Nonlinear time series analysis

Multivariate analysis

Cristina Masoller
Universitat Politecnica de Catalunya, Terrassa, Barcelona, Spain

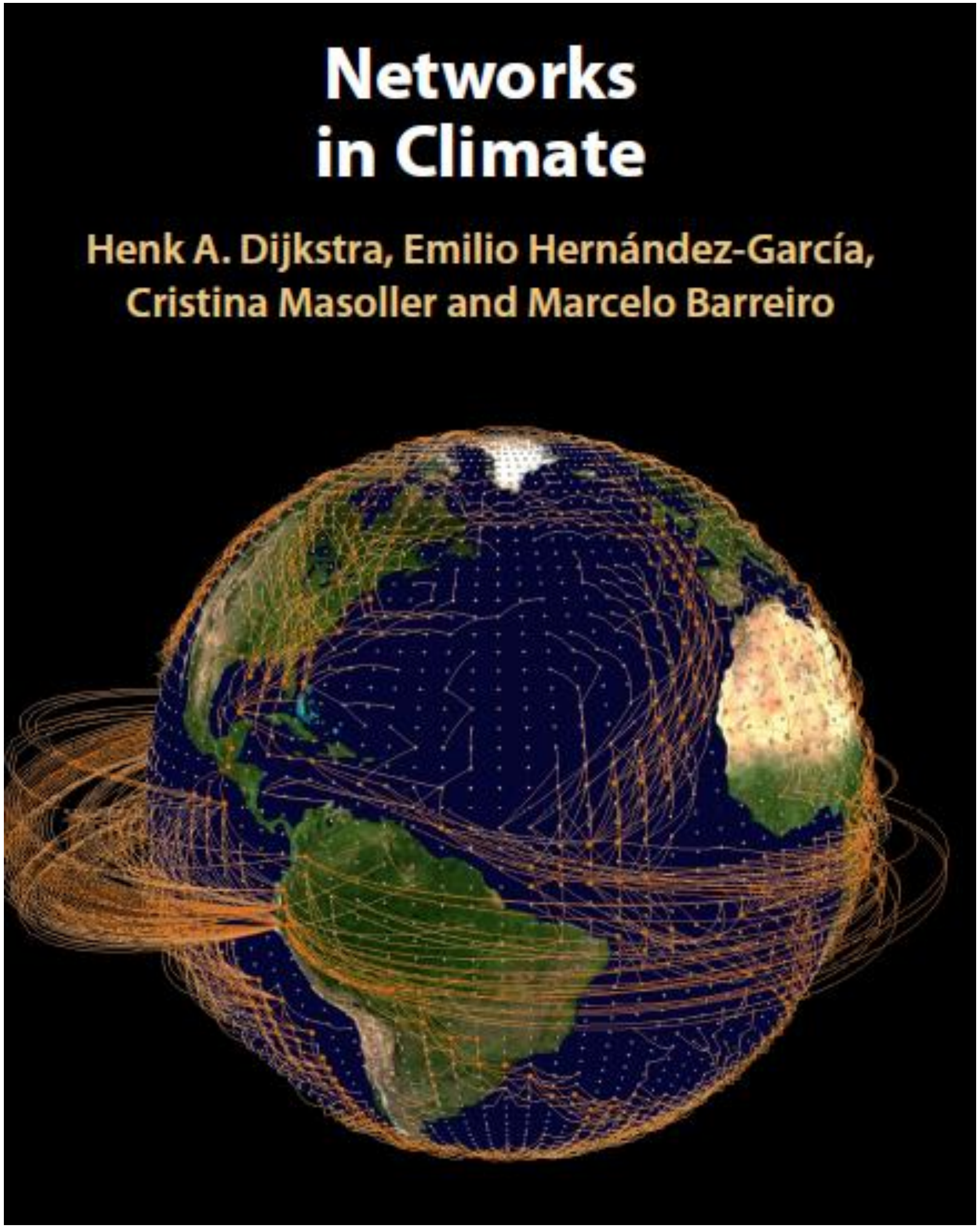
Cristina.masoller@upc.edu
www.fisica.edu.uy/~cris

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 to climate data.
- **Bivariate analysis**
  - Extracting information from two time series.
  - Correlation, directionality and causality.
  - Applications to climate data.
- **Multivariate analysis**
  - Many time series: complex networks.
  - Network characterization and analysis.
  - Climate networks.

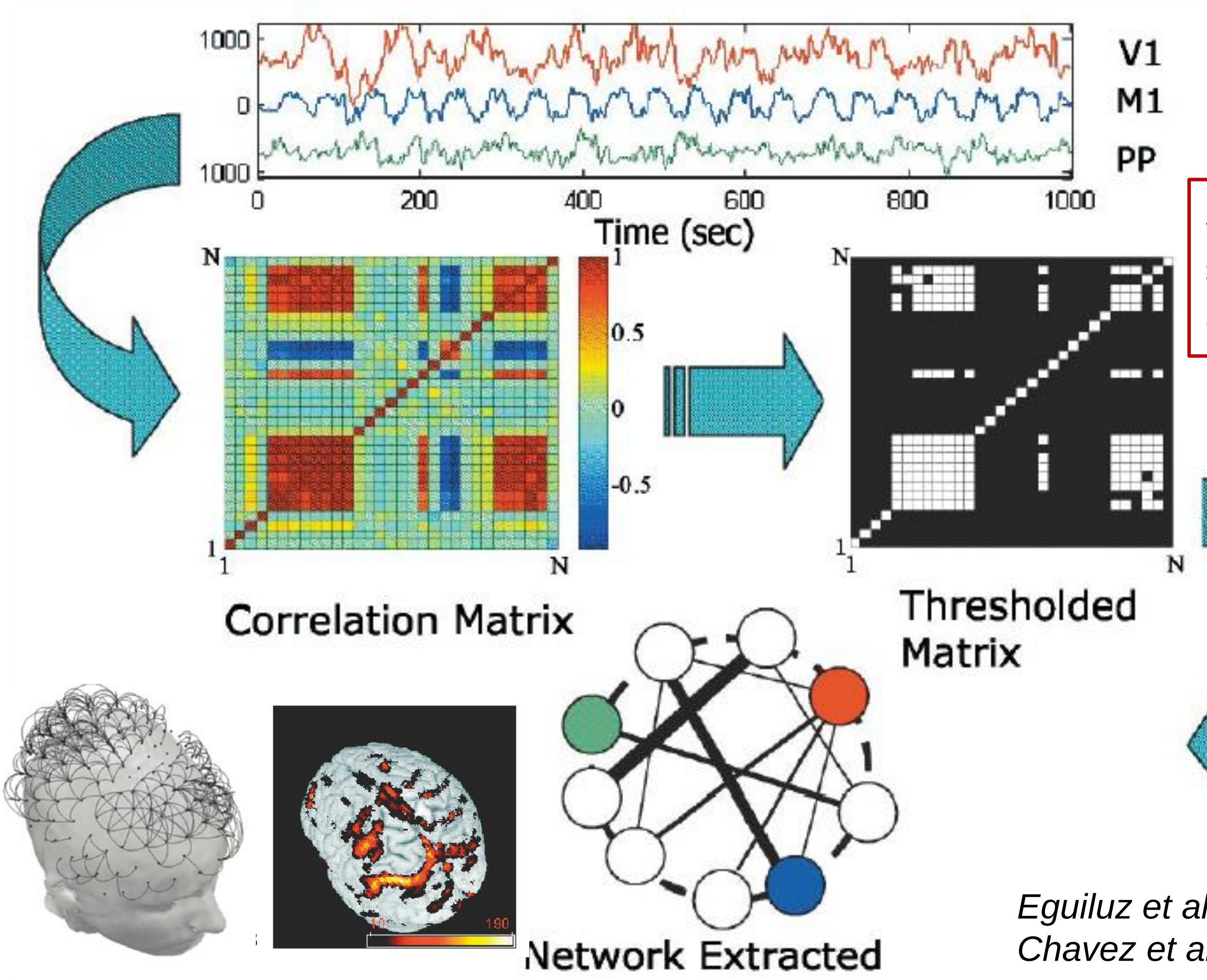


Bibliography

Cambridge University Press
February 2019

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

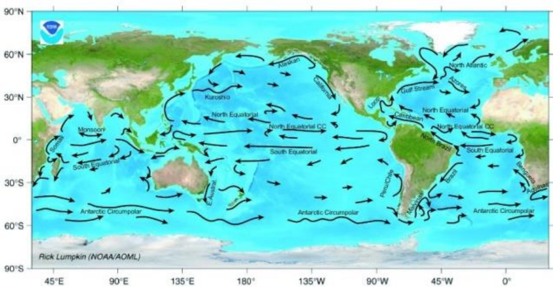
Brain functional network

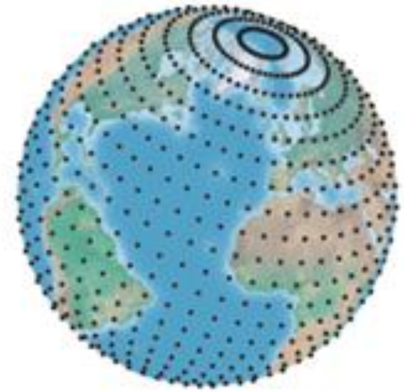


S_{ij} > Th
⇒ A_{ij} = 1,
else A_{ij}=0

Eguiluz et al, PRL 2005
Chavez et al, PRE 2008

Complex network representation of the climate system

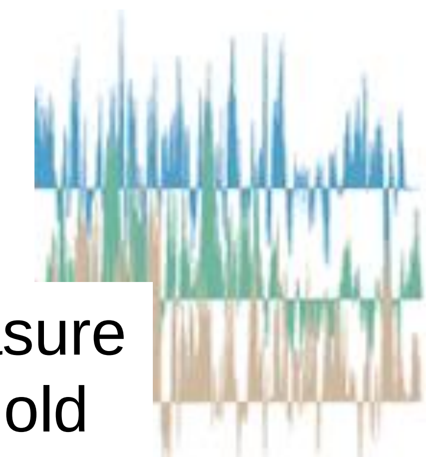




More than 10000 nodes.

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

Daily resolution: more than 13000 data points in each TS

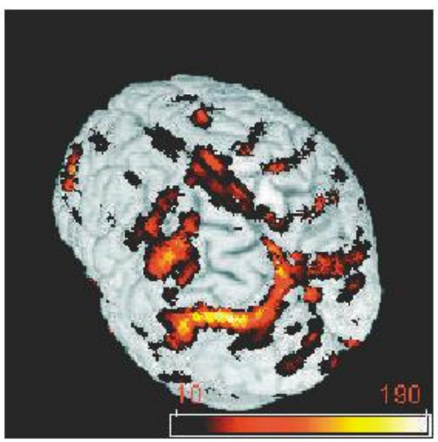


Sim. measure + threshold

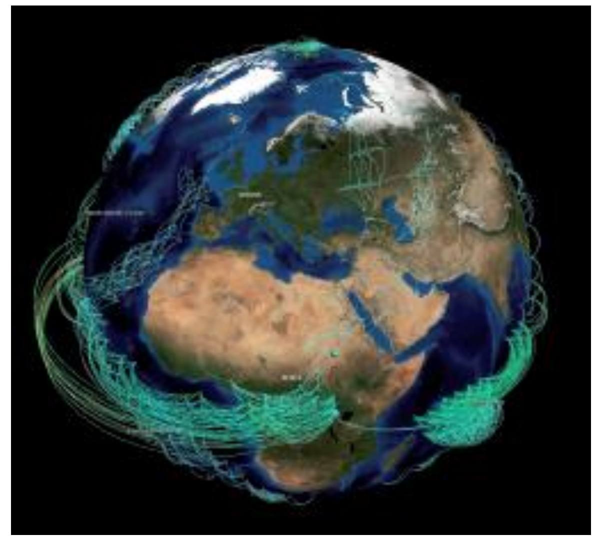
Surface Air Temperature Anomalies (solar cycle removed)

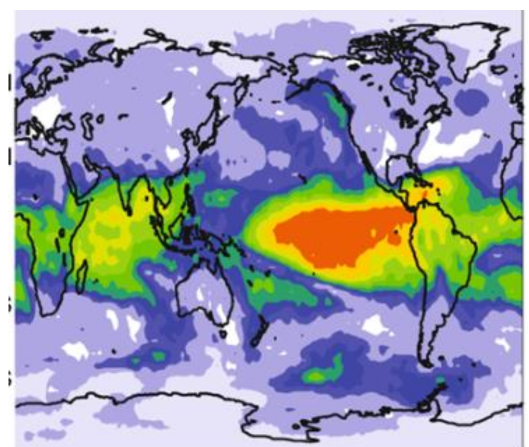
Donges et al, Chaos 2015

Brain network



Climate network



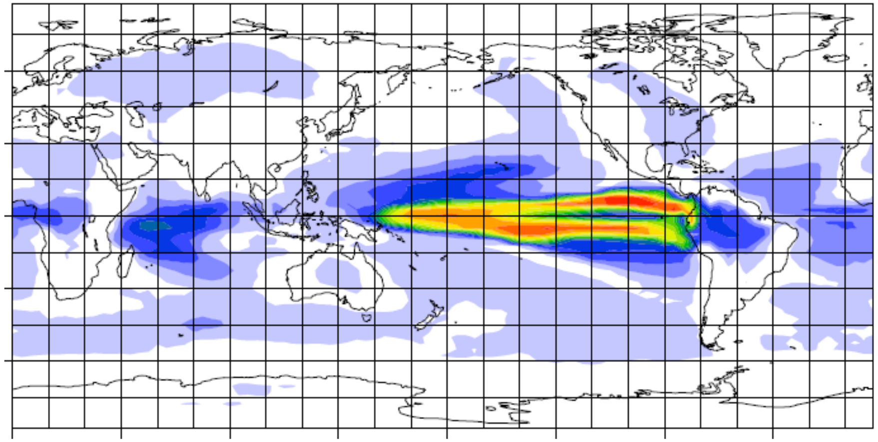


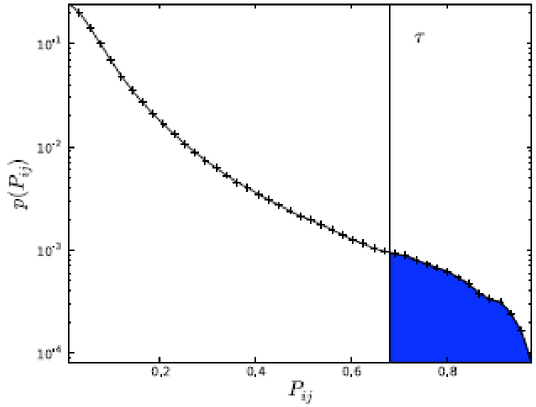
Weighted degree

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

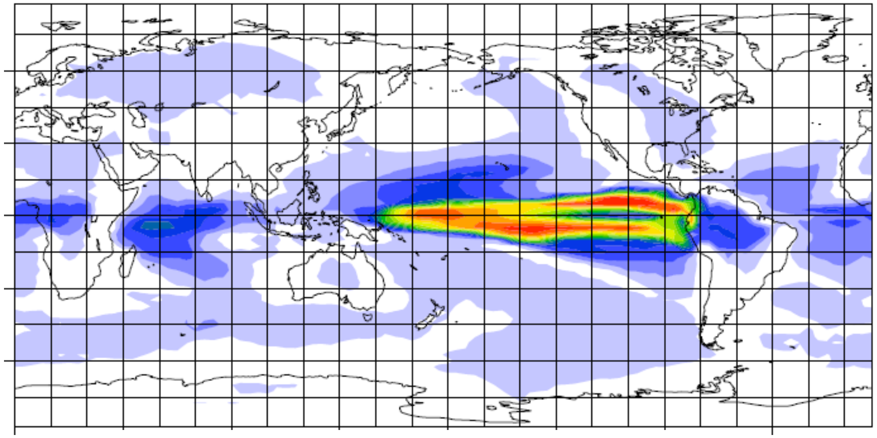
Statistical similarity measure

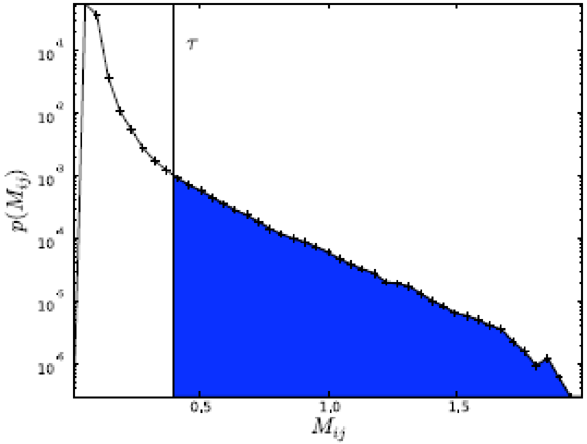
AWC computed with cross-correlation





AWC computed with mutual information



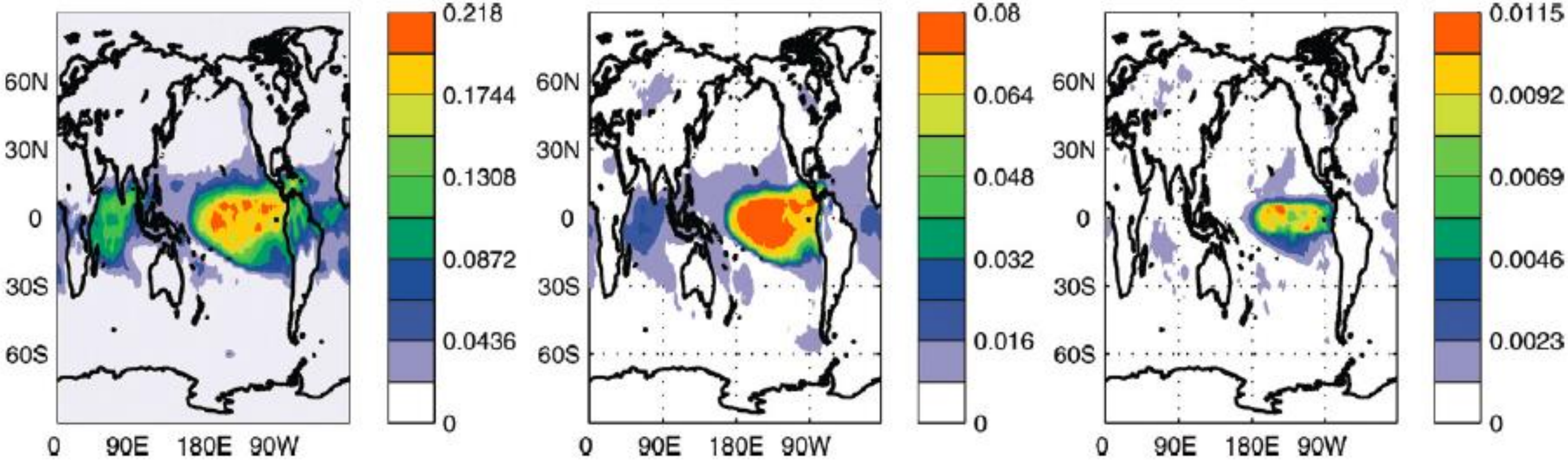


The threshold was selected to give a network with the same link density (0.005)

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)

How to select the threshold?

