Characterizing social information spreading by using event-synchronization and causality measures

Lucio Garcia[1]
Pablo Balenzuela[1]

Giulio Tirabassi[2]
Cristina Masoller[2]

(1) Universidad de Buenos Aires

(2) Universitat Politecnica de Catalunya

2nd Meeting of the Spanish Chapter of CS3
February 22, 2024

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Campus d'Excel·lència Internacional

ICREA

AGAUR

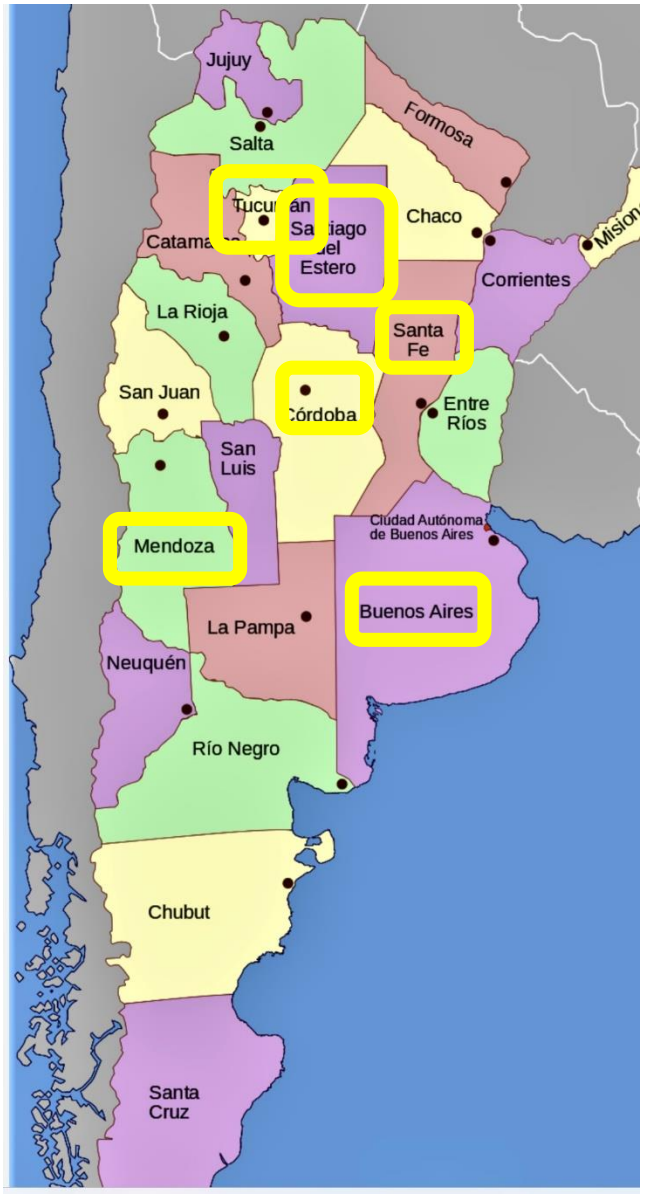
GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

Diffusion of information in Argentina

- Six observed cities:
  Buenos Aires
  Cordoba
  Tucuman
  Mendoza
  Santa Fe
  Santiago del Estero

- “Output” signals in each city: # of press articles per day in different topics.

- Goal: characterize the information diffusion process using time-series and network analysis.



cristina.masoller@upc.edu @cristinamasoll1

Data analyzed

- 28000 news articles published in Argentina in the selected six main cities.
- 4 months (26/05/2022 - 26/09/2022).
- An unsupervised matrix factorization algorithm was used to classify the articles in 20 non-orthogonal topics.

Presentan un proyecto de ley para "garantizar el abastecimiento de combustible líquido"

Política *30 de mayo de 2022* *Por LPTV*

El proyecto impulsado por el legislador Berarducci y acompañado por el bloque del PJS apunta a solucionar una de las crisis que aqueja al sector productivo.

Sergio Berni se cruzó con los transportistas en la autopista La Plata-Buenos Aires y lanzó una advertencia: "Tienen 5 minutos o me llevo los camiones"

"¿Vas a seguir haciendo show?", le preguntó el ministro de Seguridad bonaerense al líder de la protesta, antes de levantar el corte con la Policía; se quejaban por la falta de gasoil

Reunión en Tucumán

Gobernadores del norte reclamaron por la falta de gasoil

Fue durante la 10ª reunión del Consejo de la región Norte Grande. Por parte del Gobierno nacional, Juan Manzur amenazó que caerán "con todo el peso de la ley" sobre quienes alteren los valores del combustible.

S. Pinto, F. Albanese, C. O. Dorso, and P. Balenzuela. *Quantifying time-dependent media agenda and public opinion by topic modeling*. Physica A, 524:614, 2019

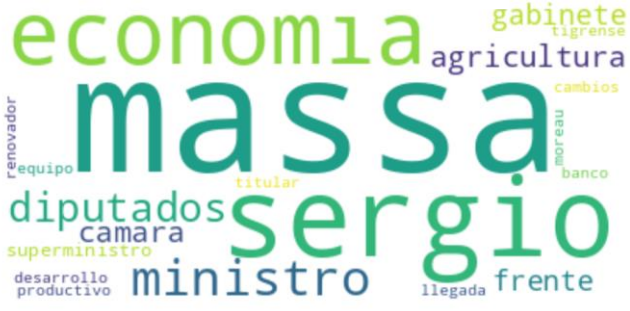
cristina.masoller@upc.edu @cristinamasoll1

Examples of topics

Alberto Fernandez

Avion Irani

Sergio Massa



Sabag Montiel

Diego Luciani

Silvina Batakis

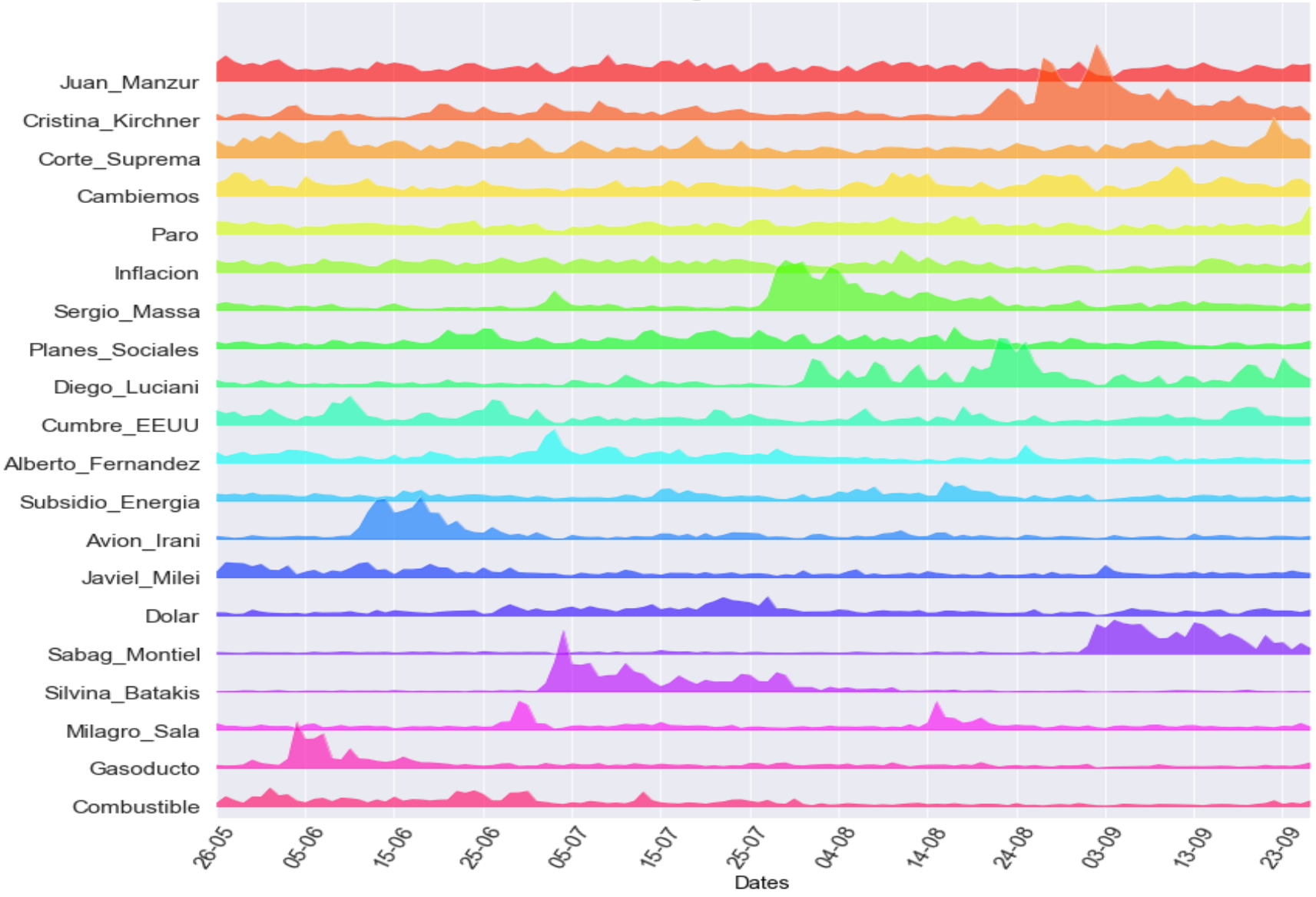
Cristina Kirchner

Inflacion

Dolar

cristina.masoller@upc.edu @cristinamasoll1

By adding the number of articles per day in the 6 cities, we obtain a time series for each topic.

