Exploiting Lag-Time Information for Optimizing Network Inference

Nicolas Rubido[1] and Cristina Masoller[2]

[1]Universidad de la República, Montevideo, Uruguay
[2]Universitat Politecnica de Catalunya, Terrassa, Barcelona, Spain

cristina.masoller@upc.edu

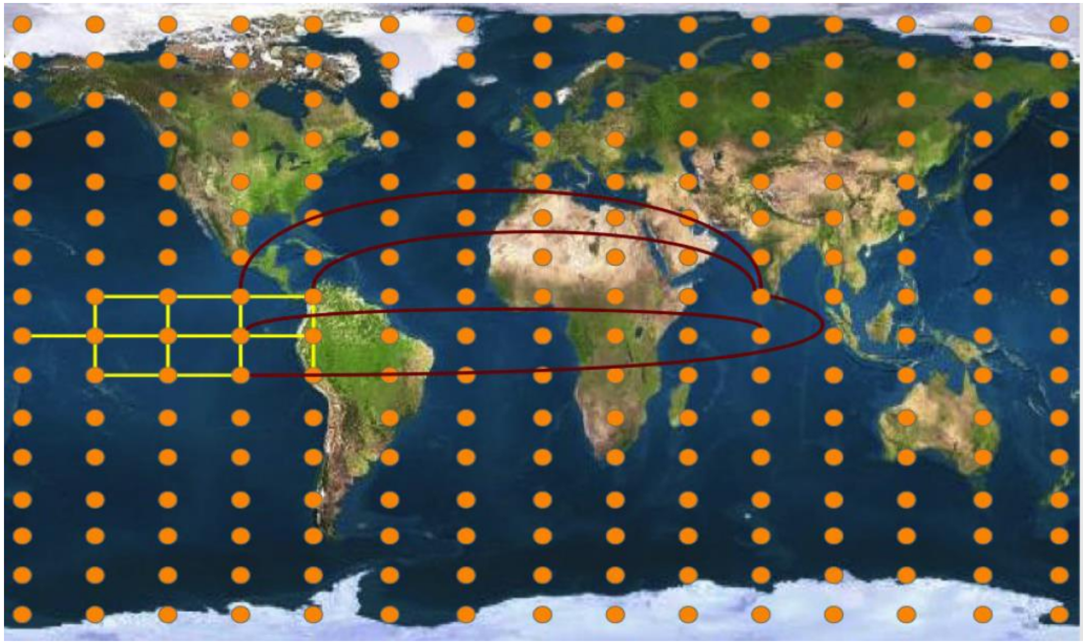
www.fisica.edu.uy/~cris

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Campus d'Excel·lència Internacional

NOLTA 2018, Tarragona, Spain

Motivation: how to infer the connectivity of a complex system from observed data?



Spatial grid points used for building climate networks

Linear and nonlinear correlation analysis are used for inferring undirected links

Time series recorded at nodes i an j: $a_i(t)$, $a_j(t)$, $t=1, \ldots, T$

- Lagged cross correlation

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

- Mutual information
  - Histograms
  - Symbolic (ordinal) patterns

$$MI_{ij}(\tau_{ij}) = \sum_{m,n} p_{ij}(m, n) \log_2 \left(\frac{p_{ij}(m, n)}{p_i(m) p_j(n)} \right)$$

$\boldsymbol{p_i}$ is the prob. that $a_i(t)$ lies in bin i

$\boldsymbol{p_j}$ is the prob. that $a_j(t+\boldsymbol{\tau_{ij}})$ lies in bin j

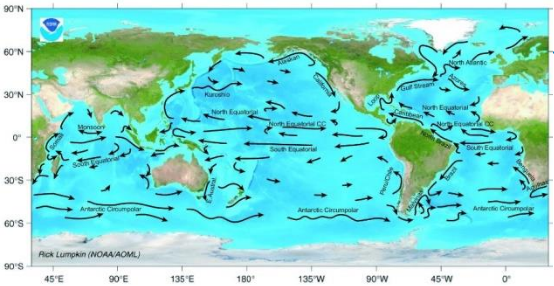
Statistical Similarity Measure (SSM): CC or MI