Three criteria are typically used:

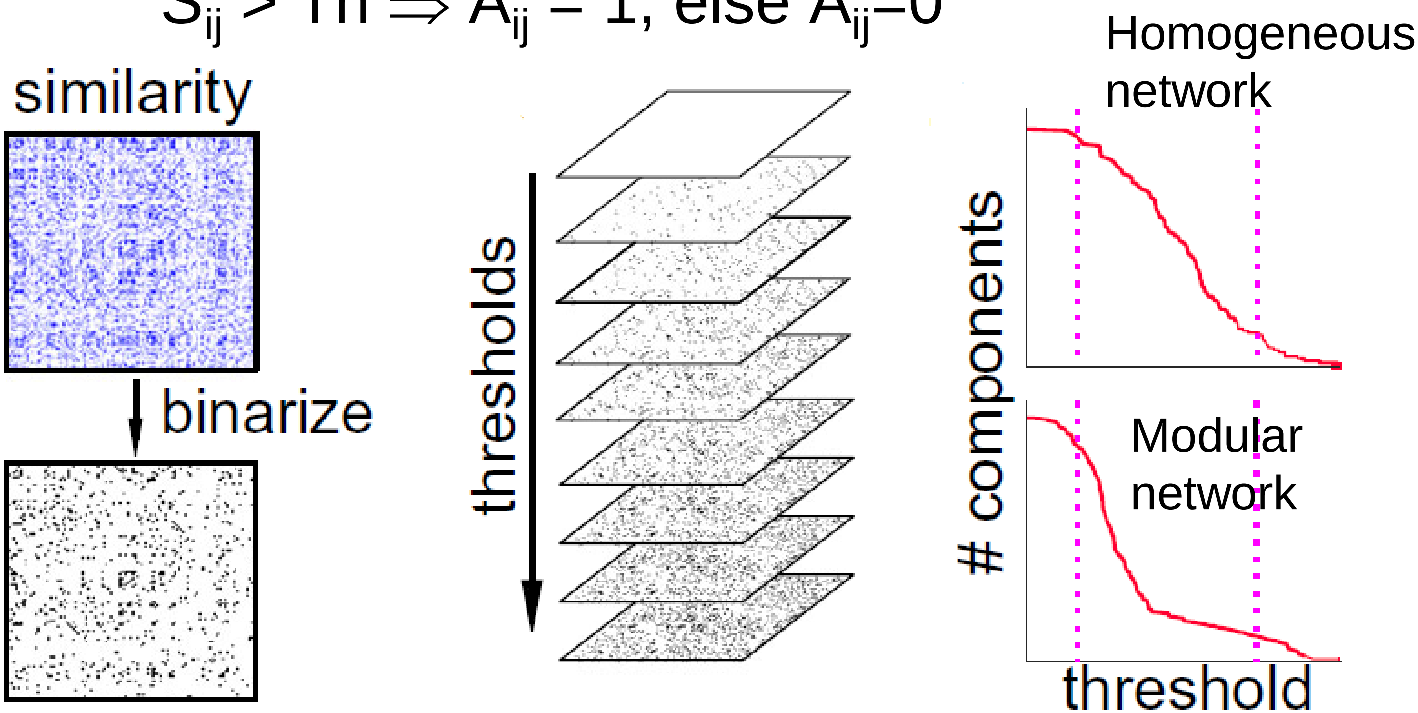
- A significance level is used (typically 5%) in order to omit connectivity values that can be expected by chance;
- We select an arbitrary value as threshold, such that it gives a certain pre-fixed number of links (or link density);
- We define the threshold as large as possible while guaranteeing that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

Problems with thresholding

- Statistical similarity measure (CC, MI, etc.)

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

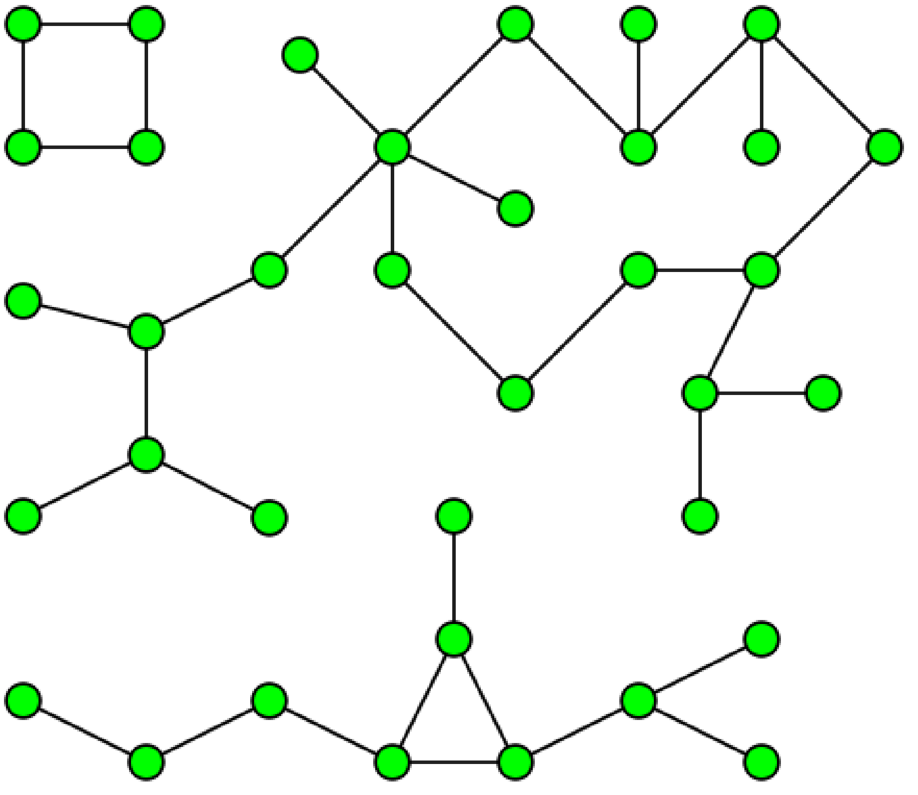


The number of *connected components* as a function of threshold reveals different structures.

- But thresholding near the dotted lines would suggest inaccurately that these two networks have similar structures.
- “Features” that persist for a wide range of thresholds are "true" features.

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

Connected components



A graph with three connected components.
Source: Wikipedia

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

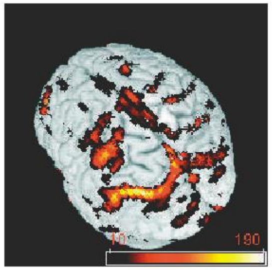
Citation: Chaos **25**, 113101 (2015); doi: 10.1063/1.4934554

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

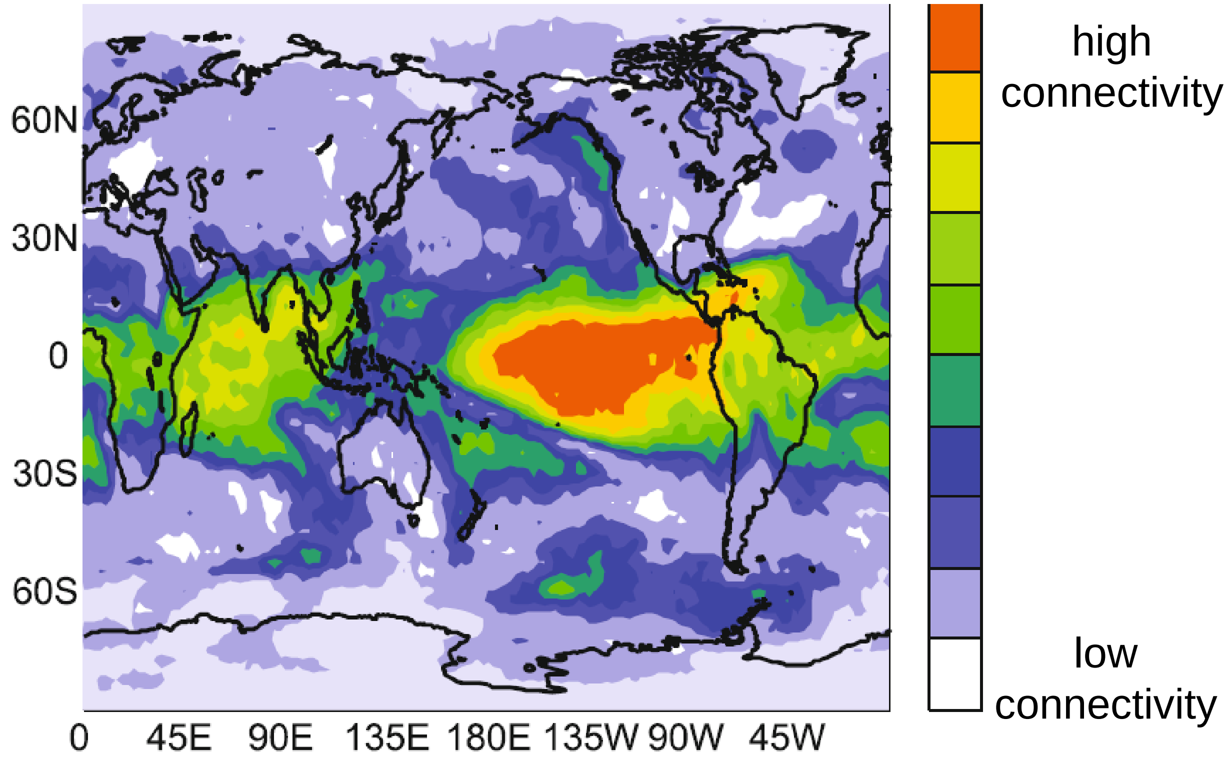
pyunicorn is available at https://github.com/pik-copan/

Graphical representation of the climate network

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



Network obtained with ordinal analysis using *__inter-annual__* *time-scale (3 consecutive years). The color-code indicates the Area Weighted Connectivity (weighted degree)*

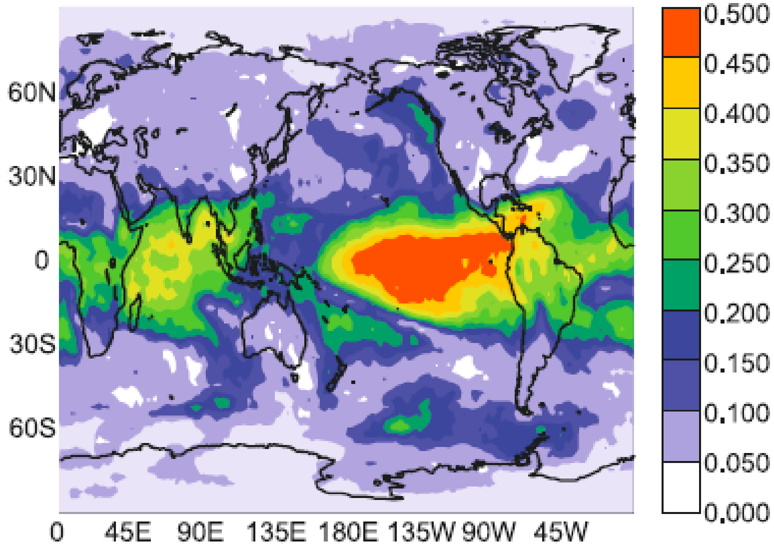


J. I. Deza, M. Barreiro, and C. Masoller, Eur. Phys. J. Special Topics 222, 511 (2013)

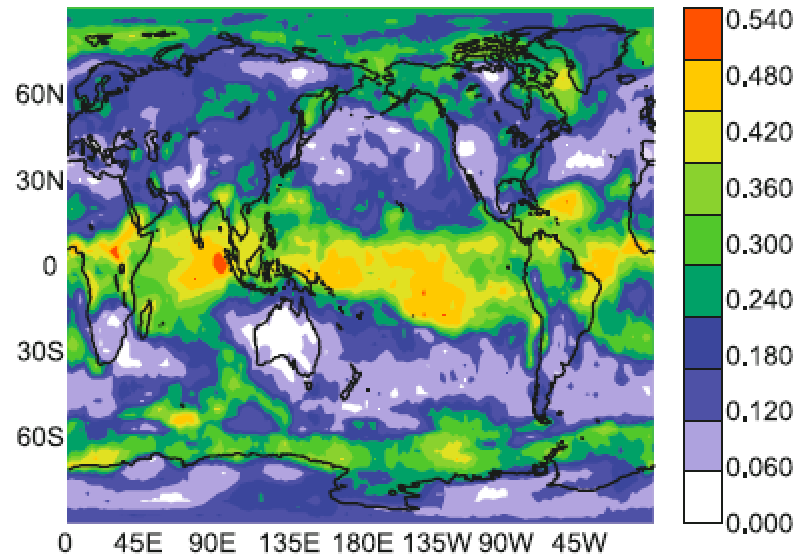
Comparison: histogram vs. ordinal mutual information

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

Network when the probabilities are computed with ordinal analysis

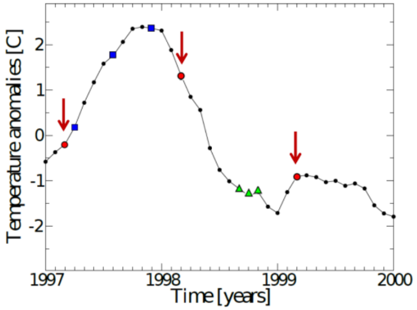


Network when the probabilities are computed with histogram of values



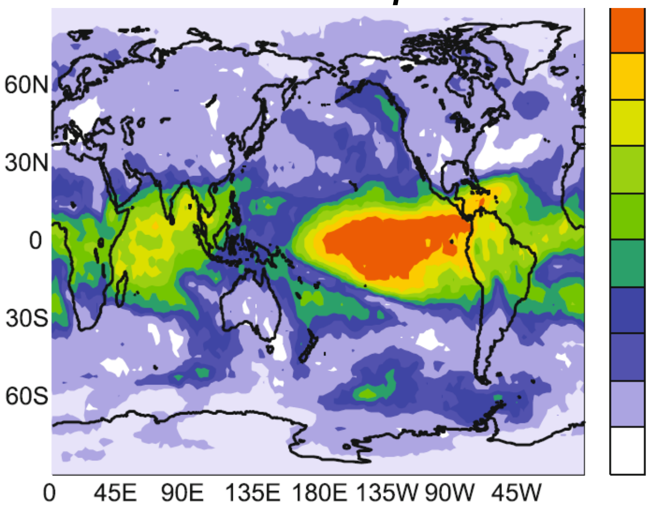
Color code indicates the area-weighted connectivity

inter-annual time scale

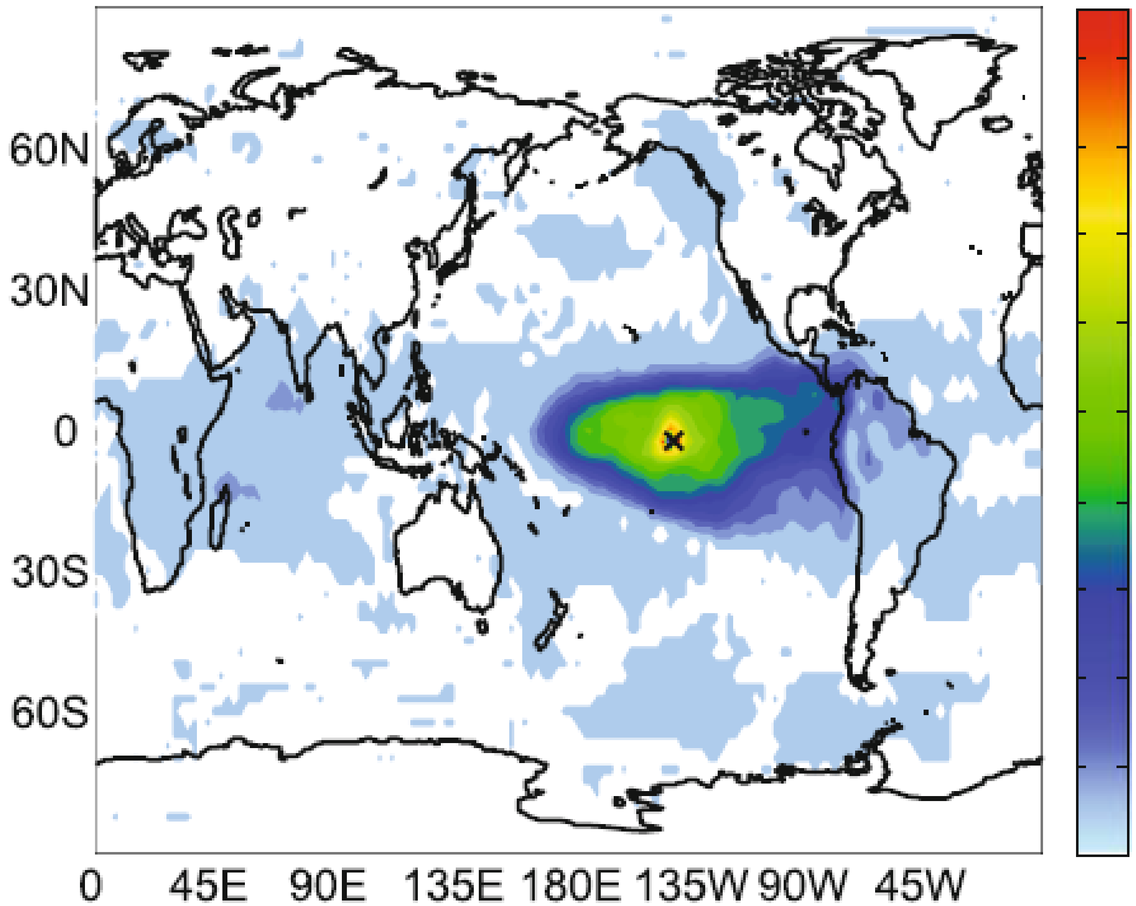


Who is connected to who?

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)

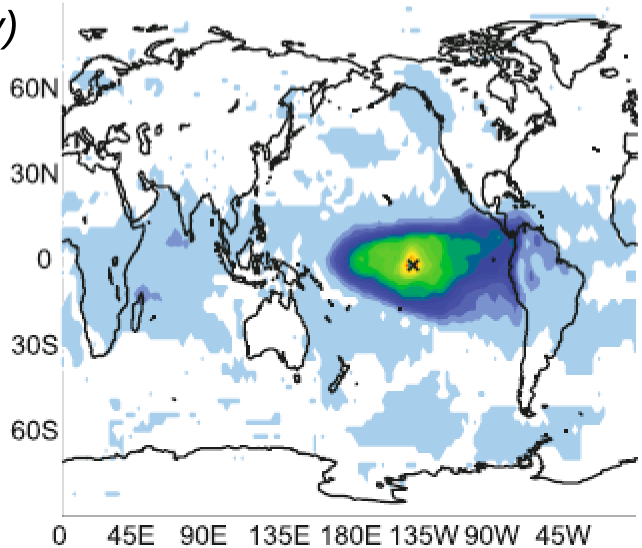
Influence of the threshold

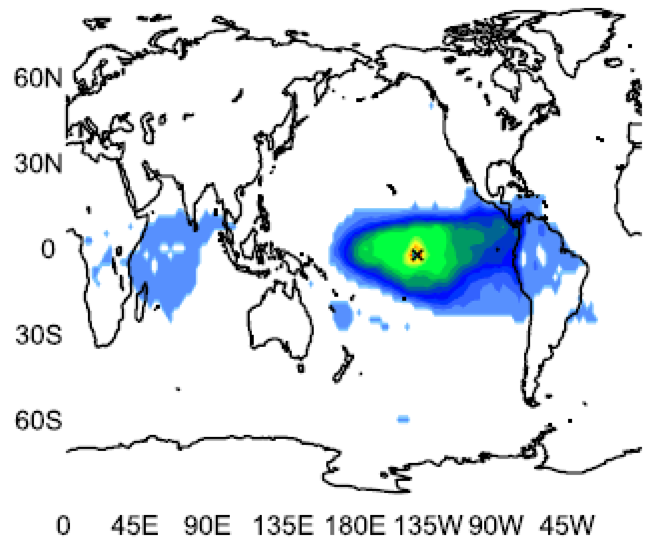
All significant links

(11% link density)

Higher threshold *(3% link density)*

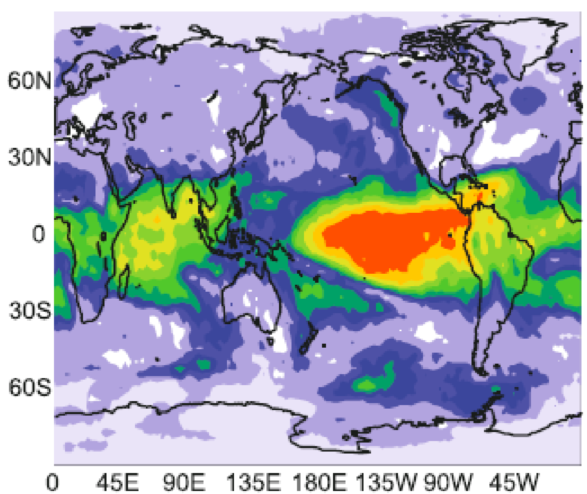
Color code: MI

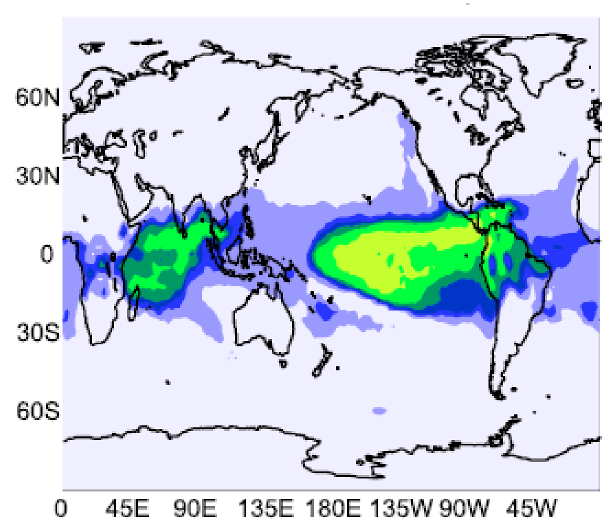




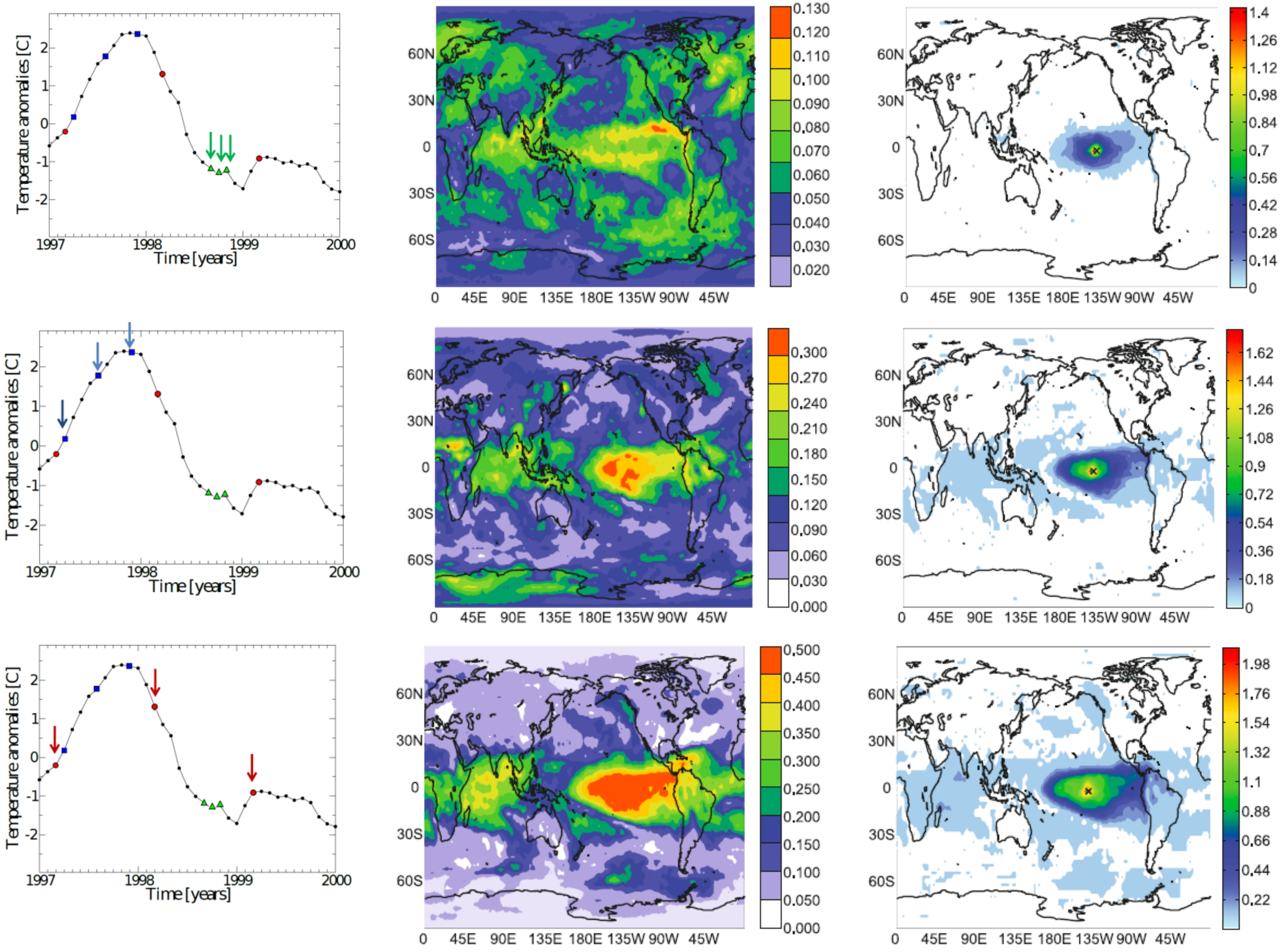
Color code: AWC

Video

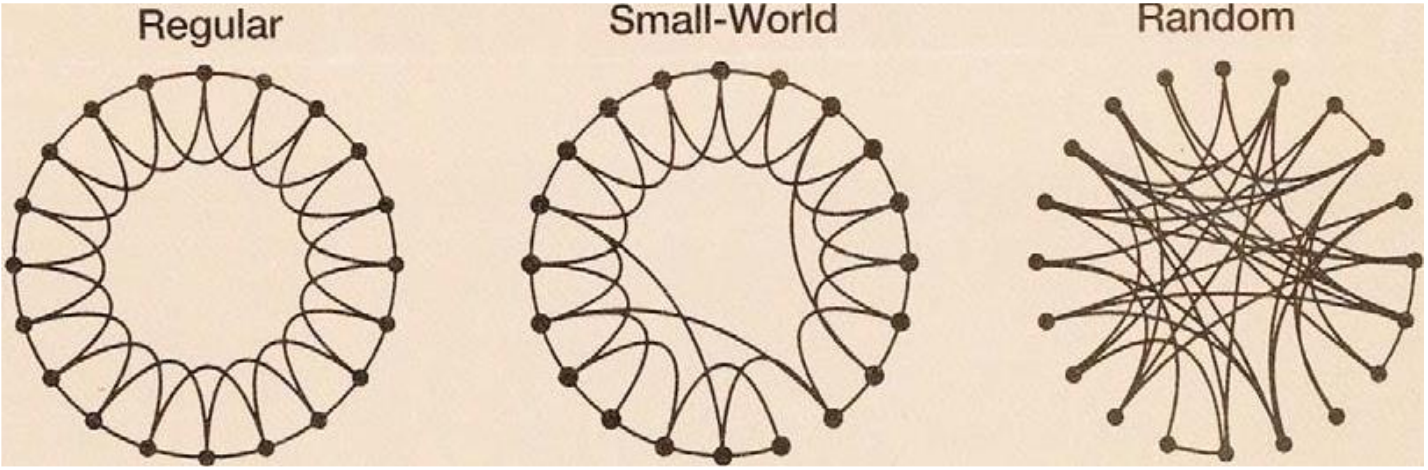




Influence of the time-scale of the pattern



Longer time-scale $\Rightarrow$ 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

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

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

- ***Assortativity***: tendency of a node to be connected to nodes with high degree

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

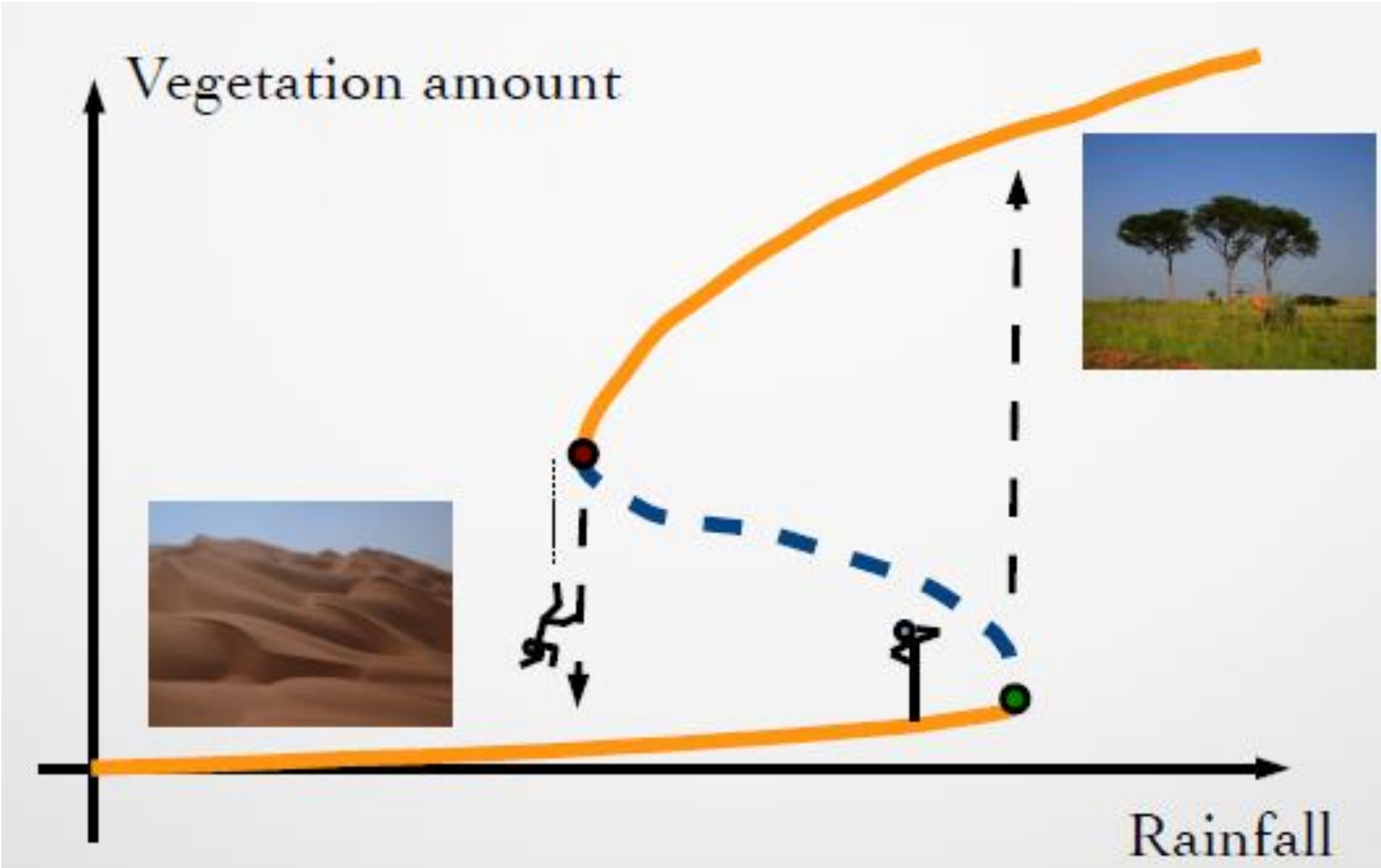
- ***Diameter***: longest shortest path

- ***Node entropy***: in weighted networks, measures the diversity of the weights of the links attached to node *i*.