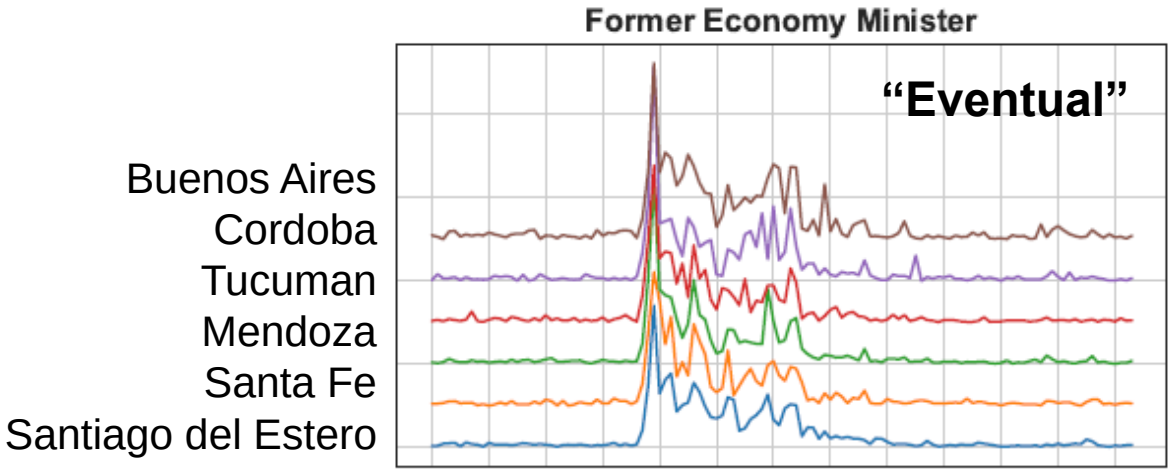


cristina.masoller@upc.edu @cristinamasoll1

We also have a time series for each topic in each city, with daily resolution.

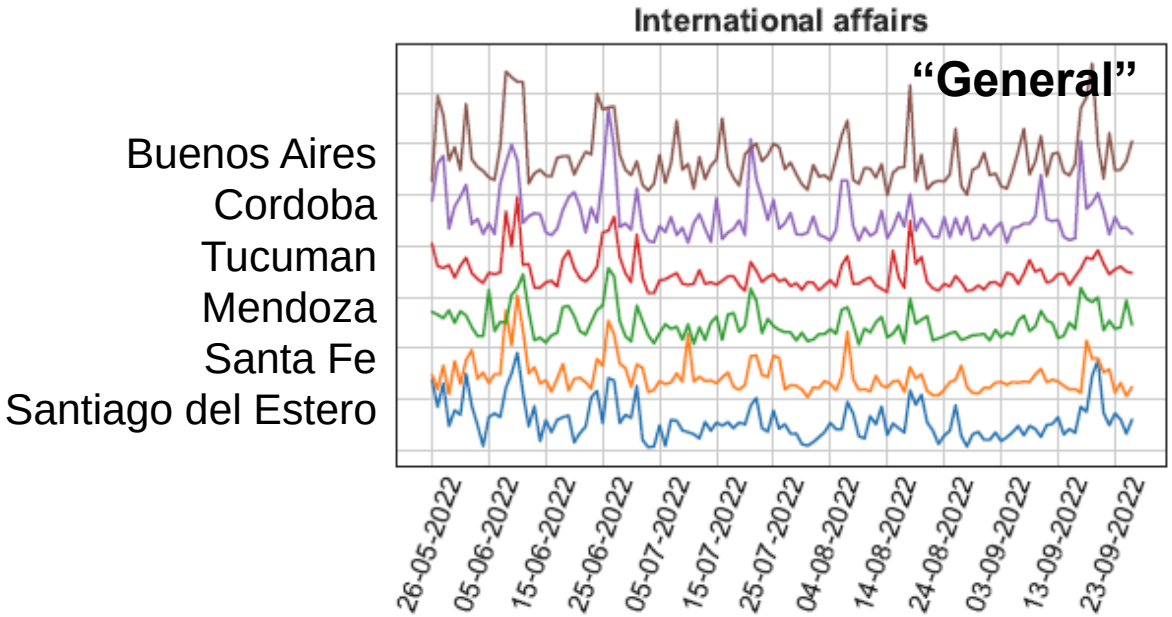
$x_{n,i}(t)$

n=topic
i = city
t = time



Normalized such that in each city, each day, the total "attention" is 1:

$$\sum_n x_{n,i}(t) = 1$$

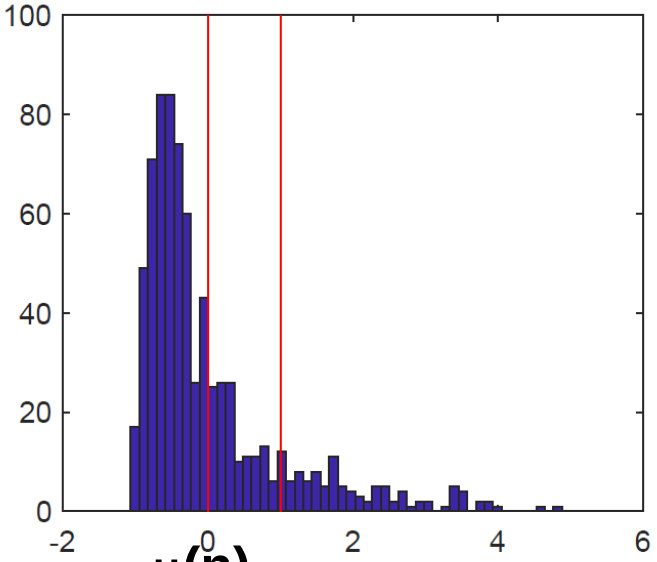


cristina.masoller@upc.edu @cristinamasoll1

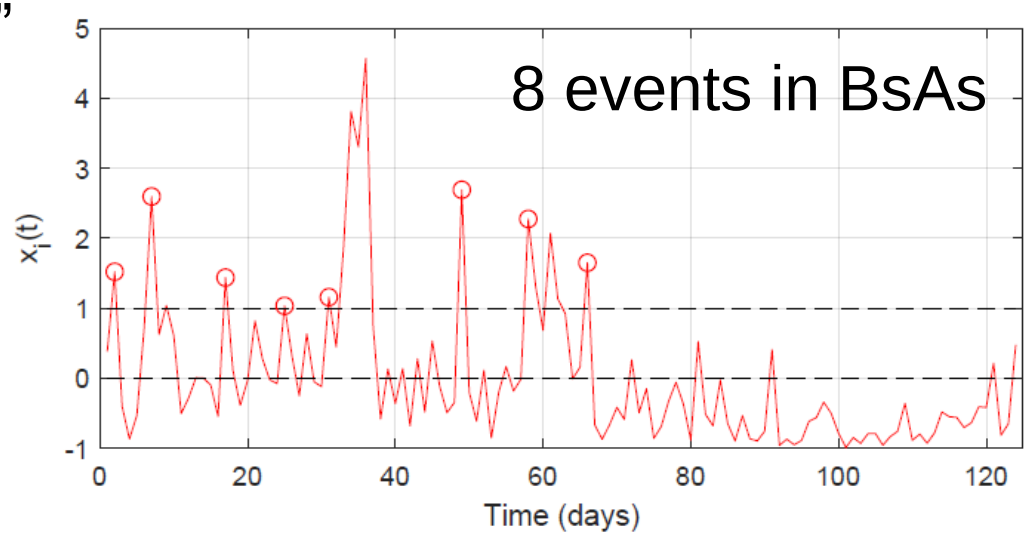
In each time series we define “events” using two thresholds