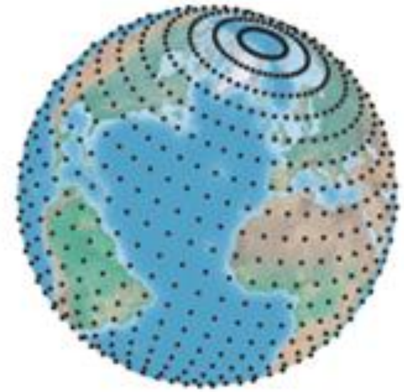
$$SSM_{ij} = \max_{\tau_{ij}} SSM_{ij}(\tau_{ij}) \qquad \boxed{\tau_{max} = T/5}$$

If **SSM $_{ij}$ > TH** the link i $\longleftrightarrow$ j exists, otherwise, it does not exist.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Complex network representation of the climate system

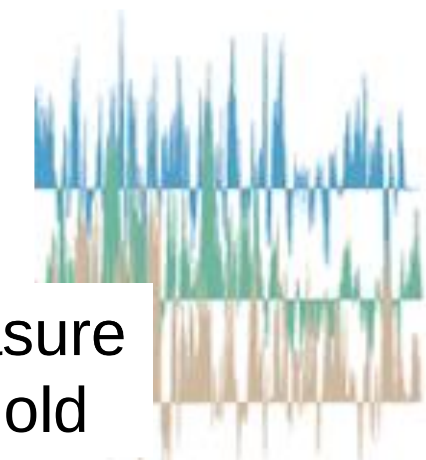




More than 10000 nodes.

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

Sim. measure + threshold



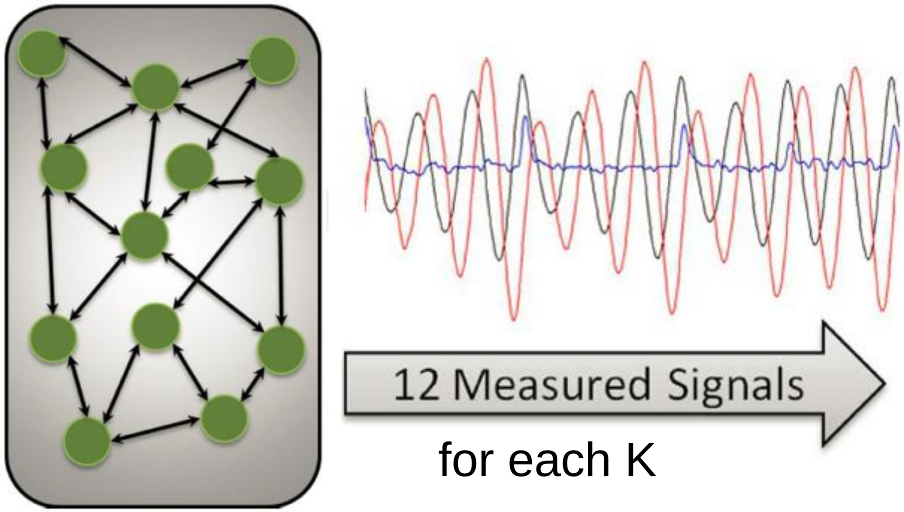
Daily resolution: more than 13000 data points in each TS

Donges et al, Chaos 2015

Surface Air Temperature
<u>Anomalies</u> (solar cycle removed)

Can we test which Statistical Similarity Measure is optimal for inferring the network connectivity?

- **12** chaotic Rossler oscillators with **known** random connectivity (**19** undirected links).
- The x variable is recorded for different coupling strengths (K)



- From the observed signals, different time series can be derived.

$$\{ x_i \} \Rightarrow \{ \varphi_i \} = HT[x_i] \Rightarrow \{ f_i \} = d\varphi_i/dt$$

Hilbert transform

Which one is the “best”?