$$H_i = -\sum_{j} p_{ij} \log p_{ij}$$

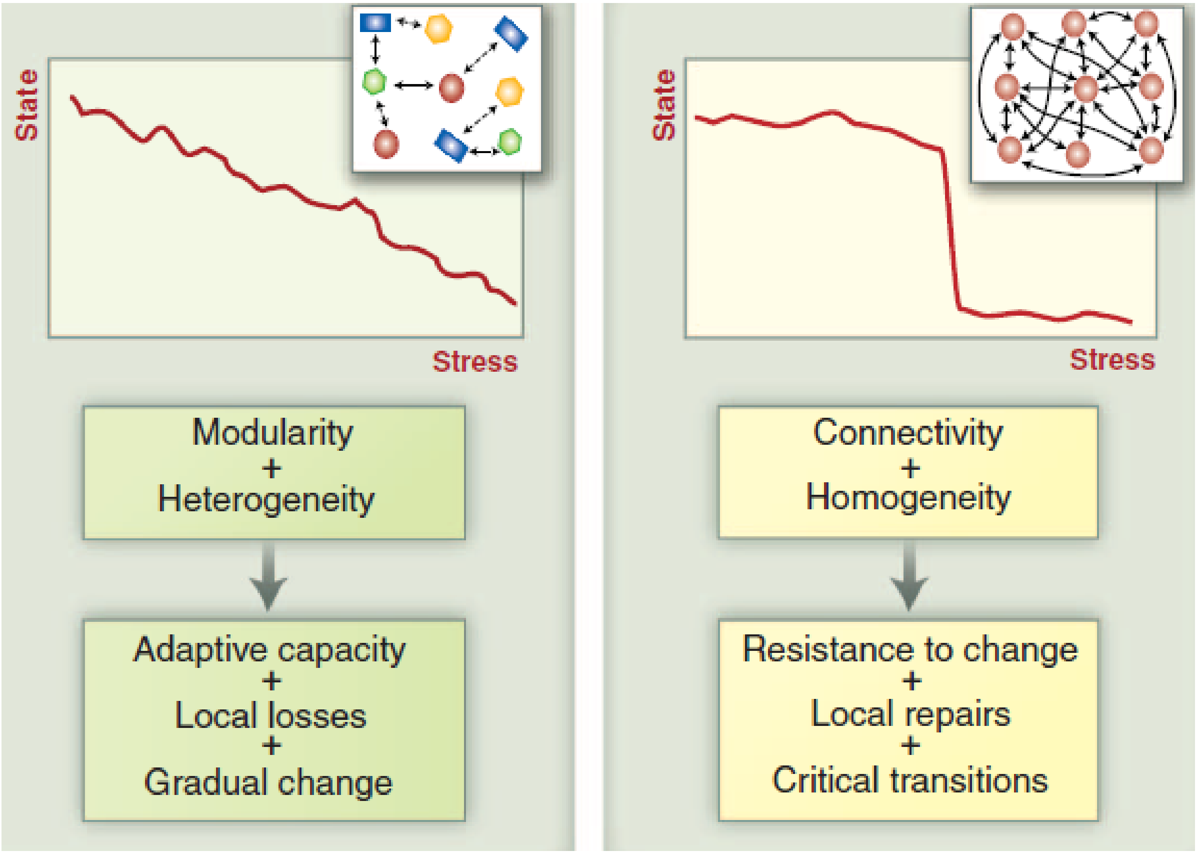
$$p_{ij} = w_{ij} / \sum_k w_{ik}$$

Example:
desertification transition under the lens of network analysis



Our goal: to develop reliable early-warning indicators

Role of the network structure



Networks in which the components are heterogeneous and where incomplete connectivity causes modularity tend to gradually adjust to change.

In highly connected networks, local losses tend to be "repaired" by subsidiary inputs from linked units until at a critical stress level the system collapses.

Scheffer et al. Science 338, 344 (2012)

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

Desertification transition: model

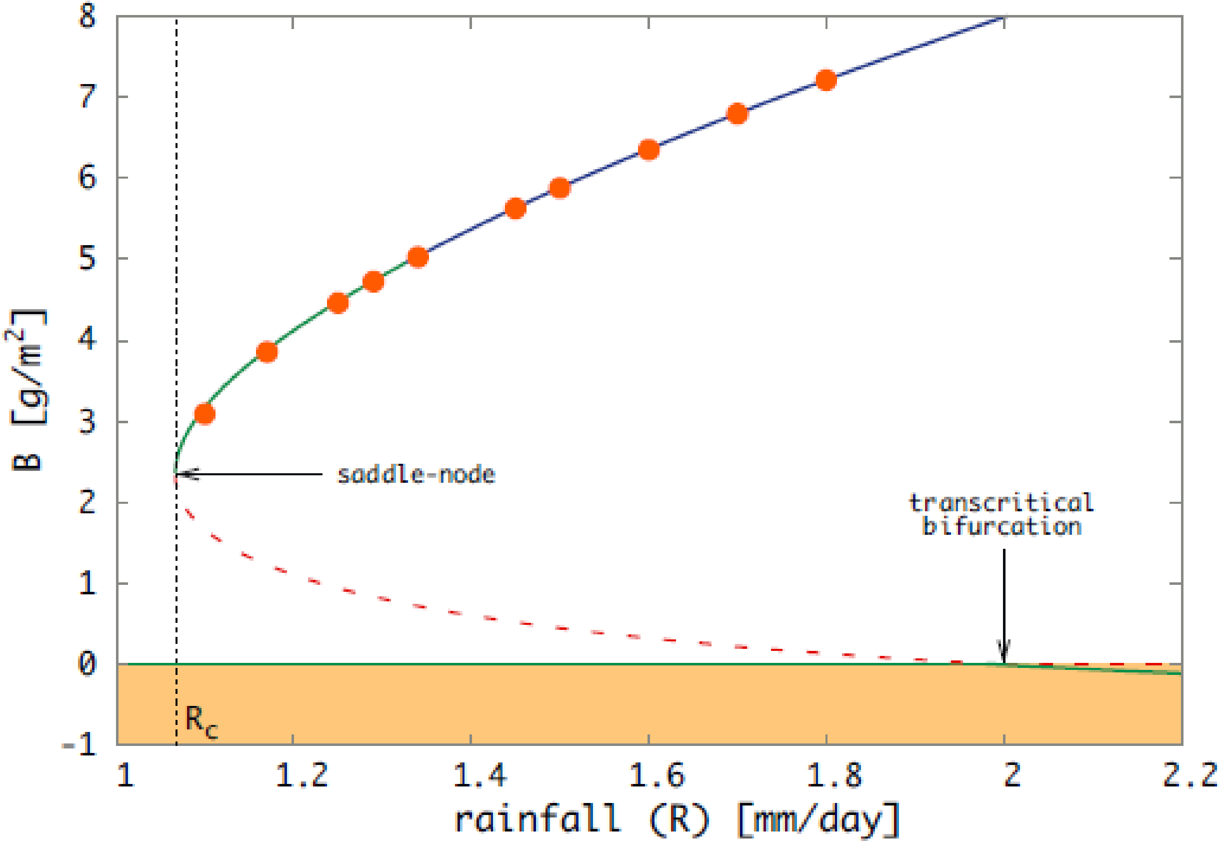
$$\frac{\partial w}{\partial t} = R - \frac{w}{\tau_w} - \Lambda w B + D\nabla^2 w + \sigma_w w_0 \xi^w(t),$$

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

- *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

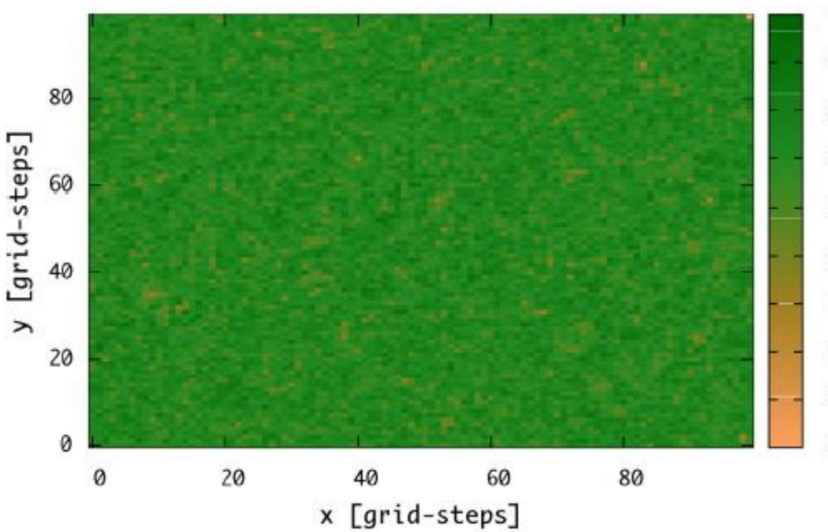


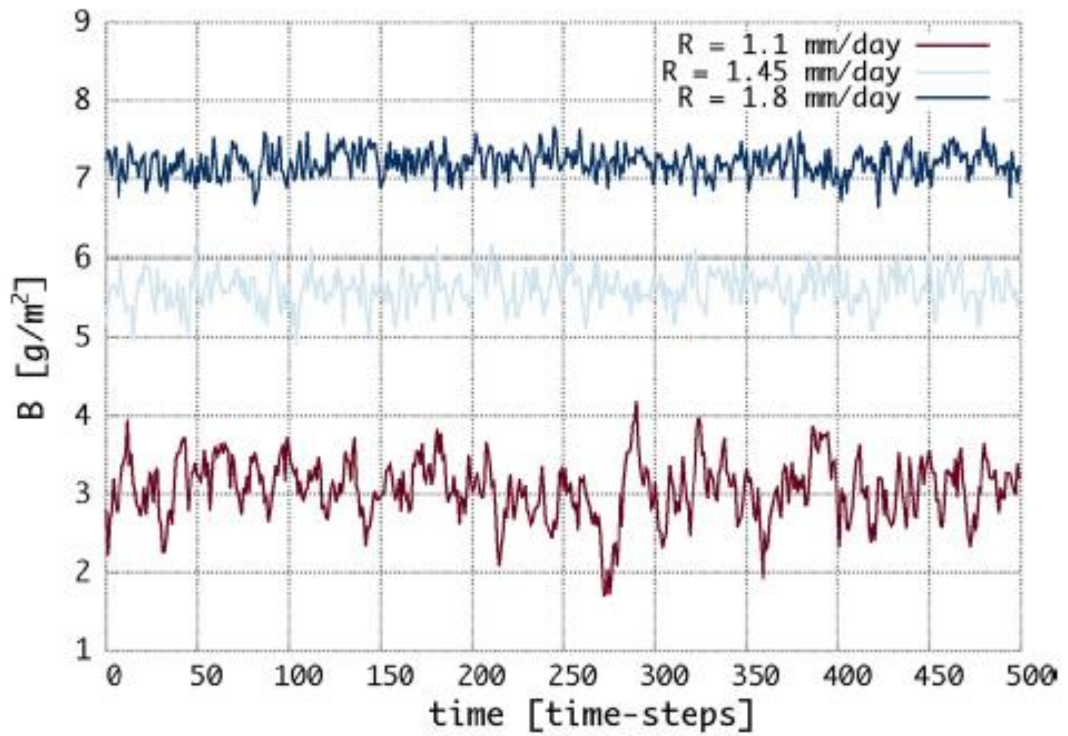
R<R_c: only desert-like solution (B=0)

R_c = 1.067 mm/day

Biomass time series

Biomass B when R=1.1 mm/day





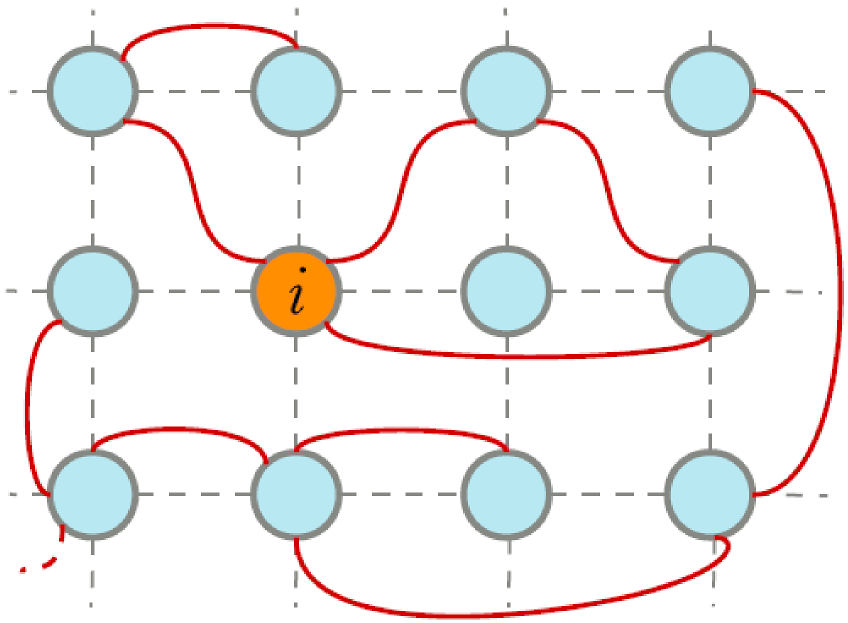
100 m x 100 m = 10^4 grid cells
Simulation time 5 days in 500 time steps
Periodic boundary conditions

Correlation Network

$$A_{ij} = \mathrm{H}(|\mathcal{C}(B_i, B_j)| - \theta)$$ Adjacency matrix

Zero-lagged cross-correlation

Threshold
θ=0.2 gives *p*<0.05



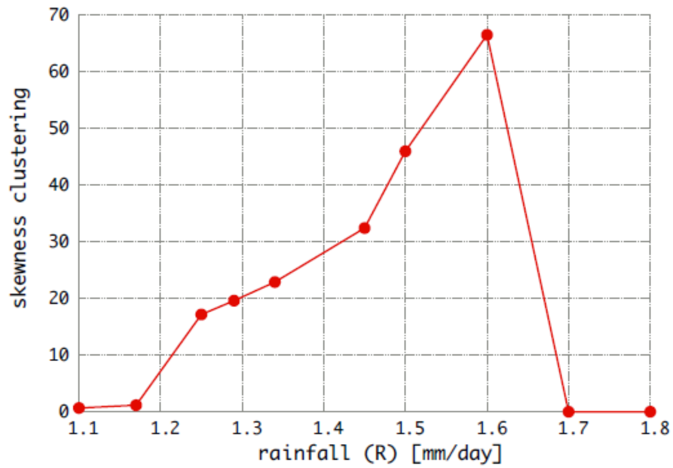
G. Tirabassi et al., Ecological Complexity (2014)

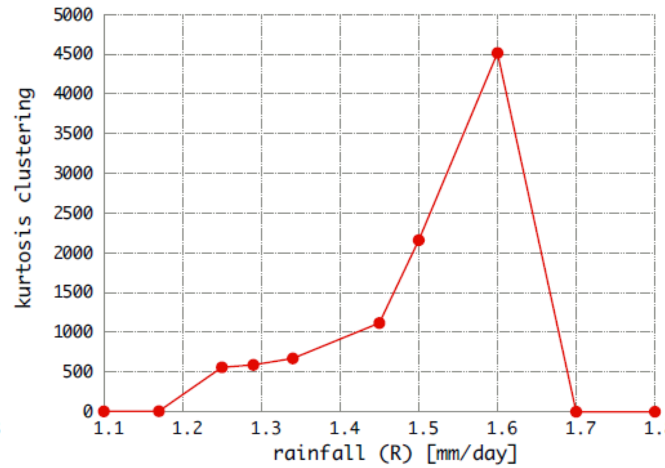
“Randomization” of the correlation network as the tipping point is approached

skewness

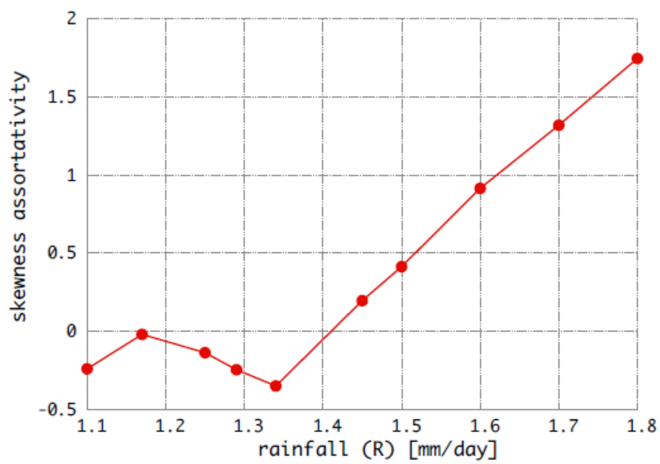
kurtosis

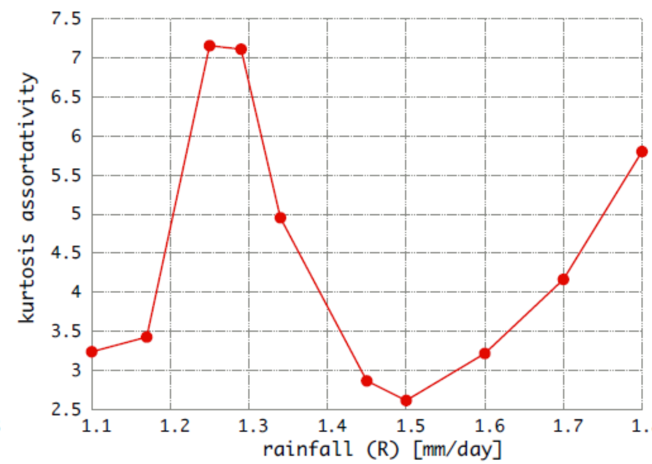
clustering



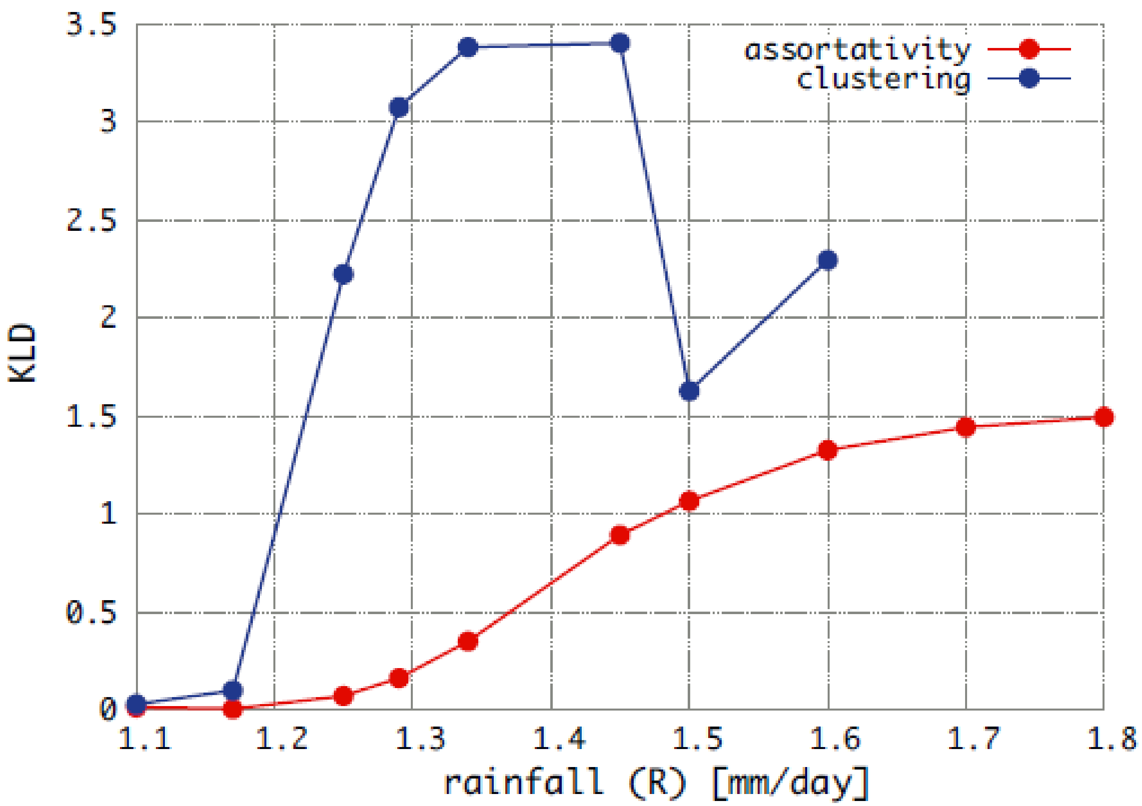


assortativity





The "Gaussianisation" of the distributions of a_i & c_i values is quantified by the Kullback–Leibler Distance



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

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

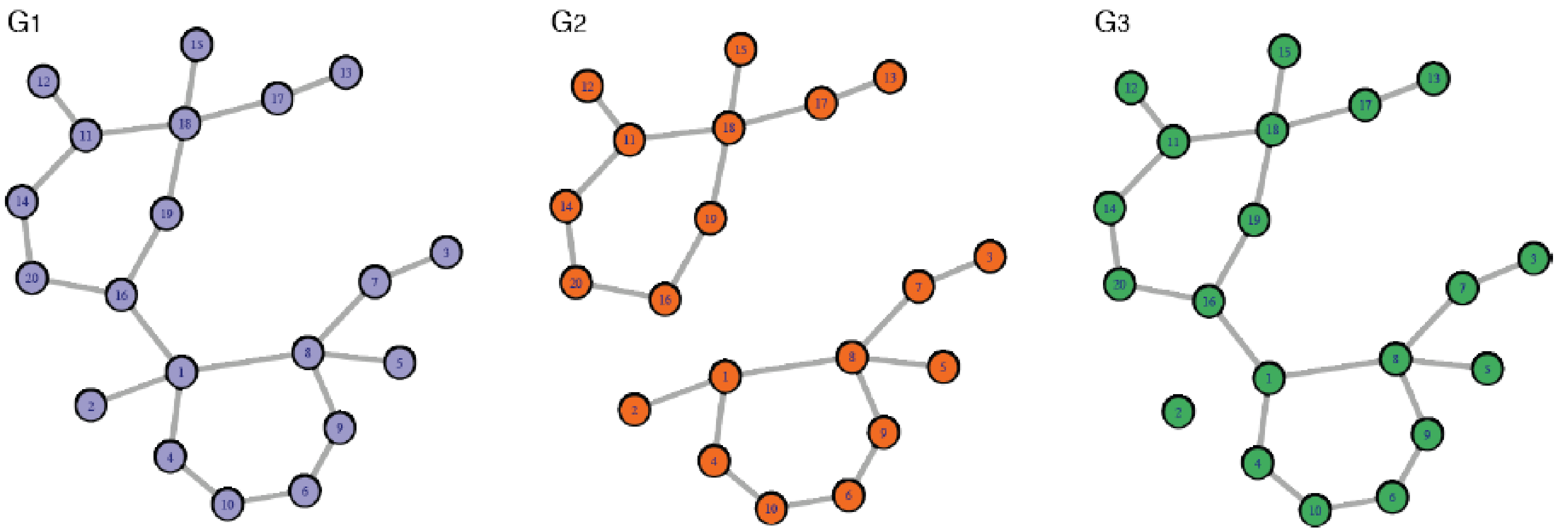
G. Tirabassi et al., Ecological Complexity 19, 148 (2014)

How to compare different networks?

Labelled networks with the same size

- Hamming distance $d_{\text{Hamming}}(y_1, y_2) = \sum_{i \neq j}^{N} \left[A_{ij}^{(1)} \neq A_{ij}^{(2)} \right]$

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



L. C. Carpi et al arXiv:1805.12350v1 (2018)

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)

- $p_i(j)$ of node “i“ is the fraction of nodes that are connected to node i at distance j
- If a network has N nodes:

$$\text{NDDs} = \text{vector of N pdfs } \{p_1, p_2, \ldots, p_N\}$$

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

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

- The *Network Node Dispersion (NND*) measures the heterogeneity of the N pdfs $\{p_1, p_2, \ldots, p_N\}$
- Quantifies the heterogeneity of connectivity distances.

$$\mathrm{NND}(G) = \frac{\mathcal{J}(\mathbf{P}_1, \ldots, \mathbf{P}_\mathrm{N})}{\log(d+1)}$$

d = diameter

$$\mathcal{J}(\mathbf{P}_1, \ldots, \mathbf{P}_\mathrm{N}) = \frac{1}{N} \sum_{\mathrm{i,j}} p_\mathrm{i}(j) \log\left(\frac{p_\mathrm{i}(j)}{\mu_\mathrm{j}}\right)$$

$$\mu_\mathrm{j} = \left(\sum_{\mathrm{i}=1}^{N} p_\mathrm{i}(j)\right) / N$$

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.

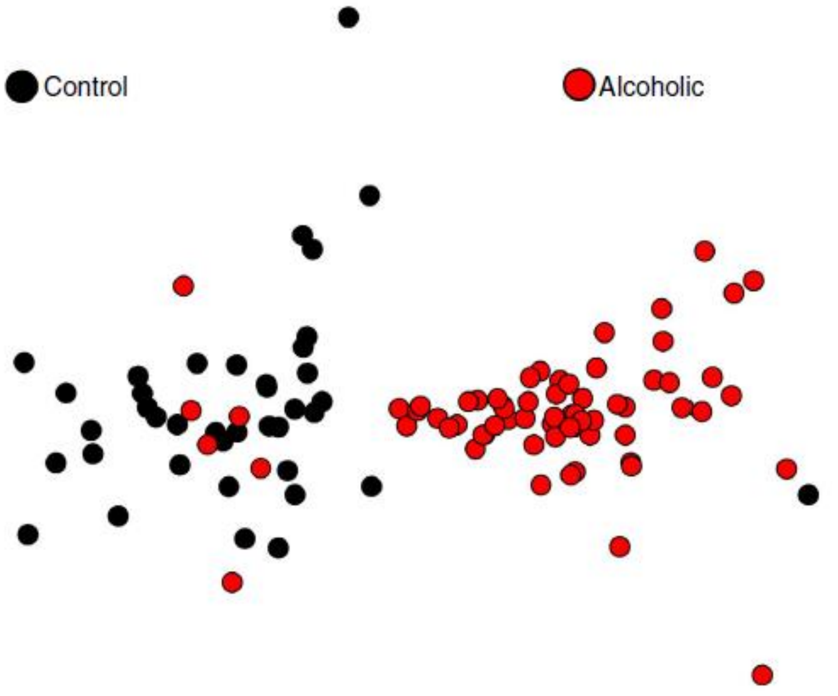
Application: comparing brain networks

- EEG data
  - https://archive.ics.uci.edu/ml/datasets/eeg+database
  - 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.

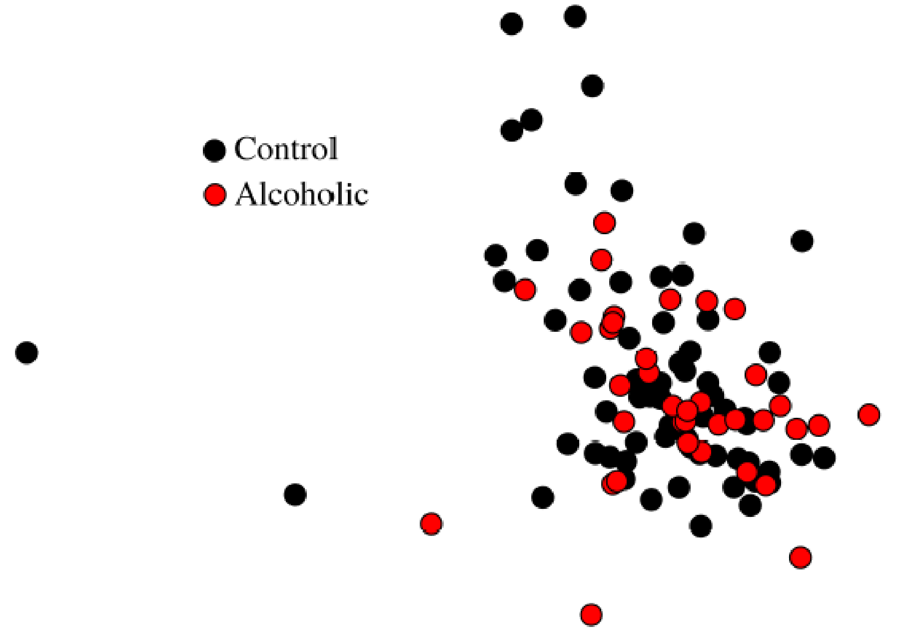
  $\Rightarrow$ We can compare the different individuals.

Two brain regions are identified ('nd' and 'y'): the weights of the links are higher in control than in alcoholic subjects

Dissimilarity measure



Hamming distance



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

Network inference: how to infer the underlying interactions from observed data?

a classification problem

Main problem:

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

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

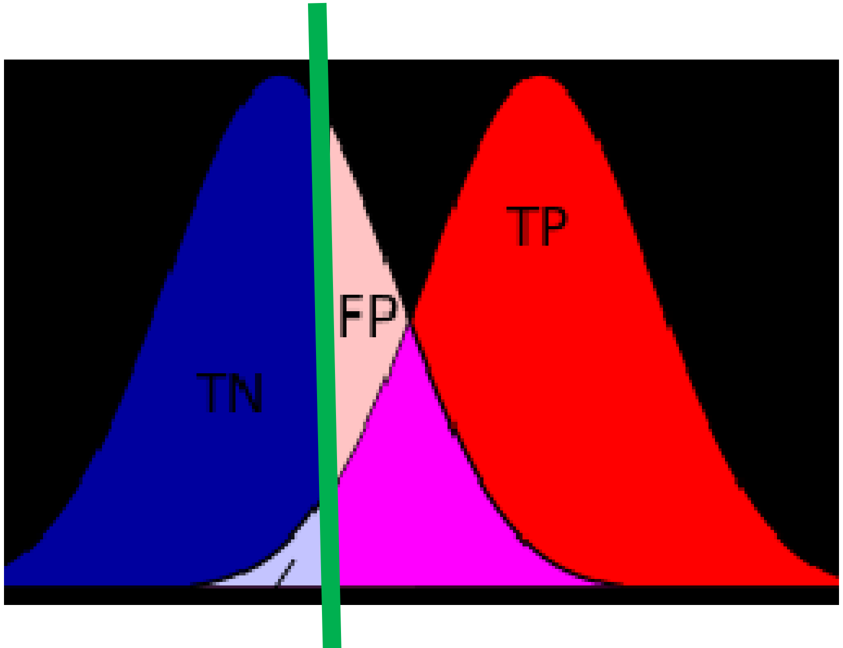
- There are many statistical similarity measures to infer bi-variate mutual interactions from observations, i.e., to classify:
  - the interaction exists (is significant)
  - the interaction does not exists (or is not significant)

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

Confusion matrix

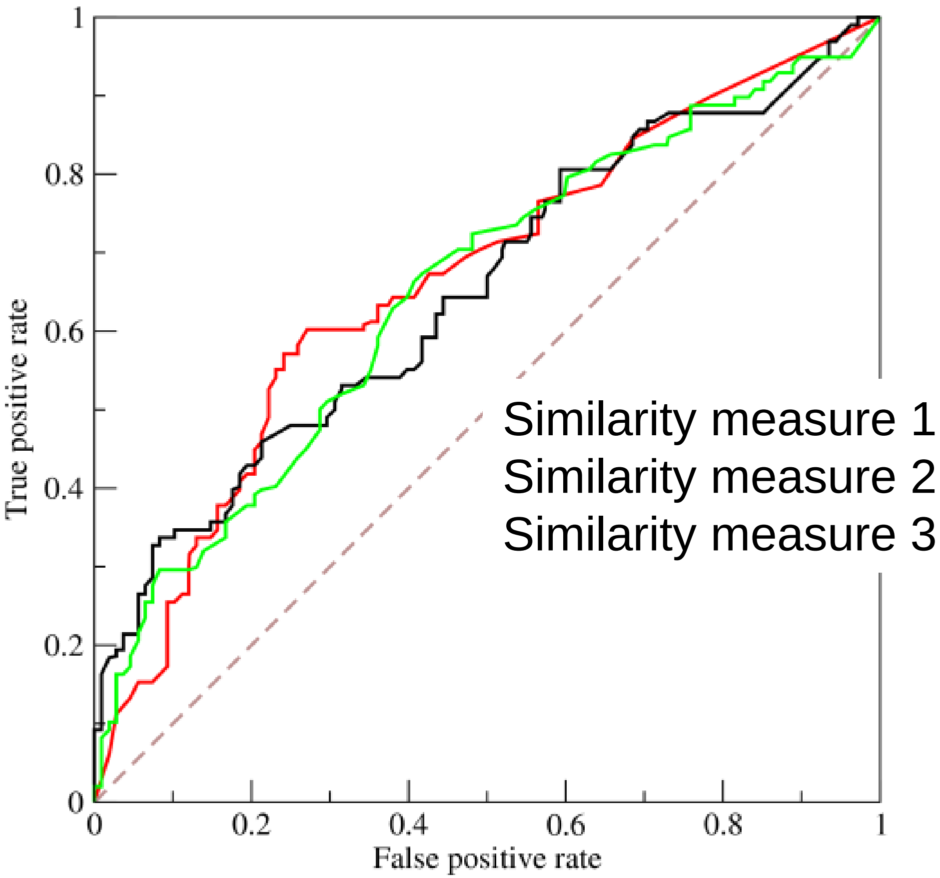
- ***Accuracy***: How often is the classifier correct? **(TP+TN)/total**
- ***Misclassification*** (Error Rate): How often is it wrong? **(FP+FN)/total**
- ***True Positive Rate*** (TPR, Sensitivity): When it's yes, how often does it predict yes? **TP/actual yes**
- ***False Positive Rate*** (FPR) : When it's no, how often does it predict yes? **FP/actual no**
- ***Specificity*** (1 – FPR) : When it's no, how often it predicts no? **TN/actual no**
- ***Precision*** (Positive Predictive Value): When it predicts yes, how often is it correct? **TP/predicted yes**
- ***Negative Predictive Value***: When it predicts no, how often is it correct? **TN/predicted no**
- ***Prevalence***: How often does the yes condition actually occur in the sample? **actual yes/total**

Receiver operating characteristic (ROC curve)



TP	FP
FN	TN

Source: wikipedia



Our goal

- To compare the performance of different statistical similarity measures for inferring interactions from observations.
- Using a “toy model” where **we know** the underlying equations and interactions and so we can check the performance of the different measures in inferring the interactions.

Kuramoto oscillators in a random network

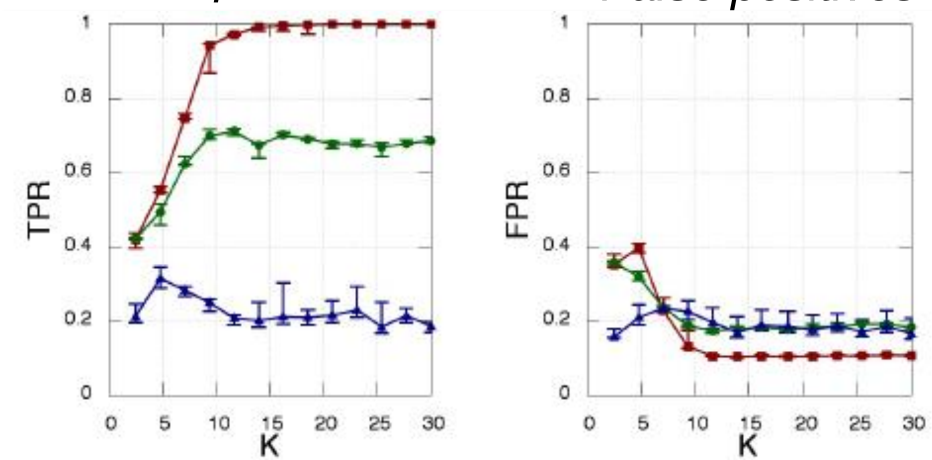
$$d\theta_i = \omega_i dt + \frac{K}{N}\sum_{j=1}^{N} A_{ij}\sin(\theta_j - \theta_i)dt + D\, dW_t^i$$

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

CC MI MIOP

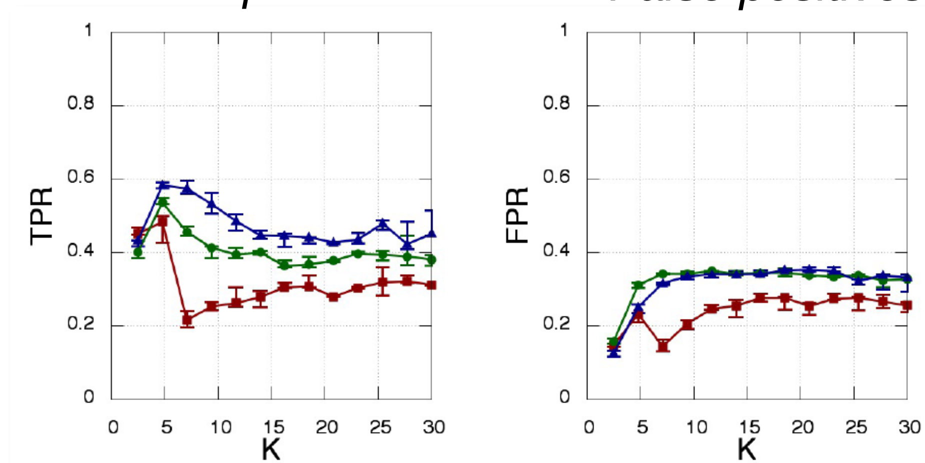
Phases (θ)

True positives *False positives*



"Observable" Y=sin(θ)

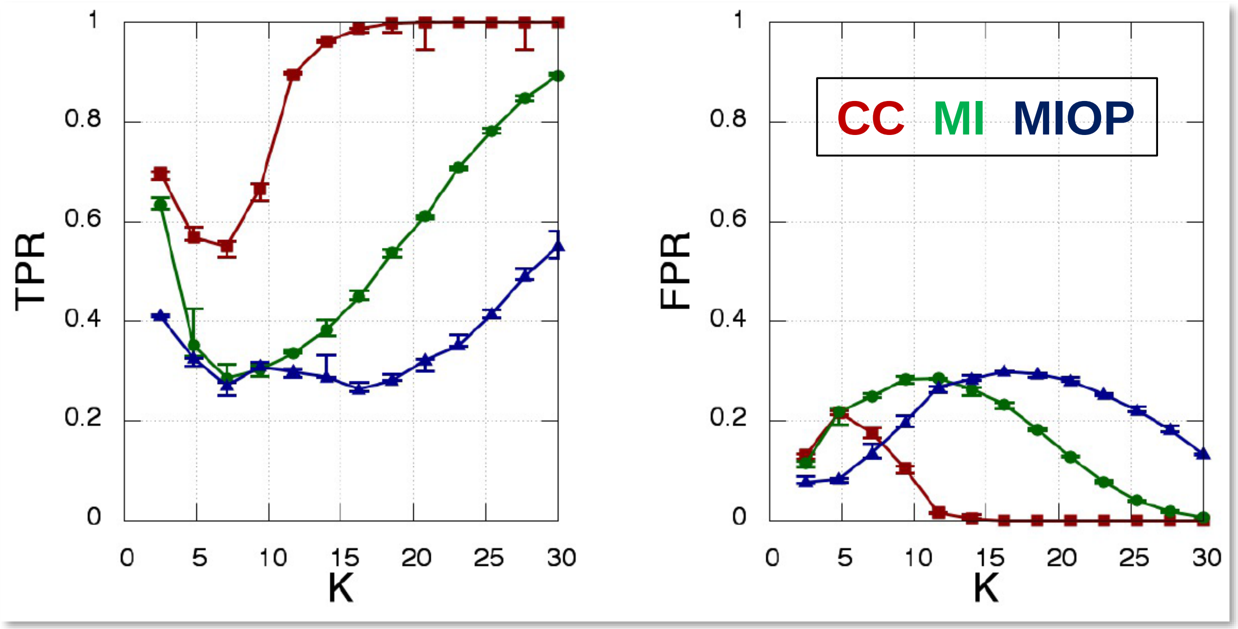
True positives *False positives*



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)



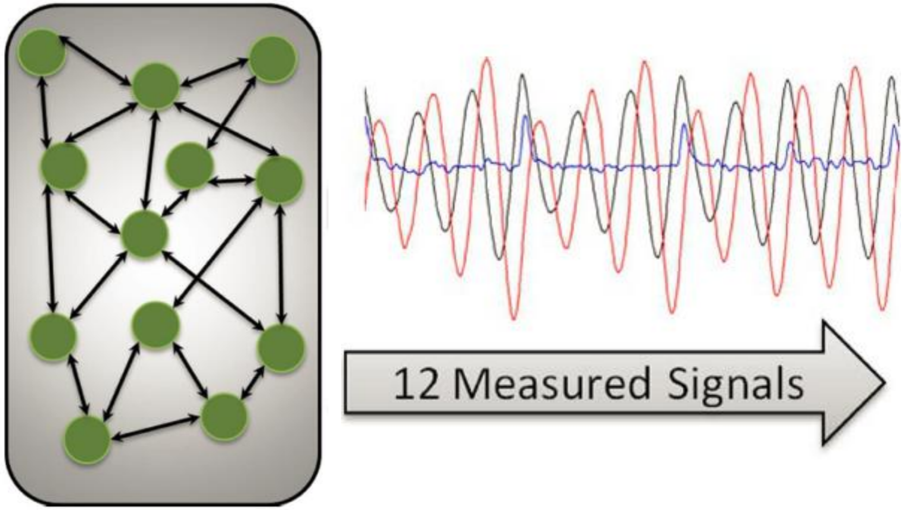
Perfect network inference is possible!

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)



The Hilbert Transform was used to obtain phases from experimental data

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

- Kuramoto Oscillators' Network

$$\theta_i$$

$$f_i = \dot{\theta}_i$$

$$Y_i = \sin(\theta_i)$$

- Rössler Oscillators' Network

$$\varphi_i = HT(x_i)$$

$$f_i = \dot{\varphi}_i$$

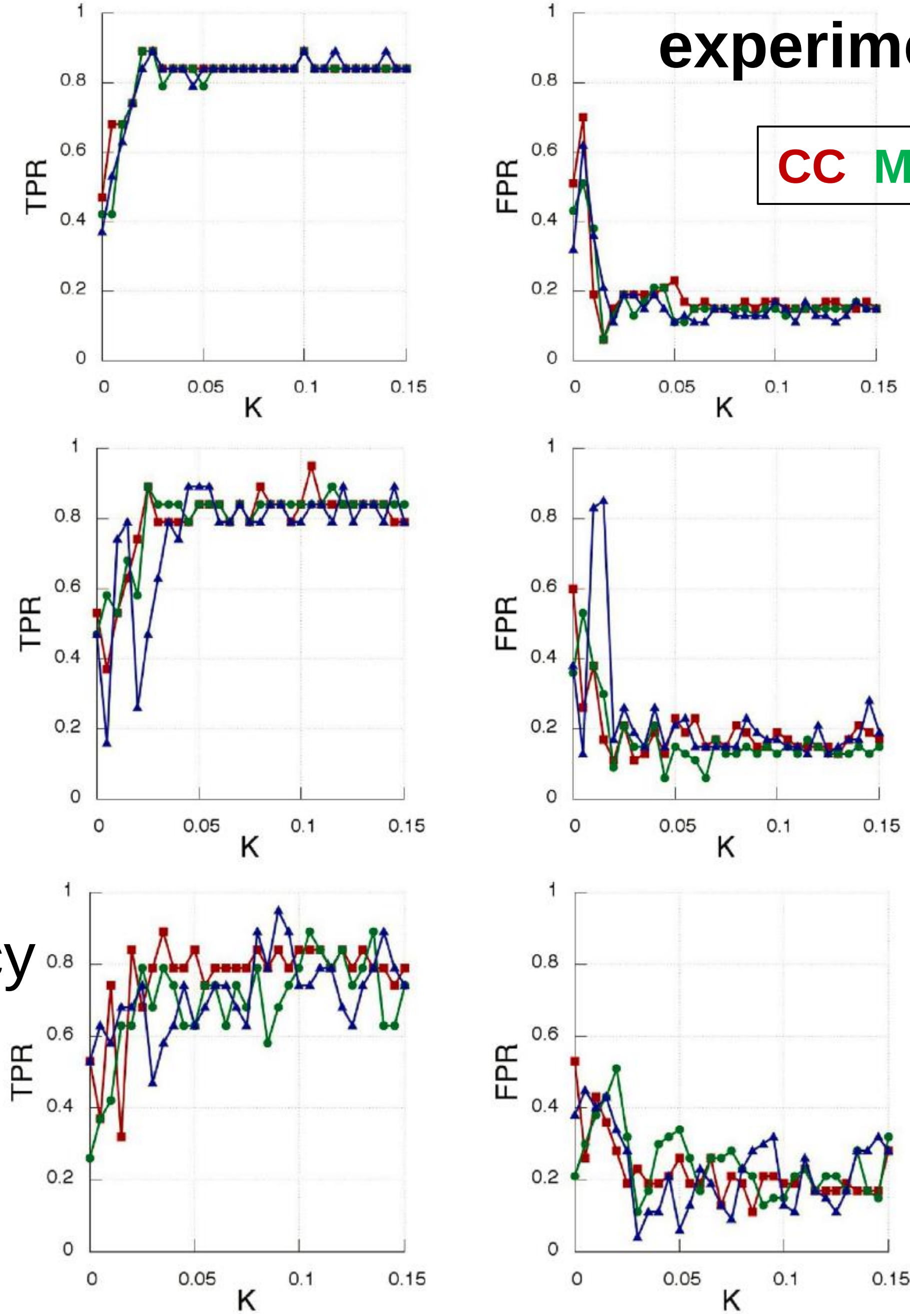
$$x_i$$

Results obtained with experimental data

Observed variable (x)

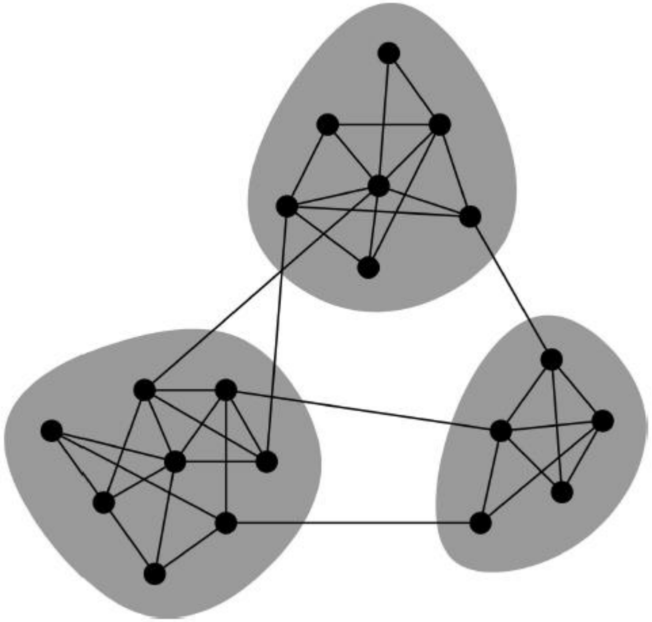
Hilbert phase

Hilbert frequency



– No perfect reconstruction

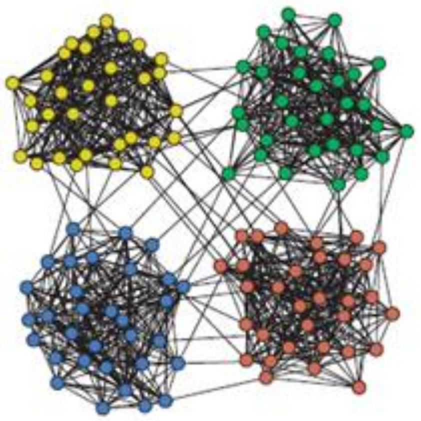
– No important difference among the 3 methods & 3 variables



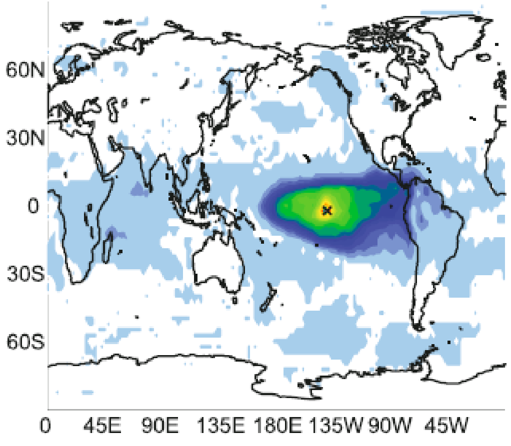
Community detection

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 $$w_{ij} = \frac{1}{\sum_{\alpha\beta}\left(TP^{i}_{\alpha\beta} - TP^{j}_{\alpha\beta}\right)^2}$$

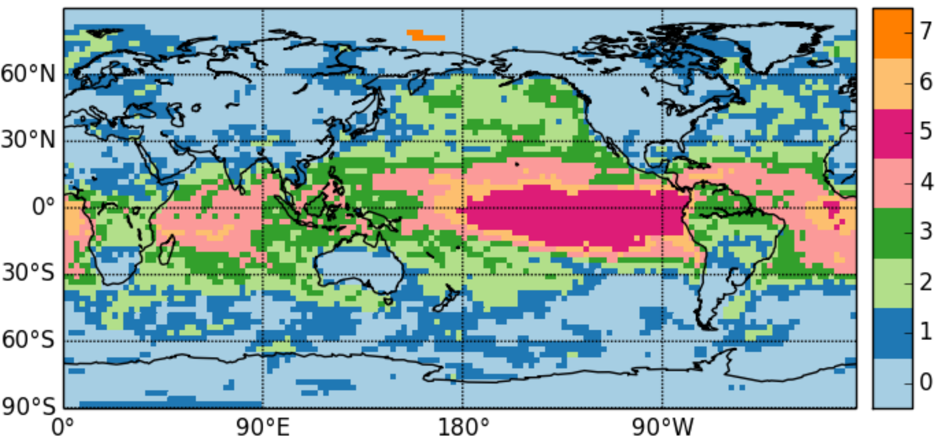
  High weight if similar symbolic "language"

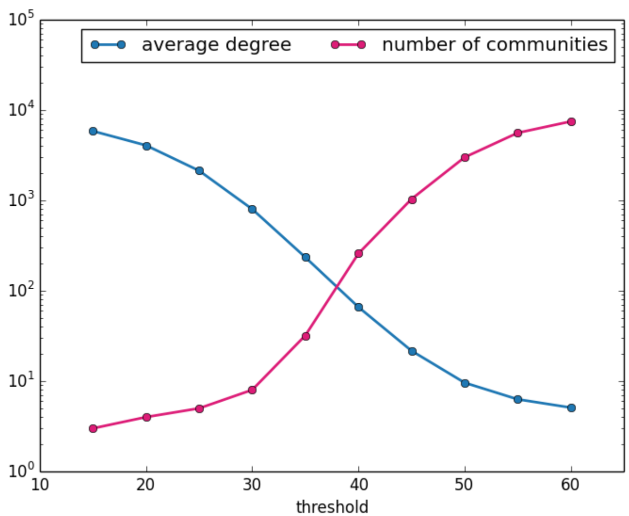
- Step 4: threshold w_{ij} to obtain the adjacency matrix.

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

Results

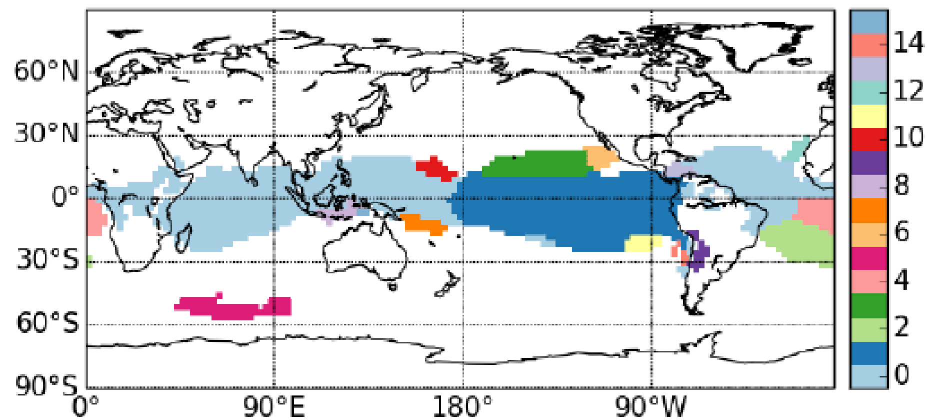
TP Network

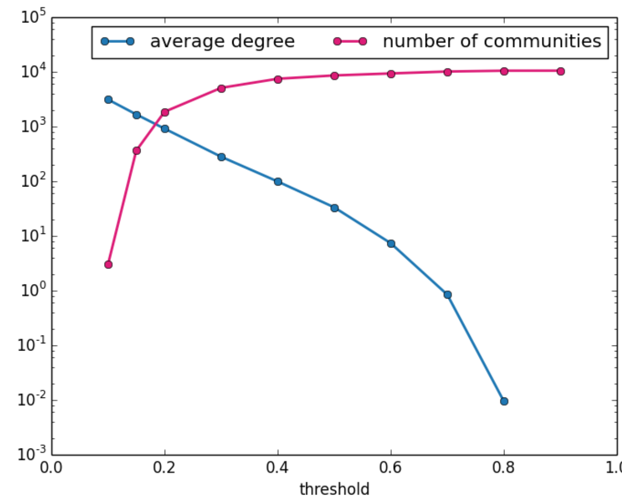




CC Network

(only the largest 16)





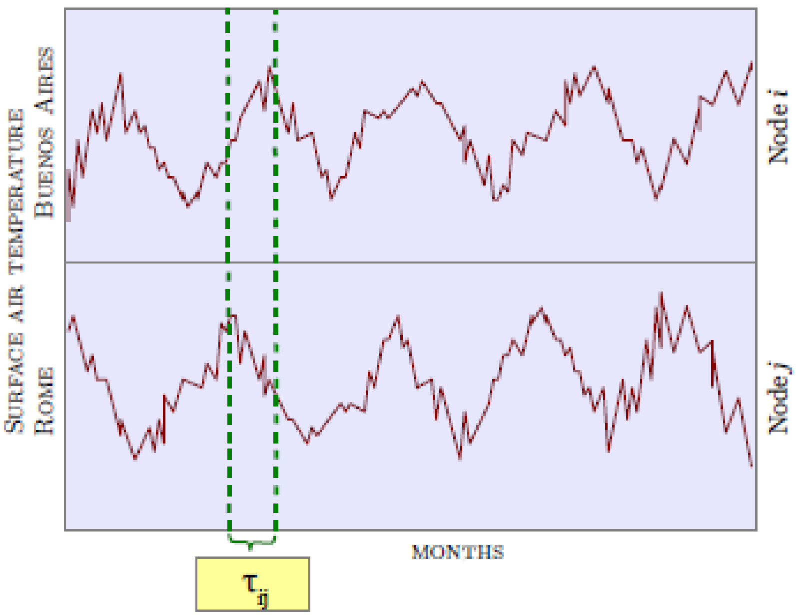
G. Tirabassi and C. Masoller, "*Unravelling the community structure of the climate system by using lags and symbolic time-series analysis*", Sci. Rep. **6**, 29804 (2016).

Community detection algorithms

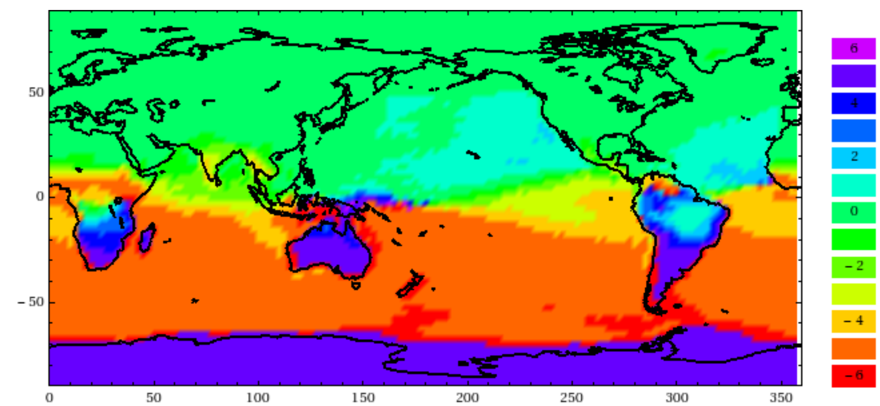
- *Infomap* (http://www.mapequation.org/code.html) and many others.
- *Infomap* clusters tightly interconnected nodes into modules and detects nested modules.
- Many other algorithms have been proposed.
- Further reading: S. Fortunato, "*Community detection in graphs*", Phys. Rep. 486, 75 (2010).

Another way to identify geographical regions with similar climate

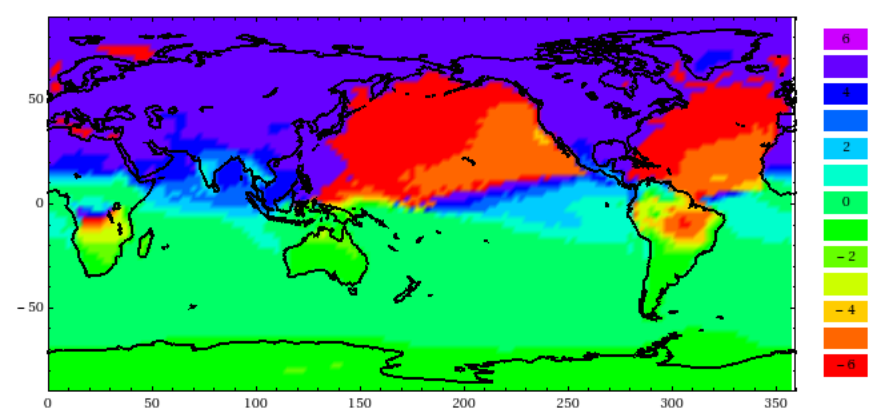
- Analyze lag-times between seasonal cycles: cross-correlation analysis of Surface Air Temperature



Rome



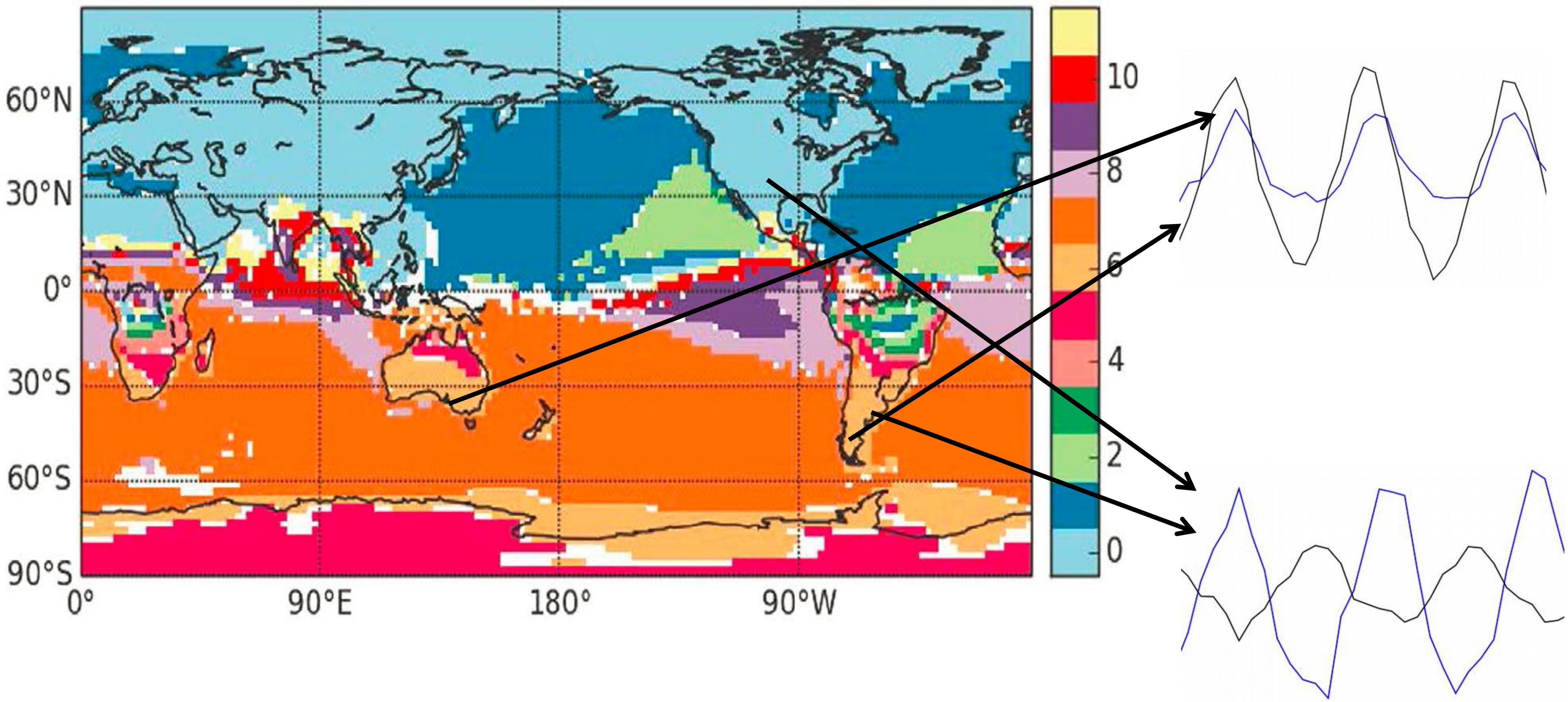
Buenos Aires



- The lags between 3 time series are well defined if

$\tau_{ij} = (\tau_{ik} + \tau_{kj})\text{mod}12$

Geographical regions with synchronous (inphase) seasonal cycles

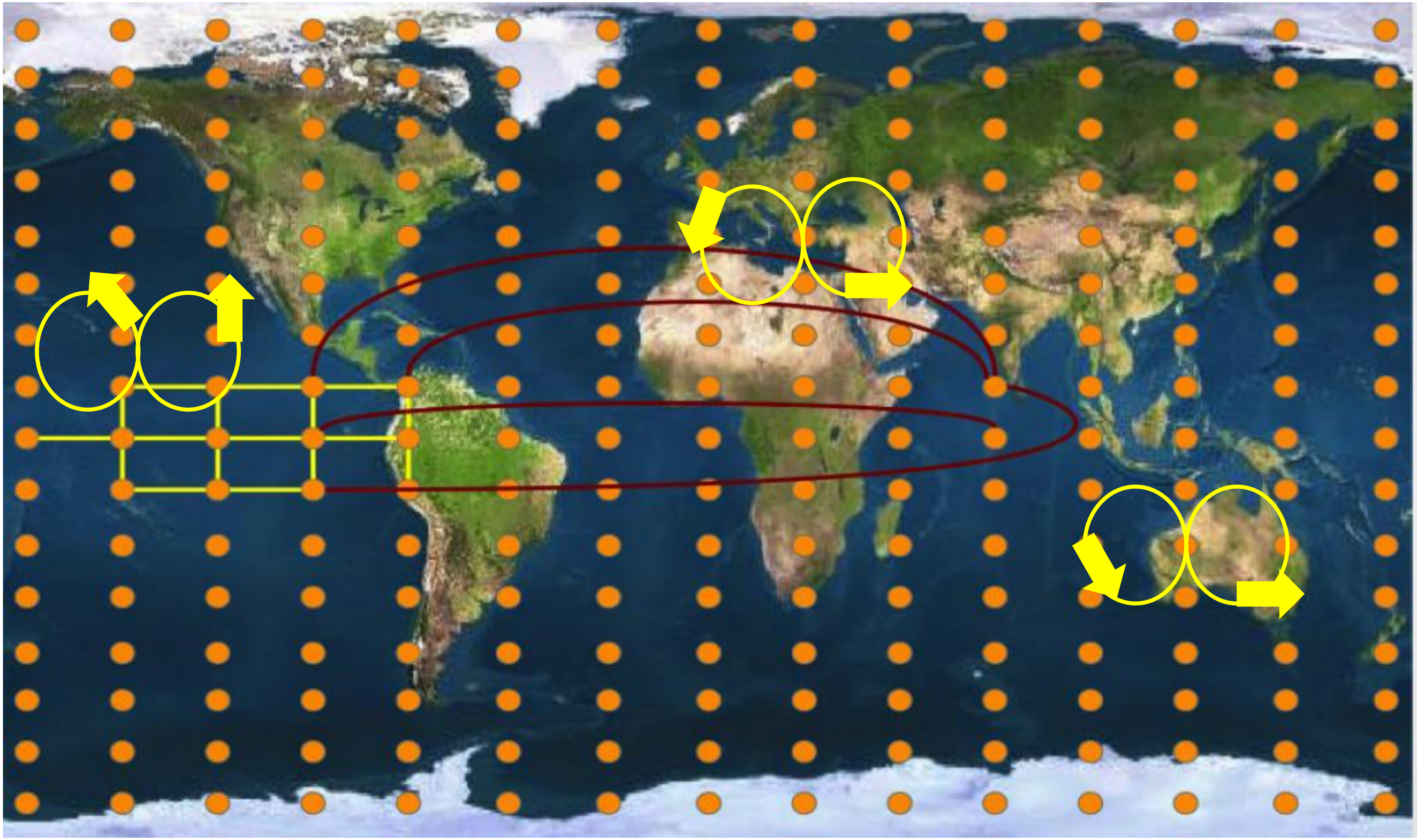


- Six-month lag between the two hemispheres.
- Oceans have a one-month lag with respect to the landmasses

[G. Tirabassi and C. Masoller, Sci. Rep. 6:29804 (2016)](#)

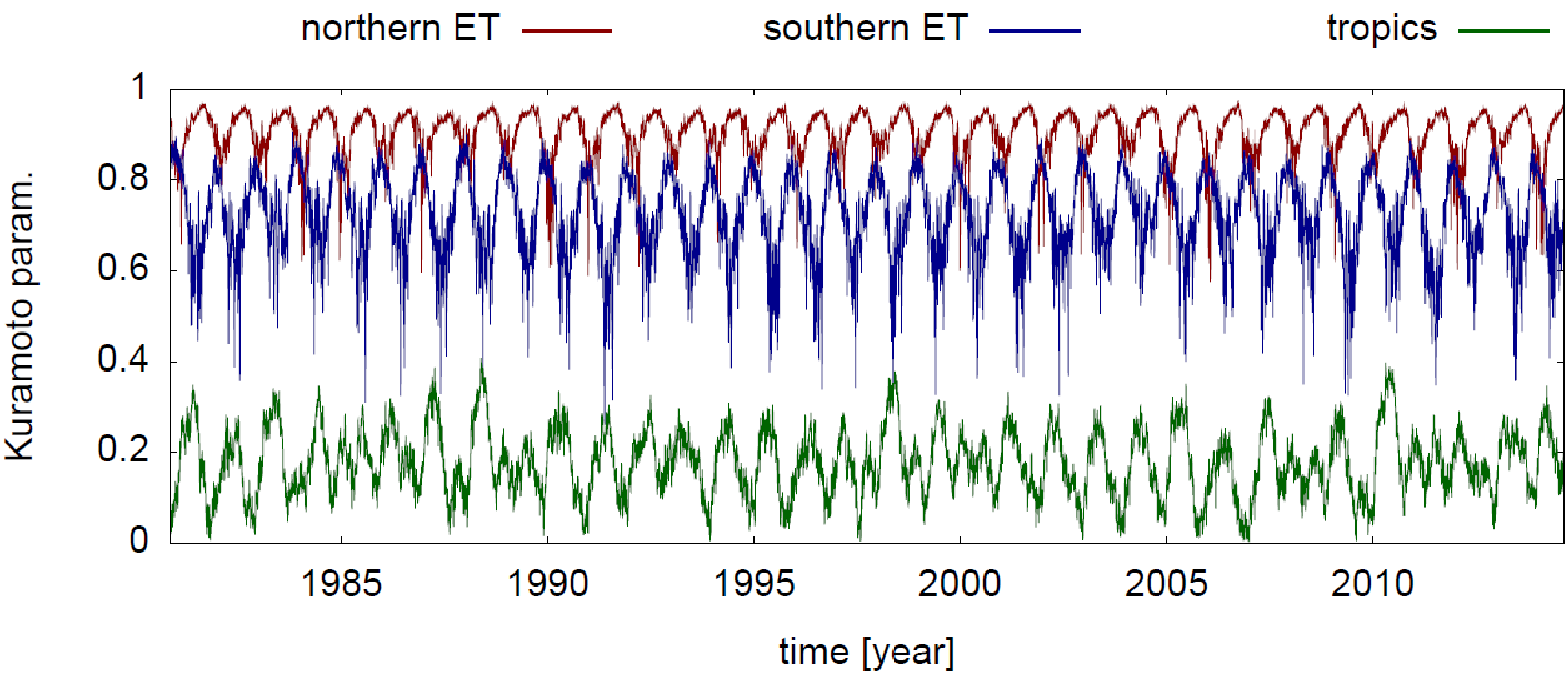
How to detect phase synchronization in climate data?

Network of individual oscillators



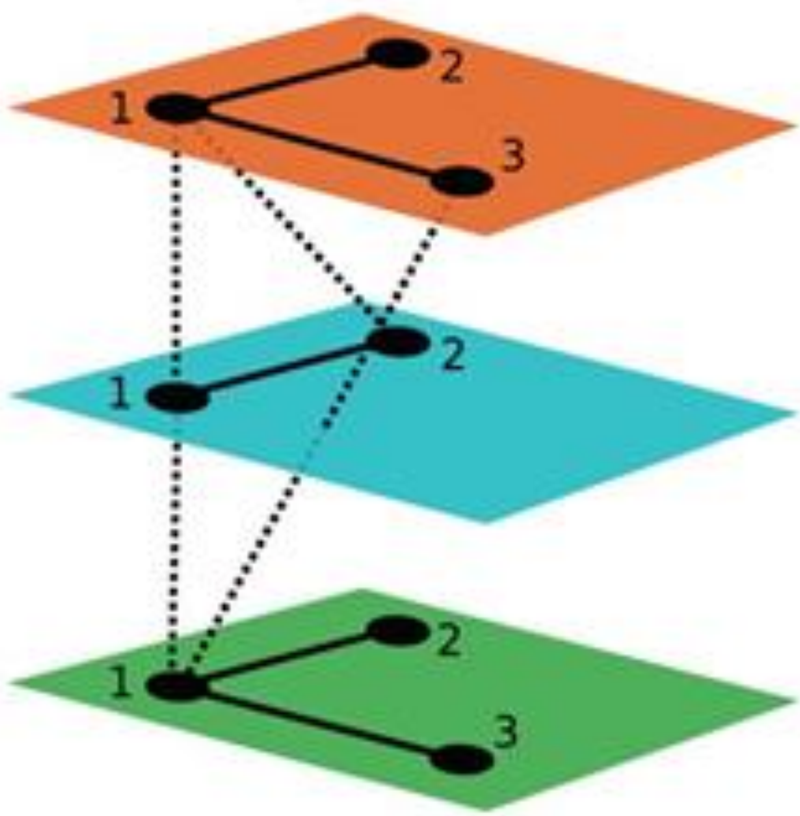
After using the Hilbert transform to obtain phase time series, we calculate the Kuramoto order parameter

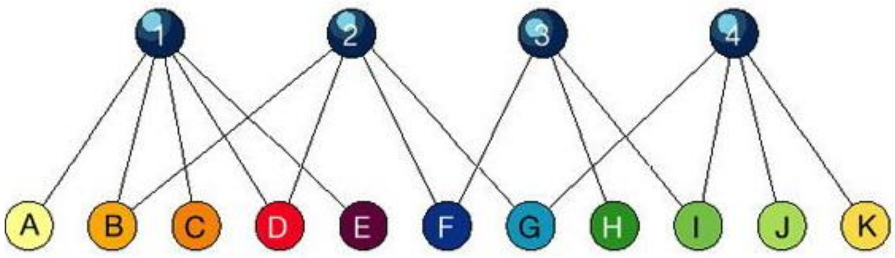
$$r(t) = \left| \frac{1}{N} \sum_{j=1}^{N} e^{i\theta_j(t)} \right|$$

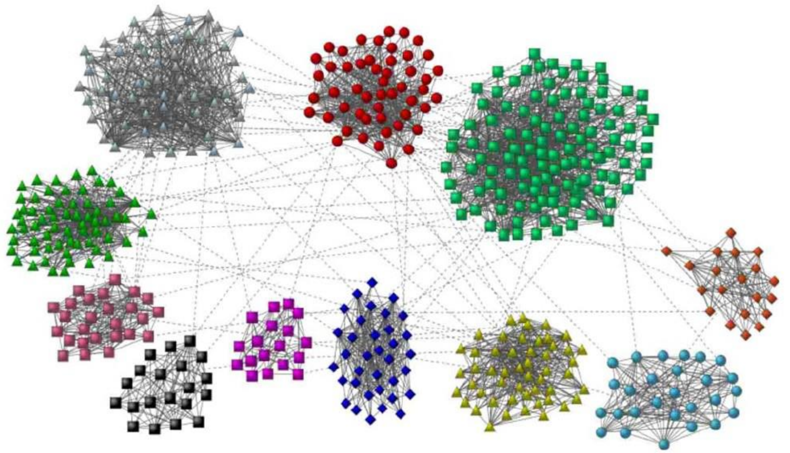


Generalizations of complex network analysis

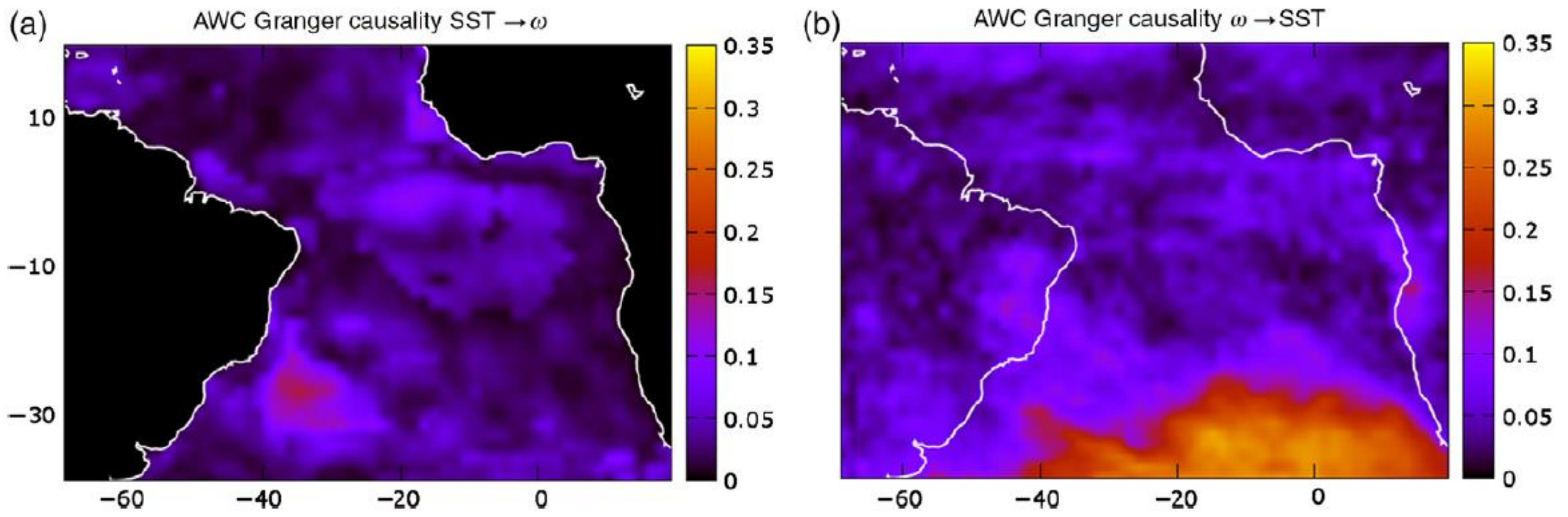
Network structures: Multilayer, multiplex, bipartite, networks of networks and many others







Example of a bilayer climate network representing ocean-precipitation interactions



Color code shows the area-weighted connectivity (weighted degree) of a bilayer network where the links are defined using Granger causality (only GCE values at 99% confidence level have been considered).

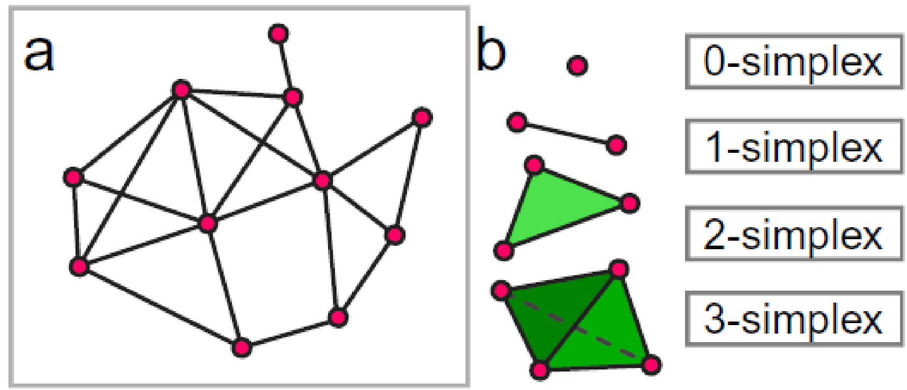
SST = Surface sea temperature

ω = vertical wind velocity at 500 hPa (precipitation proxy)

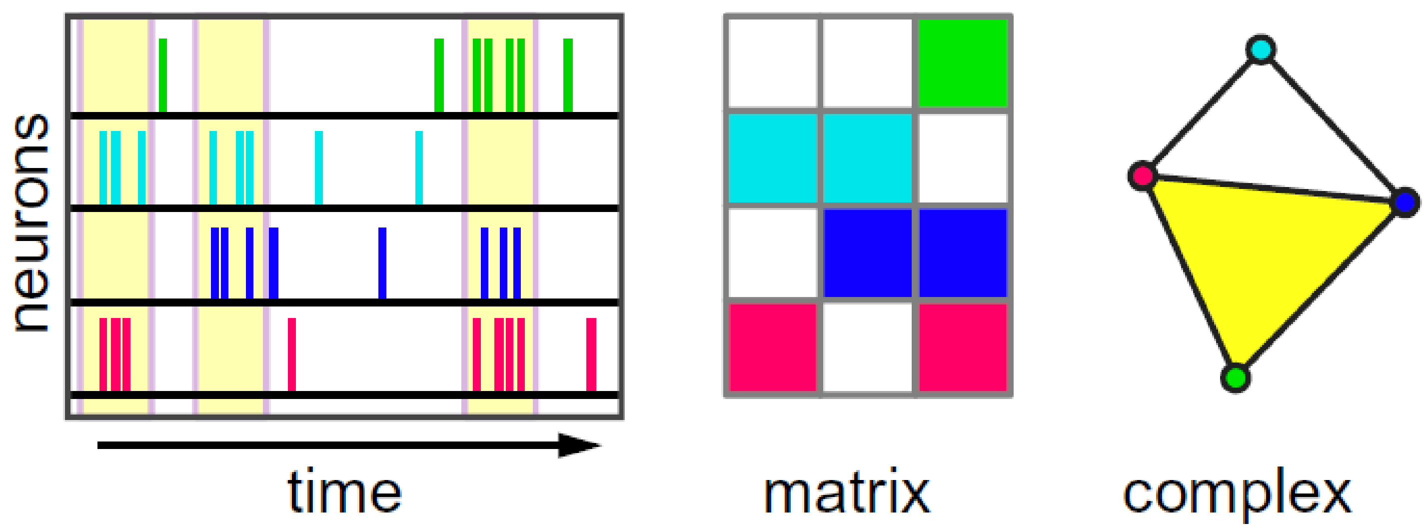
Tirabassi, Masoller and Barreiro, Int. J. of Climatology, 35, 3440 (2015)

A basic limitation of network analysis

- 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

- There are many methods for inferring the underlying connectivity of a complex system from the observed output signals.
- Different methods infer different networks.
- Comparing (quantifying differences) between networks is challenging.
- Different sets of “communities” (clusters) can be uncovered depending on the property that is analyzed.
- Network science is growing fast and has many applications!

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)
- 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)

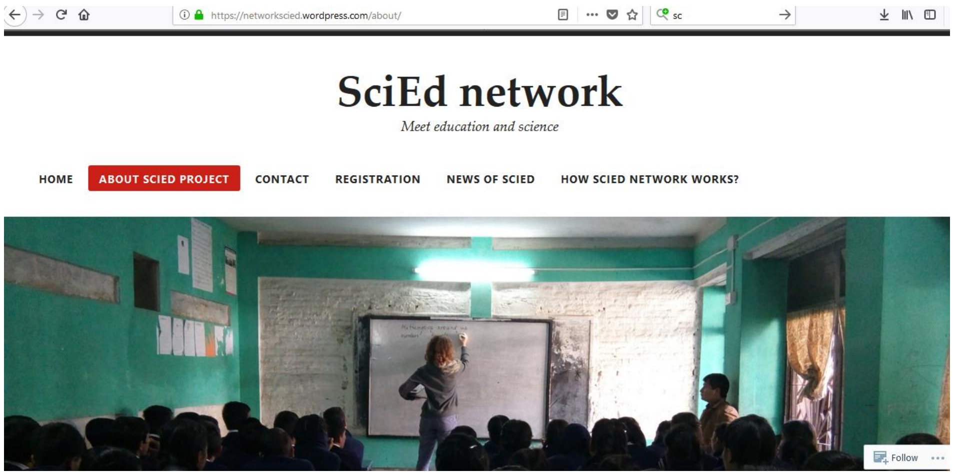
Coauthors

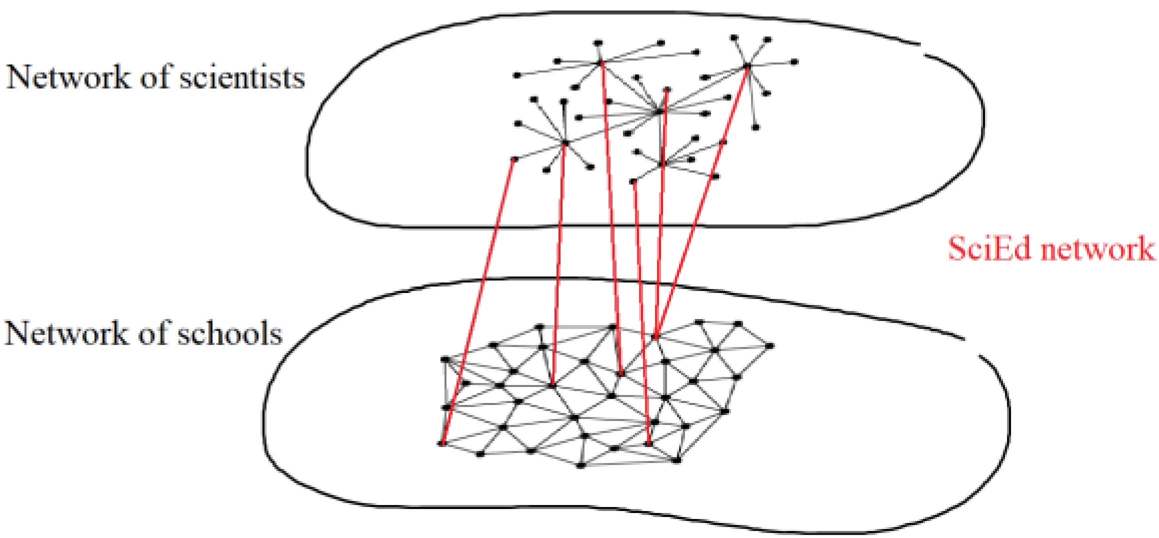
     

- Maria Masoliver, Pepe Aparicio Reinoso (*neuron models*)
- Taciano Sorrentino, Carlos Quintero, Jordi Tiana, Carme Torrent (*laser lab*)
- Andres Aragoneses, Laura Carpi (*data analysis, networks*)
- Ignacio Deza, Giulio Tirabassi, Dario Zappala, Marcelo Barreiro (*climate*)

Advertising





https://networkscied.wordpress.com/

The european project CAFÉ (*Climate Advanced Forecasting of subseasonal Extremes*) will start march 2019 and will offer several PhD positions. Interested? Contact me!

<cristina.masoller@upc.edu>

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