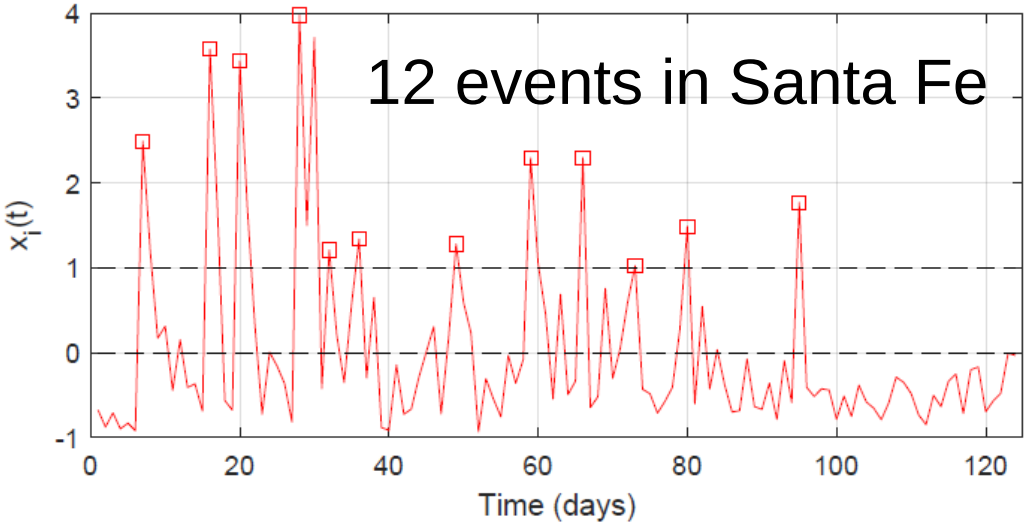
Example: topic “*Combustible*”

Histogram of values
(all cities, all times)



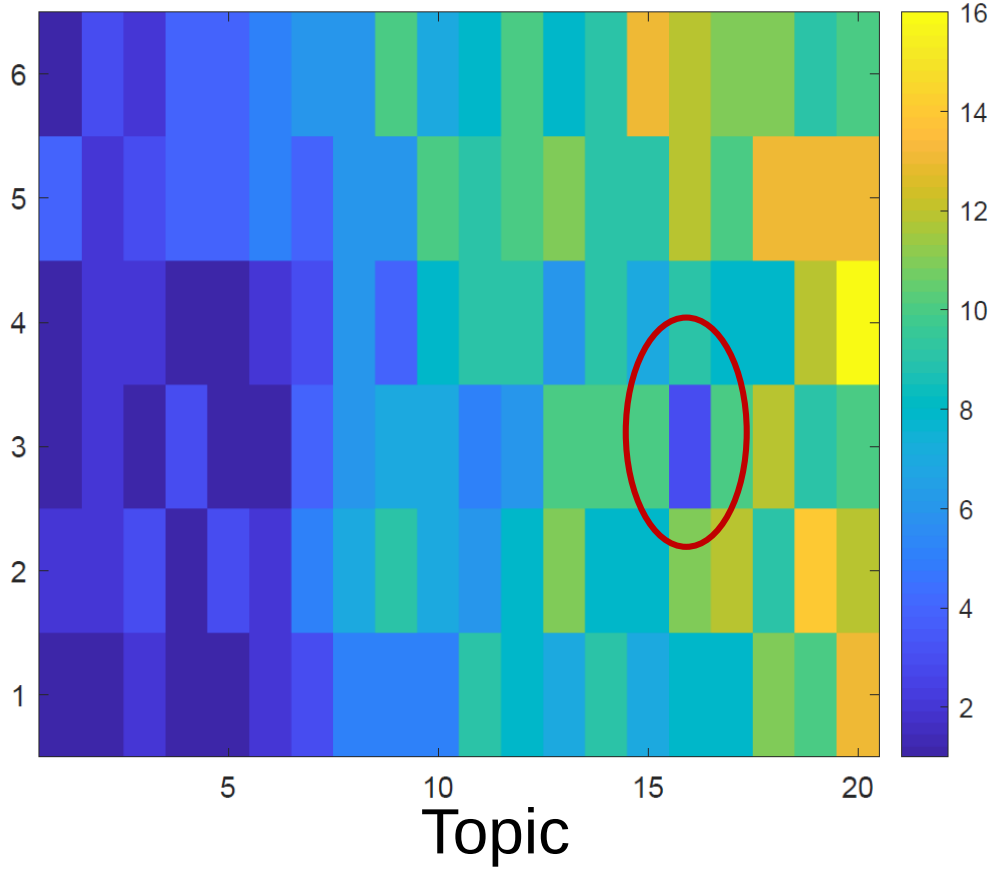
Each time series is normalized to $\mu=0$, $\sigma=1$





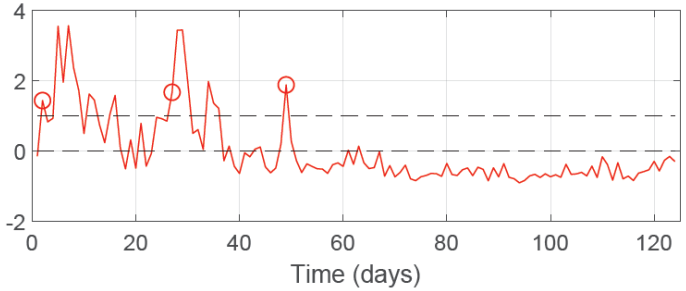
cristina.masoller@upc.edu @cristinamasoll1

Color code: Number of events

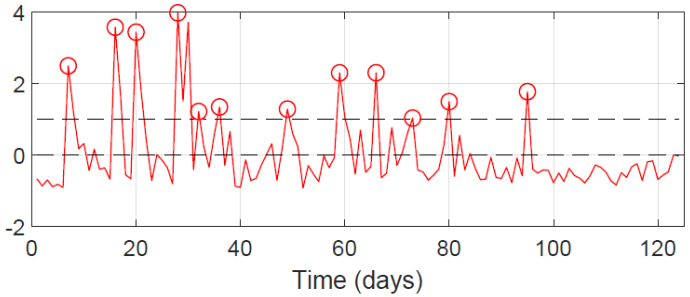


Combustible

3 events in Tucuman

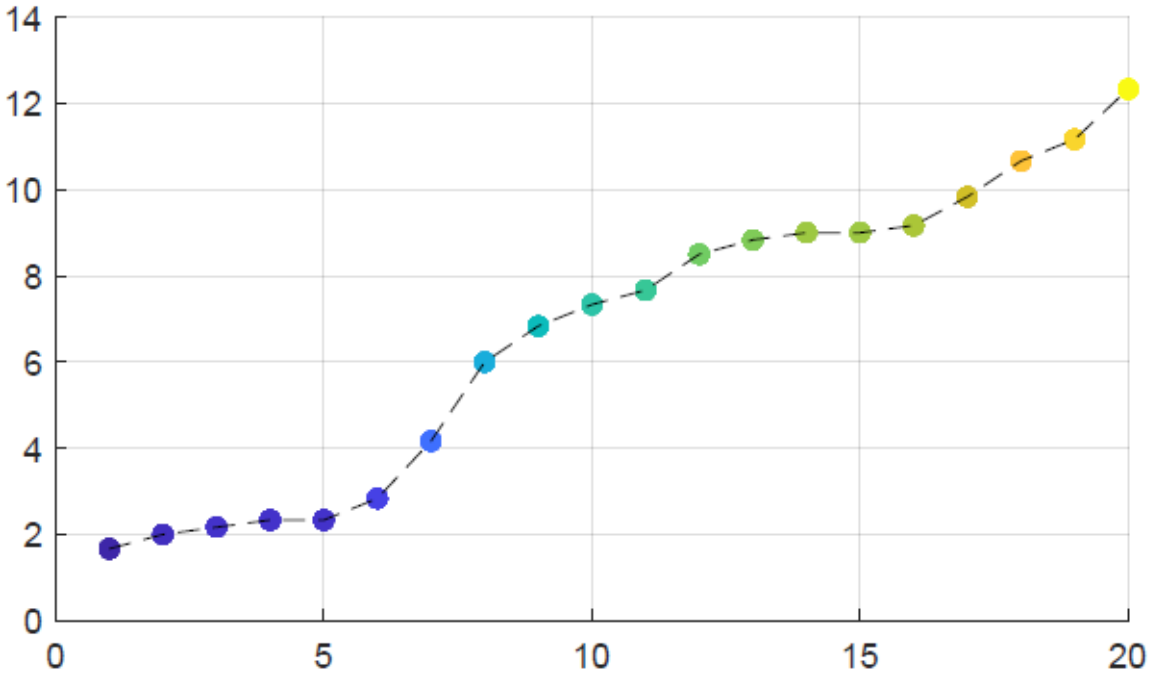


11 events in Santa Fe



“Eventual” **“General”**

cristina.masoller@upc.edu @cristinamasoll1



Cristina_Kirchner
Silvina_Batakis
Sabag_Montiel
Sergio_Massa

“Eventual”

“General”

Inflacion
Juan_Manzur
Paro
Cambiemos

- How to find “synchronized events” in different cities?
- How to find events in cities that cause events in other cities?
  (20 topics, 6 cities: 20 x 6 x 5 = 600 possible links)

cristina.masoller@upc.edu @cristinamasoll1

Event synchronization measures

- Count m_x, m_y : number of events in *x(t)* and *y(t)*; $m=[m_x+m_y]/2$
- Count $c^\tau(x|y)$ = number of times an event occurs in *x(t)* shortly after (within an interval of τ days) an event occurs in *y(t).* Idem for $c^\tau(y|x)$. Synchronized events count ½.
- Calculate: $Qs(x,y) = [c^\tau(x|y) + c^\tau(y|x)] / m$

  $Qa(x,y) = [c^\tau(x|y) - c^\tau(y|x)] / m$

$Q_s(x,y) = Q_s(y,x)$ $Q_a(x,y) = - Q_a(y,x)$

$Q_s = 1$: the events are fully synchronized.
$Q_a = 1$: events in *x* always occur before those in *y*.
$Q_a = -1$: events in *x* always occur after those in *y*.

No “causal” information

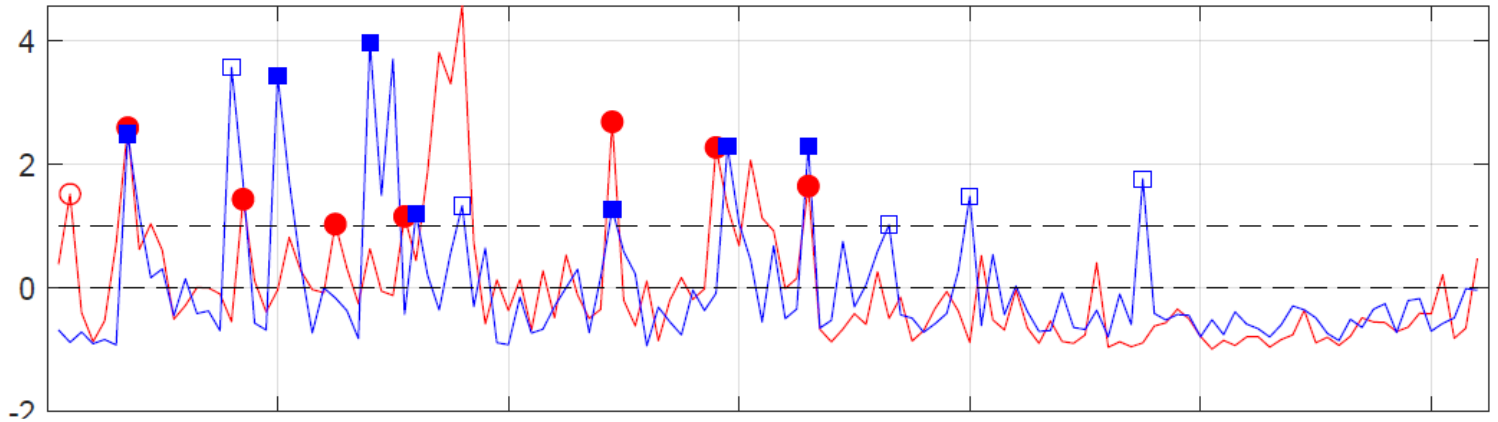
R. Quian Quiroga, T. Kreuz and P. Grassberger, PRE 66, 041904 (2002).

cristina.masoller@upc.edu @cristinamasoll1

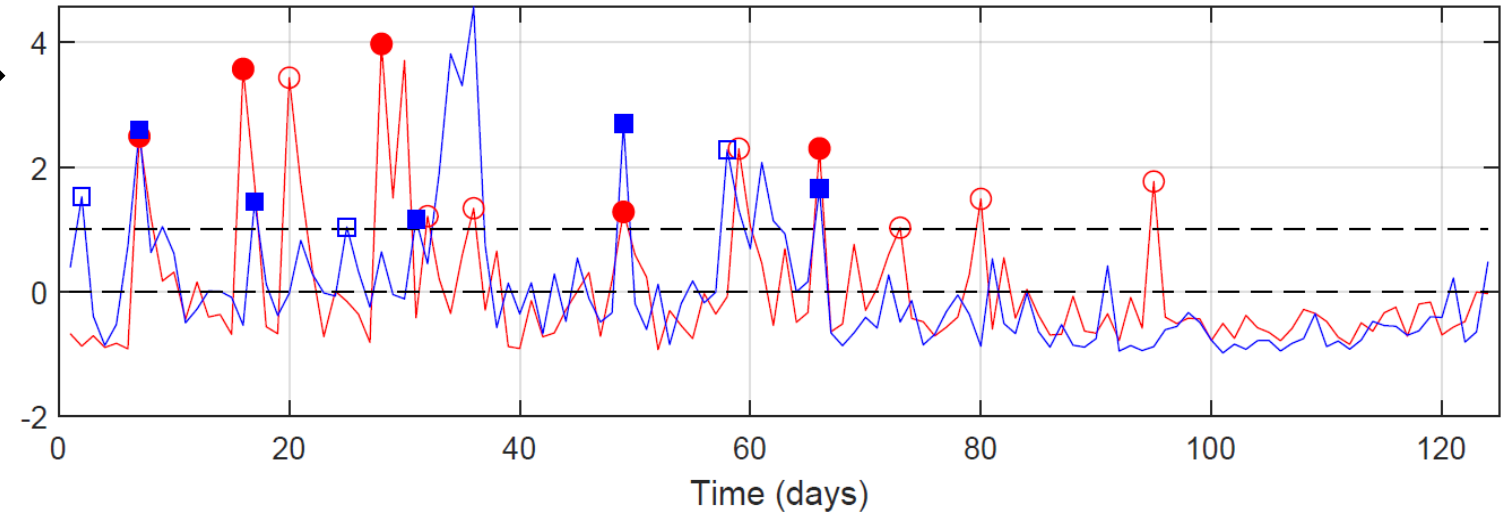
Example: topic "*Combustible*"

τ=3 days

BsAs →
Santa Fe
C(i,j)=5.5



Santa Fe →
BsAs
C(j,i)=3.5



m=[8+12]/2=10 Q_s = 9/10 = 0.9

Q_a = 2/10 = 0.2

cristina.masoller@upc.edu @cristinamasoll1

Other sync. measures: cross-correlation and mutual information, calculated from the raw data, with zero lag.

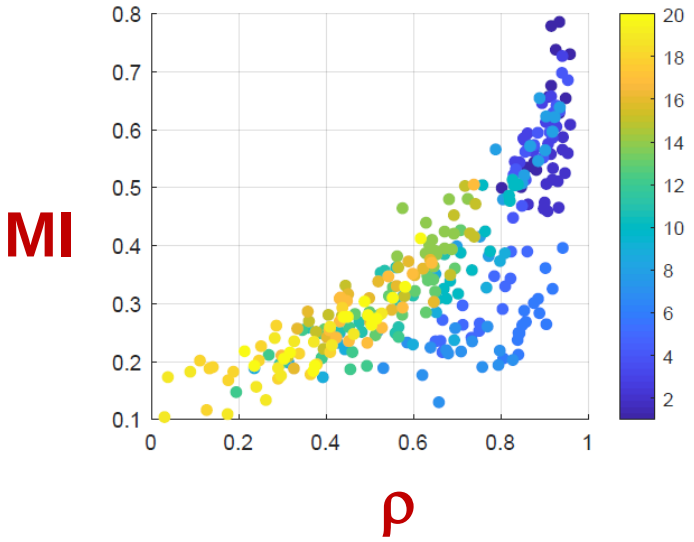
CC: $\rho=\langle|x(t)y(t)|\rangle$
(x and y zero mean and $\sigma=1$)

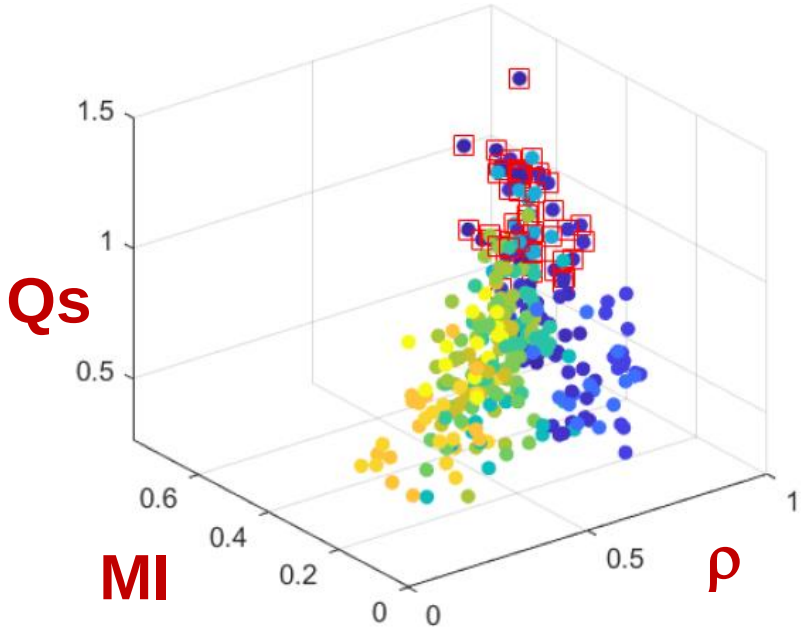
$$MI = \sum_{i\in x}\sum_{j\in y} p(x,y)\ \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

CC(i,j)=CC(j,i)

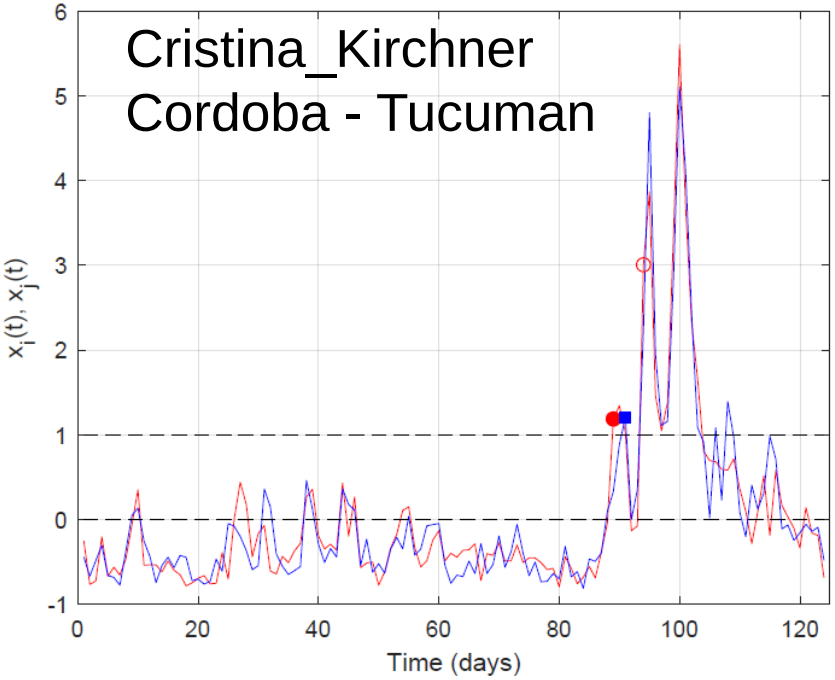
MI(i,j)=MI(j,i)

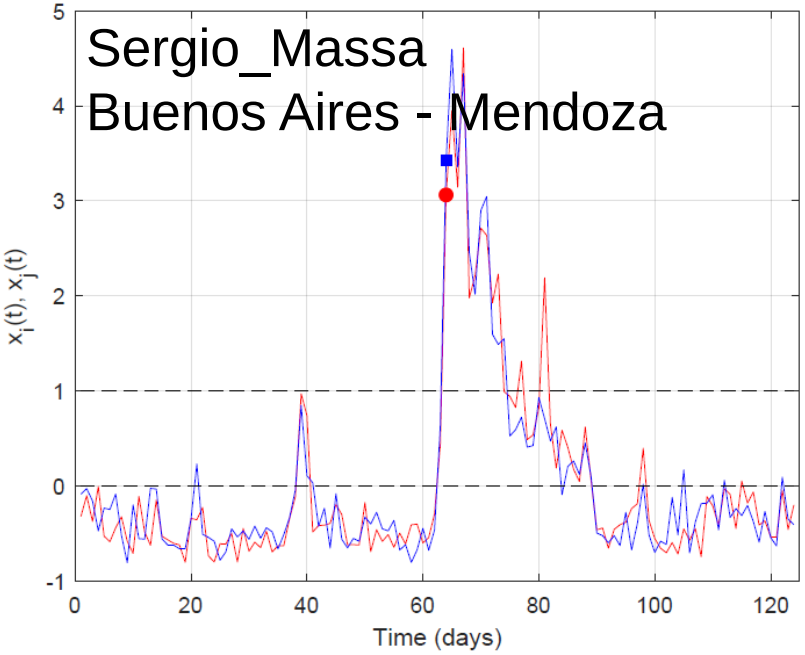
If *x* and *y* are Gaussian processes: MI = -1/2 log(1-ρ^2)

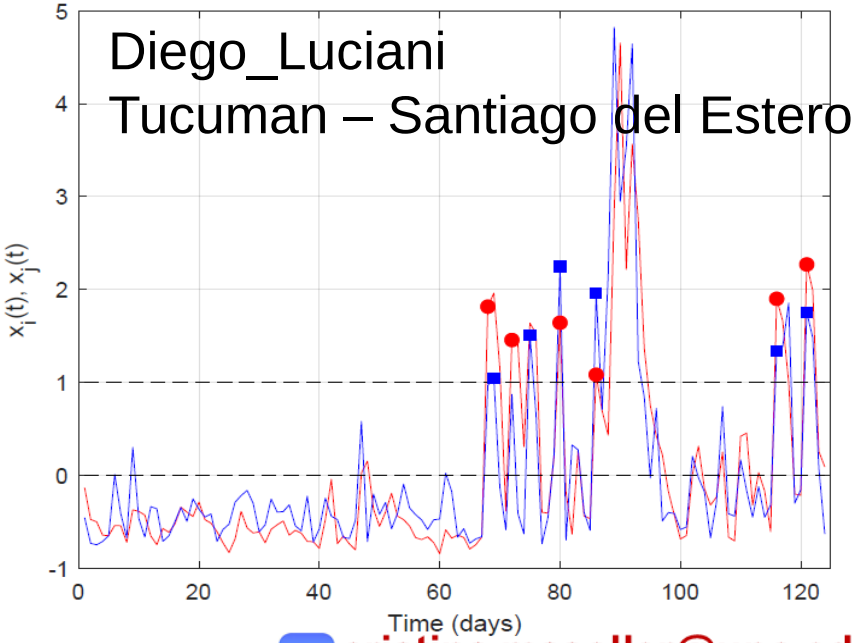


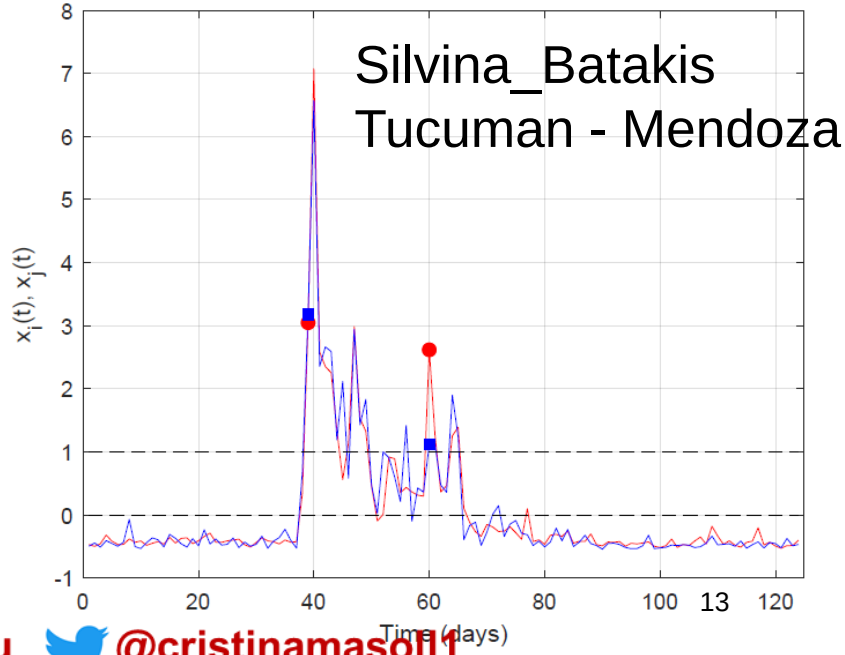


cristina.masoller@upc.edu @cristinamasoll1









cristina.masoller@upc.edu @cristinamasoll1

Causal measures: Granger Causality, Transfer entropy and Directionality index

- GC: *X* and *Y* can be described by stationary autoregressive linear models.

- TE: Conditional Mutual Information of *X* and *Y*, given the "past" of one of the variables.

- TE and GC are equivalent for Gaussian processes.

- DI = [X(i,j) - X(j,i)] / [X(i,j) + X(j,i)]
  where X=TE or GC, only values larger than a threshold.

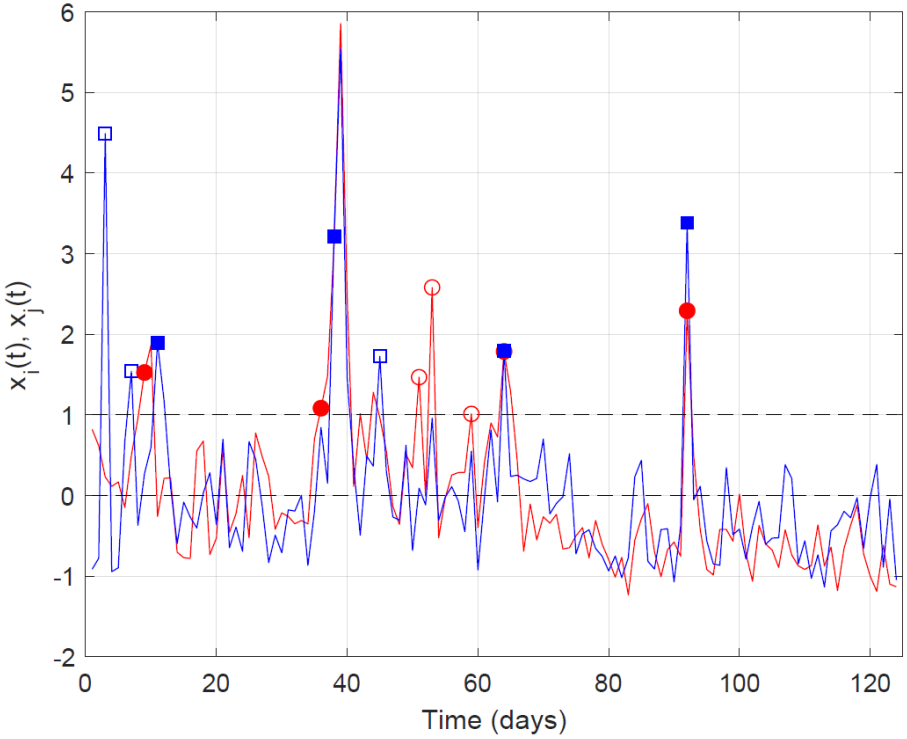
C. W. J. Granger *Investigating causal relations by econometric models and cross-spectral methods*. *Econometrica* 37, 424–438 (1969).

T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85, 461 (2000).

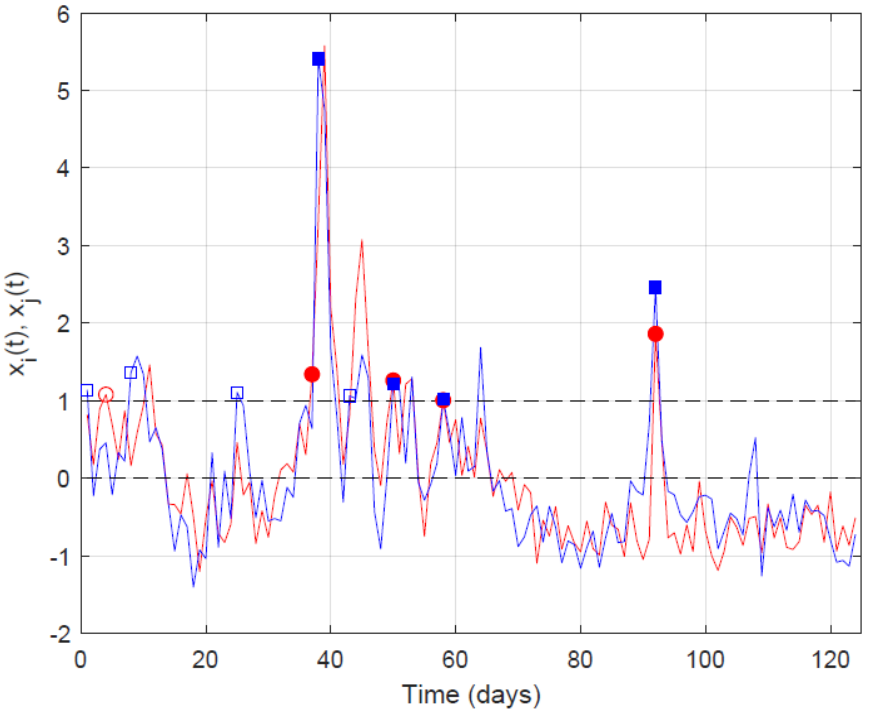
cristina.masoller@upc.edu @cristinamasoll1

Examples with high DI, topic: “Alberto Fernandez”

Cordoba – Santiago del Estero
DI(TE)=1, Qs=0.71, Qa=0.14



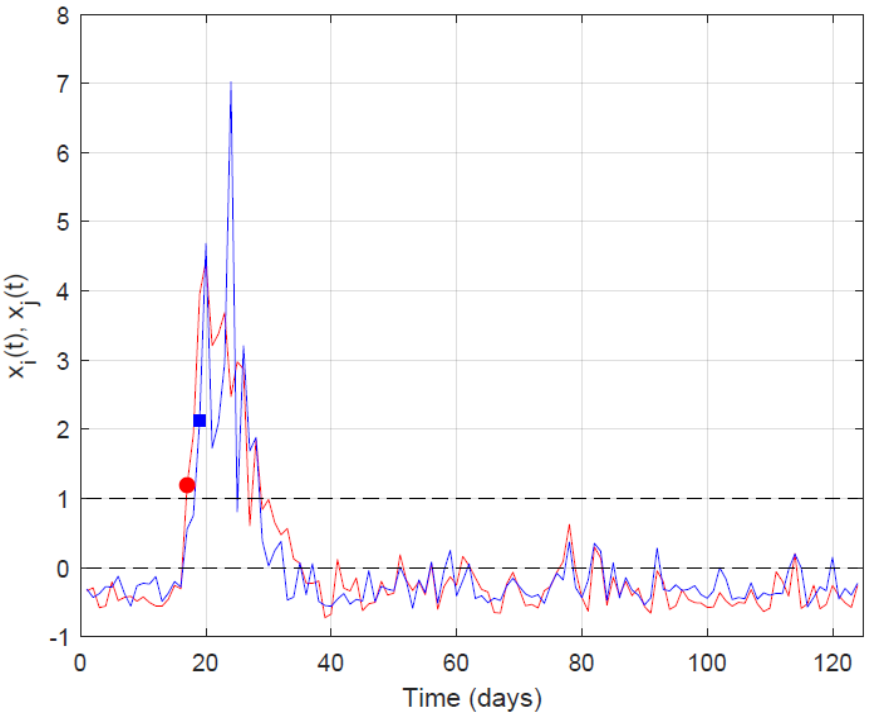
Buenos Aires – Mendoza
DI(GC)=1, Qs=0.77, Qa=0



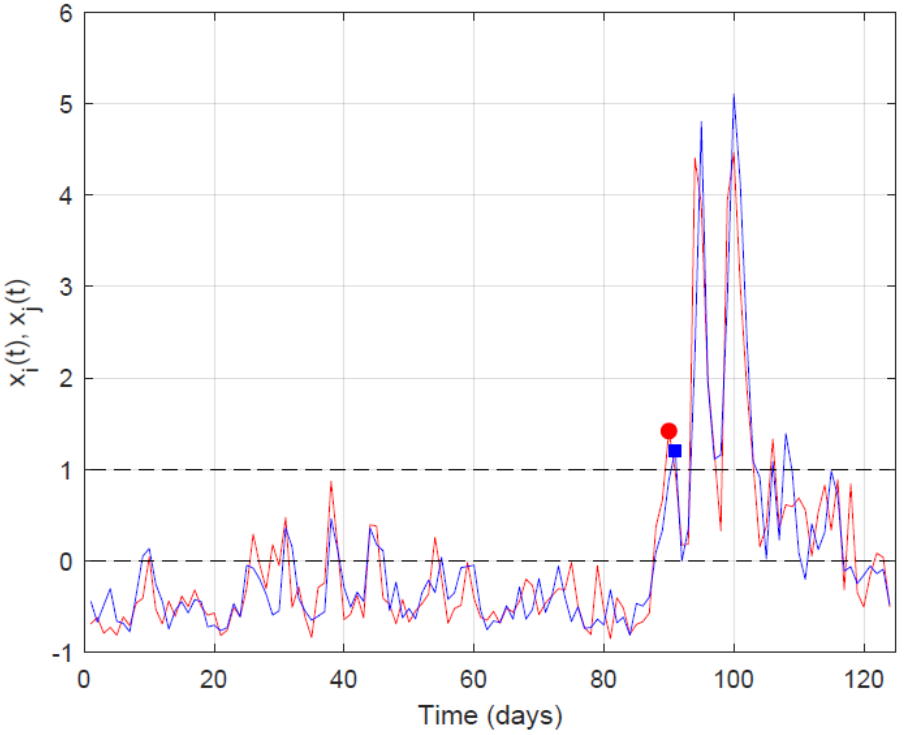
cristina.masoller@upc.edu @cristinamasoll1

Examples with high Qa

"*Avion Irani*"
Buenos Aires – Tucuman
DI(TE)=0, DI(GC)=1, Qs=1, Qa=1

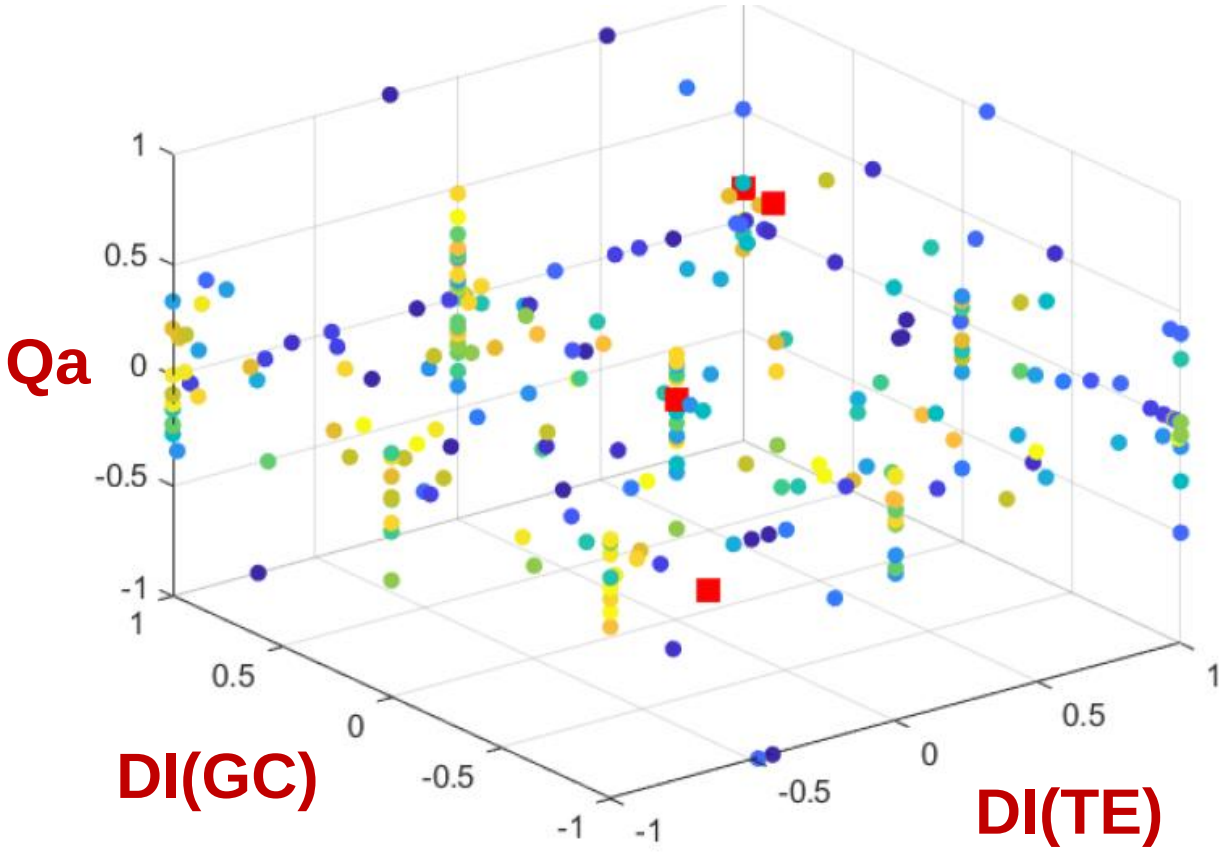


"*Cristina Kirchner*"
Santiago del Estero – Tucuman
DI(TE)=0, DI(GC)=-0.37, Qs=1, Qa=1



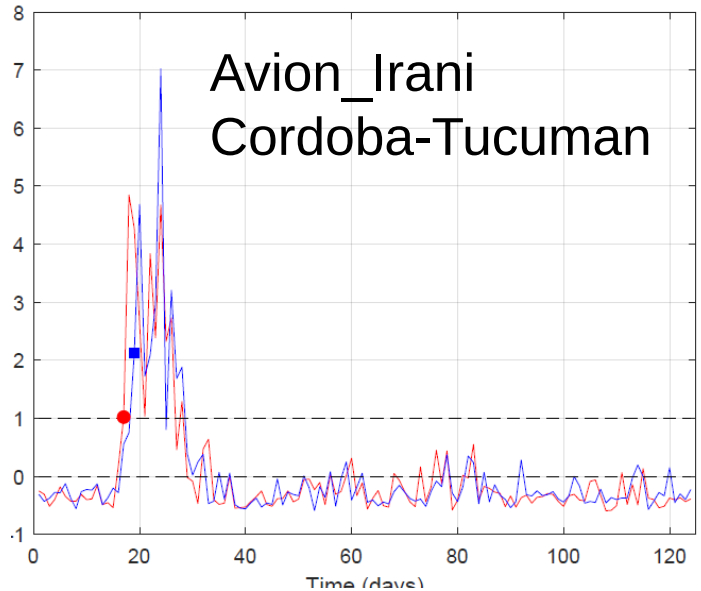
cristina.masoller@upc.edu @cristinamasoll1

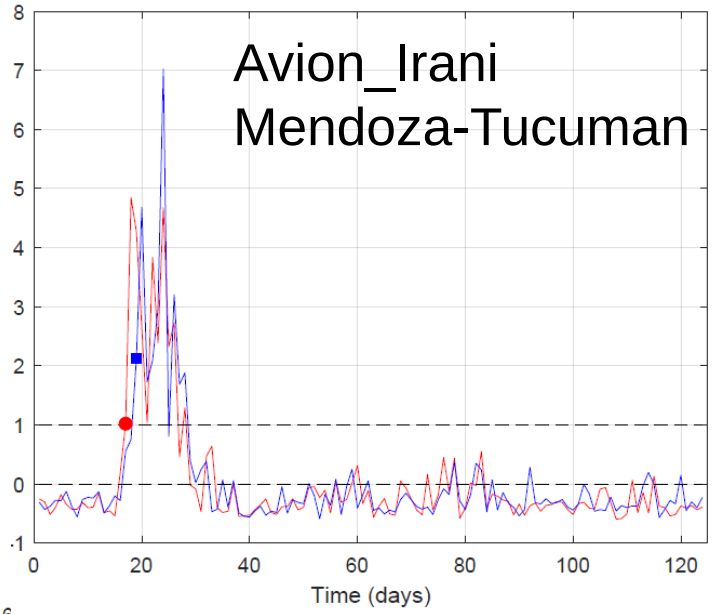
Qa, DI(GC) and DI(TE) in general do not have the same sign

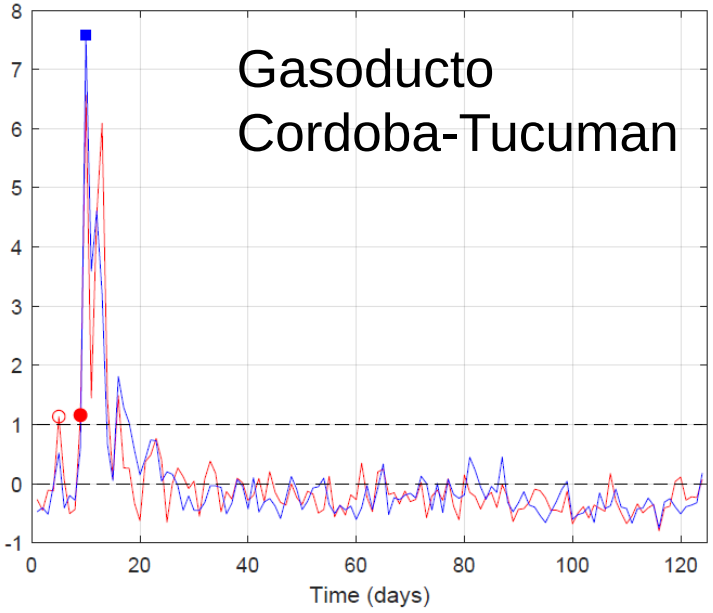


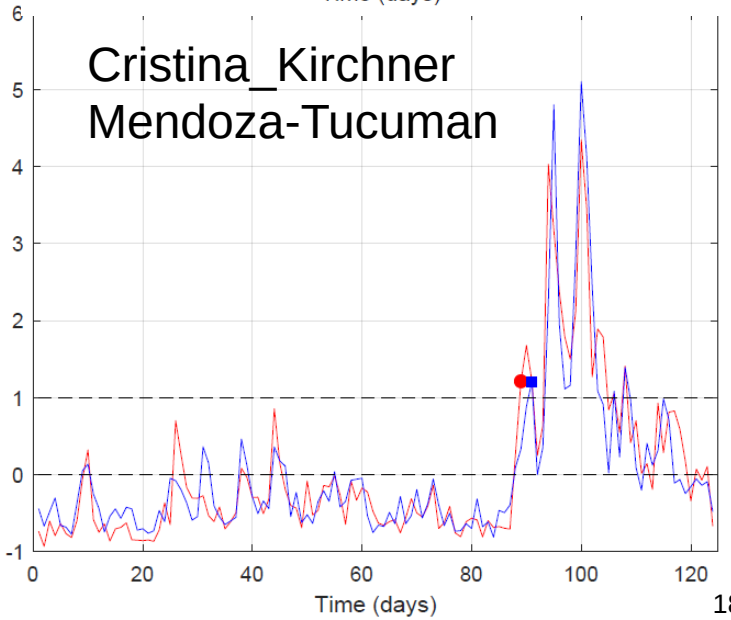
We select “links” for which Qa, DI(TE) and DI(GC) are not zero and have the same sign.

cristina.masoller@upc.edu @cristinamasoll1





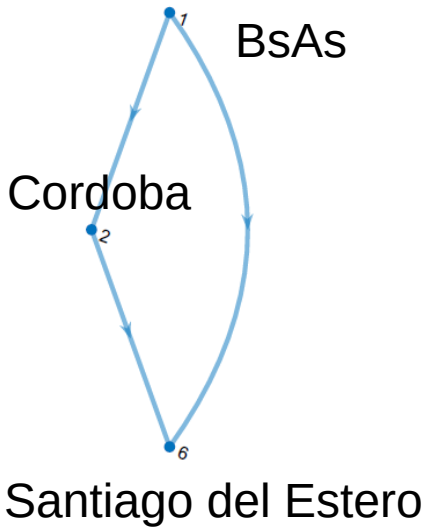




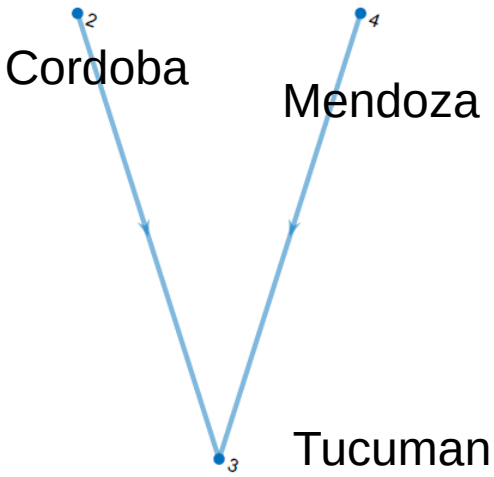
cristina.masoller@upc.edu @cristinamasoll1

Causal networks for some topics and for all the topics

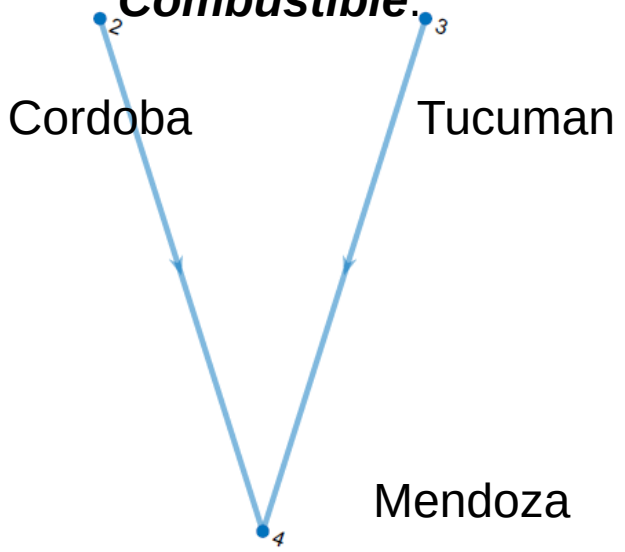
Alberto_Fernandez



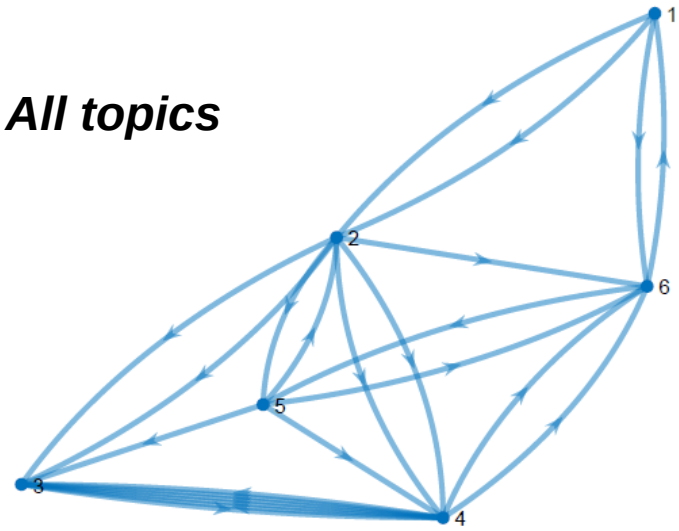
Avion_Irani



Combustible.



All topics



cristina.masoller@upc.edu @cristinamasoll1

Conclusions

- The processes of information diffusion in the press is fast, most events are simultaneous $\Rightarrow$ difficult to infer links.

- Raw-data-based measures (cross-correlation, mutual information, Granger causality and transfer entropy) not appropriated to see correlations between events (‘big news’).

- Event-based measures (Q_s, Q_a) work for “eventual” topics.

ICREA
cristina.masoller@upc.edu
@cristinamasoll1
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN