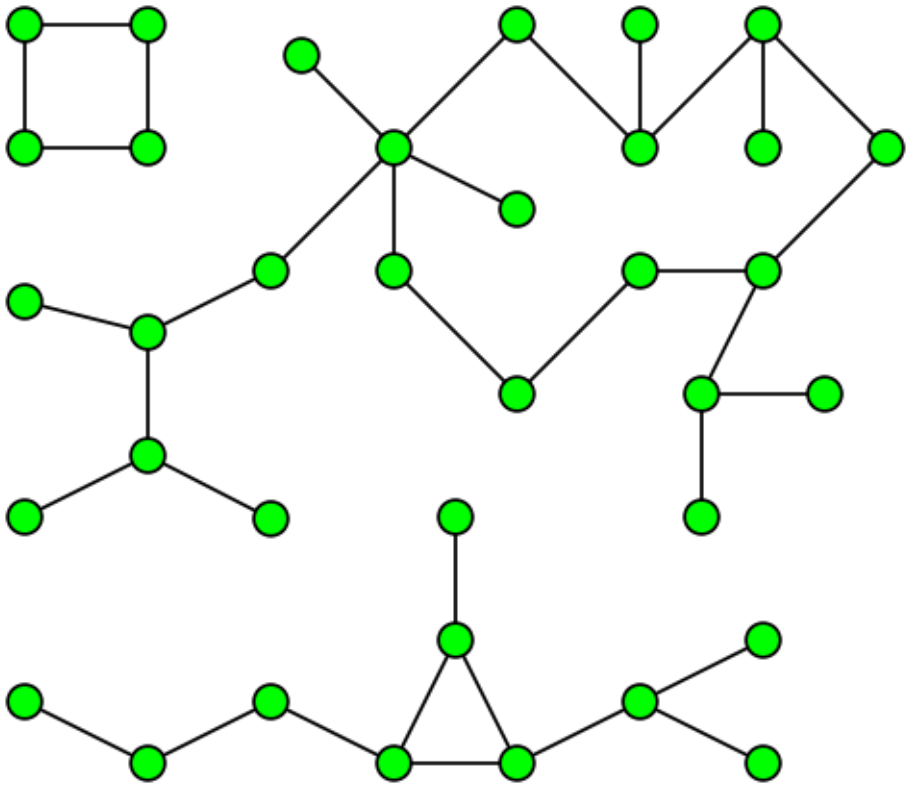
Networks or graphs

The challenge in the context of time series analysis: to infer the underlying network structure from observed signals.



Source: Strogatz
Nature 2001

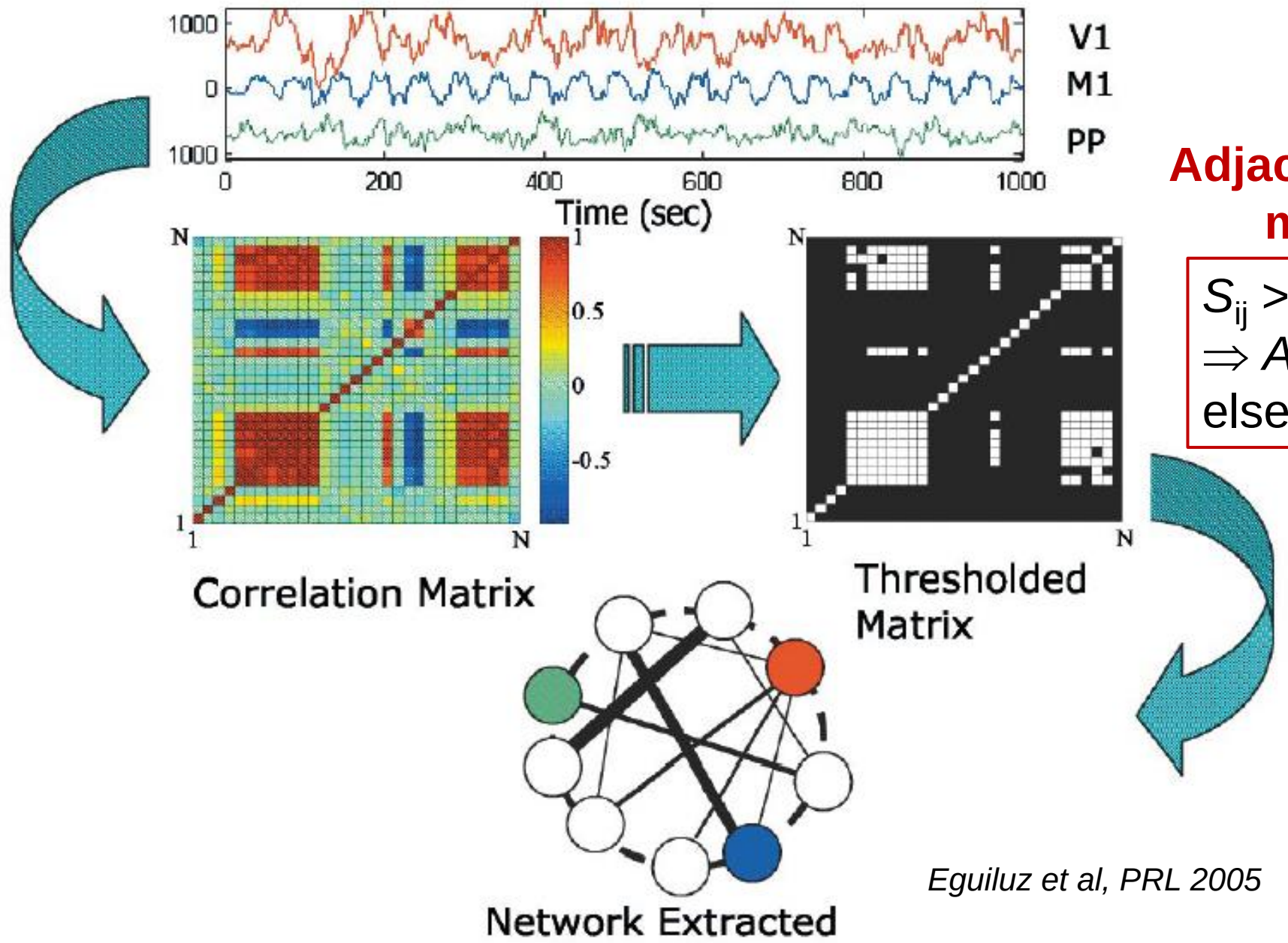
Connected components ("communities")



A graph with three connected components.
Source: Wikipedia

Using statistical similarity measures to infer interactions from data: “functional networks”

Brain functional network



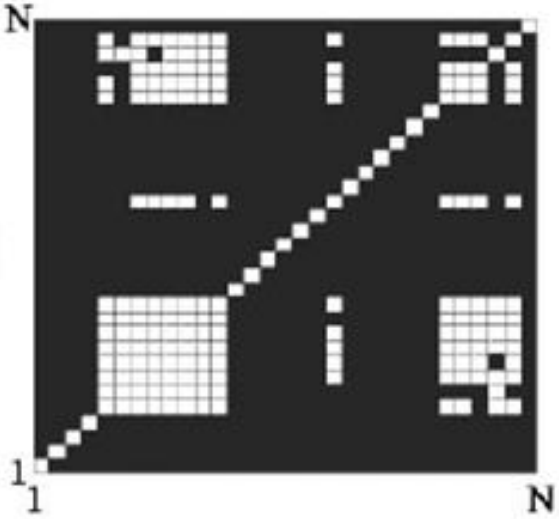
Adjacency matrix

$S_{ij} > Th$
$\Rightarrow A_{ij} = 1,$
else $A_{ij}=0$

Eguiluz et al, PRL 2005

Graphical representation

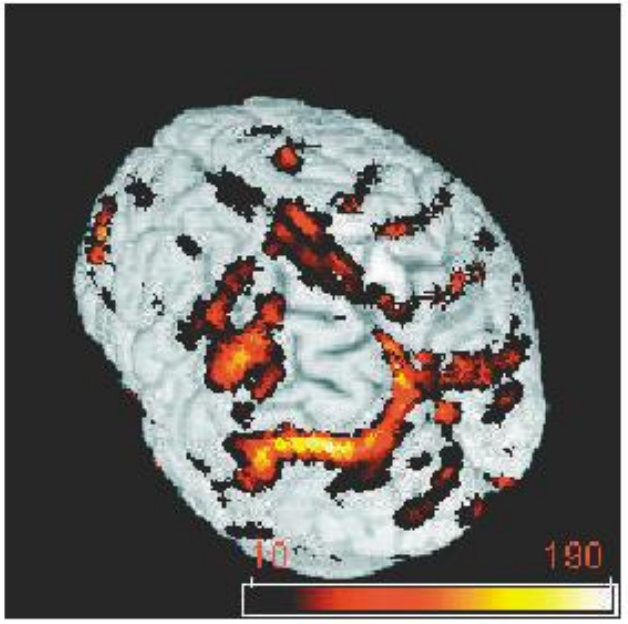
Adjacency matrix



Thresholded matrix = inferred (“functional”) network

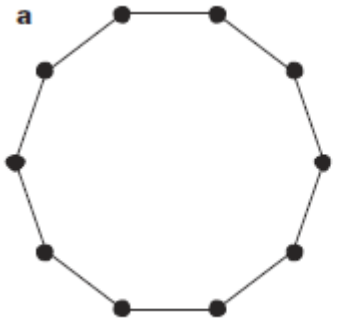
Degree of a node: number of links

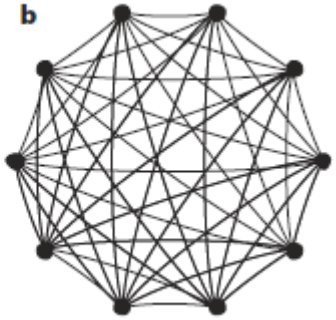
$$k_i = \Sigma_j A_{ij}$$

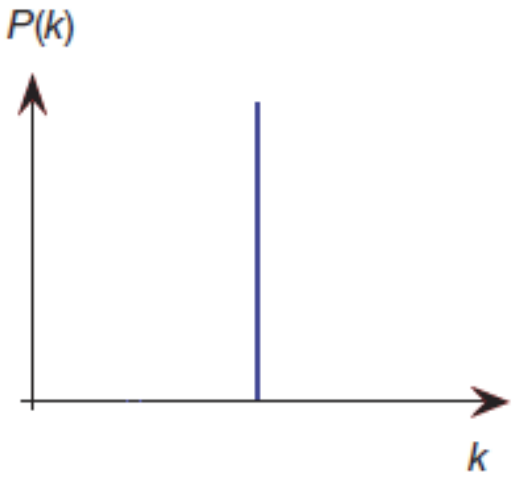


The degree distribution: usual way to characterize a graph

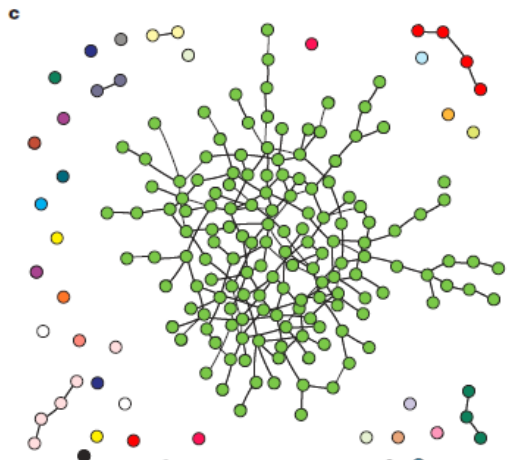
Regular

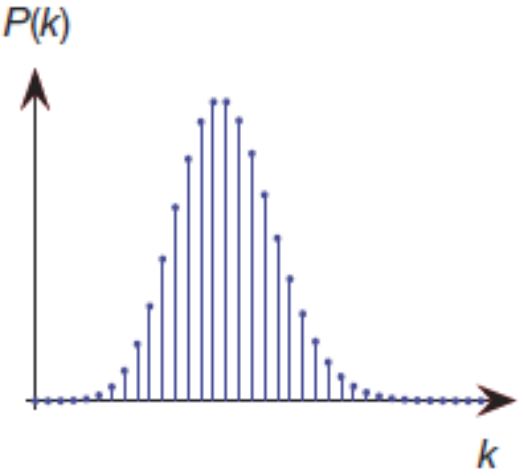




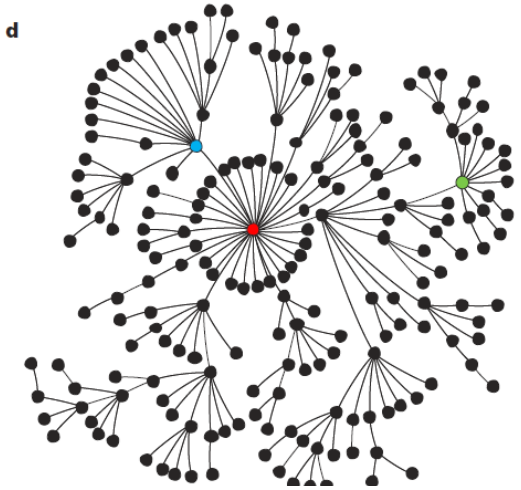


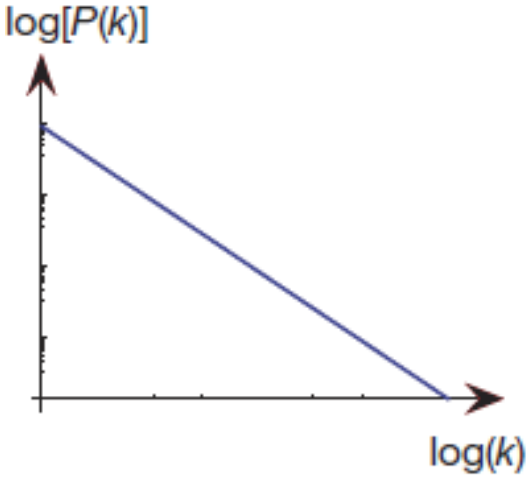
Random





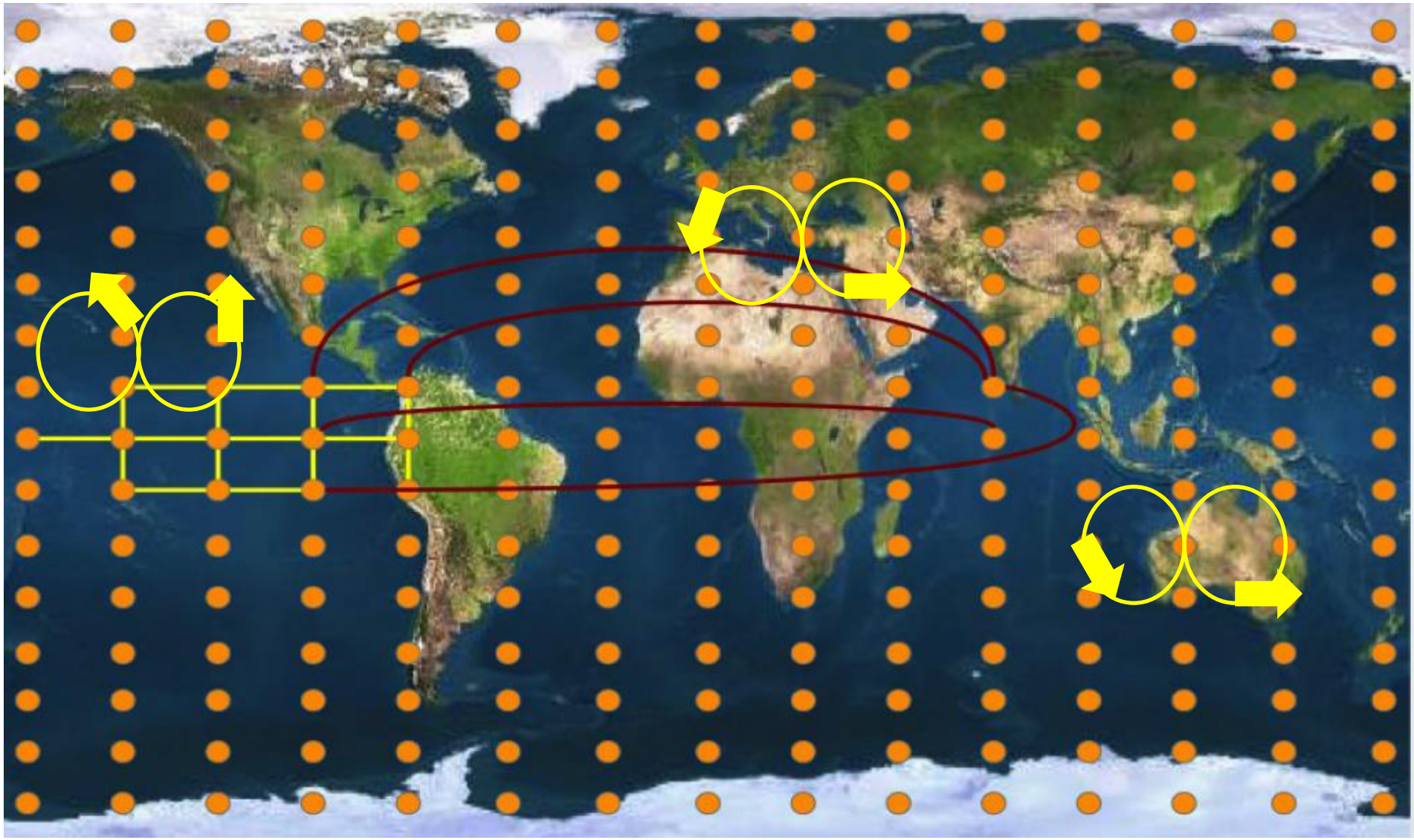
Scale-free



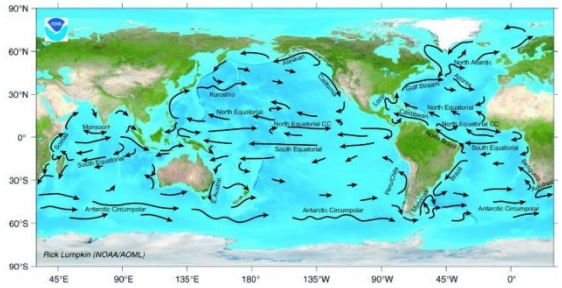


Strogatz, Nature 2001

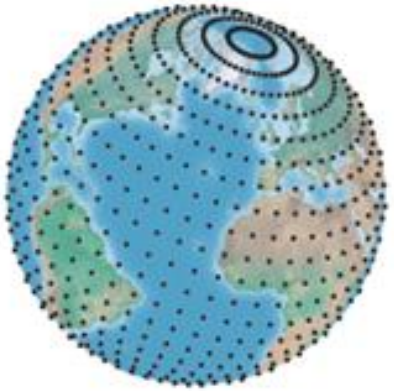
The climate system as a set of “interacting oscillators”



Complex network representation of the climate system

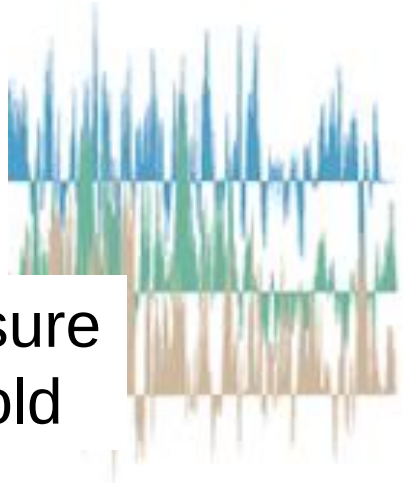


Back to the climate system: interpretation (currents, winds, etc.)



More than 10000 nodes (with different sizes).

Daily resolution: more than 13000 data points in each TS

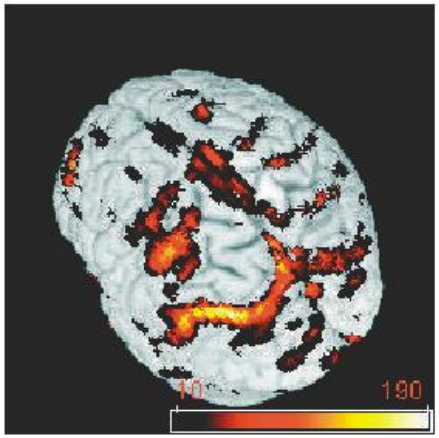


Surface Air Temperature Anomalies (solar cycle removed)

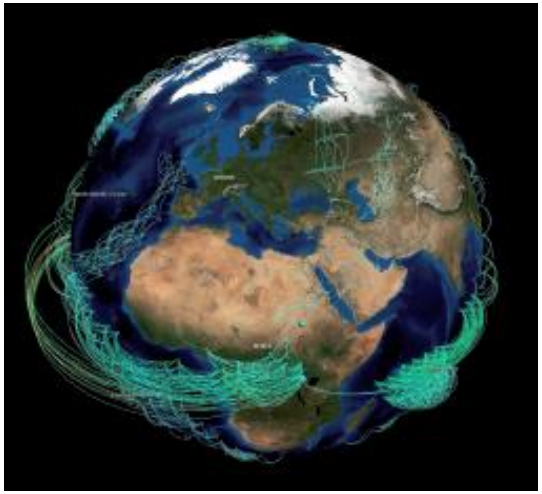
Sim. measure + threshold

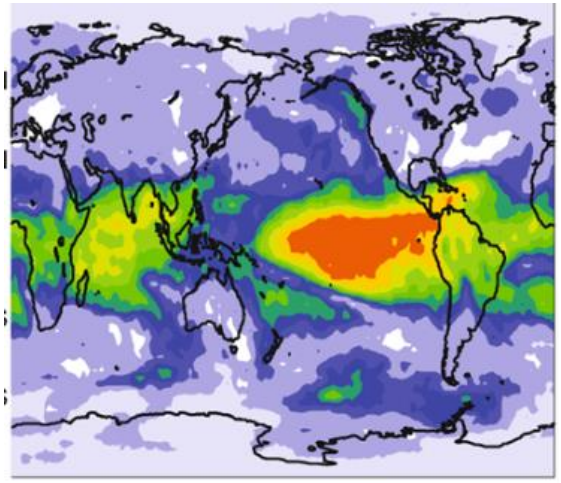
Donges et al, Chaos 2015

Brain network



Climate network





Area weighted connectivity (AWC):
weighted degree (nodes represent areas with different sizes)

$$\mathrm{AWC}_i = \frac{\sum_j^N A_{ij} \cos(\lambda_j)}{\sum_j^N \cos(\lambda_j)}$$

How to select the threshold ?

$S_{ij} > Th \Rightarrow A_{ij} = 1$, else $A_{ij}=0$

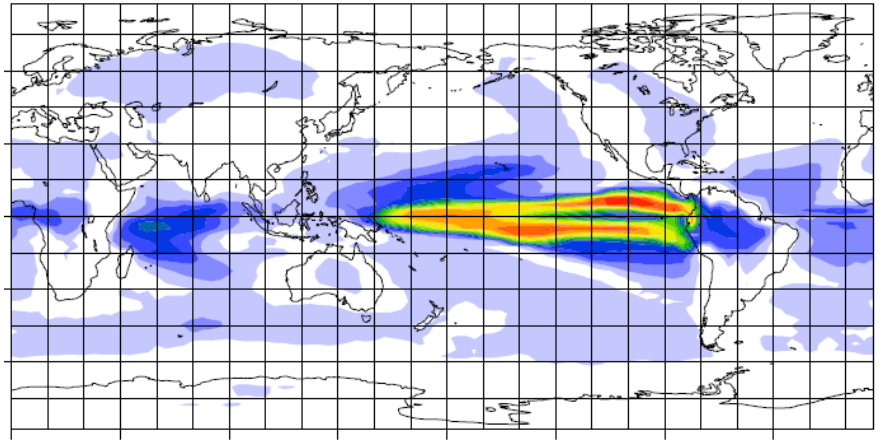
Three criteria are typically used:

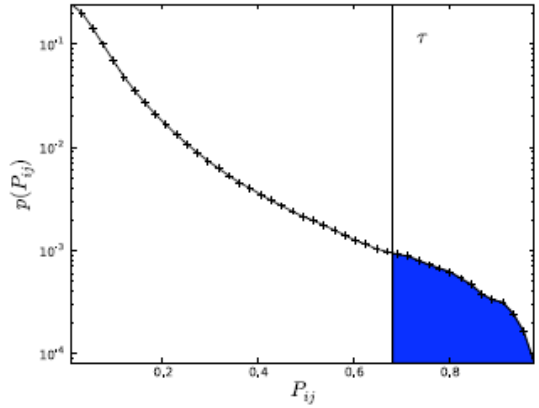
- A significance level is used (typically 5%) in order to omit connectivity values that can be expected by chance;
- We select an arbitrary value as threshold, such that it gives a certain pre-fixed number of links (or link density);
- We define the threshold as large as possible while guaranteeing that all nodes are connected (or a so-called “giant component” exists).

C. M. van Wijk et al., “*Comparing Brain Networks of Different Size and Connectivity Density Using Graph Theory*”, PLoS ONE 5, e13701 (2010)

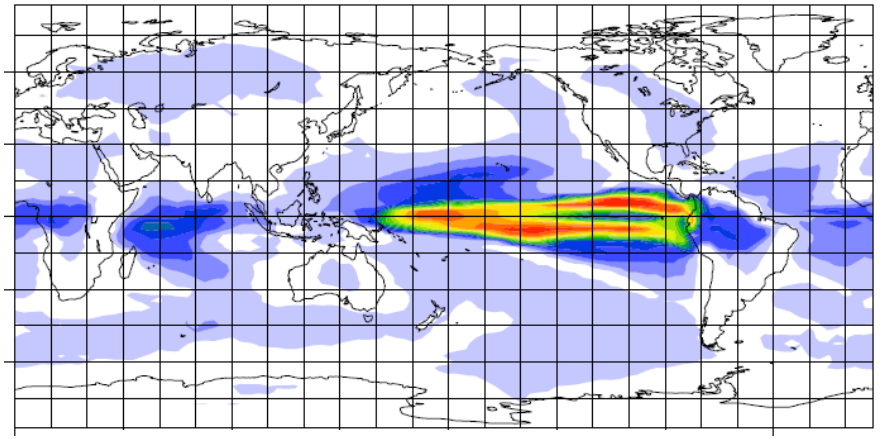
Comparison of area weighted connectivity with |CC| and MI

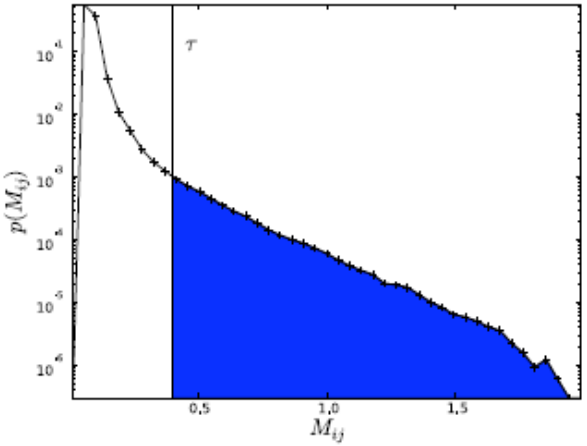
AWC computed with |cross-correlation|





AWC computed with mutual information



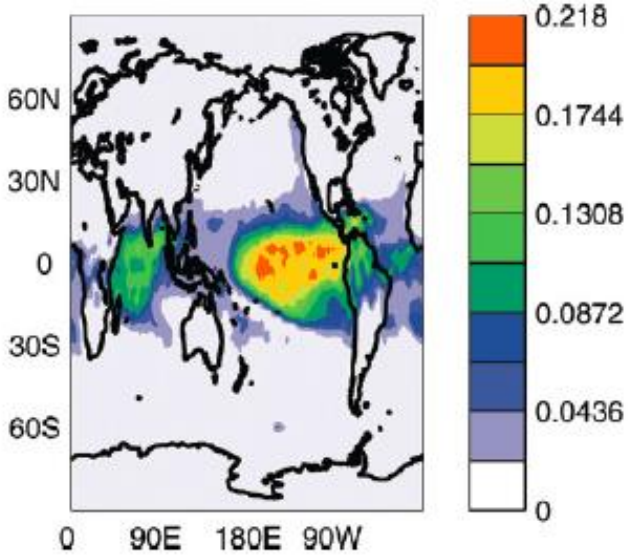


The threshold was selected to give a network with the same link density (0.005)

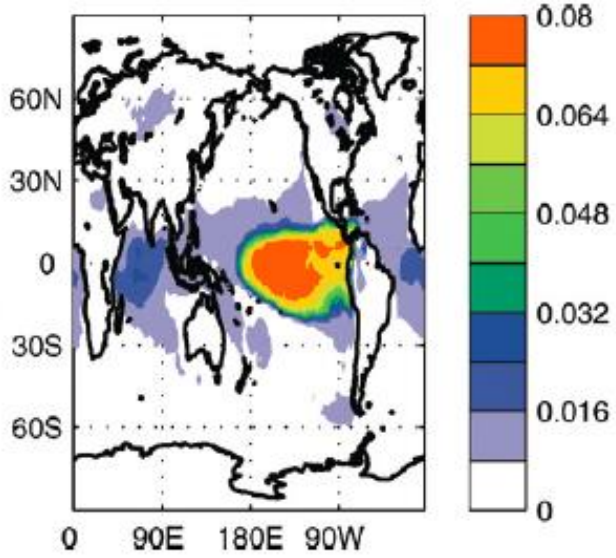
Donges et al, Eur. Phys. J. Special Topics 174, 157 (2009)

Influence of the threshold

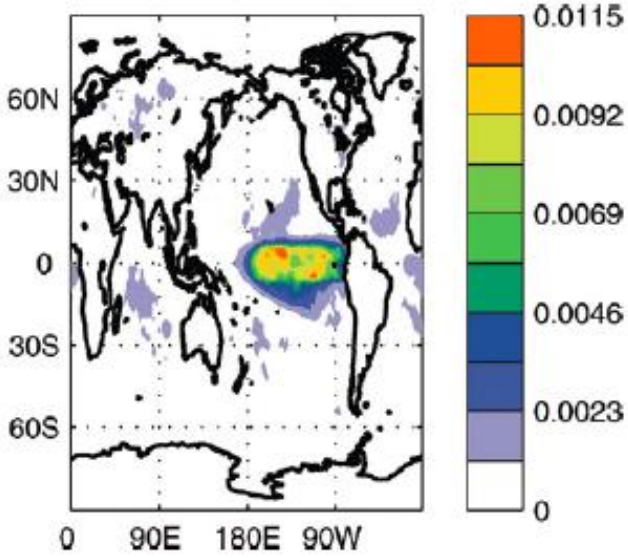
ρ=0.027



ρ=0.01

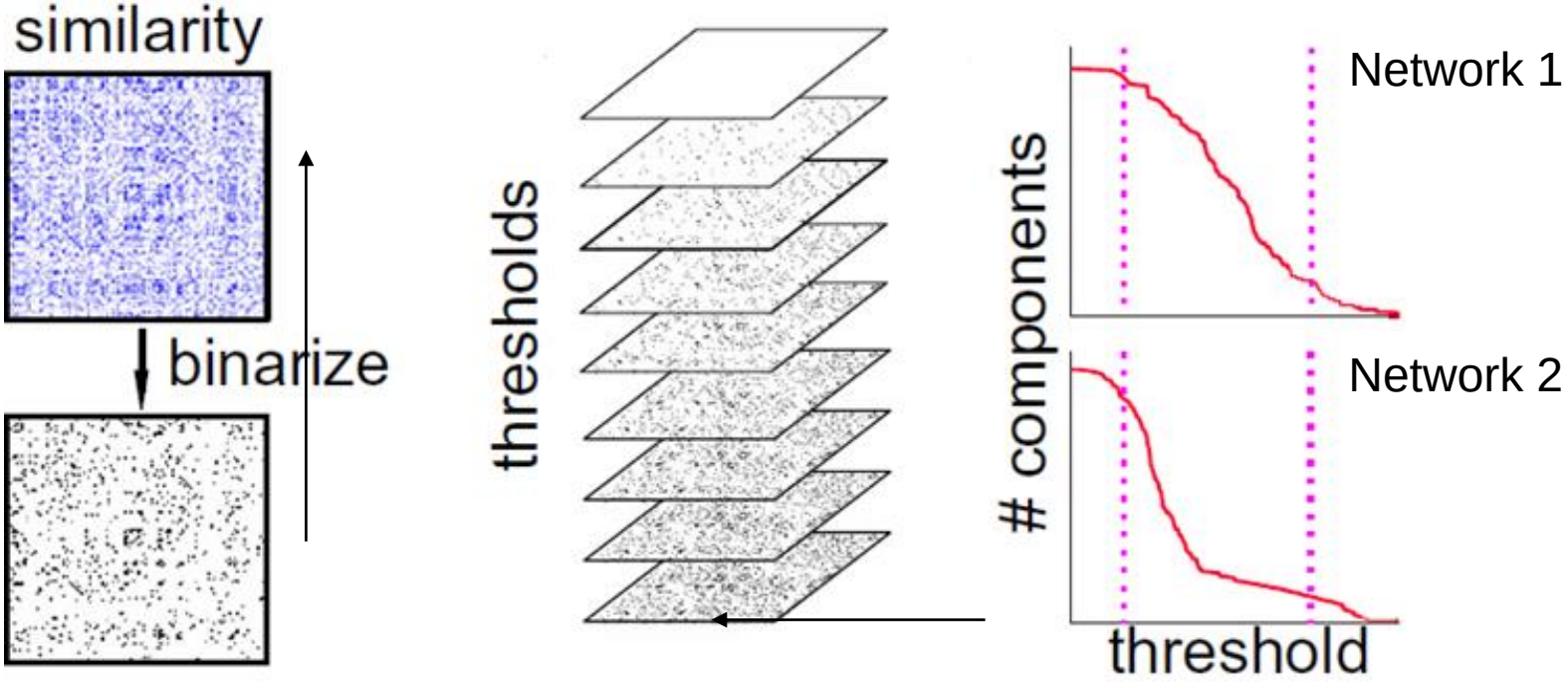


ρ=0.001



M. Barreiro, et. al, Chaos 21, 013101 (2011)

Problems with thresholding



- The number of connected components as a function of threshold reveals different structures.
- But thresholding near the dotted lines indicates (inaccurately) that networks 1 and 2 have similar structures.

Giusti et al., J Comput Neurosci (2016) 41:1–14

Software

Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package

Jonathan F. Donges' , Jobst Heitzig, Boyan Beronov, Marc Wiedermann, Jakob Runge, Qing Yi Feng, Liubov Tupikina, Veronika Stolbova, Reik V. Donner, Norbert Marwan, Henk A. Dijkstra, and Jürgen Kurths

Citation: Chaos **25**, 113101 (2015); doi: 10.1063/1.4934554

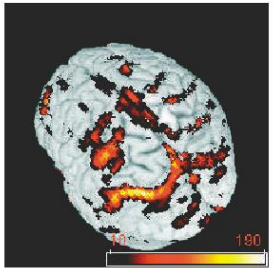
View online: http://dx.doi.org/10.1063/1.4934554

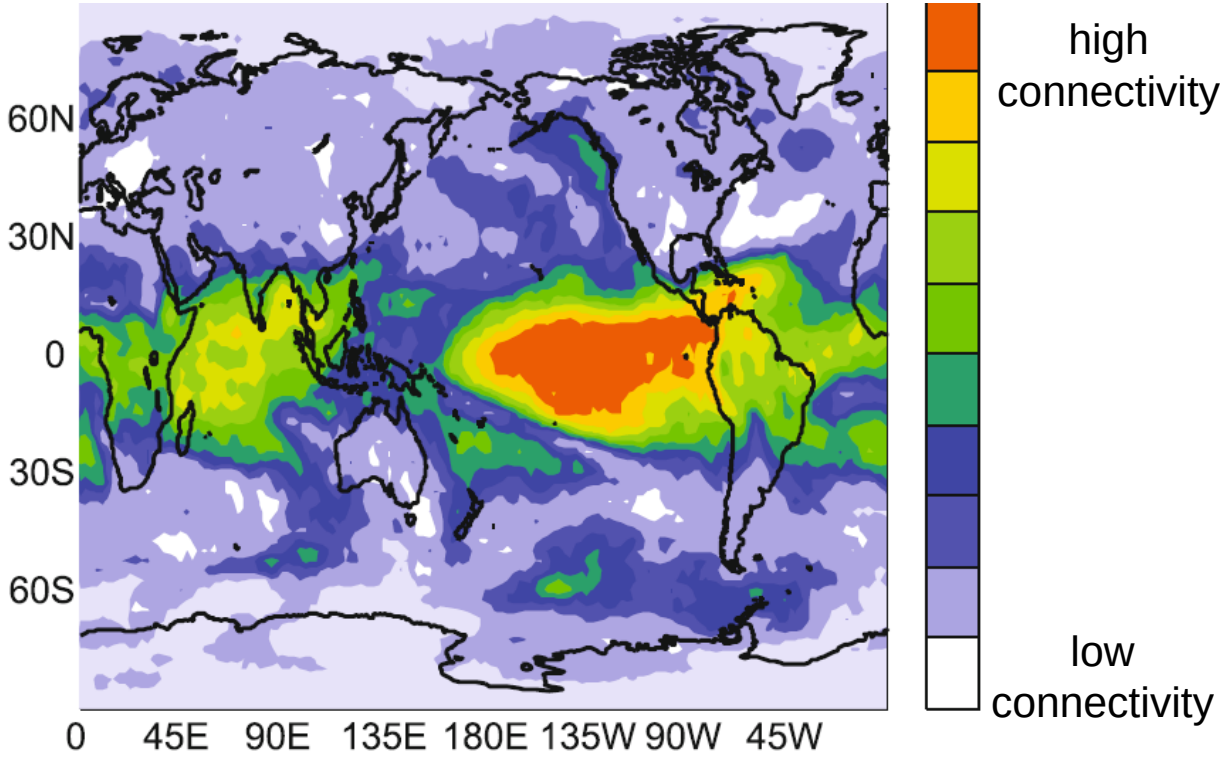
pyunicorn is available at https://github.com/pik-copan/

Climate network with mutual information computed with probabilities of ordinal patterns

$$\mathrm{AWC}_i = \frac{\sum_j^N A_{ij} \cos(\lambda_j)}{\sum_j^N \cos(\lambda_j)}$$

inter-annual time-scale (3 consecutive years). The color-code indicates the Area Weighted Connectivity (weighted degree)



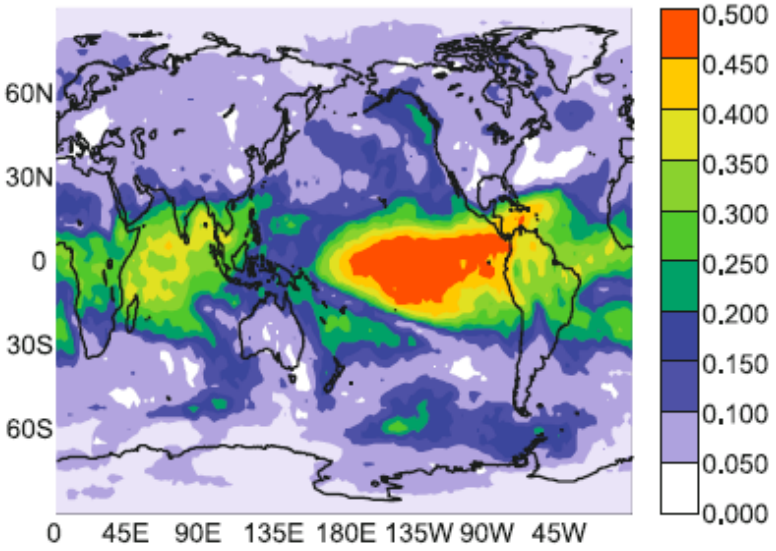


J. I. Deza, M. Barreiro, and C. Masoller, Eur. Phys. J. Special Topics 222, 511 (2013)

Comparison: ordinal probabilities vs. histogram of data values

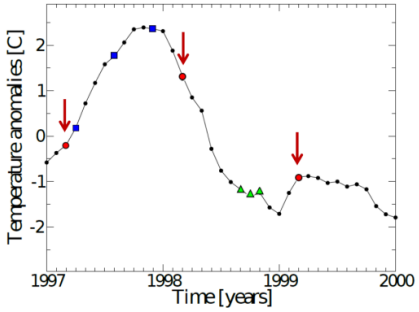
$$M_{ij} = \sum_{m,n} p_{ij}(m,n) \log \frac{p_{ij}(m,n)}{p_i(m)p_j(n)}$$

Network when the probabilities are computed with ordinal analysis

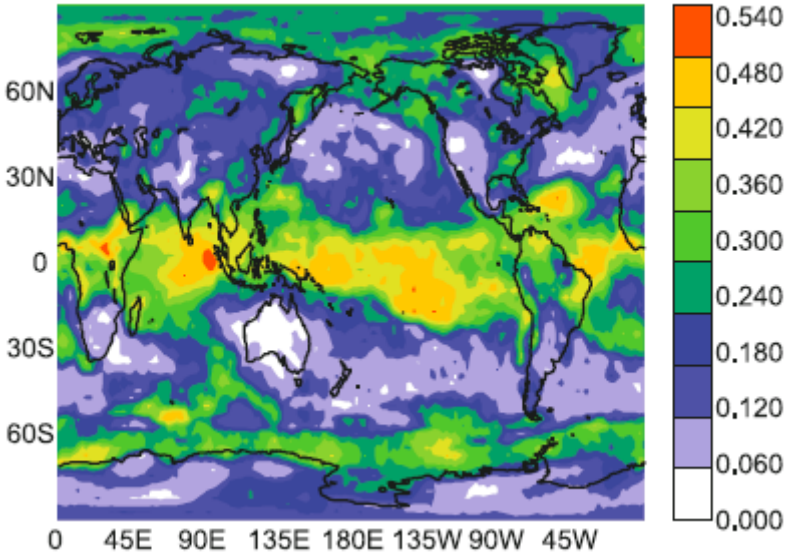


Color code indicates the area-weighted connectivity

inter-annual time scale

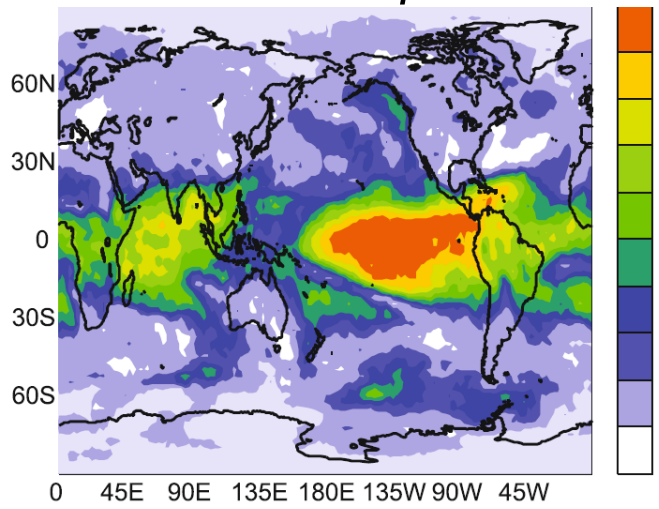


Network when the probabilities are computed with histogram of values

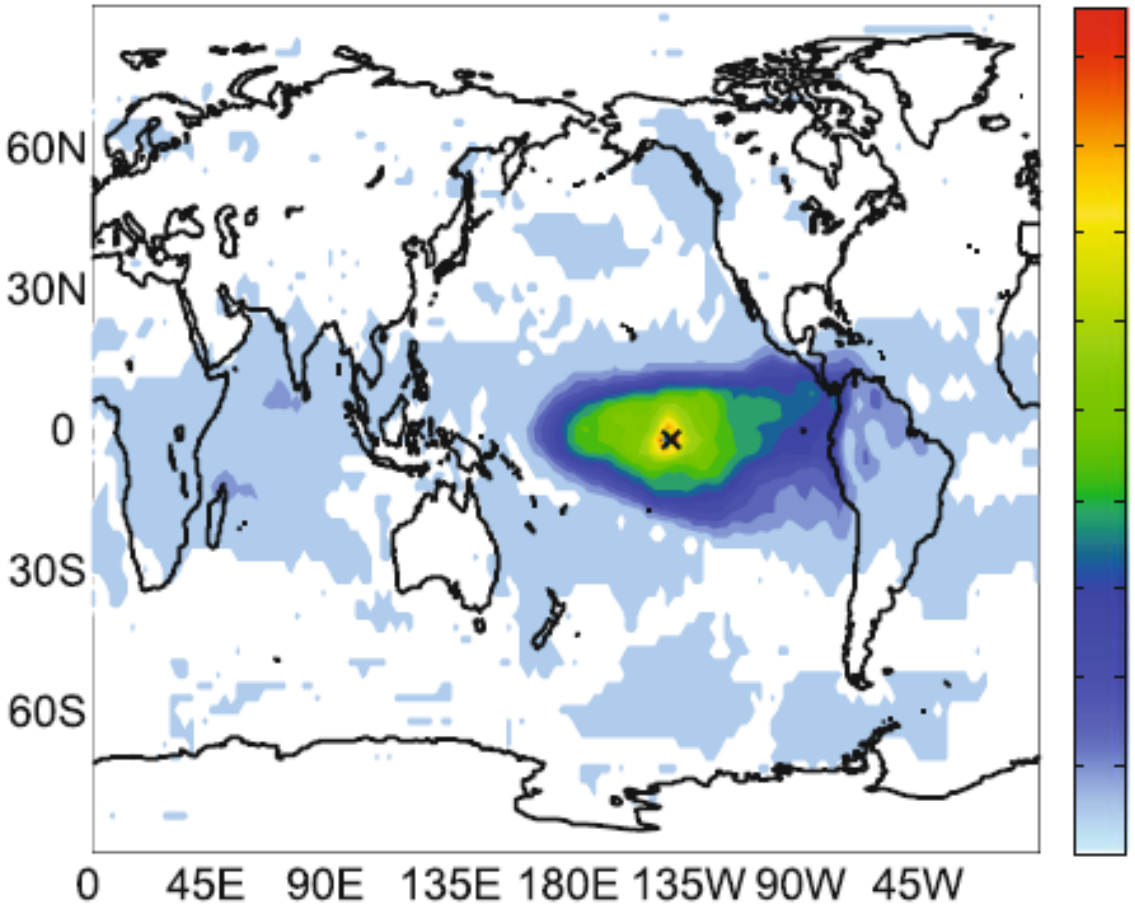


Who is connected to who?

AWC map

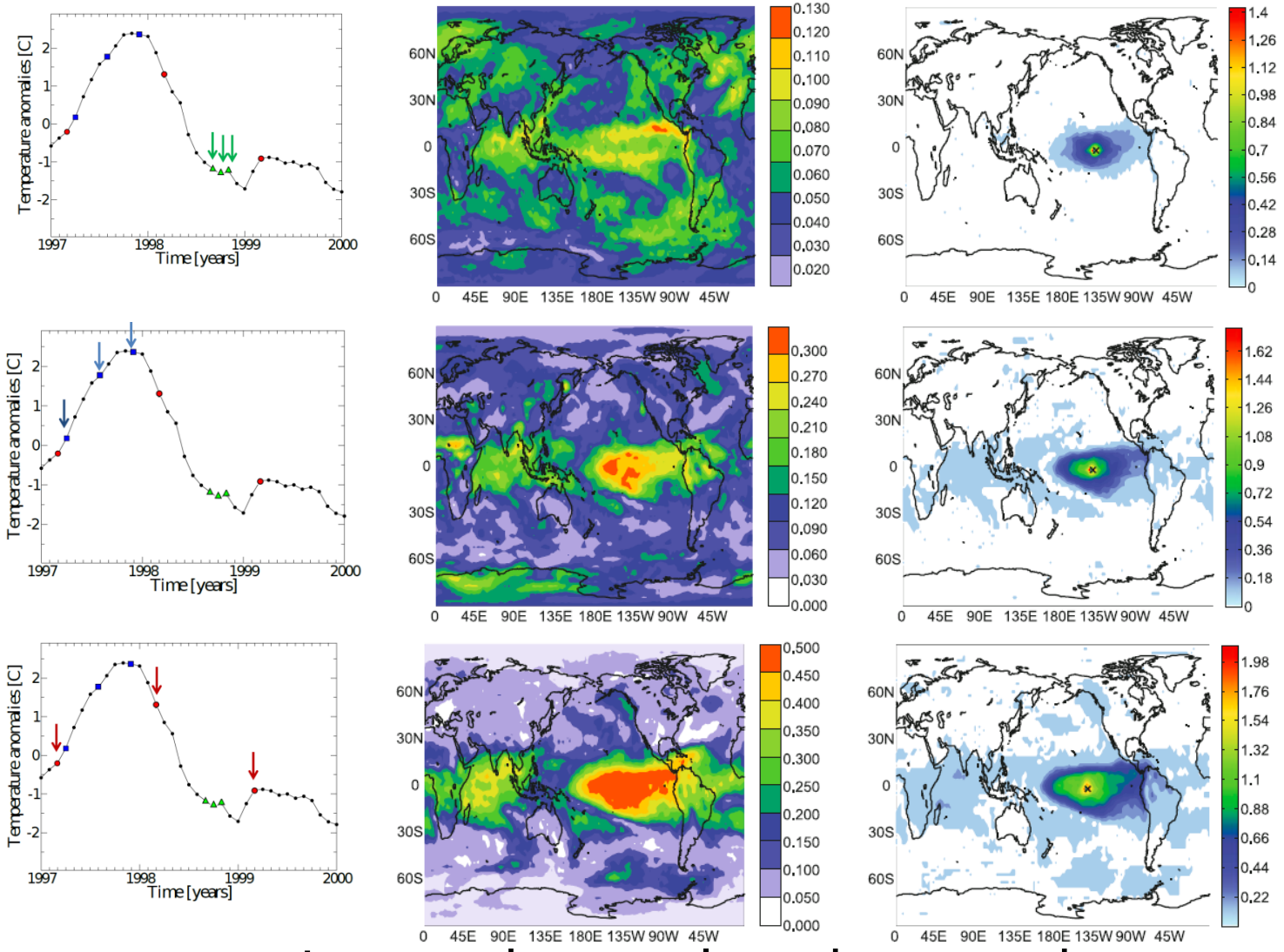


color-code indicates the MI values (only significant values)



J. I. Deza, M. Barreiro, and C. Masoller, Eur. Phys. J. Special Topics 222, 511 (2013)

Influence of the time-scale of the pattern



Longer time-scale ⇒ increased connectivity

Network characterization

Definitions (for unweighted and undirected graphs)

- ***Adjacency matrix***: $A_{ij} = 1$ if i and j are connected, else $A_{ij} = 0$.

- ***Degree*** of a node $k_i = \Sigma_j A_{ij}$

- ***Clustering coefficient***: measures the fraction of a node's neighbors that are neighbors also among themselves

$$C_i = \frac{2R_i}{k_i(k_i-1)} = \frac{1}{k_i(k_i-1)} \sum_{j=1}^{N} \sum_{l=1}^{N} \mathcal{A}_{ij}\mathcal{A}_{jl}\mathcal{A}_{li}$$

R_i is the number of connected pairs in the set of neighbors of node i

- ***Assortativity***: measures the tendency of a node with high/low degree to be connected to other nodes with high/low degree

$$a_i \equiv \frac{1}{k_i} \sum_{j=1}^{N} \mathcal{A}_{ij} k_j$$

How to characterize the degree distribution?

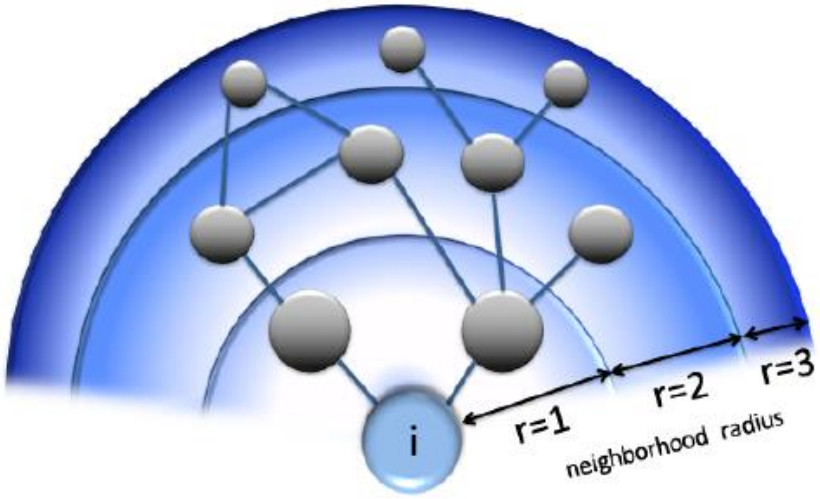
- **Mean** (expected value of X): $\mu = \mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$
- **Variance**: $\sigma^2 = Var\,(X) = \mathrm{E}[(X-\mu)^2]$
- **Skewness**: "measures" the asymmetry of the distribution

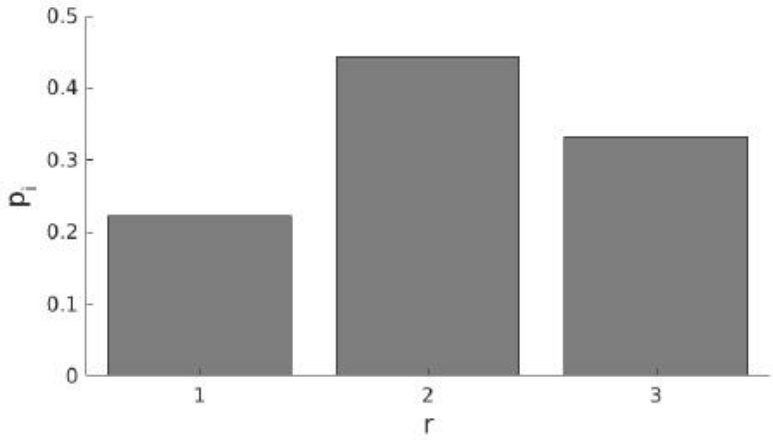
$$Z = \frac{X - \mu}{\sigma} \qquad S = \mathrm{E}[Z^3]$$

- **Kurtosis**: measures the "tailedness" of the distribution. For a normal distribution K=3.

$$K = \mathrm{E}[Z^4]$$

Diameter: longest shortest path





Node Distance Distribution (NDD) of node i: fraction of nodes that are connected to node i at distance r.

*A. Viol et al., Entropy **2019**, 21, 128*

How to compare two distributions (degree, NDD, etc.)?

Distance between two distributions *P* and P_e

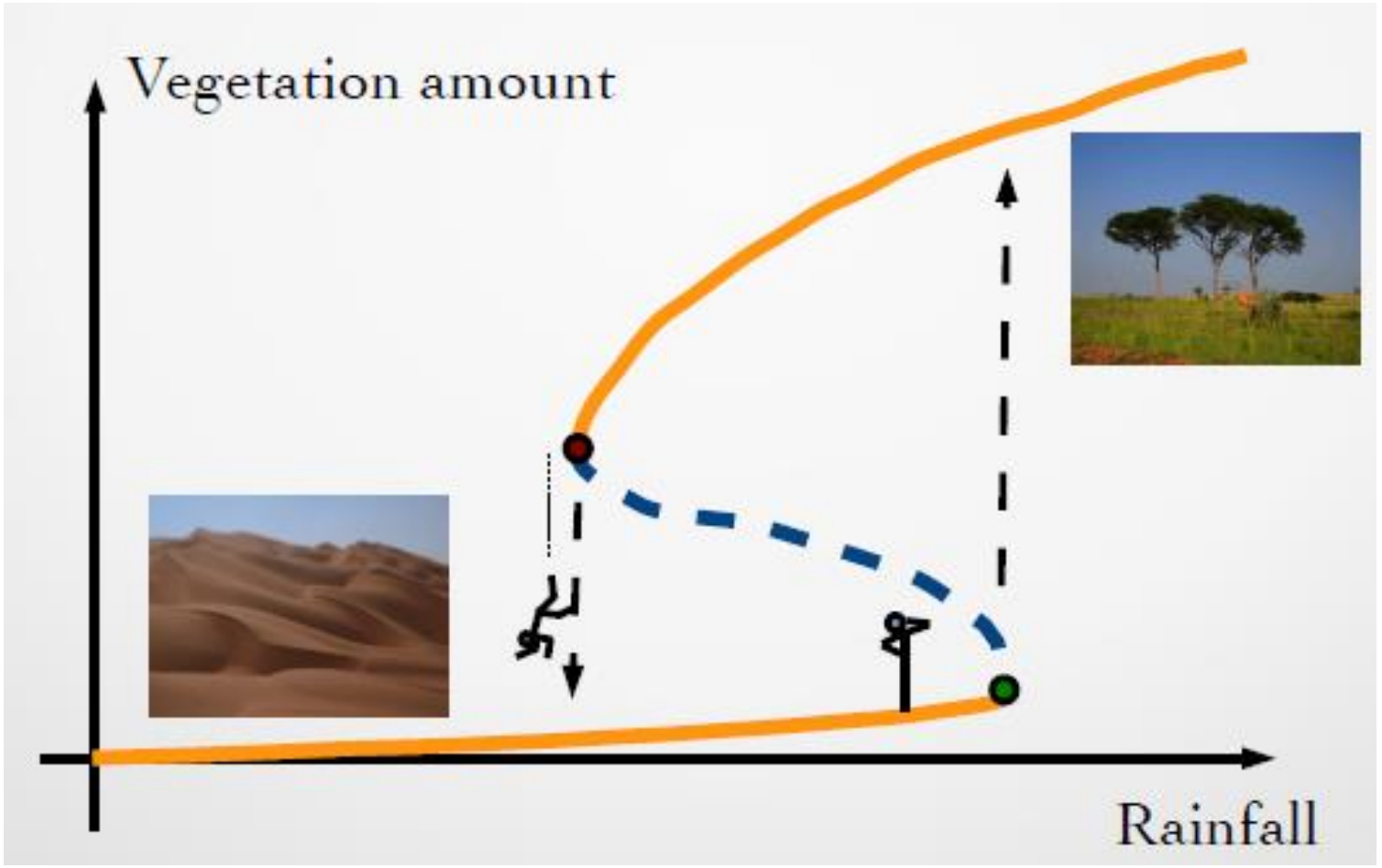
Euclidean $$D_E[P,P_e] = \|P - P_e\|_E = \sum_i \left(p_i - p_{i,e}\right)^2$$

Kullback–Leibler divergence (relative entropy) $$D_{KL}[P,P_e] = \sum_i p_i \log \frac{p_i}{p_{i,e}}$$

Jensen divergence $$D_J[P,P_e] = \frac{K\left[P\middle|P_e\right] + K\left[P_e\middle|P\right]}{2}$$

Read more: S-H Cha: *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions*, Int. J of. Math. Models and Meth. 1, 300 (2007)

Example of application: desertification transition



Our goal: to develop reliable early-warning indicators

Can we use “correlation networks” to detect the approach to a tipping point?

Model

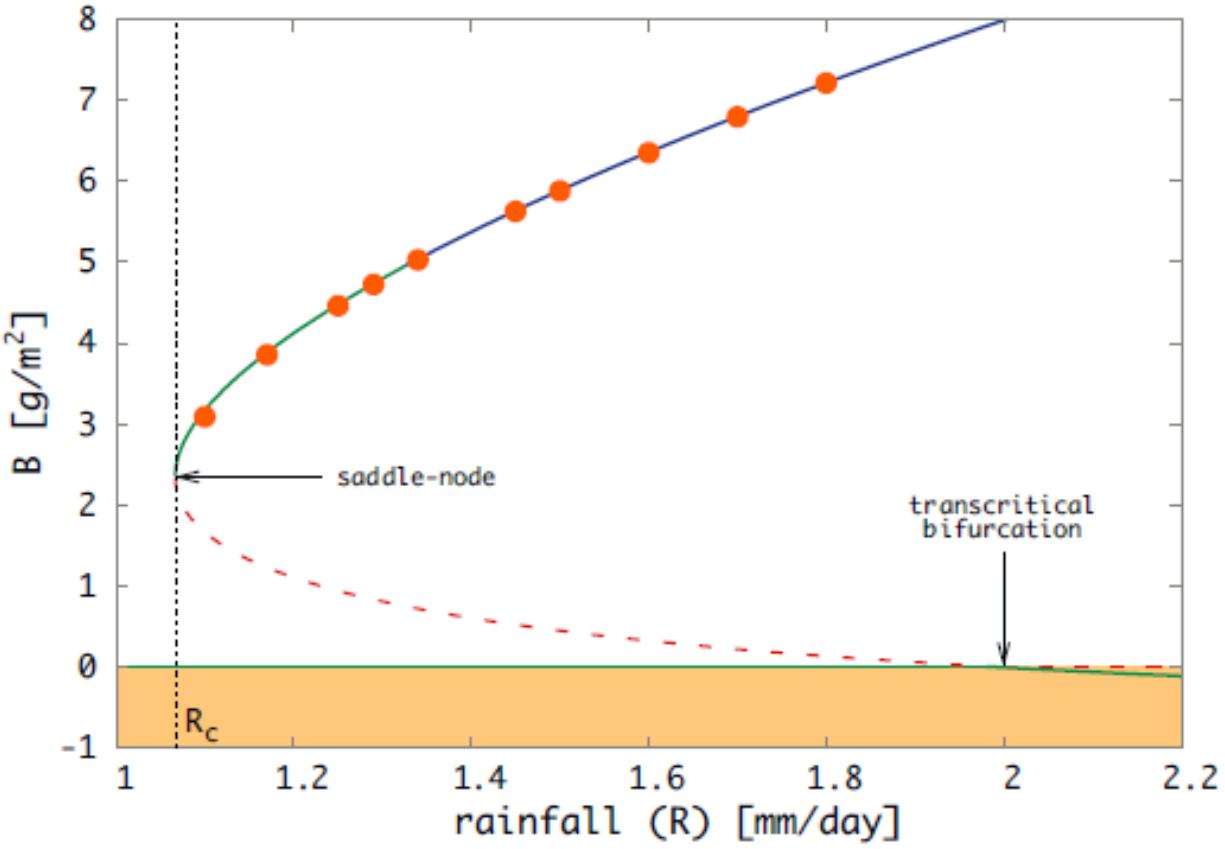
$$\frac{\partial w}{\partial t} = R + \frac{w}{\tau_w} - \Lambda w B + D\nabla^2 w + \sigma_w w_0 \xi^w(t),$$

$$\frac{\partial B}{\partial t} = \rho B\left(\frac{w}{w_0} - \frac{B}{B_c}\right) - \mu \frac{B}{B + B_O} + D\nabla^2 B + \sigma_B B_0 \xi^B(t)$$

- w (in mm) is the soil water amount
- B (in g/m^2) is the vegetation biomass
- Uncorrelated Gaussian white noise
- R (rainfall) is the bifurcation parameter

Shnerb et al. (2003), Guttal & Jayaprakash (2007), Dakos et al. (2011)

Saddle-node bifurcation

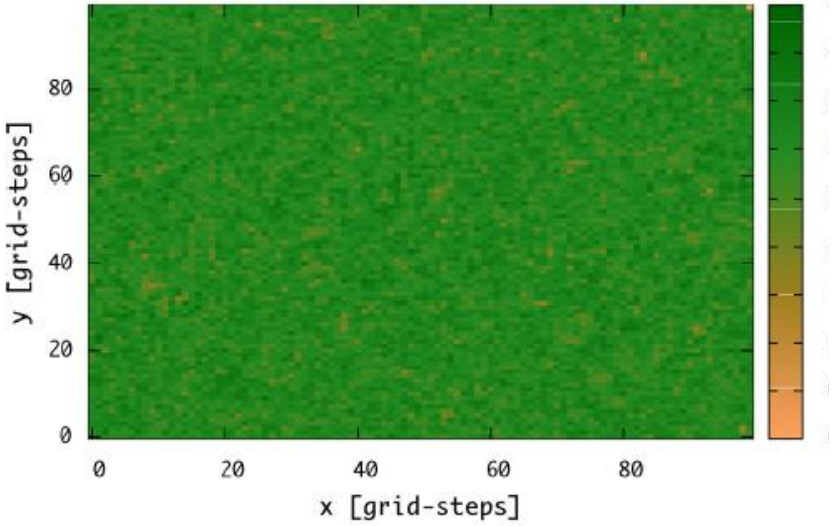


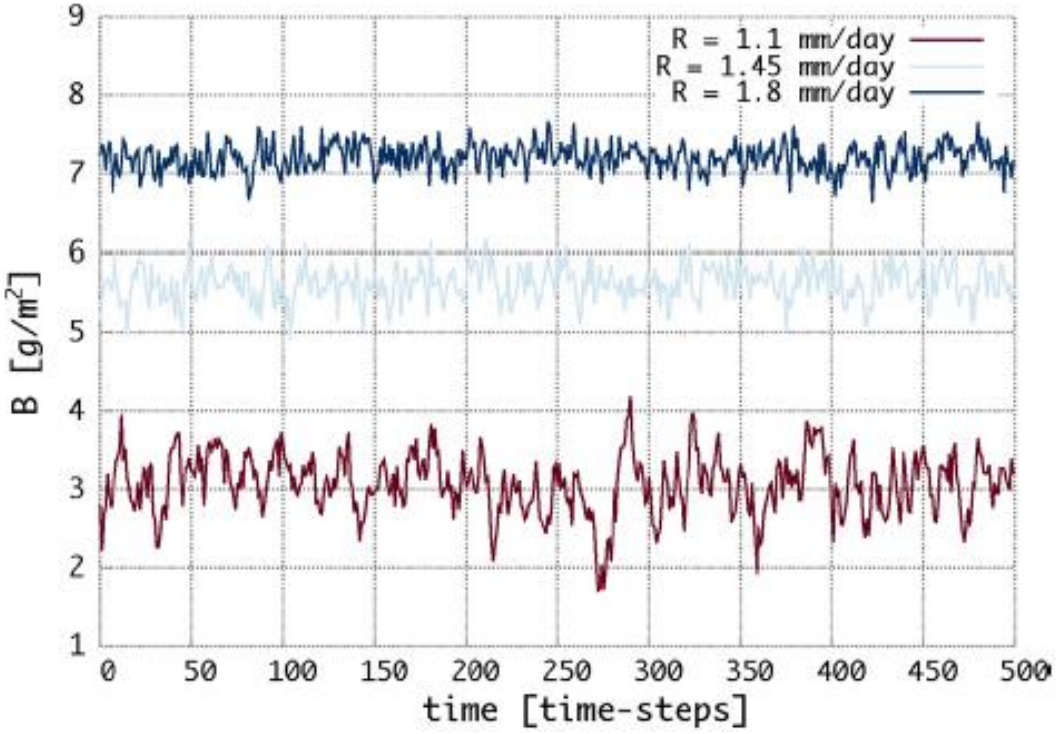
$R<R_c$: only desert-like solution (B=0)

R_c = 1.067 mm/day

Biomass time series

Biomass *B* when *R*=1.1 mm/day





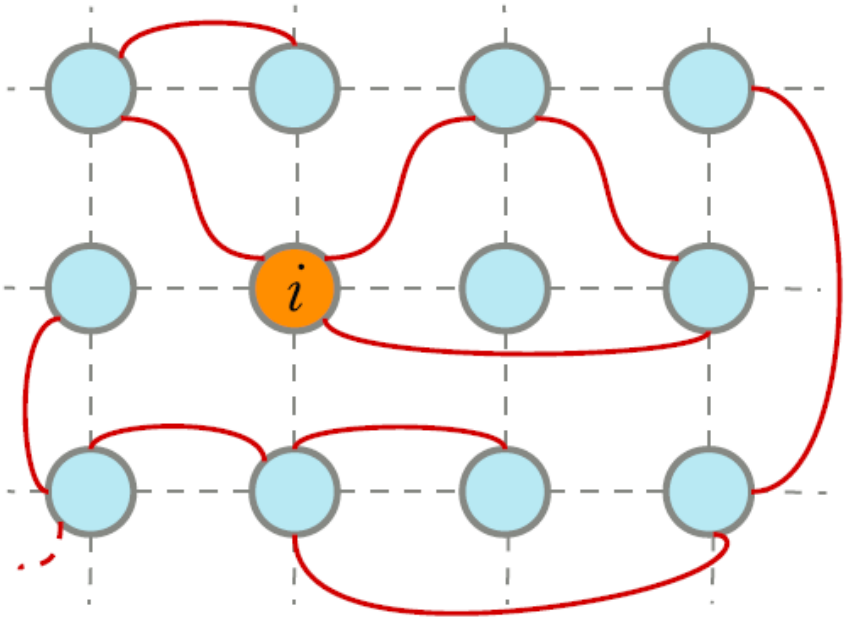
100 m x 100 m = 10^4 grid cells
Simulation time 5 days in 500 time steps
Periodic boundary conditions

Correlation Network

$S_{ij} > Th \Rightarrow A_{ij} = 1$, else $A_{ij}=0$

Statistical similarity measure: Pearson coef.= |zero-lag cross-correlation|

Threshold: Th=0.2 keeps only significant correlations (p<0.05)



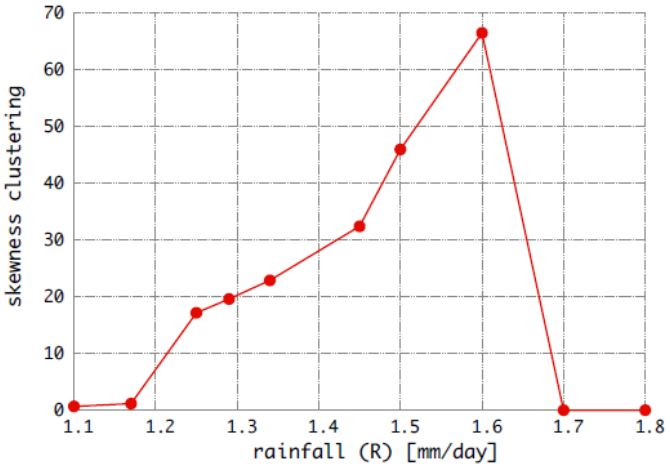
G. Tirabassi et al., Ecological Complexity (2014)

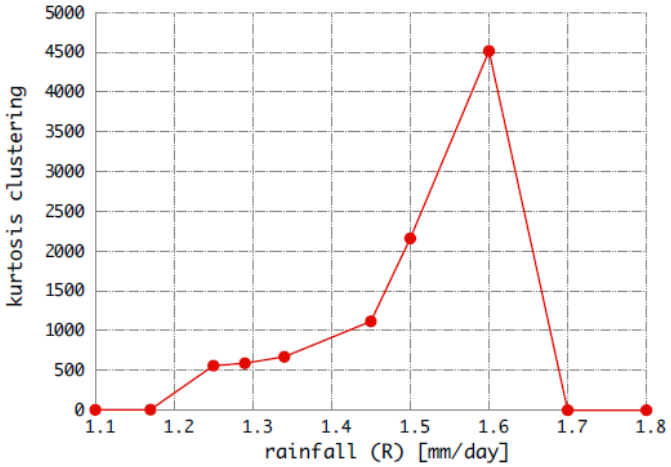
“Gaussianization” of the distributions of a_i & c_i as the tipping point is approached

skewness

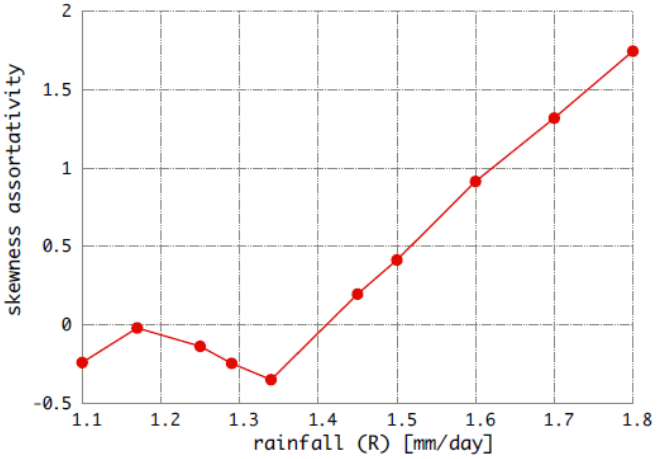
kurtosis

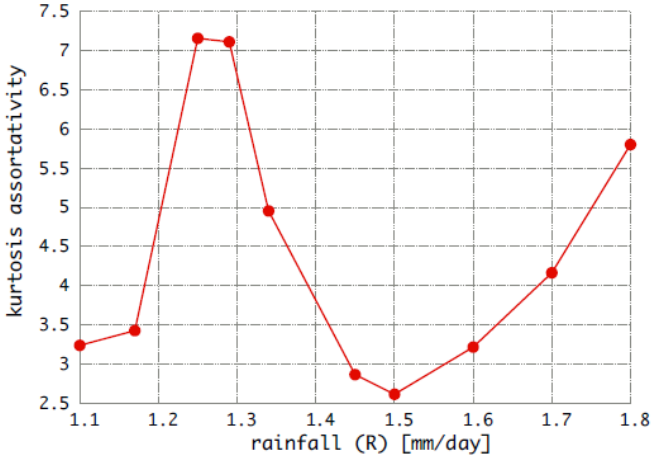
clustering



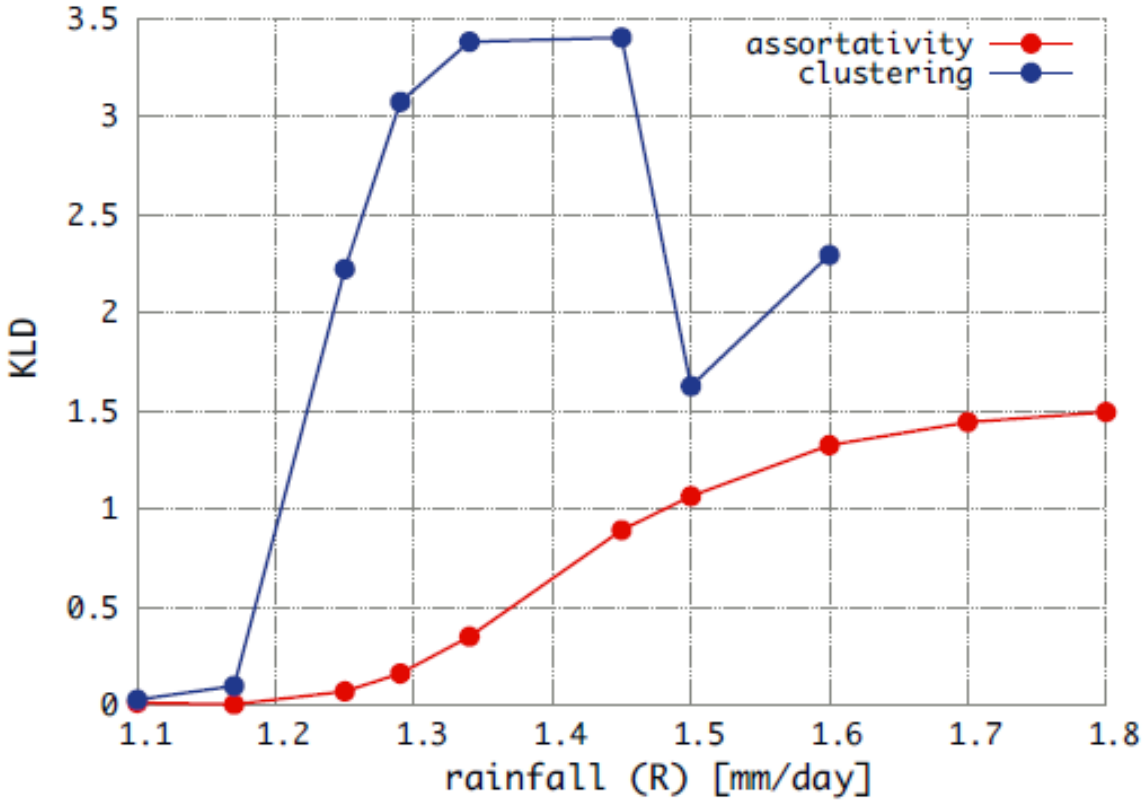


assortativity





The “Gaussianisation” is quantified by the Kullback distance to a Gaussian (Z) distribution



$$\mathrm{KLD} \equiv \int_{-\infty}^{\infty} \ln\left(\frac{P(x)}{Z(x)}\right) P(x)\, dx.$$

- Open issue: the “Gaussianisation” might be a model-specific feature.
- How to precisely quantify changes of the network?
- We need a distance to compare graphs.

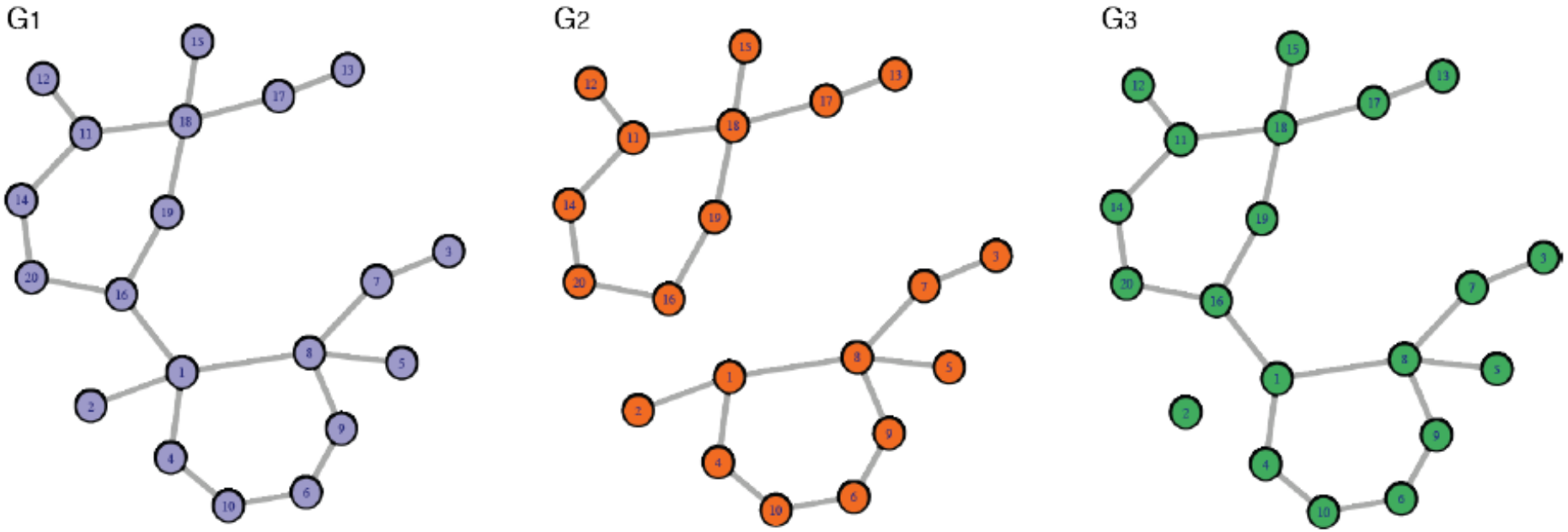
G. Tirabassi et al., Ecological Complexity 19, 148 (2014)

How to compare different networks?

Labelled networks with the same size

- Hamming distance $$d_{\text{Hamming}}(\boldsymbol{y}_1, \boldsymbol{y}_2) = \sum_{i \neq j}^{N} \left[A_{ij}^{(1)} \neq A_{ij}^{(2)} \right]$$

- Main problem: not all the links have the same importance.



L. Carpi et al, Sci. Reports 9, 4511 (2019)

In order to detect structural differences we need a precise measure to compare networks

- Degree, centrality, assortativity distributions etc. provide *partial* information.
- How to define a measure that contains detailed information about the *global topology* of a network, in a *compact* way?

$\Rightarrow$ Node Distance Distributions (NDDs)

- **$p_i(r)$** of node **i** is the fraction of nodes that are connected to node **i** at distance **r**
- A network with N nodes is characterized by N pdfs

$$\{p_1, p_2, \ldots, p_N\}$$

- If two networks have the same set of NDDs $\Rightarrow$ they have the same diameter, the same average path length, etc.

How to summarize the information contained in the node distance distributions?

The *Network Node Dispersion (NND*) of a graph G quantifies the heterogeneity of the node distance distributions $\{p_1, p_2, \ldots, p_N\}$

average NDD: $\mu = \langle p_i \rangle_i$

Kullback distance between p_i and μ: $J(p_i, \mu) = \sum_r p_i(r) \log\left(\frac{p_i(r)}{\mu(r)}\right)$

$$NND(G) = \frac{\langle J(p_i, \mu) \rangle_i}{\log(d+1)}$$

d = diameter

Dissimilarity between two networks

$$D(G,G') = w_1\sqrt{\frac{\mathcal{J}(\mu_G,\mu_{G'})}{\log 2}} + w_2\left|\sqrt{\mathrm{NND}(G)} - \sqrt{\mathrm{NND}(G')}\right|$$

$w_1=w_2=0.5$

compares the averaged connectivity

compares the heterogeneity of the connectivity distances

- Extensive numerical experiments demonstrate that isomorphic graphs return ***D=0.***
- Computationally efficient.
- Can be used to compare graphs with different number of nodes.

T. A. Schieber et al, Nat. Comm. 8, 13928 (2017)

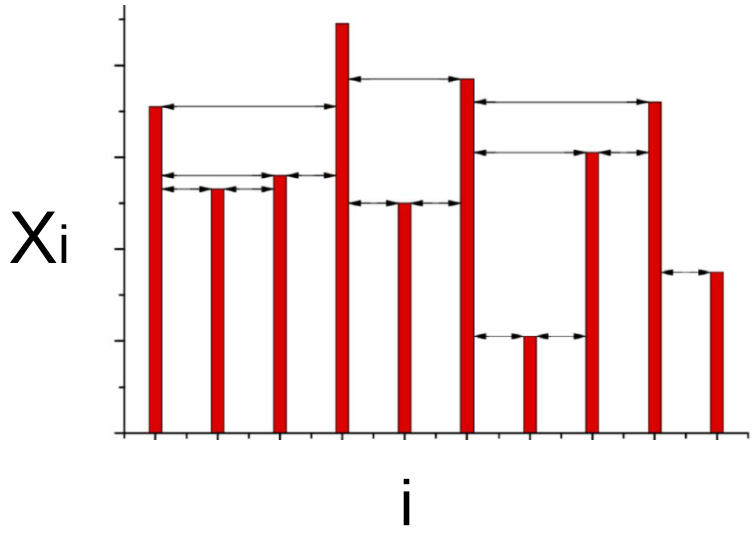
First application: comparing brain networks

- EEG data *
  - 64 electrodes placed on the subject's scalp sampled at 256 Hz during 1s
  - 107 subjects: 39 control and 68 alcoholic
- Use HVG to transform each EEG TS into a network G.
- Weight between two brain regions: **1-*D*(G,G')**
- The resulting network represents the weighted similarity between the brain regions of an individual.

  $\Rightarrow$ We can compare the different individuals.

* https://archive.ics.uci.edu/ml/datasets/eeg+database

(Reminder) How to represent a time series as a network?: the horizontal visibility graph (HVG)



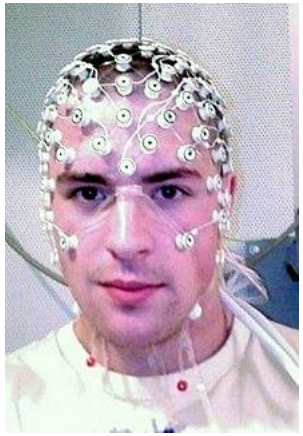
Rule: data points *i* and *j* are connected if there is "visibility" between them

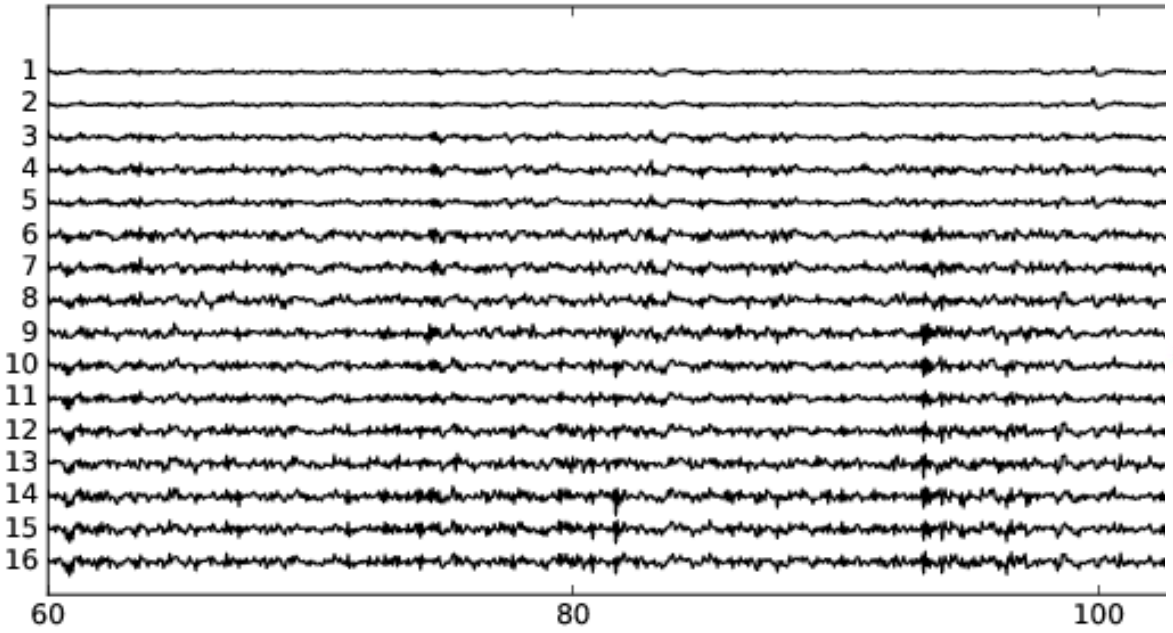
$\Rightarrow$ **Unweighted and undirected graph**

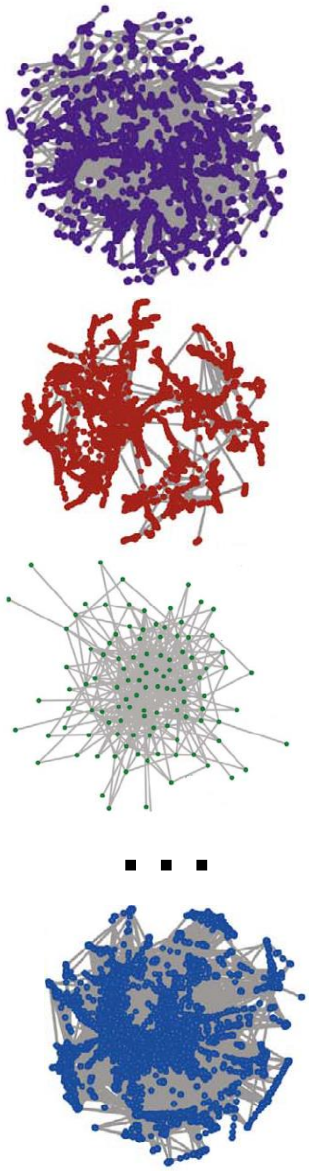
Parameter free!

Luque et al PRE (2009); Gomez Ravetti et al, PLoS ONE (2014)

For each subject, the time series recorded at each electrode is transformed into a graph



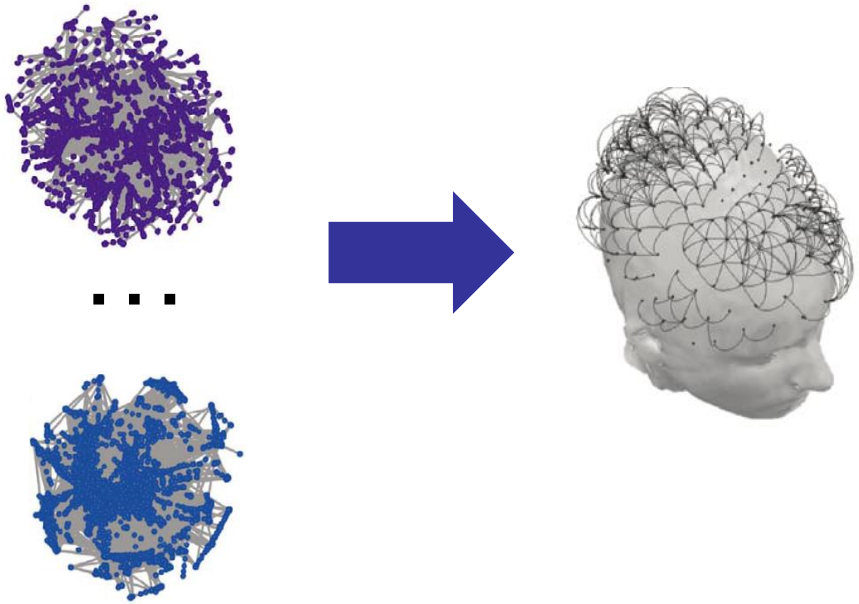




Dataset has 64 channels $\Rightarrow$ 64 networks

The brain network of each subject

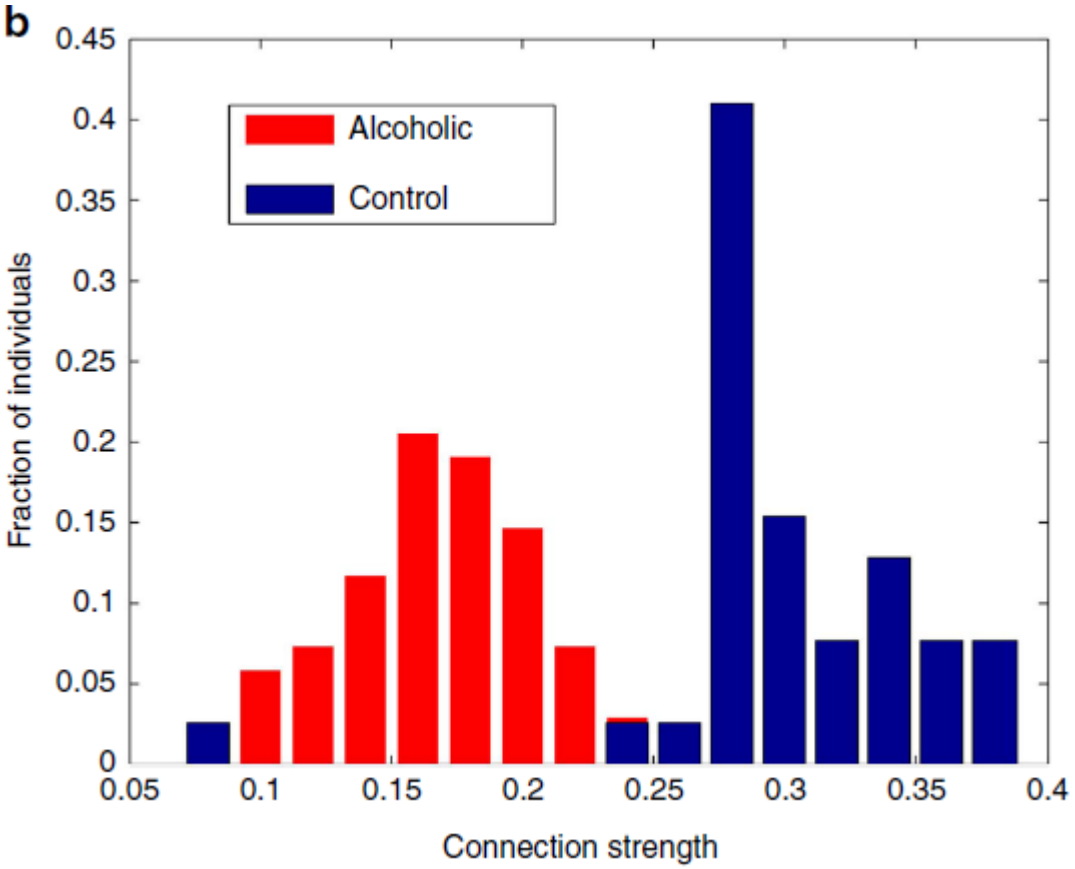
- The weight of the link between two graphs (G, G') representing two brain regions is defined as: **1-*D*(G,G')**



- The resulting network (with 64 nodes=electrodes, all-to-all coupled with weighted links) represents the similarity between the EEG signals in different brain regions of one subject.

  $\Rightarrow$ We can then compare different subjects.

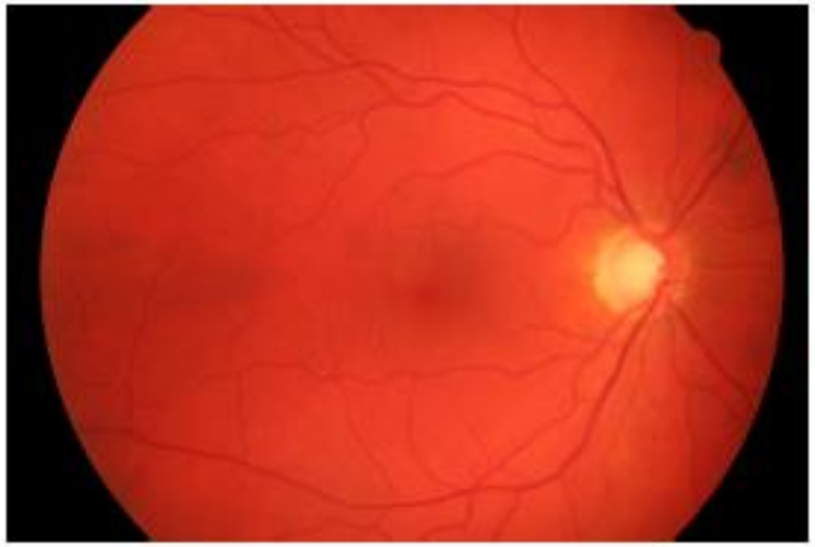
We identify two brain regions (called ‘nd’ and ‘y’), where the connection strength between these regions is higher in control than in alcoholic subjects.

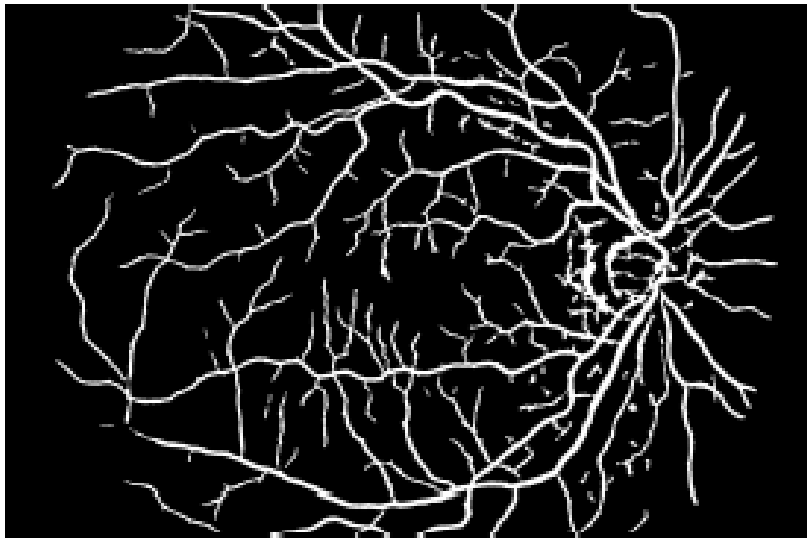


T. A. Schieber et al, Nat. Comm. 8, 13928 (2017)

Second application: classification of retina images

Pablo Amil (UPC), Fabian Reyes (Mexico) & Irene Sendiña (Madrid)





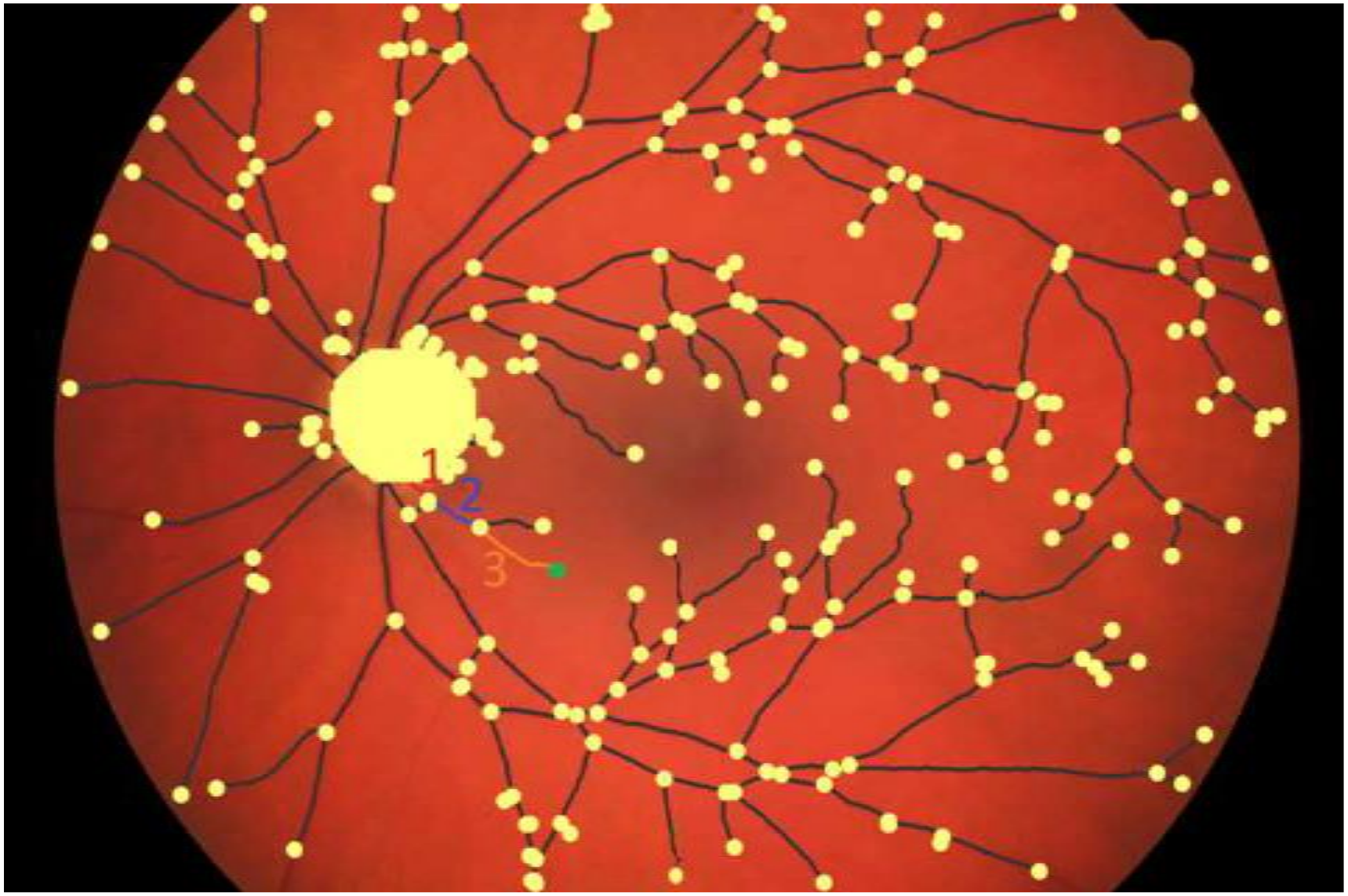
ITN

BE-OPTICAL

Advanced Biomedical Optical Imaging and Data Analysis

Beoptical.eu

Network identification and analysis of the connectivity paths to the central node

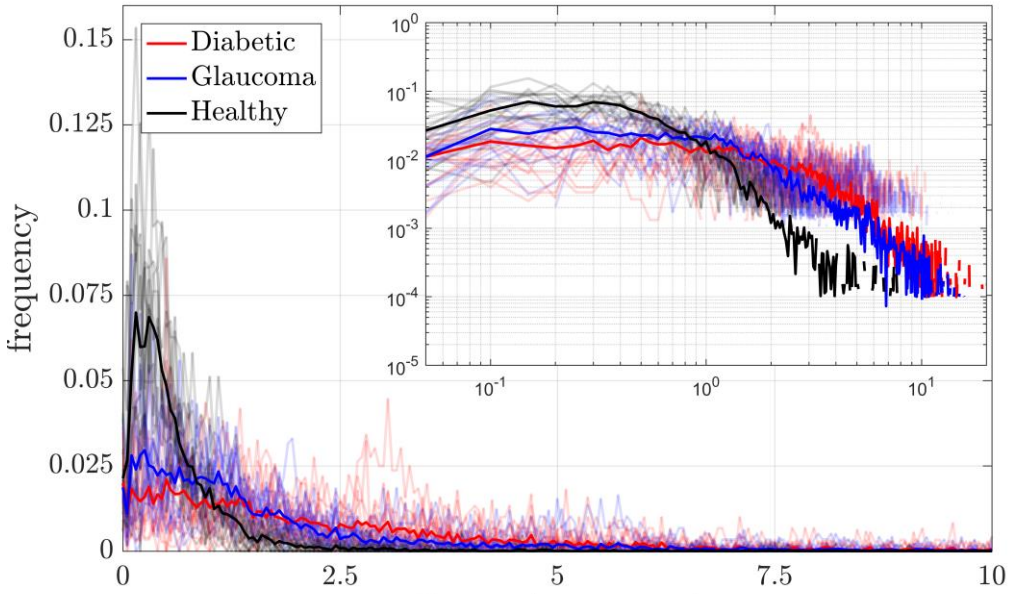


Method (1/2)

- From each image we calculate
- Distribution of distances to the central node
- Distribution of average weights along the path to central node

$$w_{i,j} = (L_{i,j})^l \, (W_{i,j})^a$$

length and width (in # of pixels) of the segment that connects nodes i and j.



distribution of average weights
(l=1, a=-2)

Method (2/2)

- We use the Jensen-Shanon (JS) divergence to compare the distributions: for each image “i“ we obtain a vector

$$\{d_{i1}, d_{i2}, \ldots d_{iN}\} \text{ (N = number of image)}$$

whose elements are the distances between the distributions extracted from image i and image j.

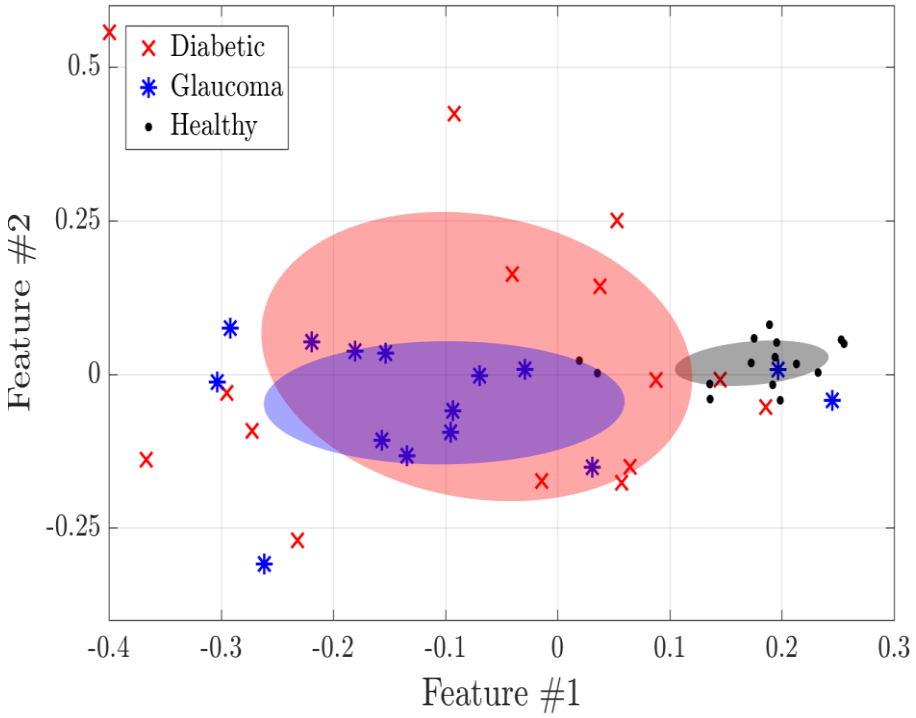
- $\{d_{i1}, d_{i2}, \ldots d_{iN}\}$ are N “features” that characterize image i.

- We apply a nonlinear dimensionality reduction algorithm (*IsoMap*) to obtain only 2 features for each image.

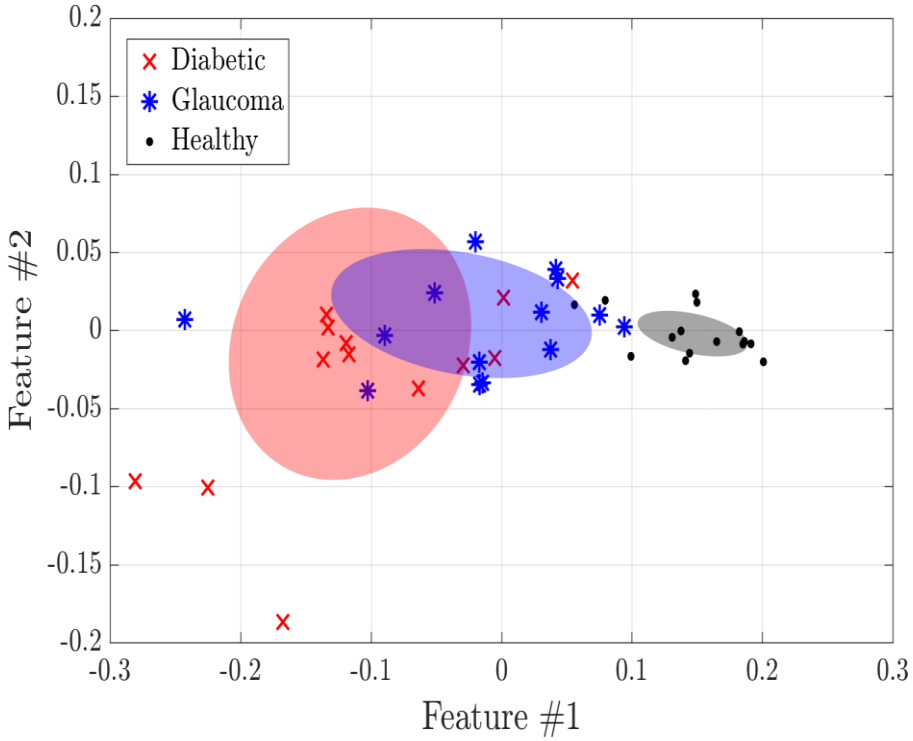
J. B. Tenenbaum et al, A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319 (2000).

Results

Distance distribution to central node



Mean weight distribution



P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE in press (2019).

How to “infer” interactions from observed data?

A classification problem

$$S_{ij} > \mathrm{Th} \Rightarrow A_{ij} = 1, \text{ else } A_{ij}=0$$

- How to select the threshold?
- In “spatially embedded networks”, nearby nodes have the strongest links.
- How to keep **weak-but-significant** links?

- There are many **statistical similarity measures** to infer interactions from observations, i.e., to classify:
  - the interaction exists (is significant)
  - the interaction does not exists (or is not significant)

Goal: use a system with known connectivity to test the performance of statistical similarity measures

Observed time series in nodes i and j: $a_i(t)$, $a_j(t)$, $t=1, \ldots, T$ *(normalized* $\mu=0$, $\sigma=1$*)*

Lagged |cross correlation|: $$CC_{ij}(\tau) = \frac{1}{T-\tau_{\max}} \left| \sum_{t=0}^{T-\tau_{\max}} a_i(t)\, a_j(t+\tau) \right|$$

Statistical Similarity Measure:

$$S_{ij} = \max | CC_{ij}(\tau) | = | CC_{ij}(\tau_{ij}) | \qquad \tau_{ij} \text{ in } [0, \tau_{max}]$$

We compare with the Mutual Information, computed from probabilities of "raw" values and from ordinal probabilities

G. Tirabassi et al., "*Inferring the connectivity of coupled oscillators from time-series statistical similarity analysis*", Sci. Rep. **5** 10829 (2015).

Kuramoto oscillators in a random network

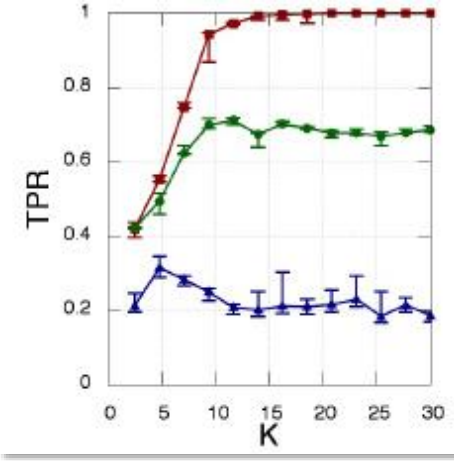
$$\mathrm{d}\theta_i = \omega_i \mathrm{d}t + \frac{K}{N}\sum_{j=1}^{N} A_{ij}\sin(\theta_j - \theta_i)\mathrm{d}t + D\ \mathrm{d}W_t^i$$

A_{ij} is a symmetric random matrix; N=12 time-series, each with 10^4 data points.

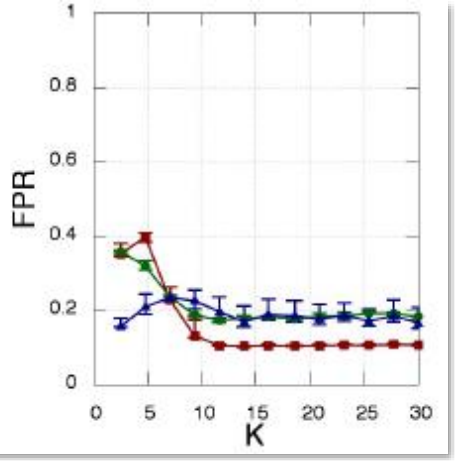
CC MI MIOP

Phases (θ)

True positives

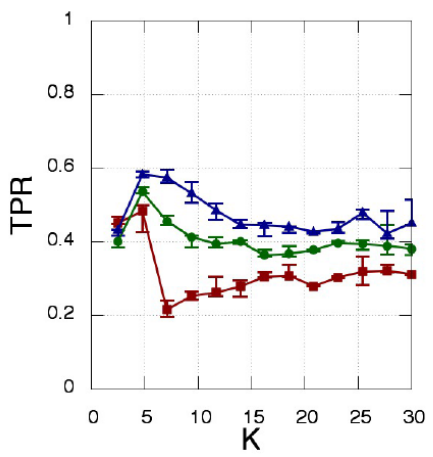


False positives

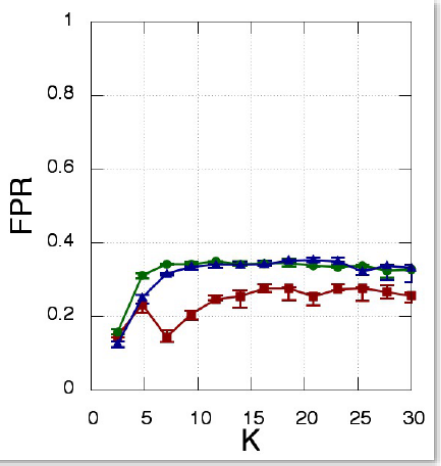


"Observable" Y=sin(θ)

True positives

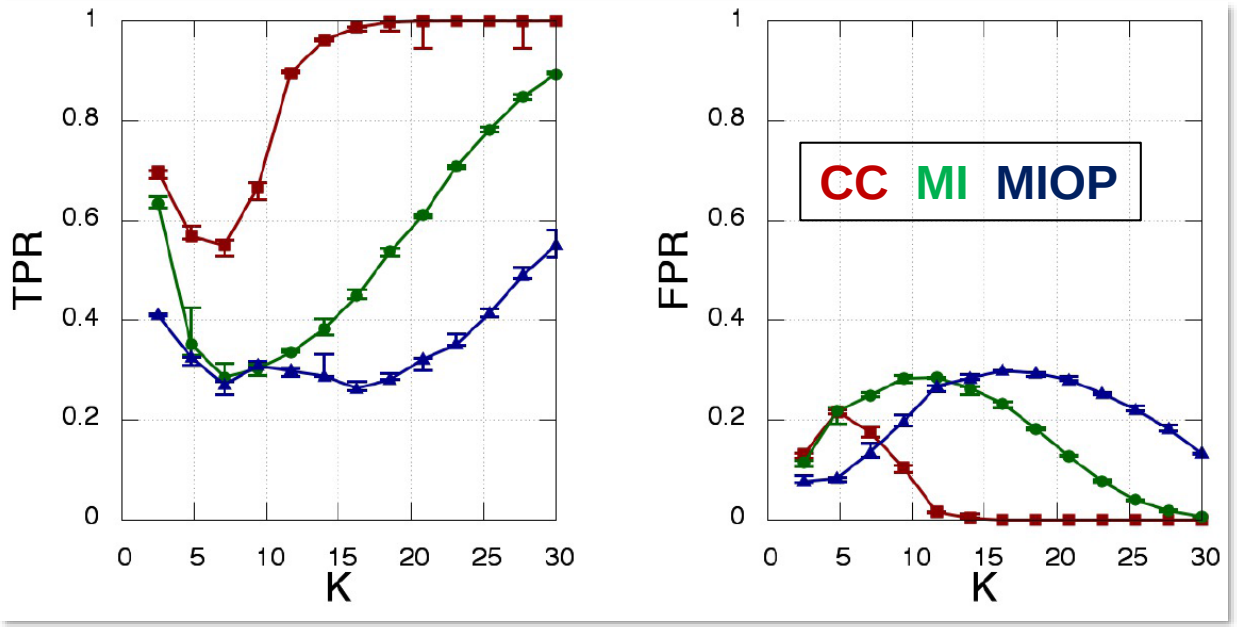


False positives



Results of a 100 simulations with different oscillators' frequencies, random matrices, noise realizations and initial conditions.
For each K, the threshold was varied to obtain optimal reconstruction.

Instantaneous frequencies (dθ/dt)



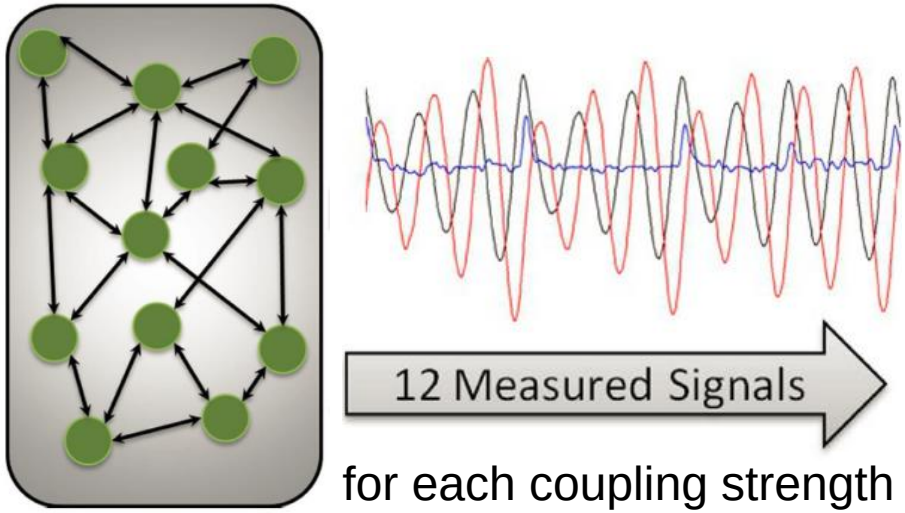
Perfect network inference is possible!

BUT

- the number of oscillators is small (12),
- the coupling is symmetric ($\Rightarrow$ only 66 possible links) and
- the data sets are long (10^4 points)

G. Tirabassi et al, Sci. Rep. **5** 10829 (2015)

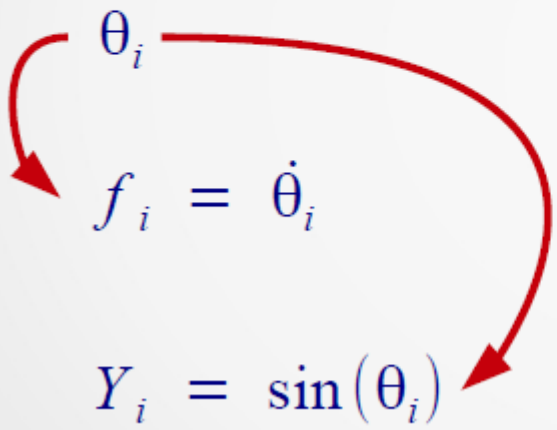
We also analyzed experimental data recorded from 12 chaotic Rössler electronic oscillators (symmetric and random coupling)



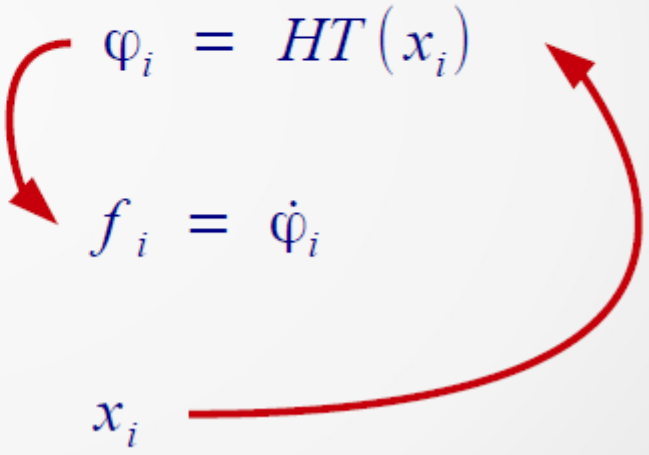
for each coupling strength

The Hilbert Transform was used to obtain phases from experimental data

- Kuramoto Oscillators' Network



- Rössler Oscillators' Network

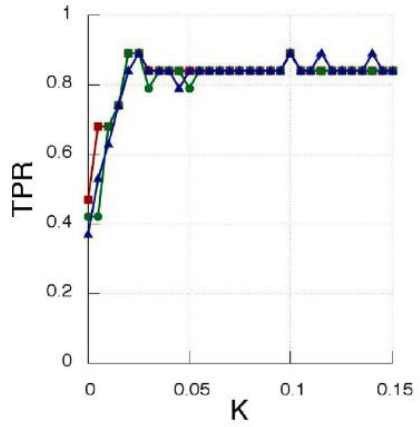


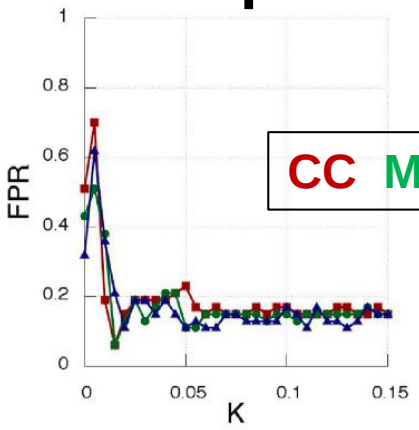
Results obtained with experimental data

Observed variable (x)

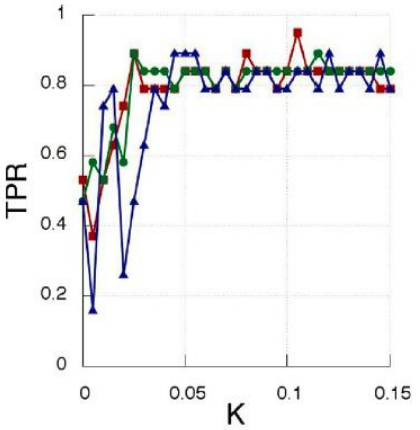
Hilbert phase

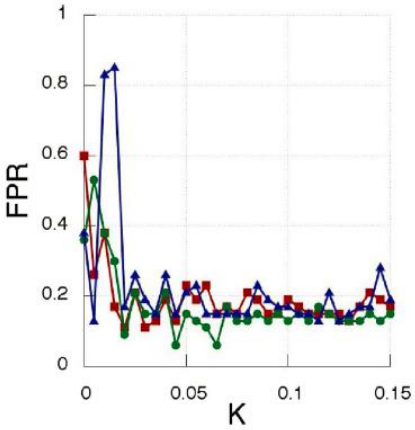
Hilbert frequency

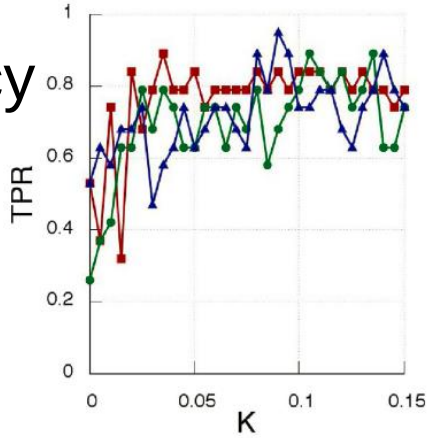


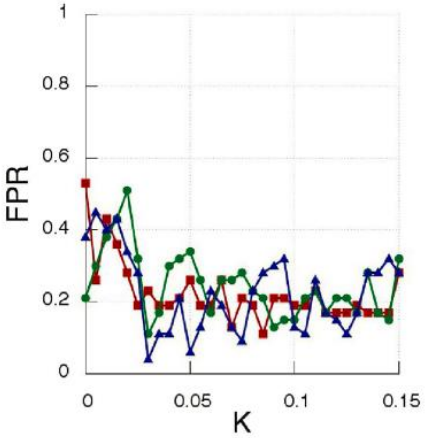


CC MI MIOP





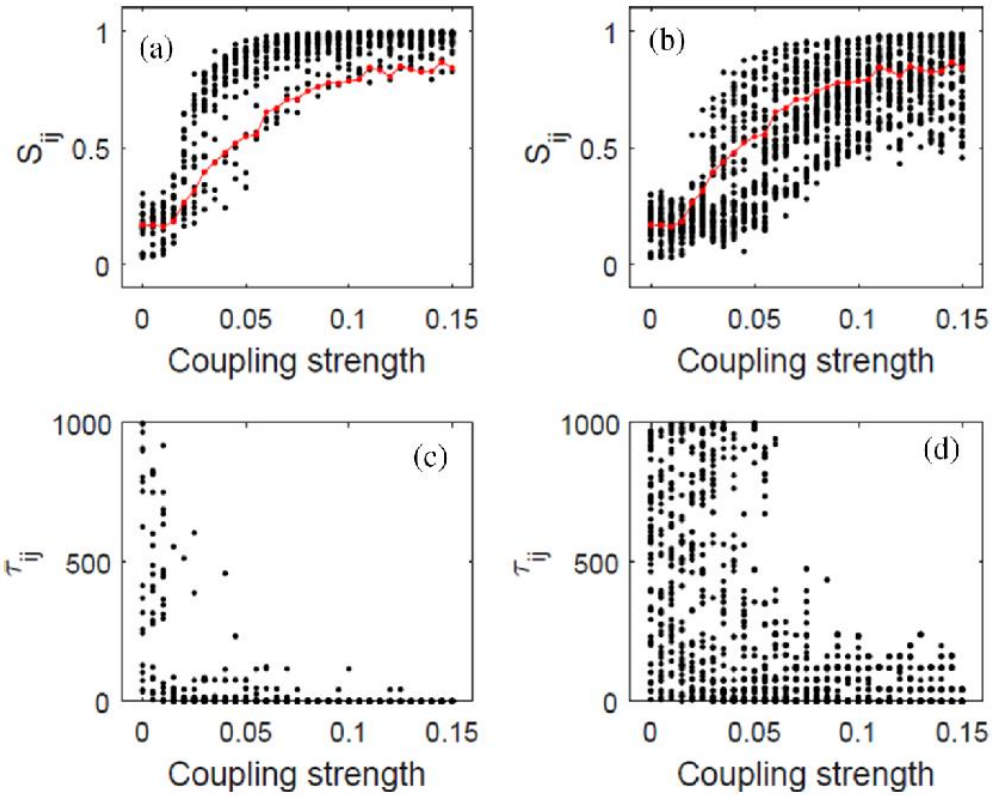




- No perfect reconstruction
- No important difference among the 3 methods & 3 variables

How the distributions of similarity values and τ_{ij} values change with the coupling strength?

12 electronic chaotic circuits



N. Rubido and C. Masoller, "*Impact of lag information on network inference*", Eur. Phys. J. Special Topics **227**, 1243-1250 (2018).

Using lag information to infer the links

If $\mathbf{S_{ij} > TH}$ the link i $\leftrightarrow$ j exists, otherwise, it does not exist

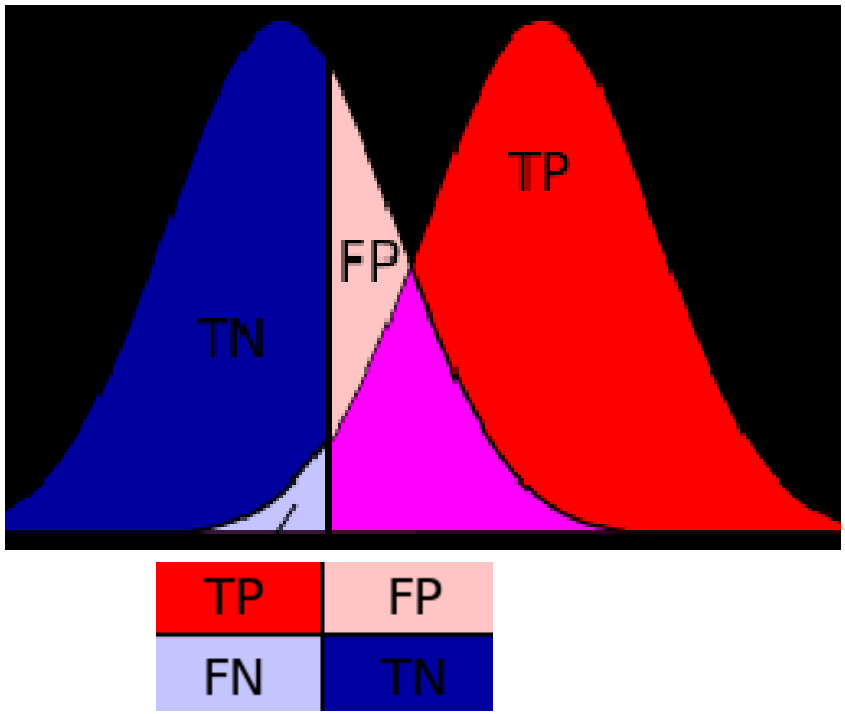
If $\tau_{ij} < \tau_{TH}$ the link i $\leftrightarrow$ j exists, otherwise, it does not exist

Three possible criteria:

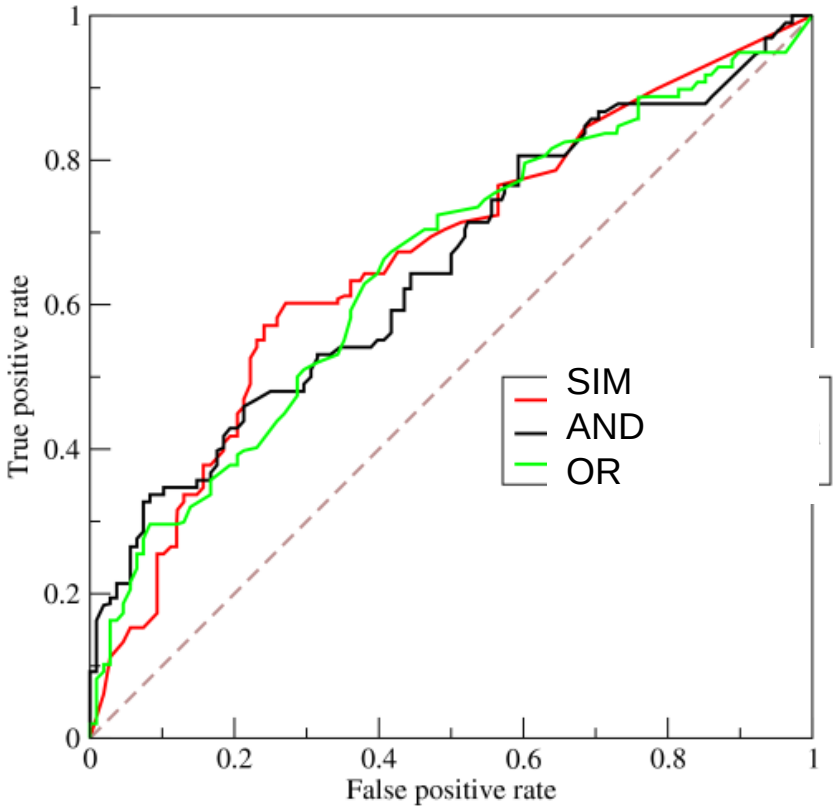
The link i $\leftrightarrow$ j exists if

- SIM : only SSM criteria holds ($\mathbf{S_{ij} > TH}$)
- AND: both criteria hold ($\mathbf{S_{ij} > TH}$ and $\tau_{ij} < \tau_{TH}$)
- OR: at least one criteria holds ($\mathbf{S_{ij} > TH}$ or $\tau_{ij} < \tau_{TH}$)

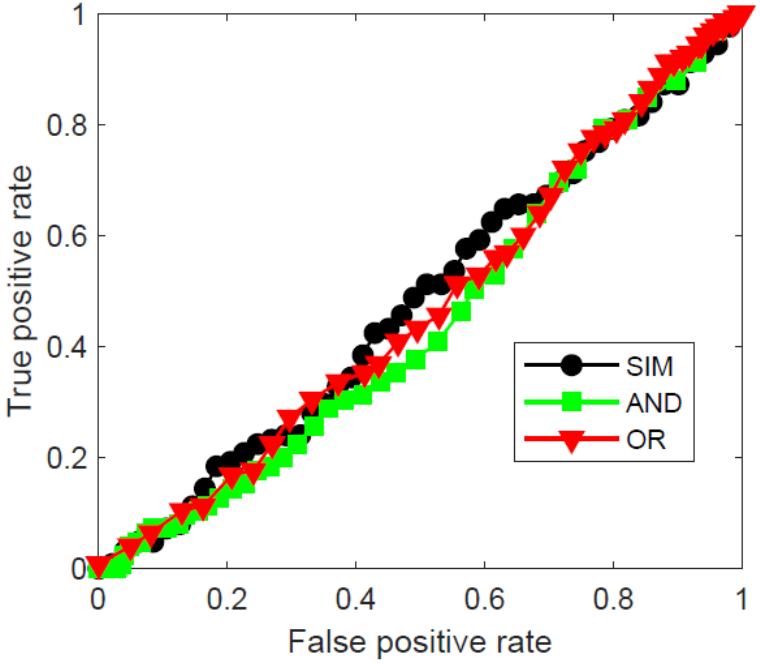
Quantifying the three criteria with receiver operating characteristic (ROC curve)



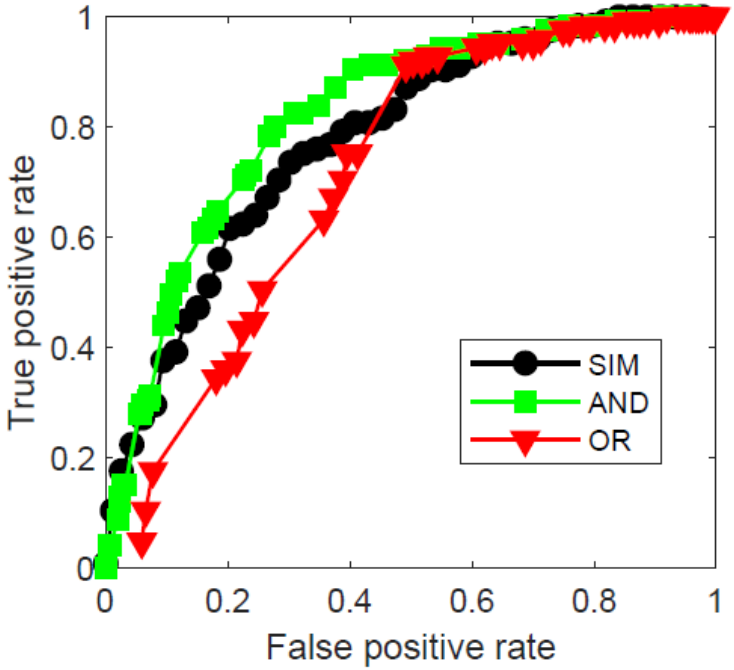
Source: wikipedia



Uncoupled oscillators



Coupled oscillators

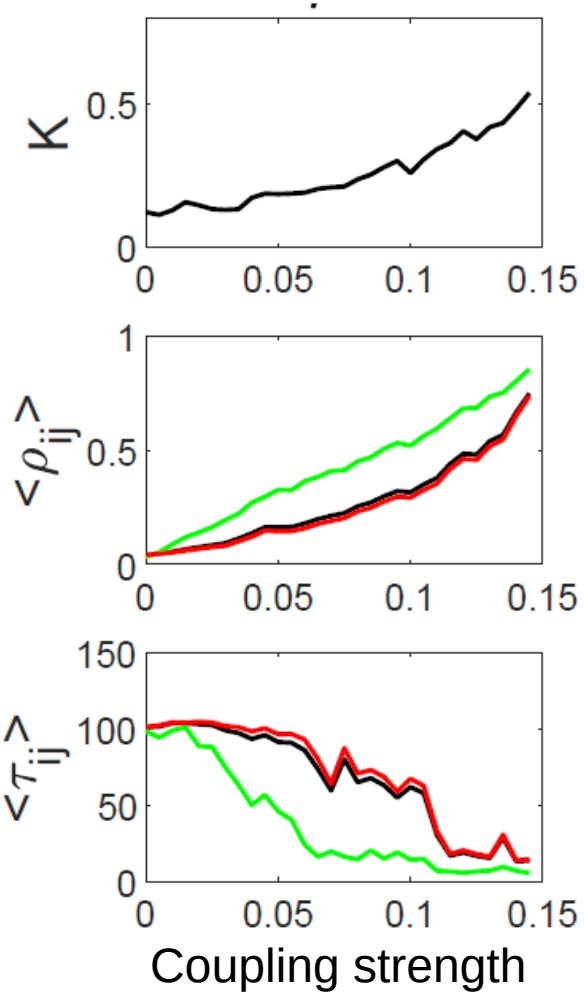


Results

50 Kuramoto phase oscillators, 10% existing links,
Similarity ρ_{ij} = max$_\tau$ cross-correlation of cos(φ_i), cos(φ_j)

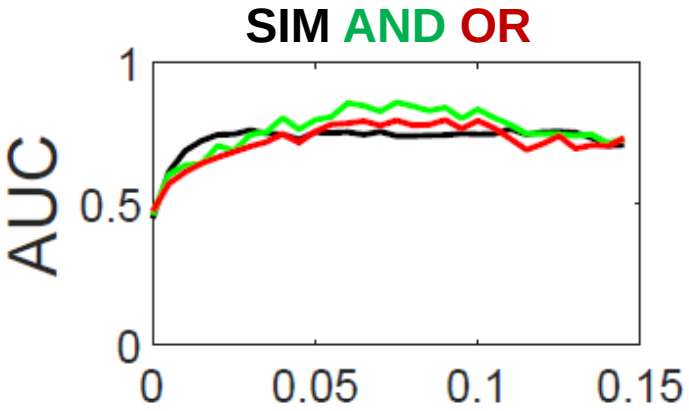
Order parameter

$$K = \left\langle \frac{1}{N} \left| \sum_{i=1}^{N} e^{i\phi_i(t)} \right| \right\rangle_T$$



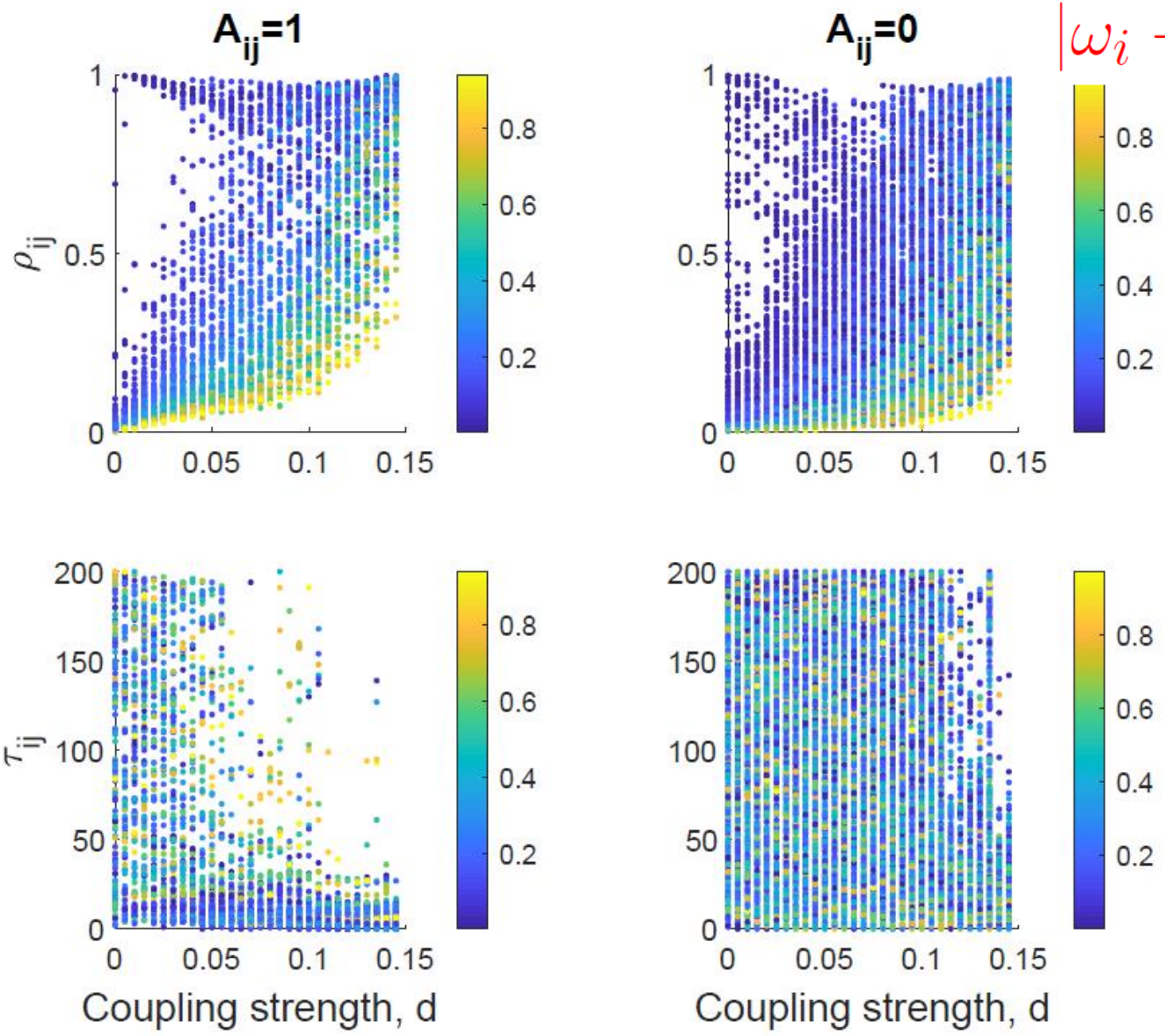
Coupling strength

All
Exist
No exist



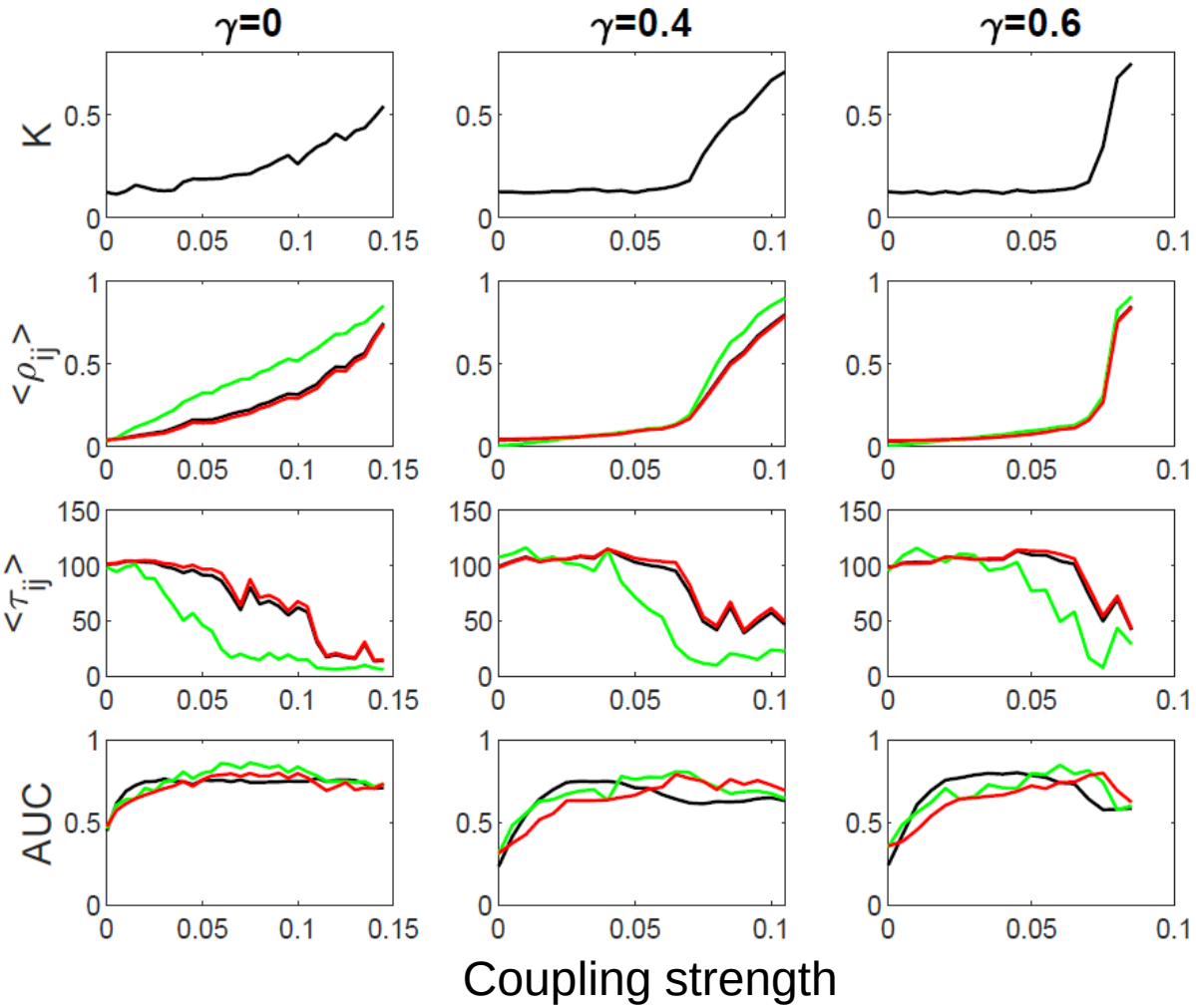
Coupling strength

Variation of similarity and τ_{ij} values with the coupling



For different parameters: explosive transition to sync

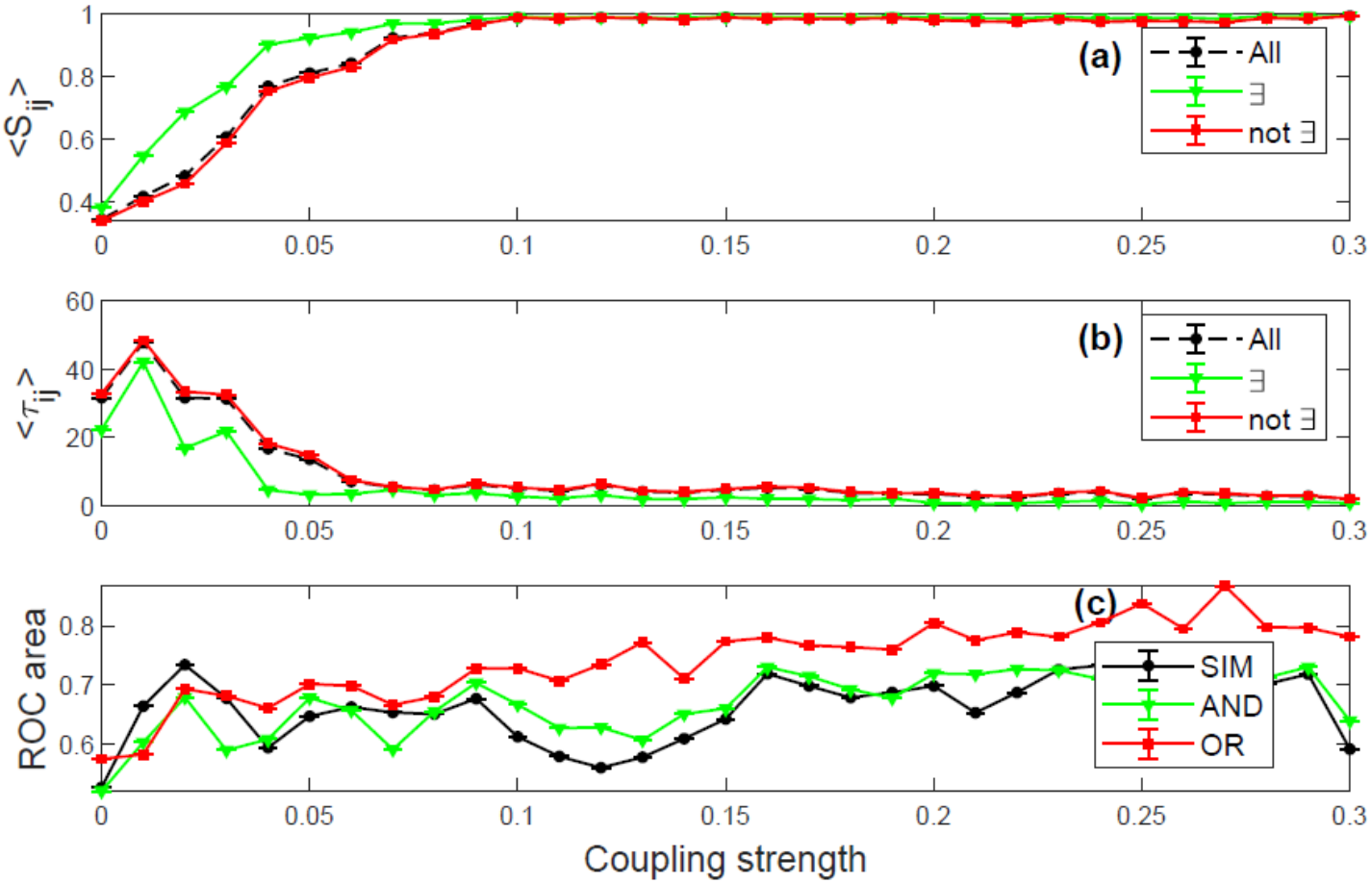
Oscillators can be linked if they have different frequencies: $|\omega_i - \omega_j| > \gamma$



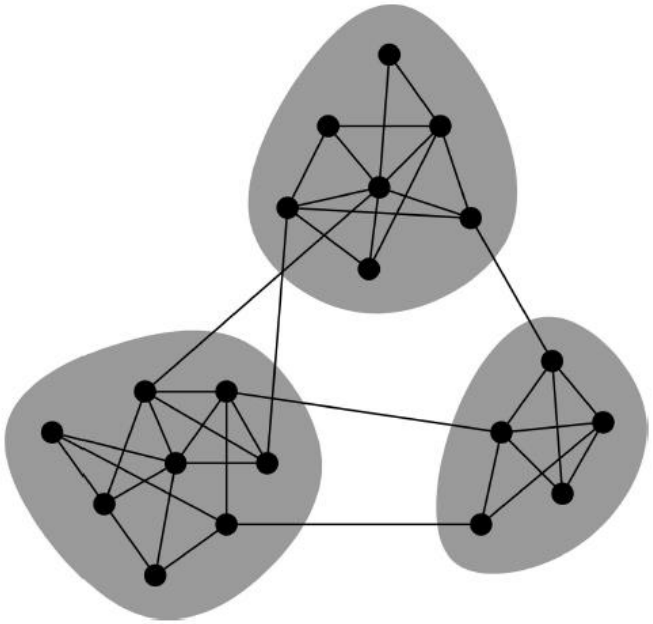
I. Leyva et al. Explosive transitions to synchronization in networked phase oscillators. Scientific Reports 3 (2013) 1281

Results obtained from experimental data

28 electronic chaotic circuits



Data from: R. Sevilla-Escoboza & J. M. Buldu, Synchronization of networks of chaotic oscillators: Structural and dynamical data sets. Data in Brief 7 (2016) 1185–1189



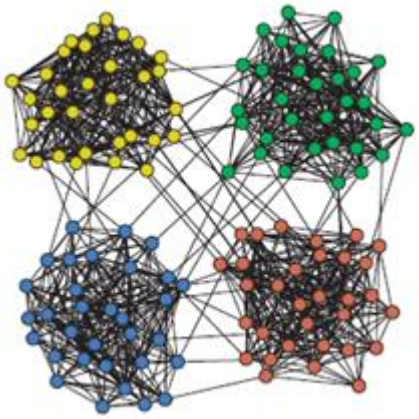
Community detection

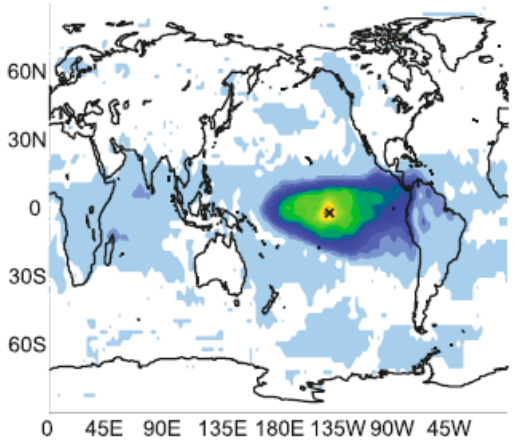
Climate “communities”

How to identify regions with similar climate?

- Goal: to construct a network in which regions with similar climate (e.g., continental) are in the same “community”.

- Problem: not possible with the “usual” correlation-based method to construct the network because NH and SH are only indirectly connected.





Network construction based on similar symbolic dynamics

- Step 1: transform SAT anomalies in each node in a sequence of symbols (we use ordinal patterns)

  s_i = {012, 102, 210, 012…} s_j = {201, 210, 210, 012, …}

- Step 2: in each node compute the transition probabilities

  $TP^{i}_{\alpha\beta} = \#(\alpha \rightarrow \beta)/N$

- Step 3: define the weights $w_{ij} = \dfrac{1}{\sum_{\alpha\beta}\left(TP^{i}_{\alpha\beta} - TP^{j}_{\alpha\beta}\right)^2}$

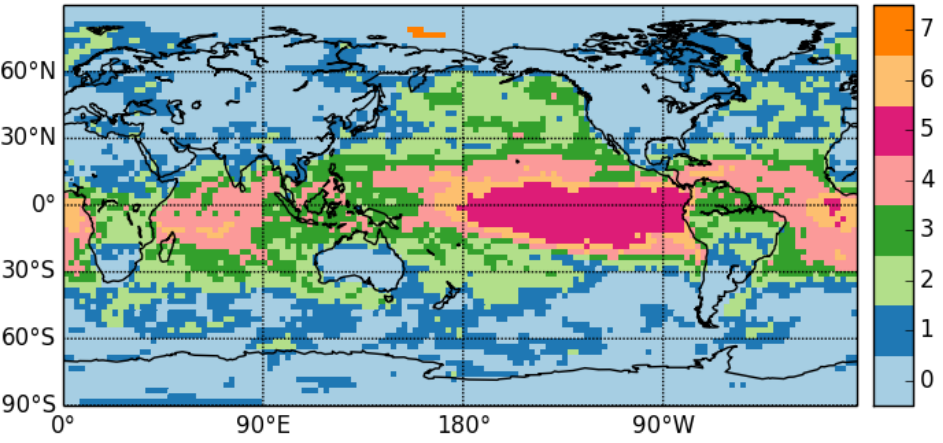
High weight if similar symbolic "language"

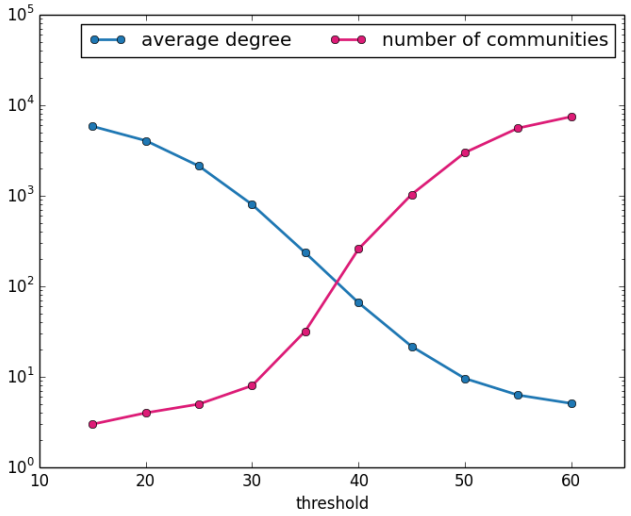
- Step 4: threshold w_{ij} to obtain the adjacency matrix.

- Step 5: run a *community detection algorithm* (*Infomap*).

Results

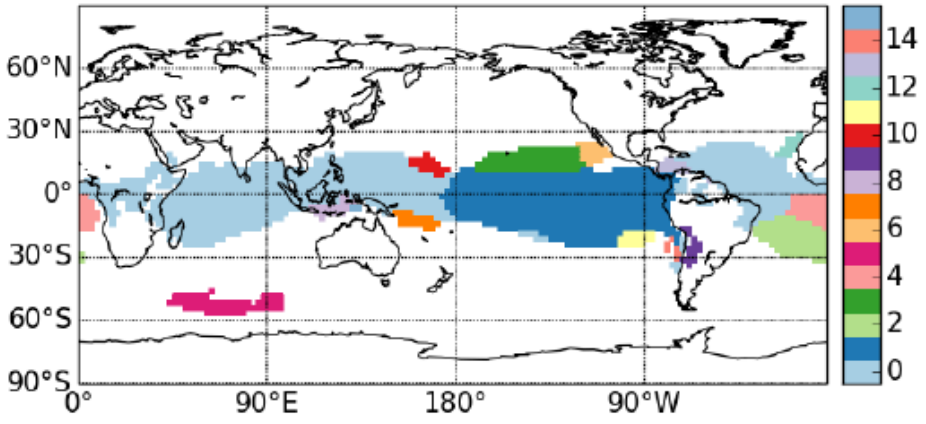
TP Network

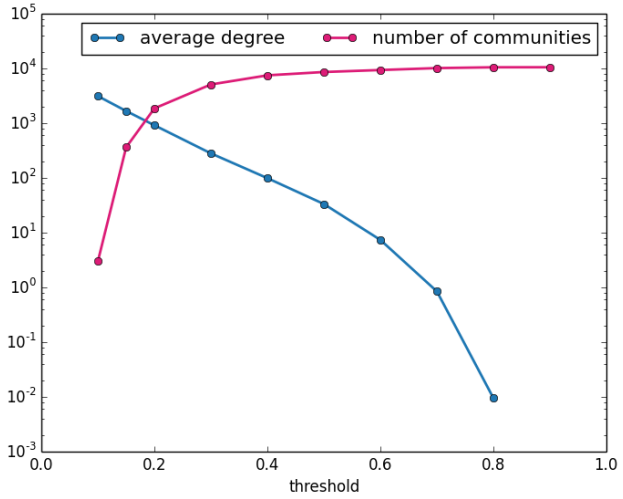




CC Network

(only the largest 16)





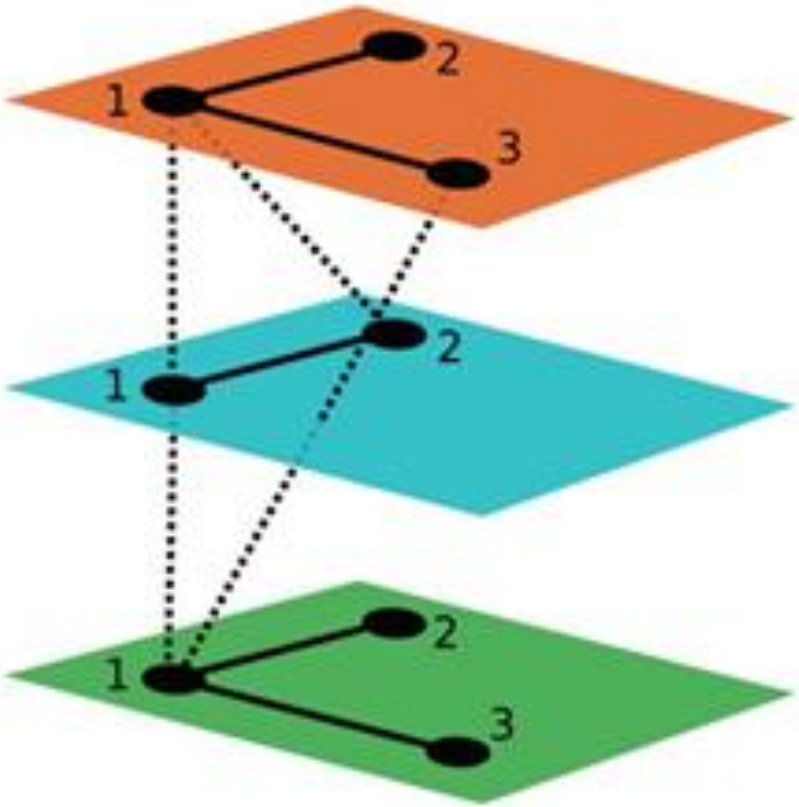
G. Tirabassi and C. Masoller, "*Unravelling the community structure of the climate system by using lags and symbolic time-series analysis*", Sci. Rep. **6**, 29804 (2016).

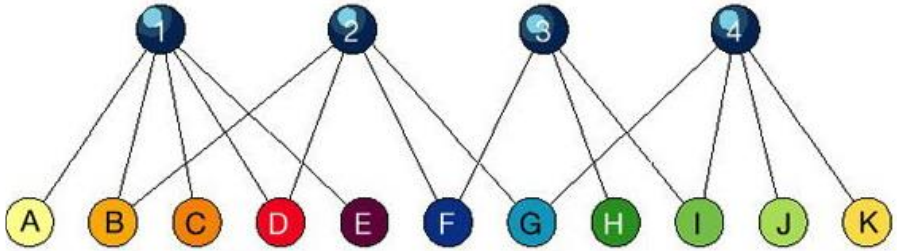
Community detection algorithms

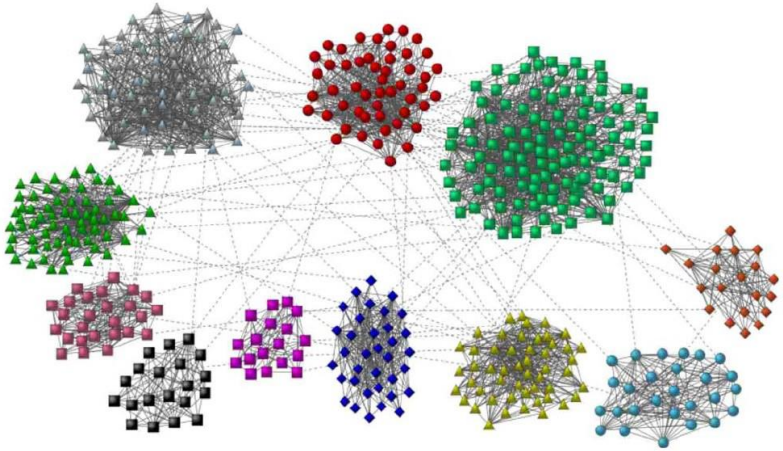
- *Infomap* (http://www.mapequation.org/code.html) and many others.
- *Infomap* clusters tightly interconnected nodes into modules and detects nested modules.
- Many other algorithms have been proposed.
- Further reading: S. Fortunato, "*Community detection in graphs*", Phys. Rep. 486, 75 (2010).

Generalizations of complex network analysis

Network structures: Multilayer, multiplex, bipartite, networks of networks and many others





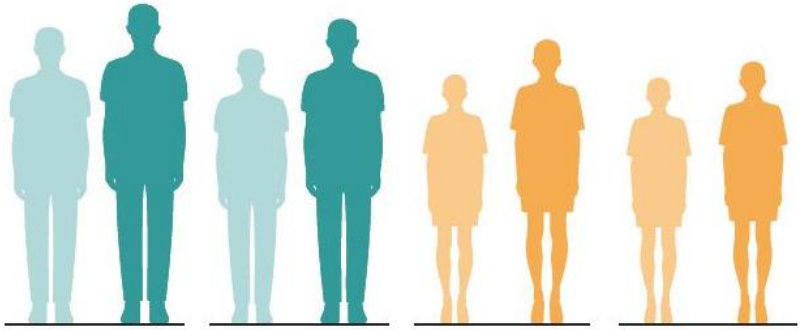


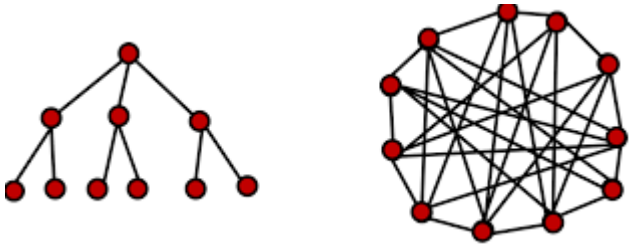
How to quantify the diversity of a complex system?

What is diversity?

Takes into account three characteristics of a population:

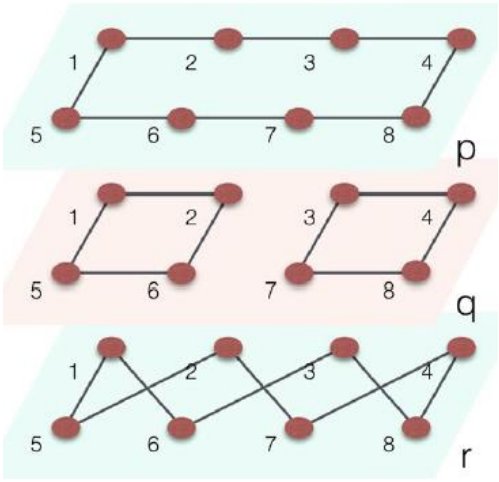
- Diversity in some **attributes** (e.g. atoms with different masses; people with different heights),
- Diversity of **types** (e.g., atoms or molecules; males or females),
- Diversity in **configuration** (e.g., configuration of atoms in a molecule; hierarchical or unstructured relations).





Diversity of complex systems represented by multiplex networks

M $(N \times N)$ adjacency matrices, $\mathcal{A} = \{A^{[1]}, A^{[2]}, \ldots, A^{[M]}\}$



- Two quantifications of diversity:
  - How diverse the connectivity paths that a node has in the different layers are;
  - How diverse the layers are.

- To quantify diversity we first need to define a “distance” to compare
  - the paths of a node in the different layers,
  - the different layers.

1) To quantify the differences in the connectivity paths of node i in layers p and q:

$$\mathcal{D}_i(\overline{p},\overline{q}) = \frac{\sqrt{\mathcal{J}(\mathcal{N}_i^{\overline{p}},\mathcal{N}_i^{\overline{q}})} + \sqrt{\mathcal{J}(T_i^{\overline{p}},T_i^{\overline{q}})}}{2\sqrt{\log(2)}}$$

Node Distance Distribution of **i** in layer **p** (global information)

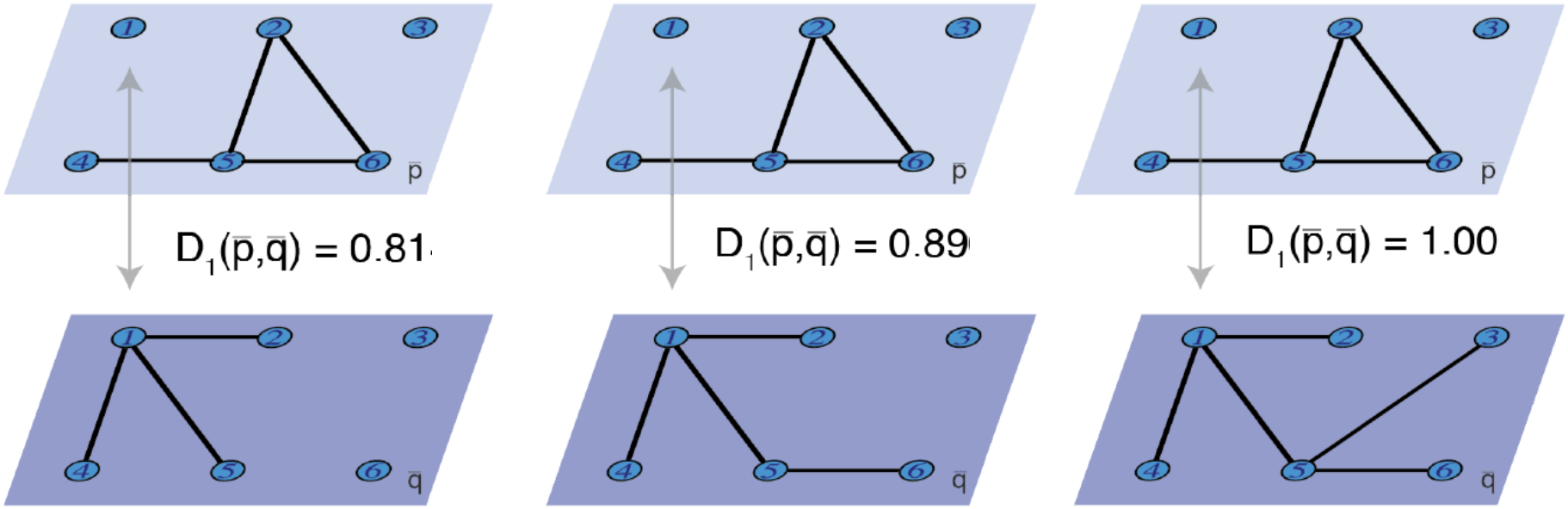
Transition Matrix of **i** in layer **p**, (adjacency matrix rescaled by the degree of **i**; local information)

- D_i=0: **i** has identical connectivity paths in layers **p** and **q**,
- D_i=1: **i** is not connected in one layer, while there are paths connecting **i** to all nodes in the other layer.

2) To quantify how different layers p and q are:

$$\mathcal{D}(\overline{p},\overline{q}) = \langle \mathcal{D}_i(\overline{p},\overline{q}) \rangle_i$$

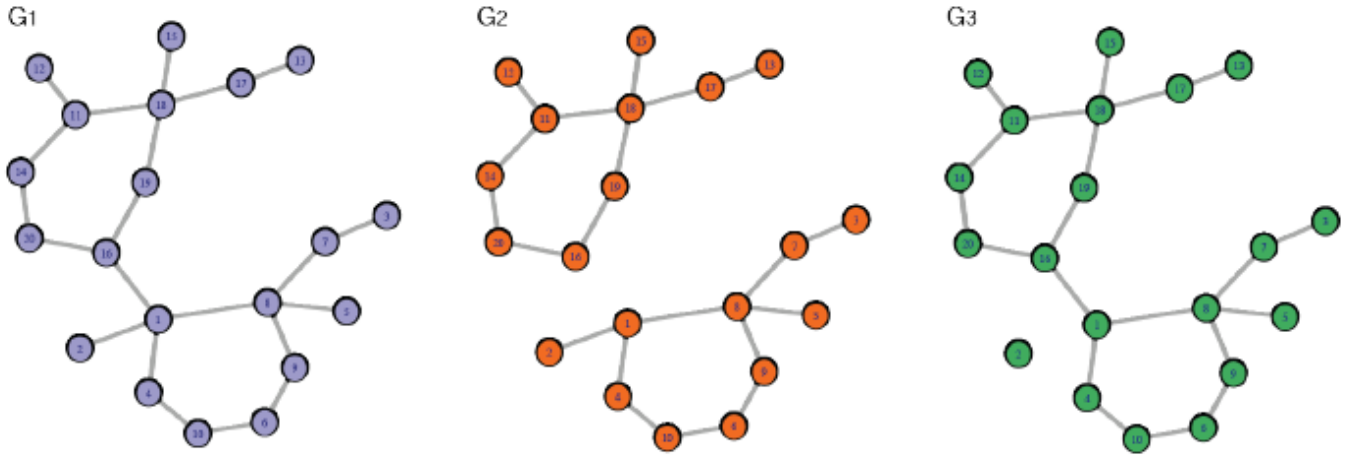
Examples



$$\mathcal{D}(\overline{G1},\overline{G3}) = 0.14$$

$$\mathcal{D}(\overline{G2},\overline{G3}) = 0.34$$

$$\mathcal{D}(\overline{G1},\overline{G2}) = 0.36$$



Diversity measure

- Main idea: the diversity of a system is defined by the distances between its elements: the larger the distances, the more different the elements are, and the more diverse the system is.

- The distance between the element **g** $\notin$ S and the set S, **D(g; S)**, is the smallest distance between **g** and any of the elements of S, $\mathcal{D}(\overline{g}, S) = \min_{\overline{s_i} \in S} \mathcal{D}(\overline{g}, \overline{s_i})$

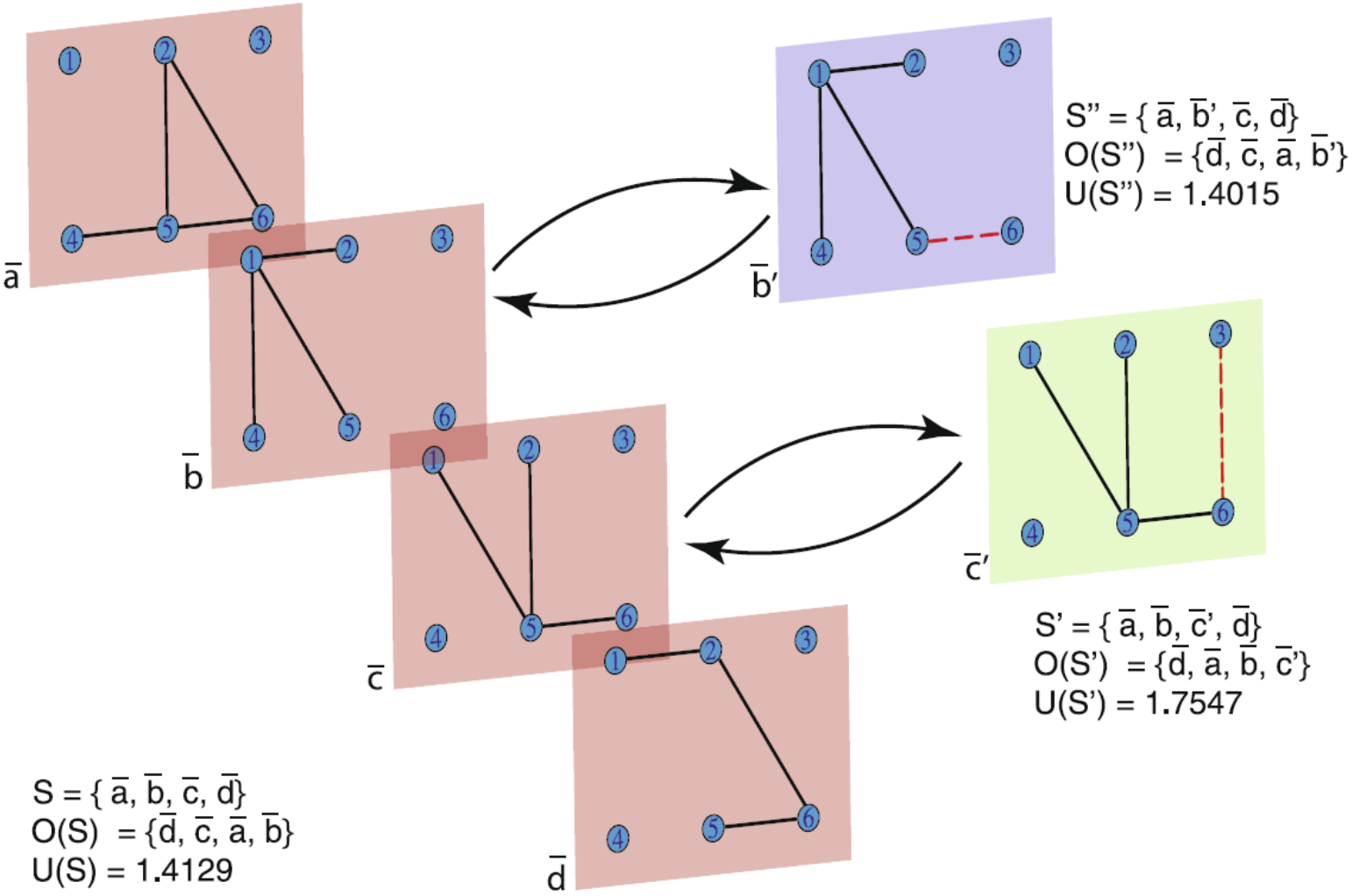
- Recursive definition: $U(S) = max_{\overline{s}_i \in S}\{U(S \backslash \overline{s}_i) + \mathcal{D}(\overline{s}_i, S \backslash \overline{s}_i)\}$

  $U(S)$=0 if |S|=1

- Diversity increases when a new element is included

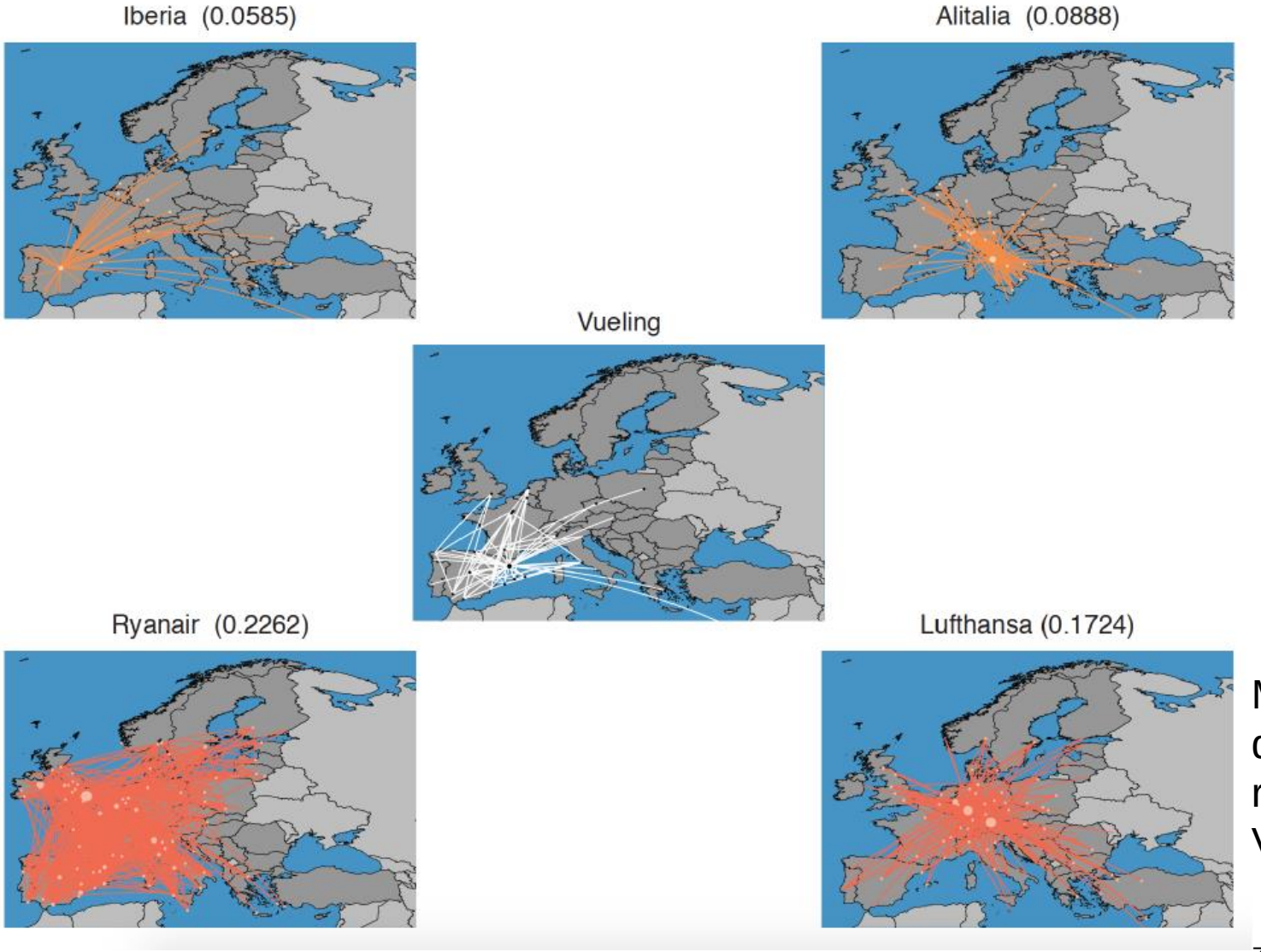
$$U(S \cup \overline{g}) \geq U(S) + D(\overline{g}, S)$$

Example: diversity decreases (increases) when we include a link present (not present) in other layers



L. Carpi et al, Sci. Reports 9, 4511 (2019)

Analysis of Europe air traffic network: distance between Vueling and other airlines

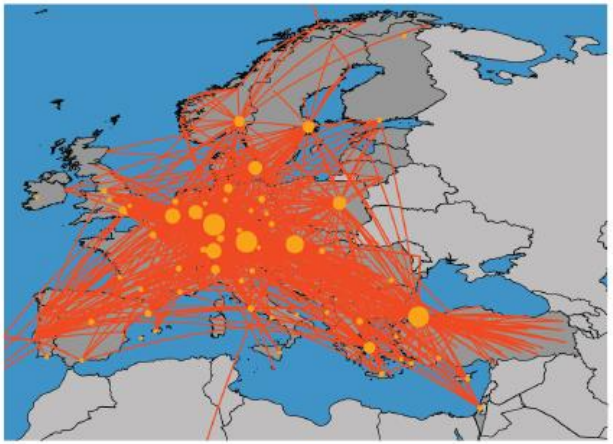


More similar routes to Vueling

More different routes to Vueling

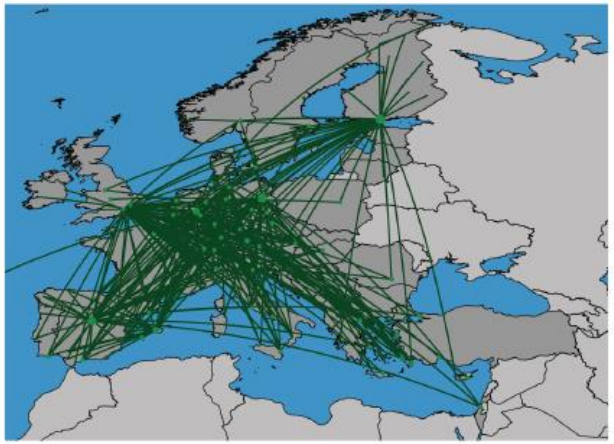
L. Carpi et al, Sci. Reports 9, 4511 (2019)

Diversity ordering: Elements can be ordered in terms of their contribution to the diversity of the set.



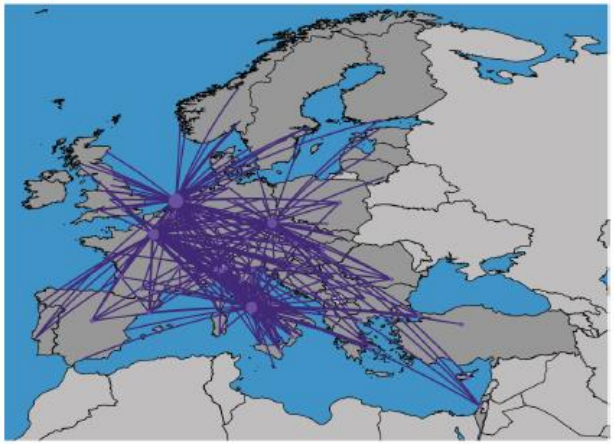
a Star Alliance

U=0.97



b Oneworld

U=0.34



c Skyteam

U=0.33

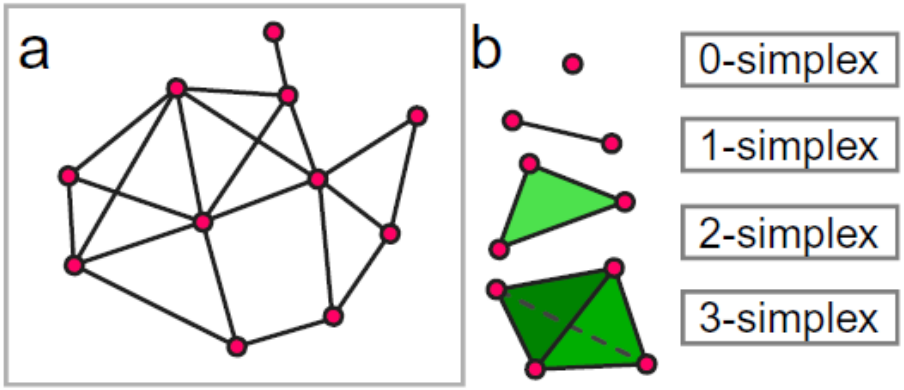
Diversity ordering of Star Alliance (from less to more contribution):

Brussels Airlines (BEL), Swiss Air (SWR), Polish Airlines (LOT), Air Portugal (TAP), Aegean Airlines (AEE), Austrian Airlines (AUA), Scandinavian Airlines (SAS), Turkish Airlines (THY), Lufthansa (DLH)

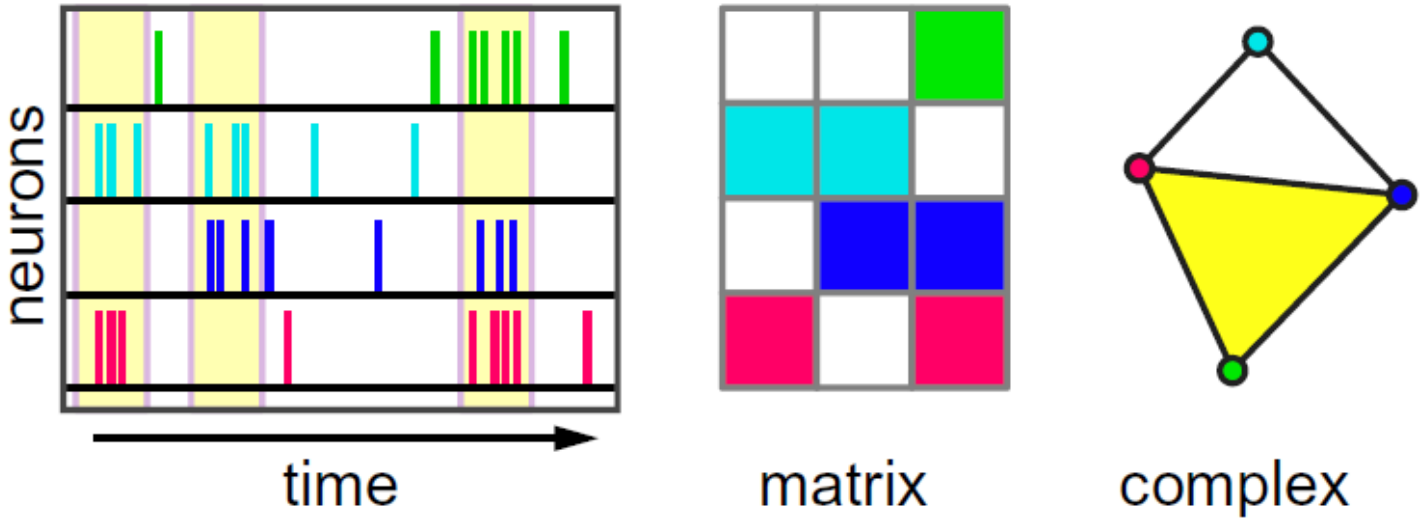
Limitations of complex network analysis

Interactions are not limited to pairs of elements

- Links represent interactions between pairs of nodes.
- ***Simplicial complexes*** represent interactions among several nodes.



Example



Giusti et al., J Comput Neurosci (2016) 41:1–14

Concluding

Take home messages

- Multivariate analysis uncovers inter-relationships in datasets
- Different similarity measures are available for inferring the connectivity of a complex system from observations.
- Different measures can uncover different properties.
- Thresholding, hidden variables, hidden “nodes” can difficult or make impossible the inference of the network structure.
- Different sets of “communities” can be uncovered depending on the property that is analyzed.
- Many many applications and challenges!

Acknowledgments












- Maria Masoliver, Pepe Aparicio Reinoso (*neurons*)
- Carlos Quintero, Jordi Tiana, Came Torrent (*laser lab*)
- Andres Aragoneses, Laura Carpi (*data analysis, networks*)
- Ignacio Deza, Giulio Tirabassi, Dario Zappala, Marcelo Barreiro (*climate*)
- Pablo Amil (*biomedical images*)

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
ICREA
BE-OPTICAL
Advanced Biomedical Optical Imaging and Data Analysis

References

- M. Barreiro, et. al, Chaos 21, 013101 (2011)
- Deza, Barreiro and Masoller, Eur. Phys. J. ST 222, 511 (2013)
- Tirabassi and Masoller, EPL 102, 59003 (2013)
- G. Tirabassi et al., Ecological Complexity 19, 148 (2014)
- G. Tirabassi et al., Sci. Rep. **5** 10829 (2015)
- G. Tirabassi and C. Masoller, Sci. Rep. 6:29804 (2016)
- T. A. Schieber et al, Nat. Comm. 8, 13928 (2017)
- L. Carpi et al, Sci. Reports 9, 4511 (2019)
- P. Amil et al, PLoS ONE in press (2019)

<cristina.masoller@upc.edu>

http://www.fisica.edu.uy/~cris/