How to quantify the success of network inference?

- **True positive rate**: number of correctly detected links / number of existing links;

- **False negative rate**: number of links which are incorrectly classified as not existing / number of existing links;

- **True negative rate**: number of correctly identified non-existing links / number of non-existing links;

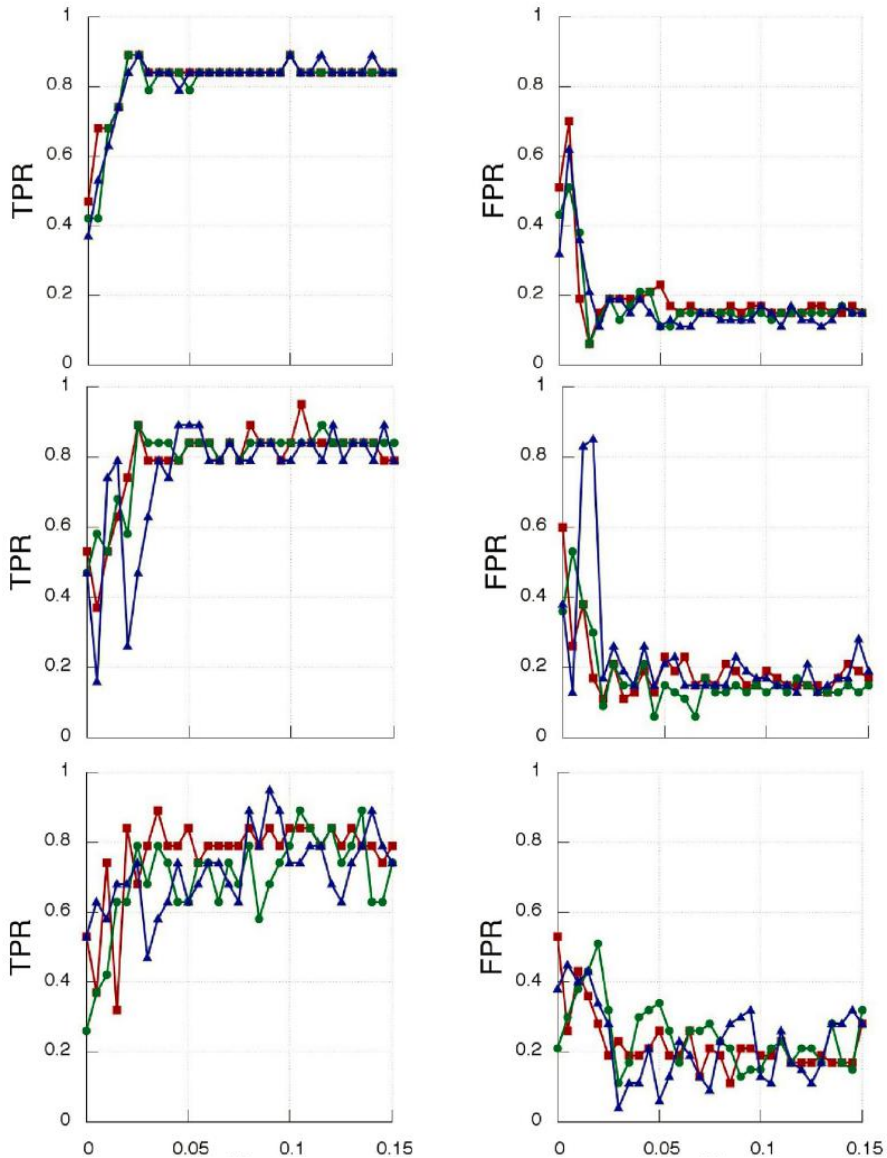
- **False positive rate**: number of non-existing links which are incorrectly classified as existing / number of non-existing links.

Results

Observed variable (x)

Hilbert phase

Hilbert frequency



Coupling strength, K

Detection threshold TH chosen to minimize errors.

Statistical similarity measures: CC MI MI(symbols)

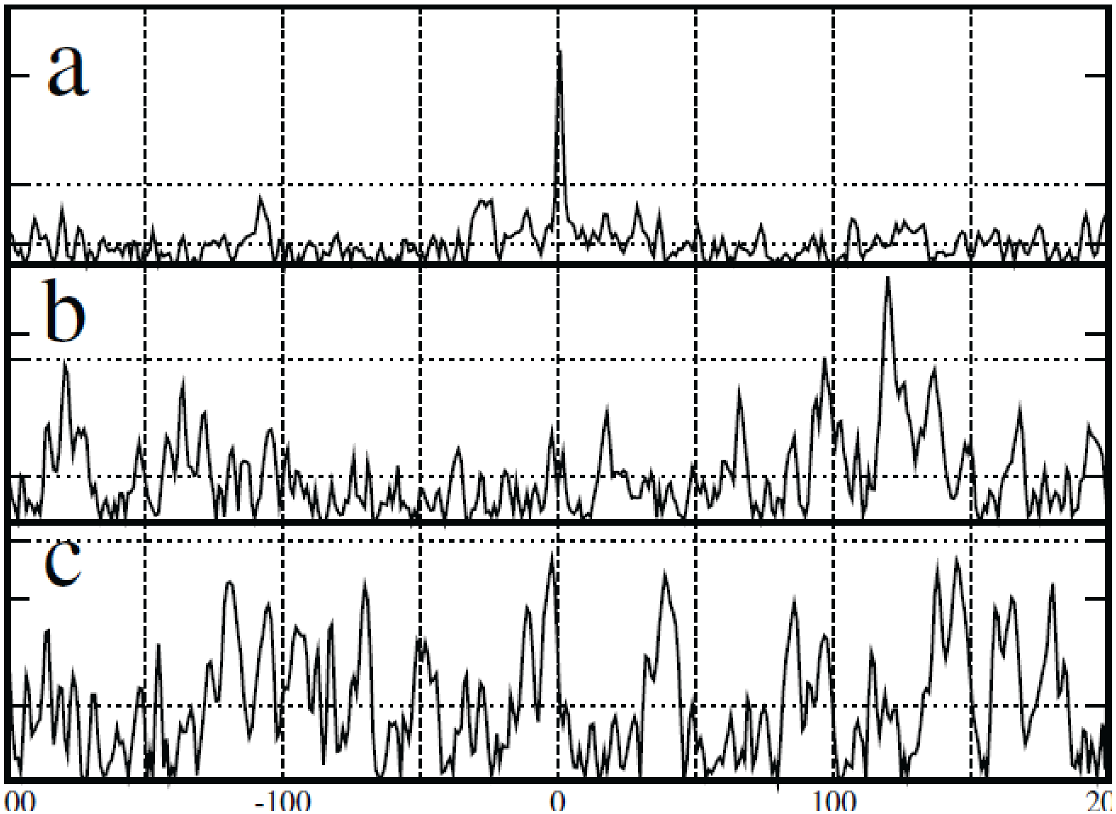
- No perfect reconstruction
- No difference between the SSMs & variables

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

Can we use the lag information to improve the inference?

$$SSM_{ij} = \max_{\tau_{ij}} SSM_{ij}(\tau_{ij})$$

An example from observed climatic data (temperature anomalies)



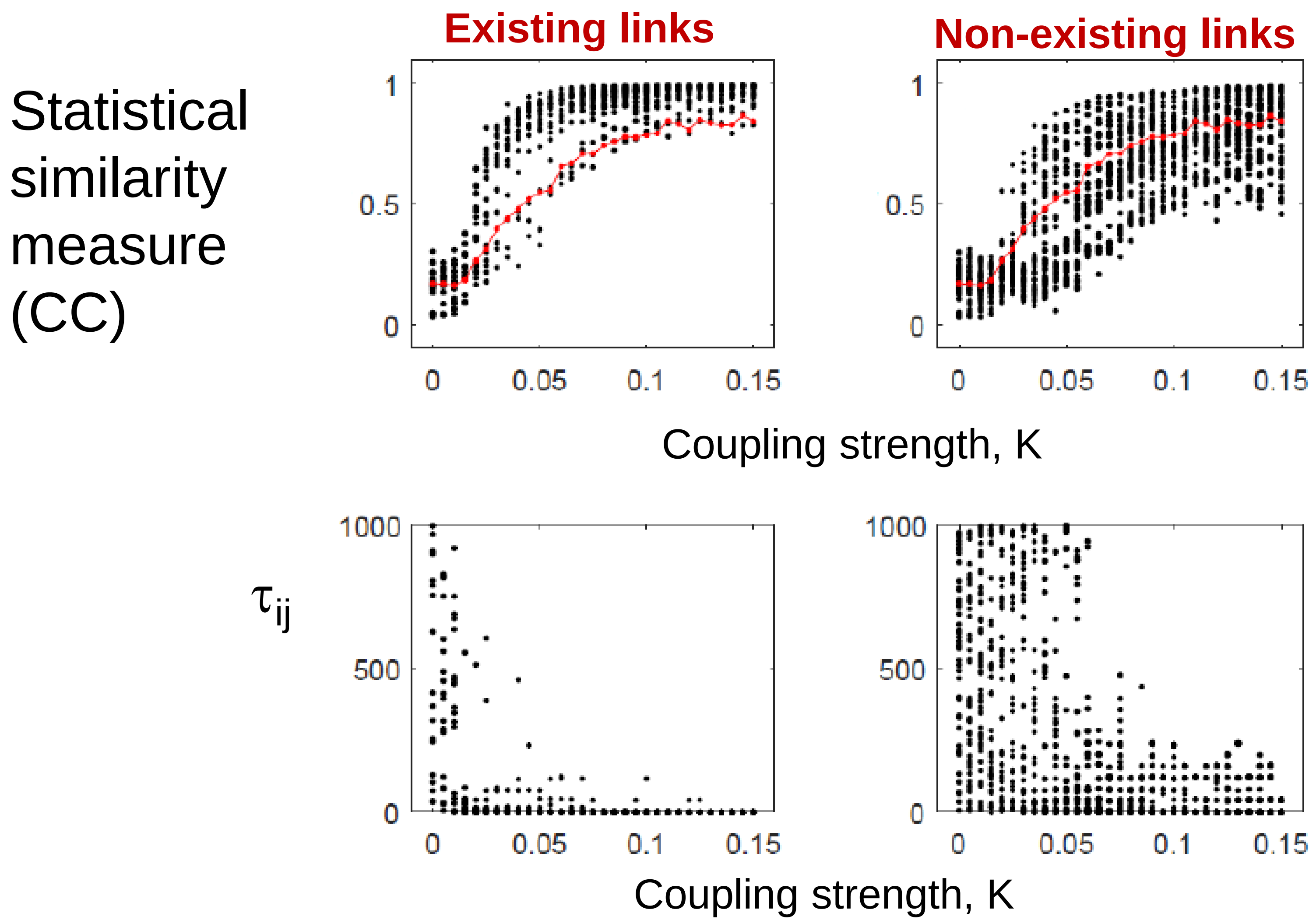
A strongly correlated link, with small time delay

An intermediately correlated link, with a few significant time delays.

A weakly correlated link, where the local maxima cannot be distinguished from noise.

Gozolchiani et al, EPL, 83 (2008) 28005

Lag analysis of the chaotic Rossler oscillators



We use two thresholds and compare three criteria to infer the network

SIM: Only the similarity measure (CC) is used to infer the links. The link between i and j exists ($A^*_{ij} = 1$) if $S_{ij} > S_{th}$, else, the link does not exist ($A^*_{ij} = 0$).

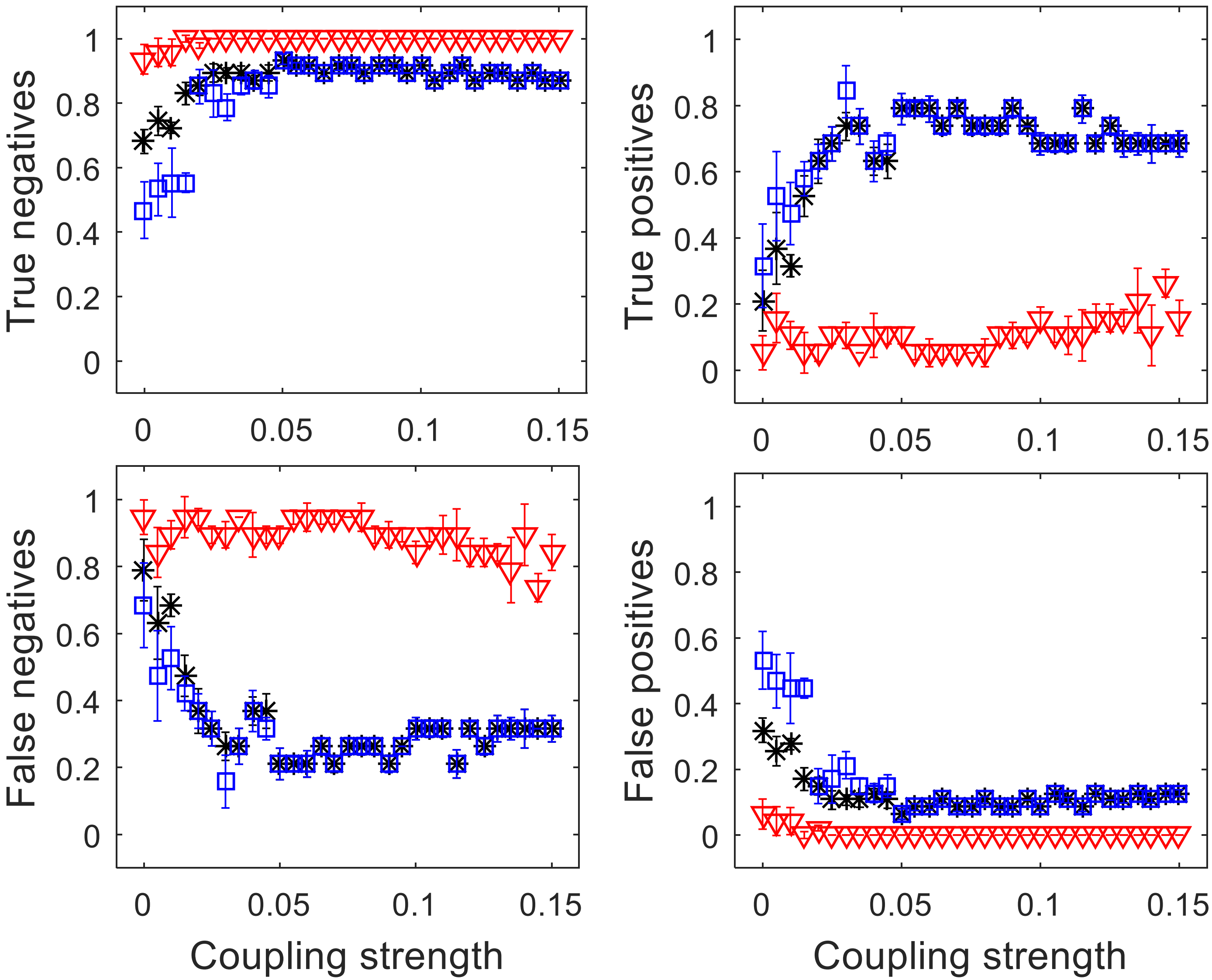
AND: The link between i and j exists ($A^*_{ij} = 1$) if $\tau_{ij} < \tau_{th}$ and $S_{ij} > S_{th}$, else, the link does not exist ($A^*_{ij} = 0$).

OR: The link between i and j exists ($A^*_{ij} = 1$) if $\tau_{ij} < \tau_{th}$ or $S_{ij} > S_{th}$, else, the link does not exist ($A^*_{ij} = 0$).

The detection thresholds S_{th} and τ_{th} are chosen to return a given number of links (we assume that we know the number of nodes and the number of links).

Results

SIM
AND
OR

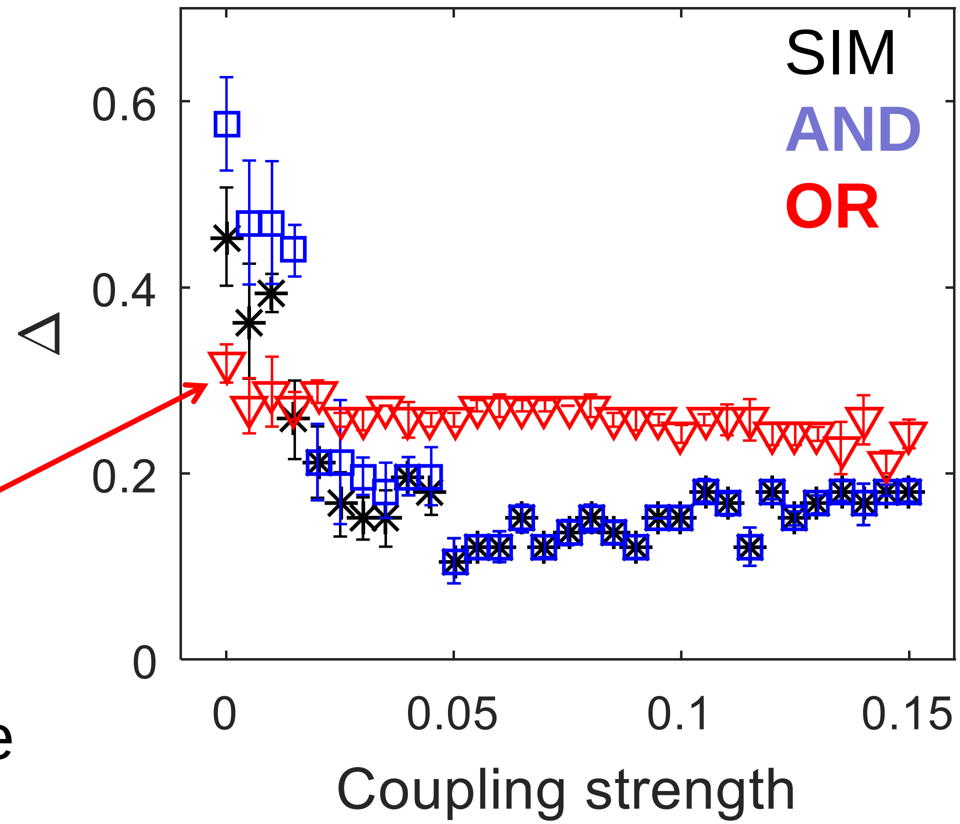


For large coupling SIM & AND give similar results.

OR minimizes the false positives but fails to detect existing links.

Total errors (% of wrongly inferred links)

$$\Delta = \frac{FN + FP}{N(N-1)/2}$$



For weak coupling, the OR criteria reduces the number of errors

What did we learn?

- For strong coupling, lag information can be used to minimize the false positives but is detrimental to detect existing links.
- For weak coupling, lag information reduces the total number of mistakes.

Future work: are these results robust to other types of oscillatory coupled systems?

Thank you for your attention

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

G. Tirabassi et al, Sci. Rep. 5 10829 (2015)
N. Rubido and C. Masoller, arXiv:1807.09636 (2018)

ICREA
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN