

Point process analysis of geographical diffusion of news in Argentina

L. L. García,^{1, 2, a)} G. Tirabassi,^{3, 4, b)} C. Masoller,^{3, c)} and P. Balenzuela^{1, 2, d)}

¹⁾*Departamento de Física, Universidad de Buenos Aires,
Facultad de Ciencias Exactas y Naturales, 1428, Buenos Aires,
Argentina*

²⁾*Instituto de Física Interdisciplinaria y Aplicada (INFINA),
CONICET - Universidad de Buenos Aires, Buenos Aires,
Argentina*

³⁾*Departament de Física, Universitat Politècnica de Catalunya,
Rambla St. Nebridi 22, Terrassa, 08222, Barcelona, Spain*

⁴⁾*Departament de Informàtica, Matemàtica Aplicada i Estadística,
Universitat de Girona, Carrer de la Universitat de Girona 6, Girona, 17003,
Spain*

(Dated: 3 December 2024)

The diffusion of information plays a crucial role in a society, affecting its economy and the well-being of the population. Characterizing the diffusion process is challenging because it is highly non-stationary and varies with the media type. To understand the spreading of newspaper news in Argentina, we collected data from more than 27000 articles published in six main provinces during four months. We classified the articles into 20 thematic axes and obtained a set of time series that capture daily newspaper attention on different topics in different provinces. To analyze the data we use a point process approach. For each topic, n , and for all pairs of provinces, i and j , we use two measures to quantify the synchronicity of the events, $Q_s(i, j)$, which quantifies the number of events that occur almost simultaneously in i and j , and $Q_a(i, j)$, which quantifies the direction of news spreading. Our analysis unveils how fast the information diffusion process is, showing pairs of provinces with very similar and almost simultaneous temporal variations of media attention. On the other hand, we also calculate other measures computed from the raw time series, such as Granger Causality and Transfer Entropy, which do not perform well in this context because they often return opposite directions of information transfer. We interpret this as due to the characteristics of the data, which is highly non-stationary, and of the information diffusion process, which is very fast and probably acts at a sub-resolution time scale.

^{a)}llgarcia@df.uba.ar

^{b)}giulio.tirabassi@upc.edu

^{c)}cristina.masoller@upc.edu

^{d)}balen@df.uba.ar

Throughout his distinguished scientific career, Professor Jason Gallas tackled a wide range of challenges related to complex systems and nonlinear dynamics, covering both fundamental and applied areas. His research encompassed diverse topics, such as economic models where he examined the simultaneous variation of multiple parameters in economic processes, social dynamic models focused on stress-related behaviors, and the analysis of science communication statistics, particularly citation patterns among physicists, to explore whether the prevalence of mathematical equations in a paper influences its citation count. In this contribution to the Focus Issue in Honor of Professor Gallas, we study news communication and public opinion dynamics in Argentina. We analyze a large number of newspaper articles published in six main provinces during four months. We find very similar and almost simultaneous temporal evolution of media attention in different provinces, which can be interpreted as due to fast information diffusion produced by social networks (Instagram, X, etc.) and the action of large media companies that act as global, external drivers of information diffusion.

I. INTRODUCTION

During his prolific scientific career, Professor Jason Gallas addressed a wide variety of problems involving complex systems and nonlinear dynamics, of fundamental and applied interest, including topics as diverse as economic models, focusing on the simultaneous variation of several parameters involved in economic processes¹, social dynamic models of people's stress-related processes², and science communication statistics, analyzing citation habits among physicists, to determine whether mathematics (i.e., the number of equations in a paper) can be blamed for the number of citations³. In this contribution to the Focus Issue in Honor of Professor Gallas, we study news communication and public opinion dynamics in Argentina.

The mass media play a preponderant role in the formation of public opinion. The systematic study of this dynamics is known as *agenda-setting theory* and began with the seminal work of McCombs and Shaw in 1972⁴. The agenda-setting theory compares the public agenda

with the issues covered by the media, i.e., the issues that the public considers a priority, and the relationship between public and media perceptions of political figures, celebrities, countries, conflicts, events, etc.^{5–7}.

The agenda-setting theory studies how the media shapes public opinion by giving particular attention to some topics. In the last few years, access to big data has enabled modeling and quantitative analyses, such as studies of correlation between media and public attention, information and opinion spreading^{6,8–15}. Recently, an unsupervised topic decomposition model on newspaper articles was developed to study the dynamics of mass media and public agendas¹⁶. In¹⁷ the topic decomposition model, combined with sentiment analysis and time series analysis methods, was used to analyze news articles during the 2016 US presidential campaign, to obtain insights into the relationship between media and public opinion, using voting intention as a proxy of public opinion.

The study of how different news media influence each other, in determining what issues to cover and how they present the news to the public, is now known as *intermedia agenda-setting*¹⁸. Studies on this line of research typically indicate that the coverage often shifts from elite news sources to other media platforms^{19–21}. Guo and Vargo²² analyzed how news media in different countries influence each other in covering international news. Mohammed and McCombs²³ studied the flow of international news and provided evidence that media such as *The New York Times* or *The Guardian* do not publish much news about peripheral countries, and when they do, it is mostly negative news. Guo and Zhang²⁴ analyzed the information flow in the United States, examining data from the 21 most populated US cities, and characterized how national news and local news coverage influence each other. Stern and collaborators²⁵ used the network approach to study how different news sources influence each other on particular topics. They found that the networks of influence depend significantly on the topic because some news sources may act as agenda-setters (i.e., central nodes) in certain topics, while they may act as followers (i.e., peripheral nodes) in others topics. Another study of information diffusion in real networks was performed by Cota and coworkers²⁶, who constructed a political communication network based on Twitter data and analyzed the diffusion of information among users using epidemiological models.

The flow of information between different regions or countries has also been studied using complex network representations. Examples, to name just a few, are the analysis of millions of geo-coded external hyperlinks, collected from Danish, Norwegian, and Swedish

news websites²⁷; the analysis of millions of news articles used to construct a network based on the sequential mention of events across various countries²⁸, and the study of geolocalized time series of tweets associated to five events (the release of a Hollywood movie, two political protests in Brazil and in Spain, the Higgs boson discovery, and the acquisition of Motorola by Google), to extract, using transfer entropy analysis, directed networks that capture the role played by different geographical units in the global exchange of information²⁹.

In this work, we characterize the diffusion of information in newspapers in Argentina by analyzing more than 27000 news articles published in six main provinces during four months. We selected a time range that included two events that attracted a lot of press attention: the arrival in Argentina of an Iranian Venezuelan plane on June 6, 2022, and the attempt to assassinate the Vice President of Argentina, on September 1, 2022. We are interested in characterizing how specific news diffuse and gain broad national attention. We extract the media agenda by decomposing the topics in 20 non-orthogonal topics, using the unsupervised topic decomposition model developed in¹⁶. Then, we use a “point process”, event-based approach³⁰ to detect the onset of coverage of different topics in different provinces. This approach is based on the hypothesis that a sudden increase in media coverage is triggered by a well-defined and highly informative event. We characterize the synchronicity of media attention events in pairs of provinces and the direction of news spreading.

We also analyze the interdependence of news coverage between different provinces by using well-known bivariate measures to detect linear and nonlinear correlations and dependencies.

The paper is organized as follows: Sec. II presents the data analyzed and the methods used, Sec. III presents the results obtained, and Sec. IV presents the discussion of the results and outlook.

II. DATA AND METHODS

A. Data

The analyzed database consists of 27,881 political news articles collected from newspapers published in the six Argentinian provinces. The study period spans from May 26, 2022, to September 26, 2022. For data collection, the two main media outlets in each region were

selected. The sources, the number of collected articles and the provinces are listed in Table I.

TABLE I. Number of articles of each province and the hyperlinks to the news media where they were published.

Number	Province	No. of articles	Media
1	Buenos Aires	6415	Clarín, La Nación
2	Córdoba	5013	La Voz, Cadena 3
3	Tucumán	5552	La Gaceta, Los Primeros
4	Mendoza	6359	Los Andes, MDZ Online
5	Santa Fe	2037	El Litoral, La Capital
6	Santiago del Estero	2505	Diario Panorama, Nuevo Diario

B. Topic decomposition model

Here, we describe how the content of the articles was analyzed and the methodology used for topic decomposition.

Pre-Processing Text: Firstly, we performed text tokenization by removing all the accents and by removing stop-words from a custom vocabulary which contained articles, verbs, and adverbs. In addition, very frequent words (present in more than 95% of the articles) and very rare words (present at most in 2 articles) were discarded.

Topic Decomposition: The texts were represented as vectors in a mathematical space using the term frequency - inverse document frequency (tf-idf) representation³¹. Each element of this vector represents words based on their weighted frequency by specificity³². A corpus composed by a group of news articles is then represented as a matrix, M , where rows correspond to articles, and columns represent the words used to describe them. Finally, from this matrix, a number of topics, N , were identified using the non-negative matrix factorization (NMF) algorithm from the Python library scikit-learn³³ that approximates M as the product of two matrices, H and W , that represent documents in the topic space and topics in the word space, respectively. The number of topics N is a parameter of the algorithm. The algorithm ensures that all the elements of H and W are non-negative, which results

in *non-orthogonal topics*. By normalizing the rows of H to the L1 norm, the elements of each row (h) can be interpreted as the degree of membership of a document (k) to the n-th topic, while the rows of W detail the characterization of each topic using the vocabulary (w_{nj} indicates the weight of term j in topic n , meaning how effectively term j encapsulates the content of topic n).

Description of the topics: Finding the optimal number of topics, N , is always a challenge in topic models. After trying several values, we decided to select $N = 20$ because it provided a set of topics with a clear interpretation (as it can be seen in the description below). Selecting a larger number of topics can reduce the topic-overlap, but can also introduce “small topics” that are not really relevant in the global agenda (but that may be locally important in different Argentinian provinces), while selecting a lower number of topics will merge different topics, making the topics more general. Clearly, the number of topics (as well as the event-detection thresholds and link-detection thresholds to be described in the next subsections), are hyper-parameters that have to be set up heuristically. Additional studies are planned to clarify the role of these parameters.

The topics are described as:

1. **Assault on Vice President:** The attempt of assassination of Vice President Cristina Fernandez de Kirchner by Fernando Sabag Montiel.
2. **Vice President:** Includes press articles that mention the vice president, including articles that mention the former president, as well as articles that mention the assassination attempt. Topic 1 and topic 2 have a large overlap because, as explained before, the topic decomposition is non-orthogonal.
3. **Diplomatic conflict:** Includes press articles that mention the arrival of a plane from Venezuela with Iranian crew.
4. **Former Economy Minister:** Includes press articles that mention the former economy minister, Silvina Batakis; the analyzed time period covers Batakis’ resignation.
5. **Economy Minister:** Includes press articles that mention the new economy minister, Sergio Massa.
6. **Gas Pipeline:** Includes press articles that mention the construction of the gas pipeline Nestor Kirchner.

7. **Social Leader:** Includes press articles that mention the health condition of social leader Milagro Sala, from the province of Jujuy (which is not one of the six main provinces included in this study).
8. **Domestic politics:** Includes press articles that mention the former governor of Tucumán, Juan Manzur, and his assumption as chief of staff.
9. **Vice President Trials:** Includes press articles that mention several legal cases of Vice President, Cristina Kirchner, particularly the “Causa Vialidad”.
10. **Dollar:** Includes press articles that mention the value of the dollar.
11. **Supreme Court:** Includes press articles that mention the Supreme Court and the proposed bill to modify it.
12. **Government:** Includes press articles that mention President Alberto Fernandez.
13. **Inflation:** Includes press articles that mention the inflation rate.
14. **Fuel Shortage:** Includes press articles that mention the shortage of diesel fuel in different regions.
15. **Strikes:** Includes press articles that mention the strikes carried out by the truckers' union, partly due to the Fuel Shortage.
16. **Libertarian Party:** Includes press articles that mention Javier Milei, the main politician of the libertarian party in Argentina (current president of Argentina).
17. **Subsidy Segmentation:** Includes press articles that mention the removal of energy subsidies in Buenos Aires and its segmentation.
18. **Opposition:** Includes press articles that mention actions and sayings of the coalition of opposition parties to the government of Alberto Fernández.
19. **International affairs:** Includes press articles that mention international politics, mainly the “Cumbre de las Américas”.
20. **Social Assistance:** Includes press articles that mention social assistance and the social organizations.

Additional information can be found in Appendix A: the word-clouds of the topics are shown in Fig. 11, and the five most relevant keywords of each topic are listed in Table II.

C. Media agenda

Here we define the global and local media agenda, which are time series that quantify the attention received by the topics described before, as a function of time, at the national level, and at a local level, in the six main provinces listed in Table I.

If the topics were mutually orthogonal, the agenda could simply be calculated as the number of news articles belonging to each topic. Since they are not, the agenda is built by adding the degree of membership (given by the elements of matrix H , $h_{k,n}$) to each topic n , of each article, k , published in a period of time, $[t, t + \Delta t]$, normalized to the total number of articles, L , published in the same time interval,

$$x^n(t) = \frac{1}{L} \sum_{k \in [t, t + \Delta t]} h_{k,n} \quad (1)$$

In this way, we obtained a set of N time series that we refer to as *global* media agenda, which quantify the total coverage received by the N topics. The data values are normalized such that each day, adding the media attention over all the topics gives 1, $\sum_n x^n(t) = 1$ for all t .

To obtain the *local* agenda of topic n in province i we added the elements of the matrix H , $h_{k,n}$, over all the documents, k , that were published within $[t, t + \Delta t]$, in a media outlet located in the province i , normalized to the total number of articles L_i published in the same time interval.

$$x_i^n(t) = \frac{1}{L_i} \sum_{k \in [t, t + \Delta t], k \in i} h_{k,n} \quad (2)$$

In this way, the data values are normalized such that in each province, each day, the media attention over all the topics is 1, that is $\sum_n x_i^n(t) = 1$ for all i , for all t .

Figure 1 shows the time series of global media agenda obtained for the 20 topics described before. It can be observed that there are topics with well-defined peaks, usually triggered by particular events, such as “Vice President” or “Diplomatic conflict” or “Social Leader”, while others remain on the agenda consistently over time, contributing to the agenda as background topics, such as “Dollar”, “Inflation” or “Government”.

It is important to remark that a noisy “background” can be just a numerical artifact. In fact, depending on the topic, no media attention was paid to it before a particular event actually occurred. A clear example of this situation are the noisy values of the agenda time series of topic $n = 3$ (Diplomatic conflict) before the plane that triggered the conflict arrived in Argentina on June 6, 2022. Therefore, all the values of the time series previous to June 6 should be considered spurious. Another example are the agenda values of topic $n = 1$ (Assault on Vice President) before the assault occurred on September 1, 2022. This motivates the “event-based” point process method used to analyze the data and it may also partially explain the bad performance of well-known methods to quantify information transfer from raw time series analysis, Granger causality and transfer entropy, as will be discussed latter.

Two examples of the local agendas are displayed in Fig. 2, where we see that when a topic is triggered by a particular event, the evolution is quite similar in the different provinces, while if the topic is consistently present in the agenda, its evolution can be quite different in the different provinces. The values of the data points vary with the particular topic and province, because a topic in a province may not be as relevant as in another.

1. Point process description of the local agendas

To define events of media attention in the local agendas, we used an activation threshold and a deactivation one. An event for the n -th topic in the i -th province is identified when the time series $x_i^n(t)$ crosses the activation threshold th_a . Then, before another event can occur, we required that the agenda value decreases below the deactivation threshold, $th_d < th_a$.

The thresholds were defined for each topic, in terms of the mean, μ_n , and the standard deviation, σ_n , of the global agenda of the n -th topic. Specifically, the thresholds were defined as:

$$\begin{aligned} th_a^n &= \mu_n + \sigma_n, \\ th_d^n &= \mu_n \end{aligned} \tag{3}$$

Figure 3 shows an example of the distribution of global agenda values of a topic, “Fuel shortage” ($n = 15$). The figure also shows the activation and deactivation thresholds of the topic, and the events detected in two provinces, Córdoba and Santa Fe.

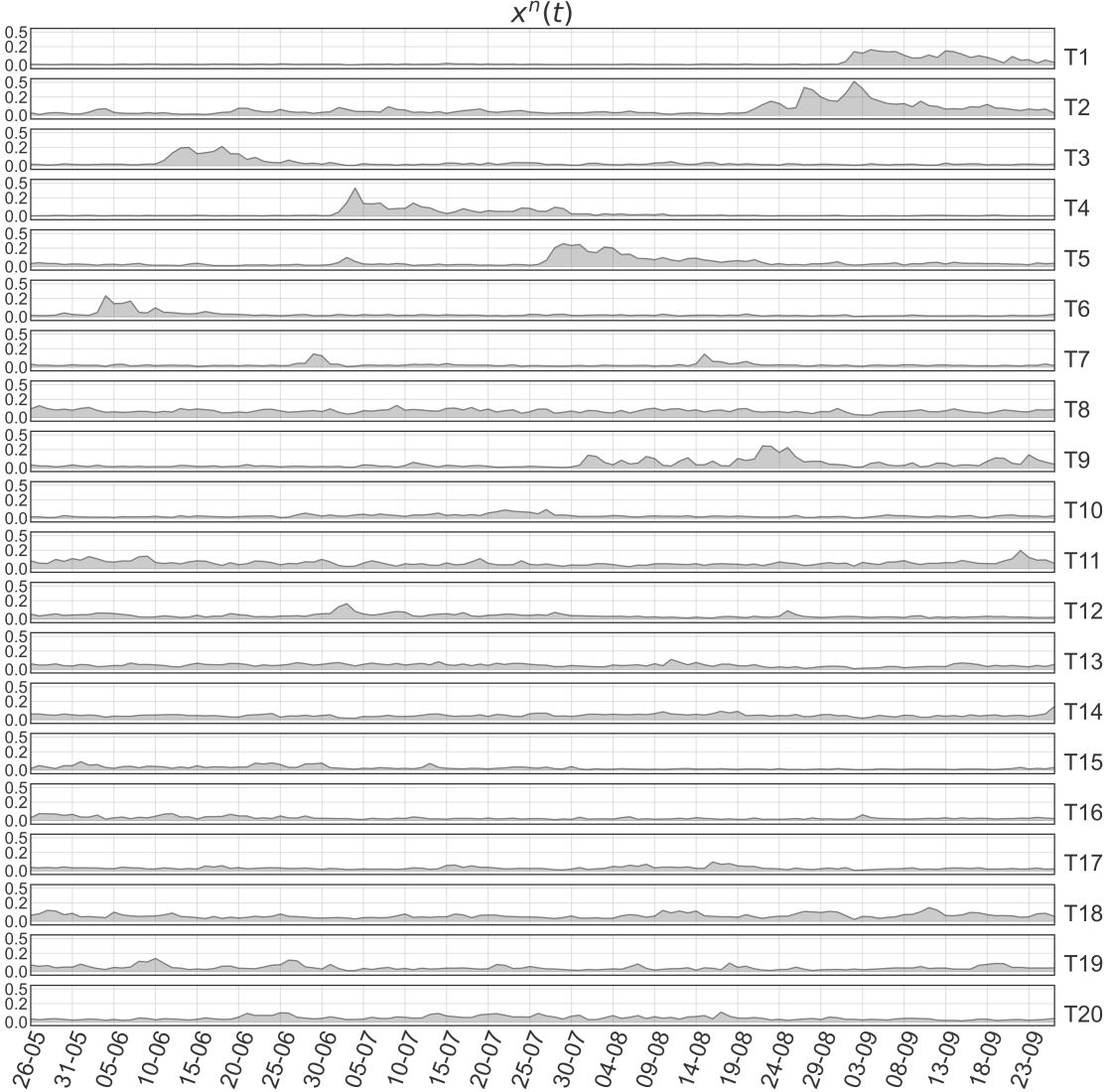


FIG. 1. **Global agendas** of the 20 topics. The notation T_i denotes the i -th topic following the labels assigned in Table II. Some topics have time series with pronounced events while other do not. The topics are numbered and displayed from more (top) to less (bottom) eventful.

The values selected for the two thresholds allow us to detect when media attention on a topic starts to grow, and also, as it will be shown later, they allow detecting events in all the time series.

Obviously, our results will depend on the event-detection strategy and thresholds employed. We tested a different strategy, based on using a single threshold to detect an event—the onset of attention in a topic in a province—and requiring that a short time interval passes, Δt (of two-three days), before another event can occur. However, with this strategy

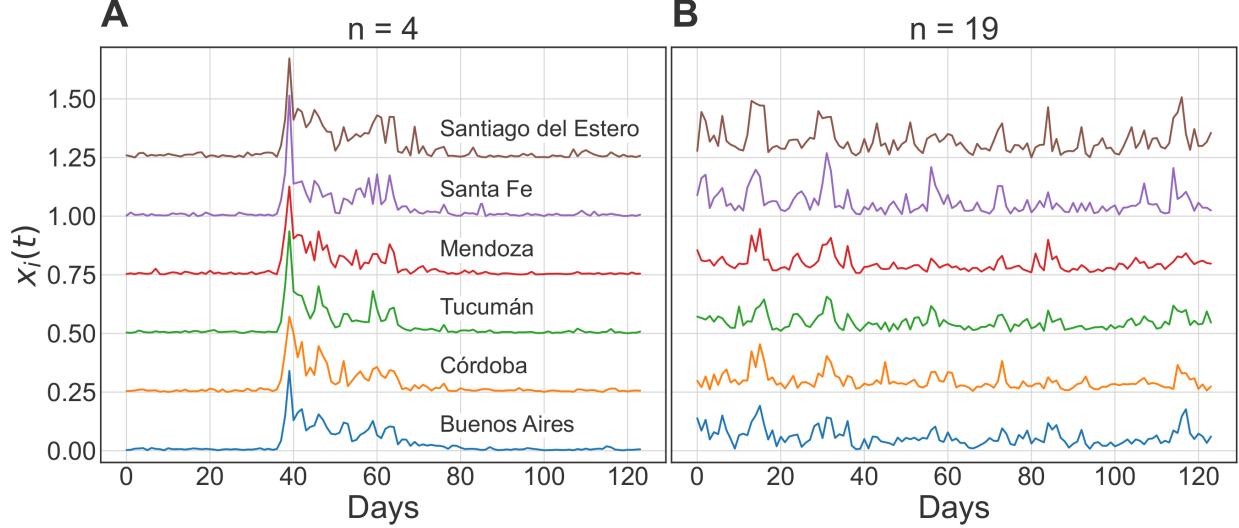


FIG. 2. **Local agendas** of two topics in the six provinces (shifted vertically for clarity). **A** Topic “Former Minister of Economy”; here we note that all time series exhibit a well-defined event that subsequently decays. **B** Topic “International affairs”; here we see a much noisier behavior, with all time series displaying numerous peaks.

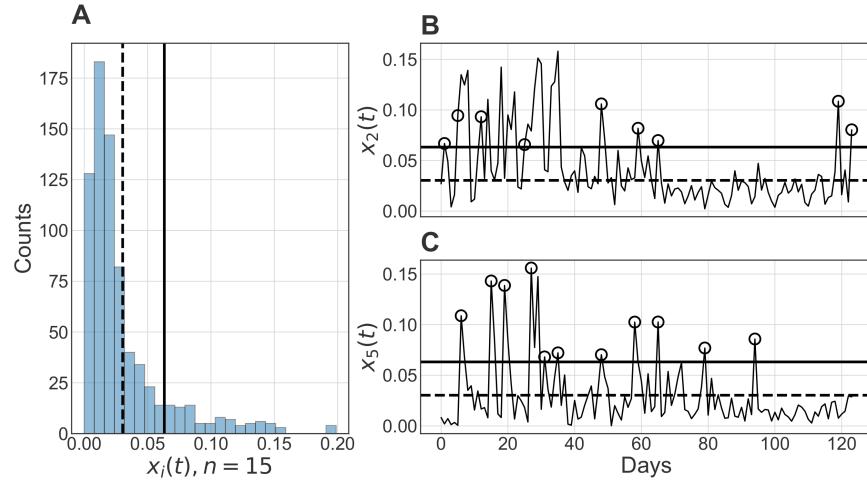


FIG. 3. **A** Histogram of all the local agenda values (for the six provinces) of topic “Fuel shortage” ($n = 15$). The vertical lines indicate the activation and deactivation thresholds, th_a (solid line) and th_d (dashed line). **B & C**, Local agendas in Córdoba ($i = 2$) and Santa Fe ($i = 5$), respectively; the circles indicate the events detected.

we found that fast and large variations in the agenda (see, for example, Figs. 3B, Figs. 3C) were not detected as different media attention events. Therefore, we decided to use two

thresholds, one to detect an increase of attention in a given topic, and the other, to detect when the attention falls (and then, another different event may occur latter). We also tested defining individual thresholds for each time series, in terms of the mean and standard deviation of the local agenda values, and found qualitatively similar results. Since already a simple definition of the thresholds, Eq.(3), leads to a set of events which appear meaningful (particularly for topics with well-defined peaks), we keep this definition. Future work will be aimed at comparing the results obtained with different event-detection strategies: with or without temporal lag to separate different events, thresholds defined for each time series (for each topic, in each province), thresholds defined for each topic (used here) and global thresholds (same for all the topics and for all the provinces).

2. Event coincidence detection

We used the method proposed in³⁰ to quantify the synchronicity of events in two different time series. The method is based on counting the number of quasi-simultaneous occurrences of events, allowing for a short delay, τ , between them. Here we use $\tau = 3$ days because, after carefully inspecting the relation among articles, it is approximately the maximum amount of time it takes for a news to appear elsewhere. While the method is generic and can be applied to any pair of time series, here we use it to analyze pairs of agenda time series of the same topic in different provinces.

Therefore, if two events in the agendas of topic n in provinces i and j , x_i^n and x_j^n , occur at times t_i and t_j with $|t_i - t_j| \leq \tau$, we detect a coincidence. We define the quantity C_{ij} as the number of coincidences for which $t_i < t_j$, and C_{ji} as the number of coincidences for which $t_i > t_j$. If the two events occur simultaneously, the coincidence is shared and we add 1/2 to C_{ij} and 1/2 to C_{ji} .

Event synchronization measures: With C_{ij} and C_{ji} we calculate the measures Q_s (symmetric/synchronous) and Q_a (asymmetric/asynchronous) proposed in³⁰.

$$Q_{s,i,j} = \frac{2}{m} [C_{ij} + C_{ji}], \quad (4)$$

$$Q_{a,i,j} = \frac{2}{m} [C_{ij} - C_{ji}] \quad (5)$$

Here, $m = m_i + m_j$, with m_i and m_j being the number of events detected in the $x_i^n(t)$ and

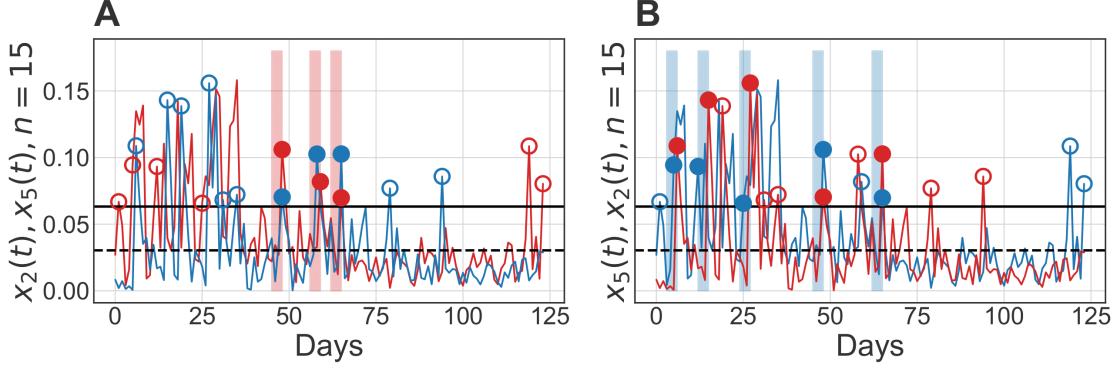


FIG. 4. Event coincidence detection and quantification. **A** Agenda time series of topic “Fuel Shortage” ($n = 15$) in Córdoba ($i = 2$, blue) and in Santa Fe ($j = 5$ red); here colored circles indicate events in Córdoba (solid blue) that precede (by a maximum of $\tau = 3$ days) or that occur simultaneously to events in Santa Fe (solid red) –empty circles indicate all the other events. Counting the number of coincidences gives $C_{2,5}(15) = 1/2 + 1 + 1/2 = 2$. **B** Time series “Fuel Shortage” in Santa Fe (blue) and in Córdoba (red). The colored circles show coincidences such that events in Santa Fe (solid blue) precede or occur simultaneously to events in Córdoba (solid red). Counting the number of coincidences gives $C_{5,2}(15) = 1 + 1 + 1 + 1/2 + 1/2 = 4$. Using Eq. 5 with $m = 20$ (total number of events) gives $Q_s^{15}{}_{5,2} = 0.6$ and $Q_a^{15}{}_{5,2} = 0.2$.

$$x_j^n(t).$$

With this definition, $Q_{s_{i,j}} = 1$ if, for each event in x_i^n (in x_j^n), there is an event in x_j^n (in x_i^n), and their timing is such that $|t_i - t_j| \leq \tau$. In addition, $Q_{a_{i,j}} = 1$ ($Q_{a_{i,j}} = -1$) if and only if the events in x_j^n (in x_i^n) always precede those in x_i^n (in x_j^n). Clearly, $Q_{s_{i,j}} = Q_{s_{j,i}}$ and $Q_{a_{i,j}} = -Q_{a_{j,i}}$.

Figure 4 illustrates how event coincidence detection works and how Q_s and Q_a are computed.

Summarizing, for each topic, n , from the events detected in the provincial agendas, we calculated the 6×6 matrices $Q_{s_{i,j}}^n$ and $Q_{a_{i,j}}^n$, which are symmetric and anti-symmetric, respectively.

3. Other analysis techniques used

We also analyzed the local agendas using well-known bivariate measures to detect correlations and dependencies, which are computed from the raw time series: two undirected measures, linear cross-correlation, CC , and nonlinear mutual information, MI , and two directed measures, linear Granger causality, GC^{34} , and nonlinear Transfer entropy, TE^{35} . GC assumes that the time series can be modeled using autoregressive processes with white noise, which implies time series with a Gaussian distribution. MI and TE do not require the distributions to be normal, however, it has been shown that the k -neighbours algorithm used to compute them works better if the time series are Gaussian-like and with similar variance³⁶. Therefore, we pre-processed the data by applying a log transformation to it, then we removed the mean to each time series and re-scaled them to unit variance.

The linear cross-correlation was calculated using Pearson's equation,

$$CC(x, y) = \frac{\sum(x - \mu_x)(y - \mu_y)}{\sqrt{\sum(x - \mu_x)^2 \sum(y - \mu_y)^2}}. \quad (6)$$

We calculated CC using different time lags between the time series, but in most cases the maximum was found at zero lag, therefore, we decided to use zero-lag CC as a synchronization metric.

MI is also a symmetric magnitude, that measures the amount of shared information between two time series, x and y ,

$$MI(x, y) = \int \int p(x, y) \log \left(\frac{p(x, y)}{p_x(x)p_y(y)} \right) dx dy, \quad (7)$$

where p_x (p_y) is the probability density function of x (y), and p is the joint probability density function of (x, y) . To estimate the integral we used a k -neighbors algorithm with $k = 4$, providing a good compromise between bias and variance³⁶.

Measuring GC consists in modeling one time series as an autoregressive process of order D and determining whether adding lagged values of another series improves the prediction of the autoregressive model, suggesting a potential causal relationship between them. In particular, to test $y \rightarrow x$ causality in Granger sense, we first fit the model

$$x_t = \sum_{i=1}^D a_i x_{t-i} \quad (8)$$

to x , obtaining residuals of variance σ_x^2 . Then, we fit the model

$$x_t = \sum_{i=1}^D a_i x_{t-i} + \sum_{i=1}^D b_i y_{t-i}, \quad (9)$$

obtaining residuals of variance $\sigma_{xy}^2 < \sigma_x^2$. Then, the prediction improvement is measured by the Granger Causality estimator

$$GC = \frac{\sigma_x^2 - \sigma_{xy}^2}{\sigma_x^2} \quad (10)$$

The order of the autoregressive process, D , was decided using the Schwartz criterion³⁷. In the dataset under study, the typical D was the order of few days, with 80% of the series having $D \leq 3$. The GC significance was evaluated using 200 time-shift surrogates. The same procedure was used to assess the significance of all the other measures. Finally, TE measures the information flow from one time series to another,

$$TE_{y \rightarrow x} = MI(y_t, x_{t-1:t-M} | y_{t-1:t-M}), \quad (11)$$

that is the mutual information between x and the past of x , conditional to the past of y . M in this case is the order of the Takens embedding of x and y . We estimated the TE using a k -neighbours algorithm³⁸, while M was chosen again using the Schwartz criterion. Also in this case we set $k = 4$.

Summarizing, for each topic, n , from the provincial agendas (raw time series), we calculated the 6×6 symmetric matrices $CC^n_{i,j}$ and $MI^n_{i,j}$, and the 6×6 matrices $GC^n_{i,j}$ and $TE^n_{i,j}$, which do not have a particular symmetry.

III. RESULTS

We present the results of the analysis of the agendas of the 20 topics. By inspecting the time series shown in Fig. 1, it is clear that when a particular event trigger attention to a topic (for instance, the arrival of the Iranian–Venezuelan plane), the agenda presents a steep rise of attention that then decays. As expected, these topics are associated to one or a very few number of events and the attention vanishes after the events (as discussed before, the small noisy values occurring several days before (and often, several days after) well-defined events are numerical artifacts). On the other hand, topics that are frequently discussed in the media do not show very pronounced peaks. These “background” or “recurrent” topics have a higher number of events due to their constant attention over time.

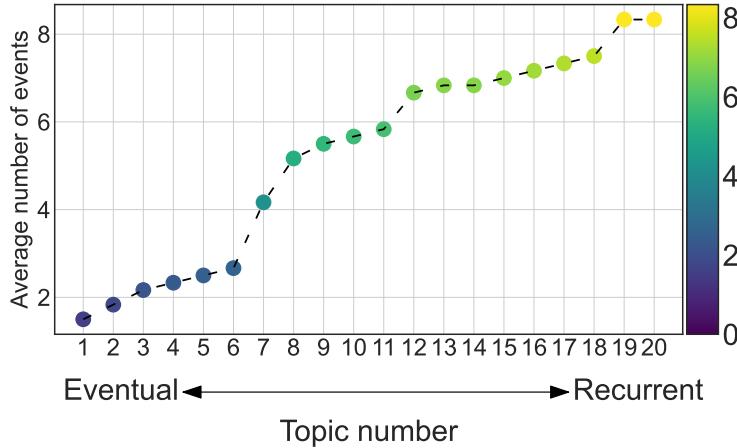


FIG. 5. **Differentiation of “eventual” and “recurrent” topics.** Average number of events per province, for each topic. The color gradient indicates which topics are more “Eventual” (dark blue) and which ones are more “Recurrent” (bright yellow).

The difference between “eventual” type and recurrent type of topic is visualized in Fig. 5 that displays the average number of detected events per province. The topics with few events (in the left part of the figure) are encapsulated by very specific words and are the most “eventual” topics, while the topics with the largest number of events (in the right part of the figure) are those that are encapsulated by less specific words and associated with broader themes.

As explained in the previous section, for each topic we calculated the symmetric 6×6 matrices Q_s^n , $CC^n_{i,j}$ and $MI^n_{i,j}$. In Fig. 6A we plot all the values (for all the topics and for all pairs of provinces). In this figure the color code indicates the topic number and is the same as in Fig. 5: dark blue (yellow) indicates topics with low (with large) average number of events, i.e., eventual (recurrent) topics. It can be seen that eventual topics tend to have large values of Q_s , CC , and MI , while recurrent topics have lower values.

The two pairs of time series with the largest synchronization level, identified as two pairs that have the largest values of $Q_s^2 + CC^2 + MI^2$, are shown in Figs. 6B and 6C. They correspond to the topics “Vice President Trials” and “Vice President”. The first one is for provinces Córdoba and Tucumán, while the second one, for Buenos Aires and Santiago del Estero.

The synchronization between provinces can be visualized by using a network representation. To define the network, we selected the pairs of agenda time series that have largest

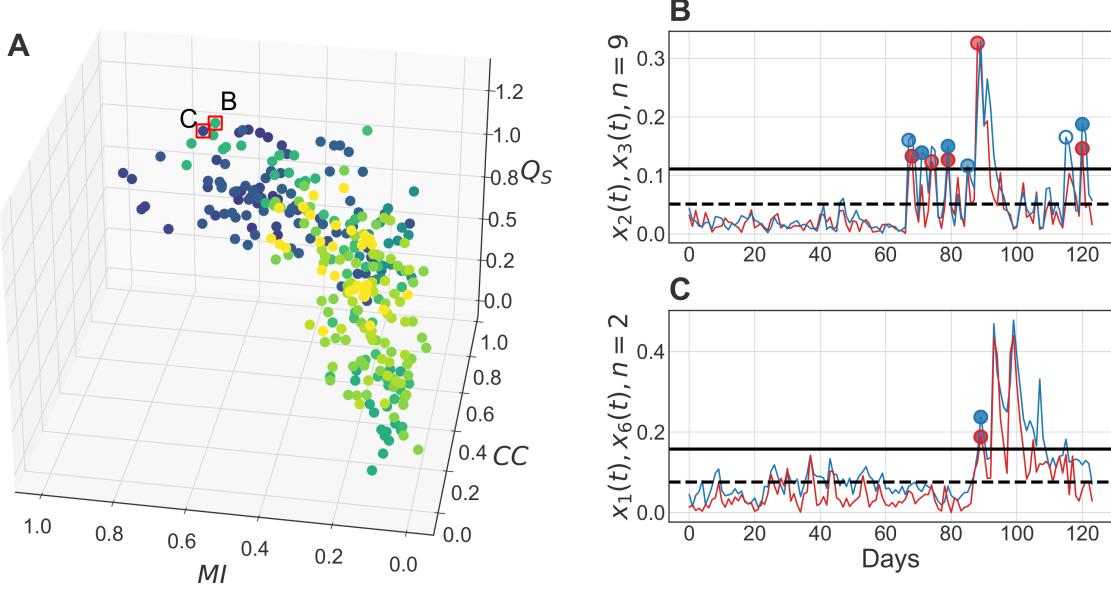


FIG. 6. Analysis of the symmetric quantifiers. Panel **A** displays the 3D plot of $Q_{s,i,j}^n$, $CC_{s,i,j}^n$ and $MI_{s,i,j}^n$ for all topics and all pairs of provinces. The color indicates the topic, as in Fig. 5. The dots marked **B** and **C** have the maximum value of the sum of squares of the three metrics. Panels **B** and **C** display the corresponding time series: the agendas of topic “Vice President Trials” in Córdoba and Tucumán, and of topic “Vice President” in Buenos Aires and Santiago del Estero.

similarity, measured in terms of the largest values of the sum $Q_s^2 + CC^2 + MI^2$. We chose this criterion because the three measures can be expected to provide complementary information, in the sense that they quantify different aspects in which two time series can be statistically similar (CC detects linear correlations, MI detects nonlinear correlations, and Q_s detects synchronicity in the events defined in the time series).

Because these matrices are symmetric, we have $6 \times 5/2$ pairs of provinces and 20 topics, therefore, 300 possible pairs. We selected the top 10% (30 pairs) with largest similarity. We remark that by selecting the largest values, we are also choosing the most significant ones, because, the larger the CC or MI value is, the lower the p-value is (as shown in Fig. 12 in the Appendix).

Figure 7A displays the network obtained in this way. The number of links that correspond to eventual topics (that have less than 4 events per province –20 links) doubles the number of links that correspond to recurrent topics (that have 4 or more events events per province –10 links), showing again that the most synchronized links are those that correspond to

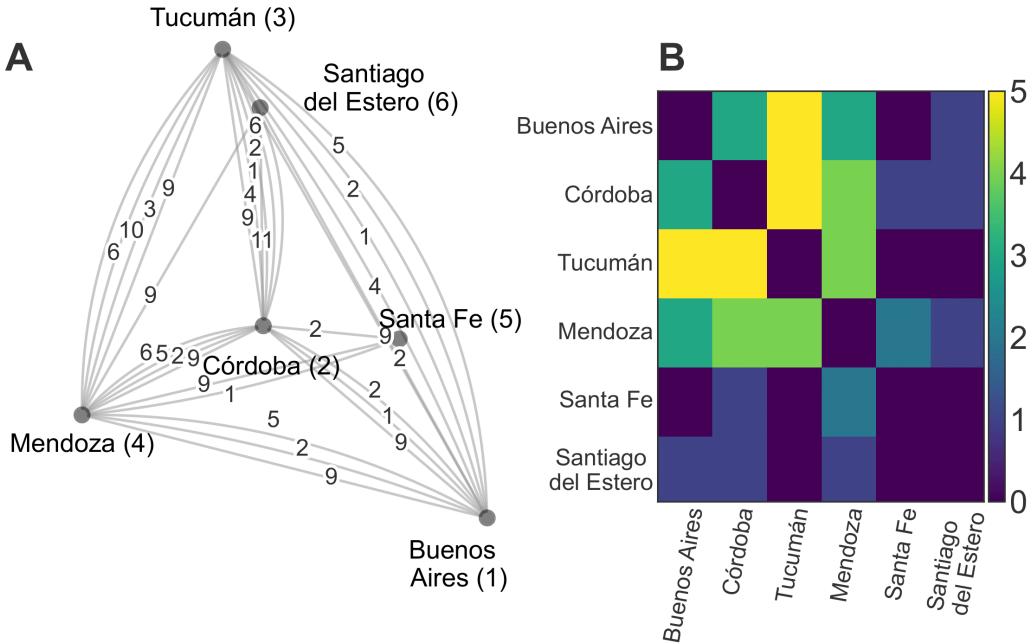


FIG. 7. Network representation of the similarities uncovered between pairs of provincial agendas. **A** Network formed by the 30 links (top 10%) that have the largest values of the sum $Q_s^2 + CC^2 + MI^2$. The labels on the links indicate the topic. **B** Weighted adjacency matrix where the weights (in color code) represent the number of (non-directed) links that are shared between pairs of provinces.

topics with few and well-defined events. In Figure 13 we show the time series corresponding to the top ten links with the largest similarity.

Figure 7B displays the adjacency matrix of the same network without distinguishing topics. Here the color code indicates the number of links between pairs of provinces. From this matrix one could try to establish if there is a province that serves as a news broadcaster; however, no province has a clear news dissemination role. However, we find two pairs of provinces, Tucumán and Buenos Aires, and Tucumán and Córdoba, that share a large number of links, which indicates that the local agendas in various topics follow quite similar temporal evolution (an example was shown in Fig. 6B that displays the agendas of topic “Vice President Trials” in Tucumán and Córdoba).

As a final step, we study the direction of information diffusion by analyzing the matrices $Q_{a,i,j}^n$, $GC_{i,j}^n$ and $TE_{i,j}^n$.

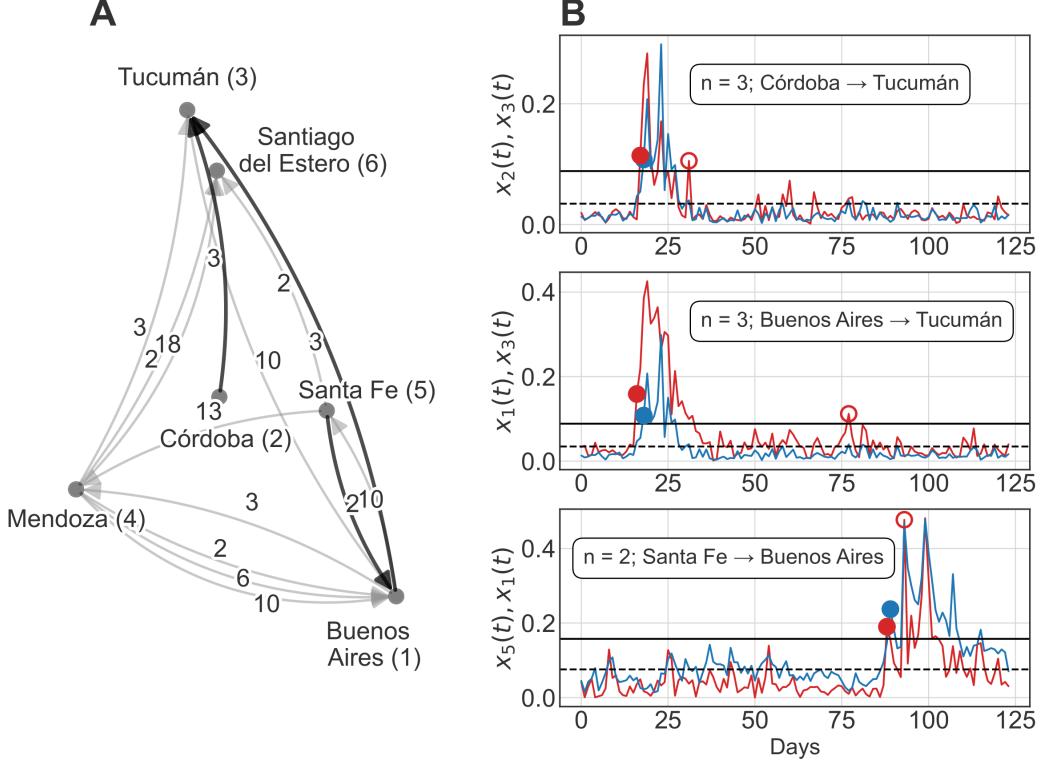


FIG. 8. Analysis of directed links identified from event-based analysis. **A** Directed network constructed by selecting the links that satisfy the conditions $|Q_a| > 0.6$ and $Q_s > 0.6$, which gives 14 directed links. **B** Time series corresponding to the three links marked with different line styles (the time series associated to all the links are shown in Fig. 14). In each panel we can observe that the event on the red time series precedes the event on the blue time series by one day.

The absolute values of the elements of Q_a provide information about the level of asymmetry of the synchronization between two provinces, and their sign indicate which province is the leader, and which one lags behind (low $|Q_a|$ values indicate that the events in the two provinces alternate, while high $|Q_a|$ values indicate that events in one province tend to precede the events in the other province).

To represent the most relevant links in a directed network, we selected pairs of provinces whose values of $|Q_a|$ and Q_s are both larger than 0.6. The network obtained in this way, shown in Fig. 8A, has 14 links. In Fig. 8B, the time series of the three bold links are shown. Here, we can observe that the event of the red time series precedes the event of the blue one with one day of difference. It is important to remark that the network obtained depends on the threshold used to select the largest $|Q_a|$ and Q_s values (and also, on the thresholds

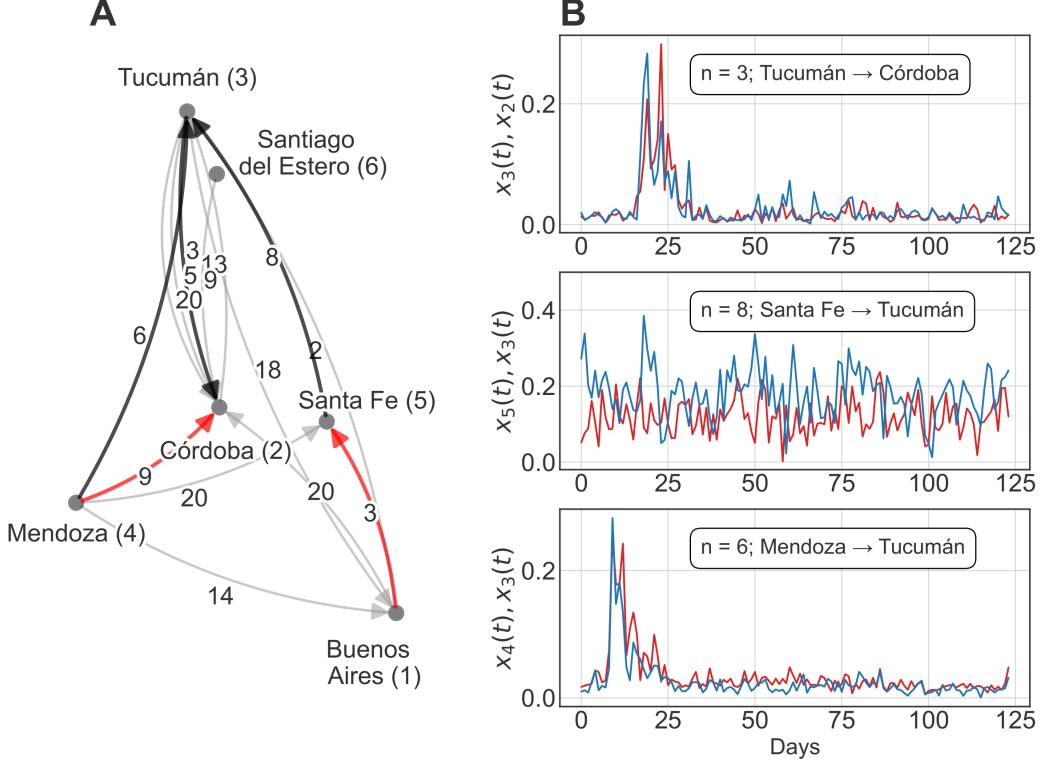


FIG. 9. Analysis of directed links identified from transfer entropy analysis. **A** Directed network constructed by selecting the 14 links (the same number of links as in the network shown in Fig. 8) with highest TE values. The black edges are the ones presented in **B**, while the red edges are the ones that match with the GC network. **B** Time series corresponding to the three links marked with different colors (the time series associated to all the links are shown in Fig. 15). The upper panel shows the series for Tucumán and Córdoba where the TE detected an inverse direction of the flow compared to the Q_a method.

used to detect the events). We remark that, while the links that are selected vary with these thresholds, the topics that are associated to these links are always the most eventual ones.

The networks obtained from TE and GC are displayed in Figs. 9 and 10 respectively. In both cases, we selected the 14 links with largest value, to make them comparable with the previous case. The TE network doesn't share any links with the Q_a network. Even though the two methods detect an interaction between Córdoba and Tucumán for the topic "Diplomatic Conflict", in the Q_a network it goes from Córdoba to Tucumán (as can be observed in the temporal series of top panel of Figure 8B), while the TE method sets the inverse relation. When comparing the TE and GC networks, they only share two links

(colored in red in both figures): between Buenos Aires and Santa Fe for topic 3 (Diplomatic Conflict) and between Mendoza and Córdoba for topic 9 (Vice President Trials).

Moreover, the two networks do not share any link with the network shown in Fig. 8, obtained by thresholding Q_s and $|Q_a|$. We believe that the discrepancy can be due to three main reasons: First, the event-based approach filters out small noisy fluctuations (which, as discussed before, are usually an artifact), but GC and TE do not filter them because GC and TE are computed from the raw time series. Secondly, the time series are strongly non-stationary, and one of the implicit assumptions of both GC and TE is that the time series under study are stationary. This non-stationary nature, which is observed in both types of topics, is due to the complex temporal variation of the mean and the variance, therefore, it cannot be modified by detrending. The quantities Q_s and Q_a , instead, do not have this kind of restriction. Thirdly, Q_s and Q_a are symmetric and asymmetric measures respectively, thus capture the strength and the net direction of information flow between two provinces, whereas GC and TE are not asymmetric measures and do not return a net direction of information flow.

While asymmetric quantities could have been defined by using the differences $GC_{i,j}^n - GC_{j,i}^n$ and $TE_{i,j}^n - TE_{j,i}^n$, this easy solution has the drawback that the pairs of values $(GC_{i,j}^n, GC_{j,i}^n)$ and $(TE_{i,j}^n, TE_{j,i}^n)$ may not have a similar level of significance (i.e., they may have very different p-values). Therefore the largest values of $|GC_{i,j}^n - GC_{j,i}^n|$ and $|TE_{i,j}^n - TE_{j,i}^n|$ may not capture the most significant links.

IV. CONCLUSION AND DISCUSSION

We have analyzed data collected from Argentinian newspapers published in six main provinces during a period of four months. We classified the articles into 20 non-orthogonal topics, and obtained a set of 120 time series representing the evolution of the media agenda in the provinces. To capture statistical similarities and dependencies, the time series were analyzed using six bivariate quantifiers, which can be classified into two main types: those computed from the raw time series (the cross-correlation, the mutual information, Granger causality, and transfer entropy) and those computed from events defined in the time series (symmetric and asymmetric measures, Q_s and Q_a , respectively).

To define events in the time series, we used two thresholds, one to identify when the media

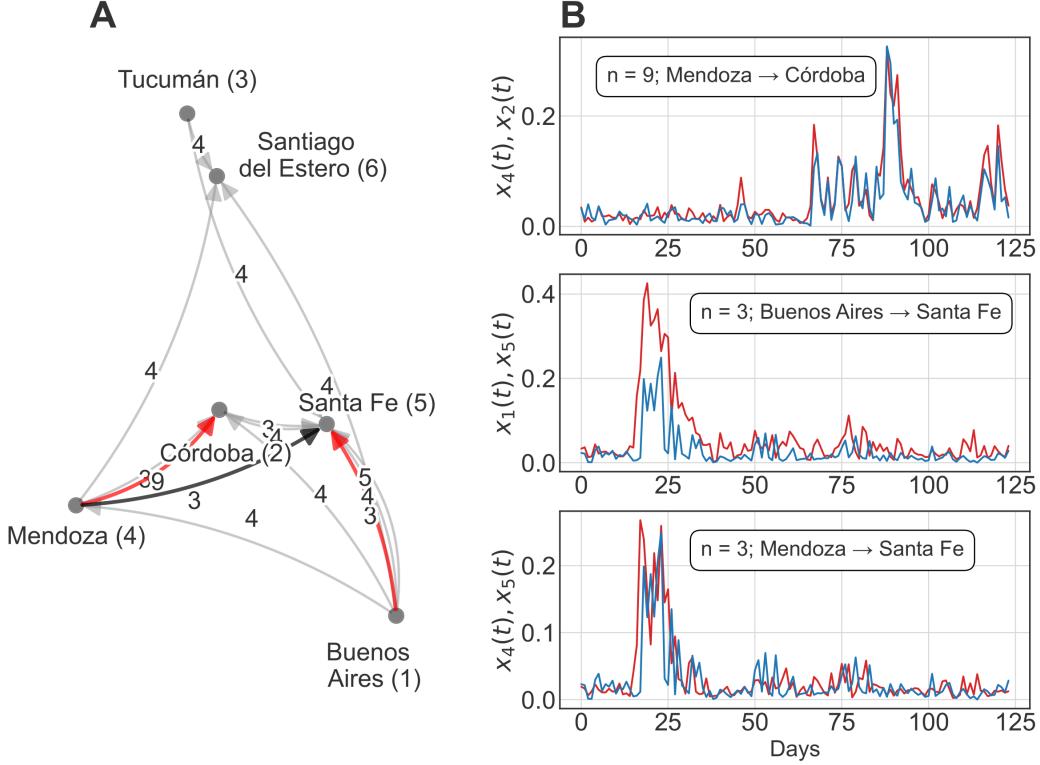


FIG. 10. Analysis of directed links identified from Granger causality analysis. **A** Directed network with the 14 links that have highest *GC* values. The time series corresponding to coloured edges are displayed in panel **B**, the red edges [for Mendoza and Córdoba for the topic “Vice President Trials” and for Buenos Aires and Santa Fé for the topic “Diplomatic Conflict”] are also present in the network shown in Fig. 9A.

attention picks up and another to detect when the attention decays and another event may take place latter.

By analyzing the number of events for the different topics in the different provinces, we were able to qualitatively distinguish them into two categories (Fig 5): “Recurrent” topics, which are continuously active in the press agenda and typically display several events in each province, and “Eventual” topics, which are topics that have attracted the media attention at particular times, and often show only one or few events in each province. Examples of eventual topics include “Assault on Vice President” ($n = 1$), “Vice President” ($n = 2$) and “Diplomatic conflict” ($n = 3$), while examples of generalist topic include “Opposition” ($n = 18$), “International Affairs” ($n = 19$) and “Social Assistance” ($n = 20$).

The six bivariate measures used to analyze pairs of times series for a given topic in

different provinces can also be classified into two types: undirected (CC , MI , and Q_s) and directed (GC , TE , and Q_a). We used the undirected measures to identify statistical similarities between pairs of time series and the directed ones to determine the direction of information flow. To analyze the performance of these measures in detecting meaningful connections and the main paths of information diffusion, we tested different strategies to identify, for each topic, the most significant links between provinces.

Because the three undirected measures provide complementary information, in the sense that they quantify different aspects in which two time series can be statistically similar, we selected the links whose sum, $CC^2 + MI^2 + Q_s^2$, was higher than a given threshold. The network obtained, shown in Fig. 7, certainly depends on the threshold used, however, an inspection of the time series of the links identified in this way, revealed clear similarities in the temporal variation of the media attention (two examples are shown in Fig. 6B).

The three directed measures we used, GC , TE , and Q_a , can also be expected to provide complementary information. GC and TE are computed from raw data, while Q_a results from the relative timing of media attention events. We found that GC and TE often gave contradicting results, with the dependency returned by GC being the opposite of that returned by TE . In this sense, we found that thresholding GC and TE was not useful for identifying clear dependencies in the process of information dissemination, at least with the hyper-parameters used in this study. Thus, we used a simple pruning strategy to identify directed links: we selected links whose Q_s and $|Q_a|$ were both above a threshold.

Again, the network obtained depended on the threshold used, and an inspection of the time series corresponding to the links identified revealed situations in which the attention on a given topic in two provinces began simultaneously or almost simultaneously. As it can be seen in Fig. 8, using a threshold of 0.6 allowed us to identify some topics and provinces where the attention started first, followed by the growth of attention in other provinces.

It is important to remark that the event-based measures, Q_s and Q_a , have, besides the thresholds used to define events, another hyper-parameter that plays an important role, which is the time interval allowed between two events that occur in the same topic in different provinces, to count them as related. In general, increasing τ results in more values of Q_a and Q_s that are above the thresholds. This may be because, when the attention on topic n begins in province i , in province j the attention is on a different topic, and we need to allow for more than one day to be able to detect the fact that in province j , topic n

gains attention. Of course, using values of τ that are too long leads to count as synchronous events that can be unrelated.

Taken together, our results show that the process of information diffusion in newspapers in Argentina is very fast, and genuinely related “events” in different provinces are most often simultaneous. It is worth noticing that, in the networks found using the different thresholding strategies, no particular province was found to play a clear role of “information hub”. The quasi-simultaneous activation of a topic in different provinces can be understood because all of them are triggered by the same events and looks almost synchronous within a one day resolution time. Furthermore, information disseminated on social media or by large news agencies also has the effect of acting as global drivers.

For future work, it will be interesting to test the generality of our conclusions, by analyzing data extracted from articles in newspapers in other countries and seize the public aspect of this dynamic. This would allow us to compare the rates of propagation between the public agenda and the media and test their influences over each other. This methodology can be used also to understand how news propagate in social media and how this process depends on social media selected features as well as news content and emotional reaction to them. It will be also interesting to correlate media attention on socio-economic topics with financial time series that effectively capture the temporal evolution of the economy, at the provincial level, national level, and global international level.

ACKNOWLEDGMENTS

C.M. acknowledges the support of ICREA ACADEMIA, AGAUR (2021 SGR 00606), and Ministerio de Ciencia e Innovación Spain (Project No. PID2021-123994NB-C21). P.B acknowledges the support of the Universidad de Buenos Aires (UBA) through Grant UBA-CyT, 20020220100181BA and the Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación through Grant No. PICT-2020-SERIEA-00966.

DATA AVAILABILITY

The time series analysed, the code to calculate the events, coincidences and the matrices are available at https://github.com/luciolgarcia/News_diffution

REFERENCES

- ¹J. A. C. Gallas and H. E. Nusse, “Periodicity versus chaos in the dynamics of cobweb models,” *Journal of Economic Behavior and Organization* **29**, 447–464 (1996).
- ²R. J. Field, J. A. C. Gallas, and D. Schuldberg, “Periodic and chaotic psychological stress variations as predicted by a social support buffered response model,” *Communications in Nonlinear Science and Numerical Simulation* **49**, 135–144 (2017).
- ³J. E. Kollmer, T. Pöschel, and J. A. C. Gallas, “Are physicists afraid of mathematics?” *New Journal of Physics* **17**, 013036 (2015).
- ⁴M. E. McCombs and D. L. Shaw, “Public opinion quarterly,” *Public Opinion Quarterly* **36**(2):176-187 (1972), 10.1086/267990.
- ⁵T. O, D. Calacci, and D. Lazer, “A frame of mind: Using statistical models for detection of framing and agenda setting campaigns,” *ACL* **1**, 1629–1638 (2015).
- ⁶L. Guggenheim, S. Jang, S. Bae, and W. Neuman, “The dynamics of issue frame competition in traditional and social media,” *Ann. Am. Acad. Political Soc. Sci.* **659**, 207–224 (2015).
- ⁷M. G. Cosenza, M. E. Gavidia, and J. C. González-Avella, “Against mass media trends: Minority growth in cultural globalization,” *PLOS ONE* **15**, 1–14 (2020).
- ⁸G. Ferraz de Arruda, F. Aparecido Rodrigues, P. Martín Rodríguez, E. Cozzo, and Y. Moreno, “A general Markov chain approach for disease and rumour spreading in complex networks,” *Journal of Complex Networks* **6**, 215–242 (2017).
- ⁹M. N. Kuperman, “Cultural propagation on social networks,” *Phys. Rev. E* **73**, 046139 (2006).
- ¹⁰M. G. E. da Luz, C. Anteneodo, N. Crokidakis, and M. Perc, “Sociophysics: Social collective behavior from the physics point of view.” *Chaos, Solitons and Fractals* **170**, 113379 (2023).
- ¹¹N. Crokidakis, “Effects of mass media on opinion spreading in the sznajd sociophysics model,” *Physica A: Statistical Mechanics and its Applications* **391**, 1729–1734 (2012).
- ¹²S. Soroka, M. Daku, D. Hiaeshutter-Rice, L. Guggenheim, and J. Pasek, “Negativity and positivity biases in economic news coverage: Traditional versus social media,” *Commun. Res.* **45** (2017), <https://doi.org/10.1177/0093650217725870>.

- ¹³M. McCombs, “A look at agenda-setting: Past, present and future,” *Journalism Studies* **6**, 543–557 (2005).
- ¹⁴A. El Ali, T. C. Stratmann, S. Park, J. Schöning, W. Heuten, and S. C. Boll, “Measuring, understanding, and classifying news media sympathy on twitter after crisis events,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2018) p. 1–13.
- ¹⁵W. R. Neuman, L. Guggenheim, S. M. Jang, and S. Bae, “The dynamics of public attention: Agenda-setting theory meets big data,” *J. Commun.* **64**, 193–214 (2014).
- ¹⁶S. Pinto, F. Albanese, C. O. Dorso, and P. Balenzuela, “Quantifying time-dependent media agenda and public opinion by topic modeling,” *Physica A: Statistical Mechanics and its Applications* **524**, 614–624 (2019).
- ¹⁷F. Albanese, S. Pinto, V. Semeshenko, and P. Balenzuela, “Analyzing mass media influence using natural language processing and time series analysis,” *Journal of Physics: Complexity* **1**, 025005 (2020).
- ¹⁸M. McCombs and S. Valenzuela, *Setting the Agenda: The Mass Media and Public Opinion* (Wiley, New Jersey, NJ, 2020).
- ¹⁹S. Reese and L. Danielian, “Intermedia influence and the drug issue: Converging on cocaine,” in *Communication Campaigns about Drugs: Government, Media, and the Public*, edited by P. J. Shoemaker (Lawrence Erlbaum Associates, New York, 1989) pp. 29–46.
- ²⁰G. Golan, “Inter-media agenda setting and global news coverage,” *Journalism Studies* **7**, 323–333 (2006).
- ²¹M. J. Kushin, *Tweeting the issues in the age of social media? Intermedia agenda setting between The New York Times and Twitter*, Ph.D. thesis, Washington State University, Edward R. Murrow College of Communication, Pullman, Washington (2010).
- ²²L. Guo and C. Vargo, “Global intermedia agenda setting: A big data analysis of international news flow,” *J Commun* **67**, 499–520 (2017).
- ²³S. Mohammed and M. McCombs, “Intermedia agenda setting or international news flow? cross-lagged comparison of elite international newspapers,” *International Journal of Communication* **15**, 3948–3969 (2021).
- ²⁴L. Guo and Y. Zhang, “Information flows from local to national: Evidence from 21 major us cities,” *Journalism* **24**, 2651–2667 (2023).

- ²⁵S. Stern, G. Livan, and R. E. Smith, “A network perspective on intermedia agenda-setting,” *Applied Network Science* **5**, 31 (2020).
- ²⁶W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, “Quantifying echo chamber effects in information spreading over political communication networks,” *EPJ Data Science* **8** (2019), <https://doi.org/10.1140/epjds/s13688-019-0213-9>.
- ²⁷H. Sjøvaag, E. Stavelin, M. Karlsson, and A. Kammer, “The hyperlinked scandinavian news ecology,” *Digital Journalism* , 507–531 (2018).
- ²⁸S. Alipour, N. Di Marco, M. Avalle, G. Etta, M. Cinelli, and W. Quattrociocchi, “The drivers of global news spreading patterns,” *Sci. Rep.* **14** (2024), <https://doi.org/10.1038/s41598-024-52076-6>.
- ²⁹J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno, and A. Vespignani, “The dynamics of information-driven coordination phenomena: A transfer entropy analysis,” *Science Advances* **2**, e1501158 (2016).
- ³⁰P. Grassberger, R. Q. Quiroga, and T. Kreuz, “Event synchronization: A simple and fast method to measure synchronicity and time delay patterns,” *Phys. Rev. E* **66**, 041904 (2002).
- ³¹E. L. Bird, Steven and E. Klein, “Natural language processing with python.” (2019).
- ³²E. Nguyen, “Data mining applications with r,” Text mining and network analysis of digital libraries in R , 95–115 (2014).
- ³³F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- ³⁴C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica* **37**, 424–438 (1969).
- ³⁵T. Schreiber, “Measuring information transfer,” *Phys. Rev. Lett.* **85** (2000), [10.1103/PhysRevLett.85.461](https://doi.org/10.1103/PhysRevLett.85.461).
- ³⁶A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical review E* **69**, 066138 (2004).
- ³⁷G. Schwarz, “Estimating the dimension of a model,” *The annals of statistics* , 461–464 (1978).

³⁸J. Zhu, J.-J. Bellanger, H. Shu, and R. Le Bouquin Jeannès, “Contribution to transfer entropy estimation via the k-nearest-neighbors approach,” Entropy **17**, 4173–4201 (2015).

Appendix A



FIG. 11. Wordclouds of the 20 topics

TABLE II. Topics labels assigned and their most relevant five keywords which are used to describe it.

Number	Topic label	Keywords
1	Assault on Vice President	sabag, montiel, uliarte, brenda, capuchetti
2	Vice President	vicepresidenta, cristina, kirchner, atentado, odio
3	Diplomatic conflict	avion, irani, iranies, tripulantes, aeronave
4	Former Economy Minister	batakis, silvina, ministra, economia, guzman
5	Economy Minister	massa, sergio, economia, ministro, diputados
6	Gas Pipeline	gasoducto, kulfas, licitacion, off, matias
7	Social Leader	sala, milagro, jujuy, tupac, amaru
8	Domestic politics	jaldo, manzur, tucuman, obras, gobernador
9	Vice President Trials	luciani, baez, fiscal, juicio, vialidad
10	Dollar	dolar, blue, dolares, pesos, mercado
11	Supreme Court	corte, proyecto, suprema, ley, diputados
12	Government	fernandez, alberto, presidente, guzman, cristina
13	Inflation	inflacion, precios, argentina, economia, 2022
14	Strikes	paro, salarial, trabajadores, sute, docentes
15	Fuel Shortage	gasoil, combustible, transportistas, faltante, abastecimiento
16	Libertarian Party	milei, javier, libertad, avanza, libertario
17	Subsidy Segmentation	subsidios, energia, segmentacion, gas, usuarios
18	Opposition	larreta, pro, cambio, bullrich, macri
19	International affairs	cumbre, estados, unidos, biden, argentina
20	Social Assistance	sociales, planes, grabois, organizaciones, social

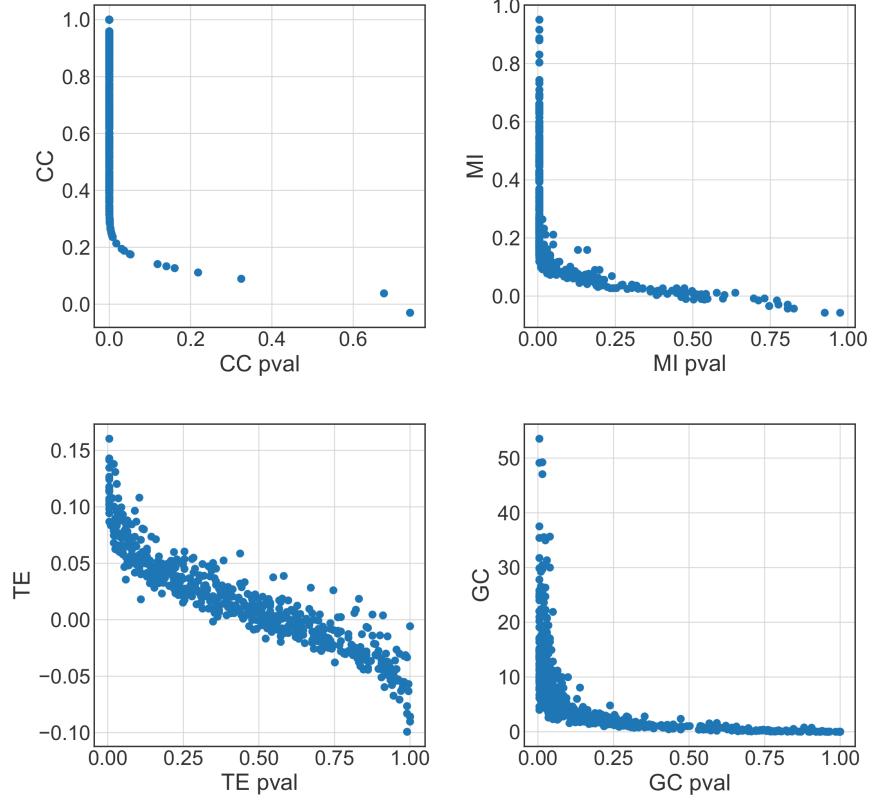


FIG. 12. Relationships between the metrics computed from the raw time series (cross-correlation, mutual information, Granger causality and transfer entropy) with their p-values. It can be seen that the lower p-values correspond to the largest values of the metrics. For this reason, selecting the largest values implies selecting the most significant ones.

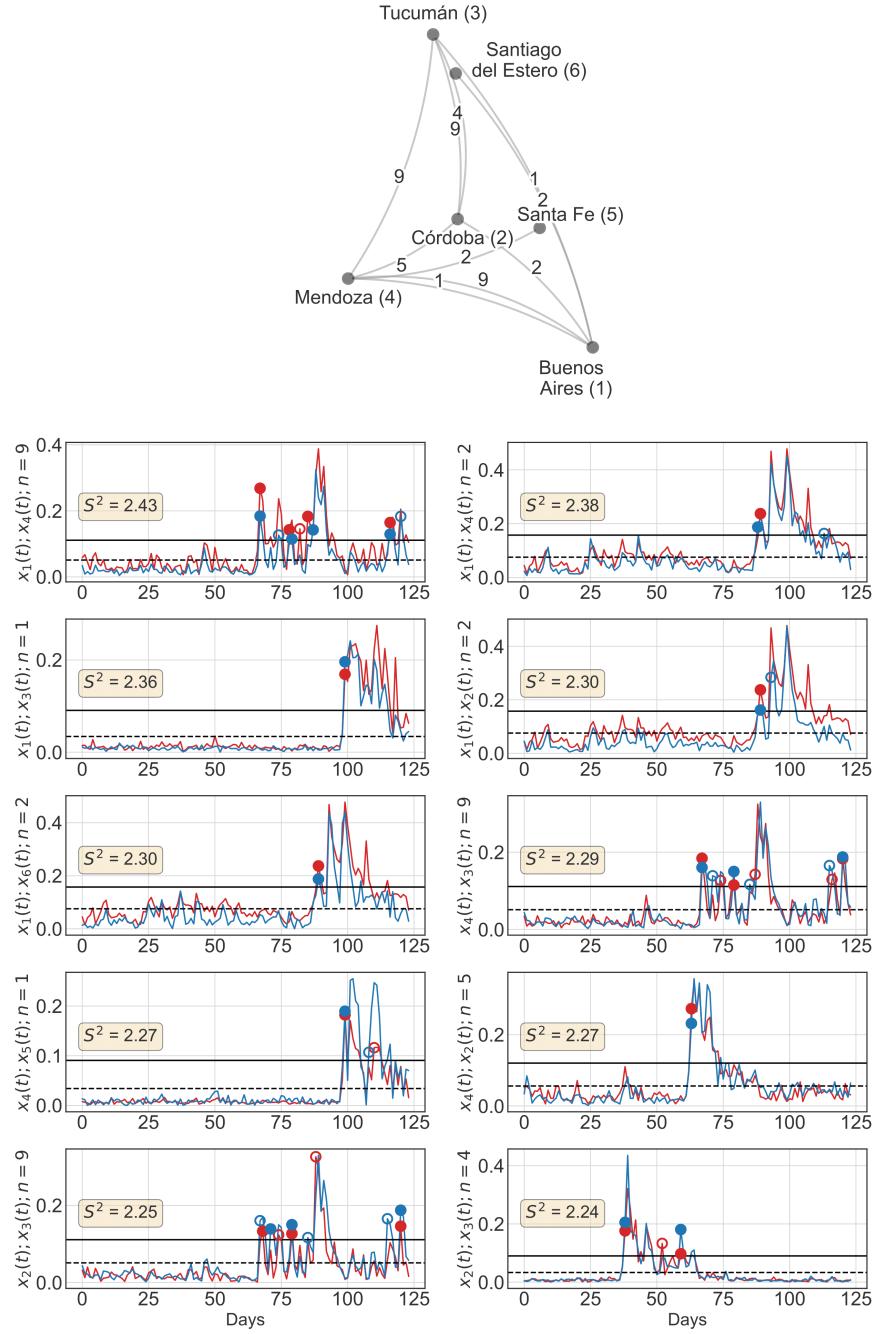


FIG. 13. **Top panel:** Network constructed by selecting the 10 top links which maximizes the squared sum of the symmetric measures. **Bottom panels** display the two time series of each link.

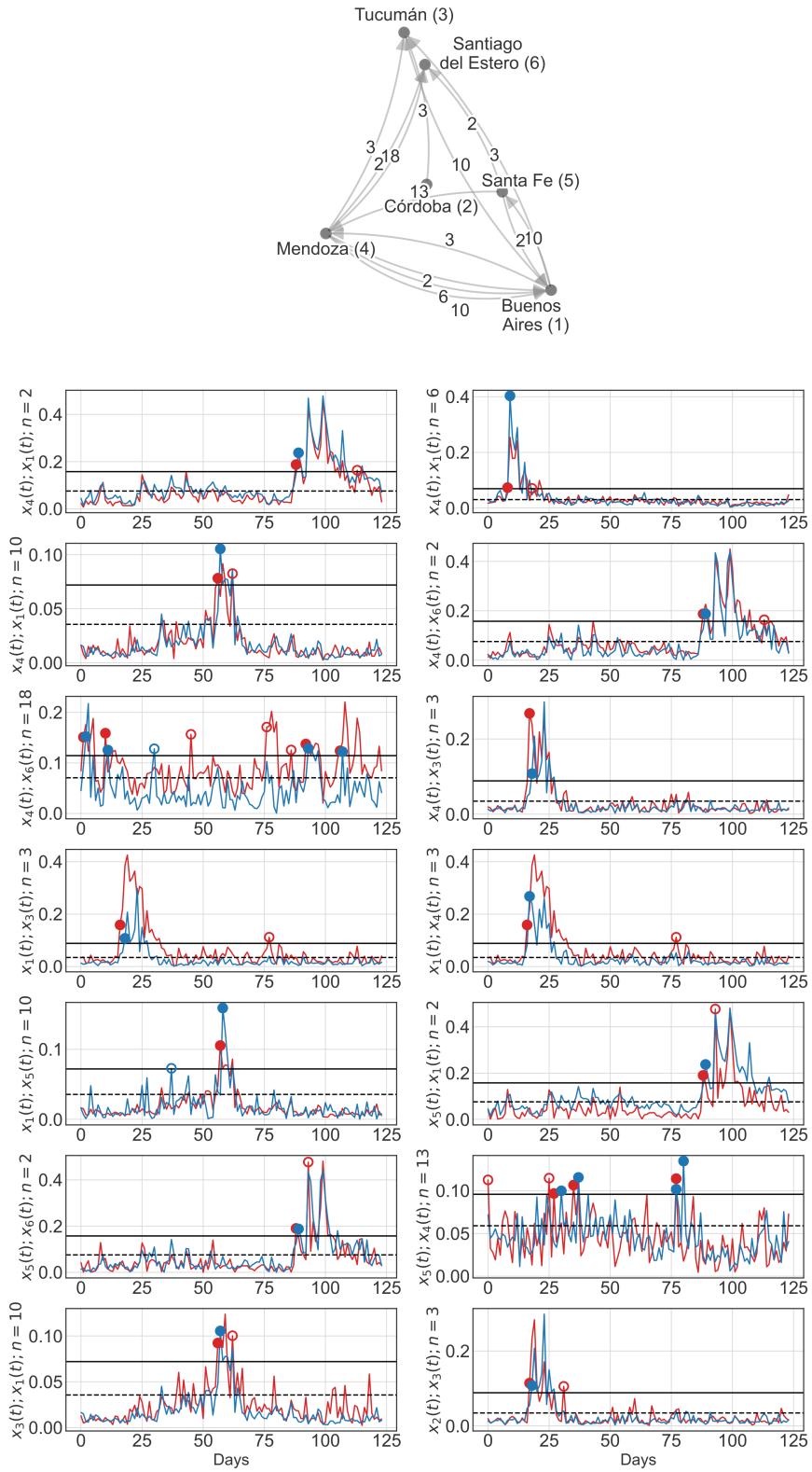


FIG. 14. **Top panel:** Network constructed by selecting the links with Q_a and $Q_s > 0.6$. This yields 14 directed links. **Bottom panels** display the two time series of each link.

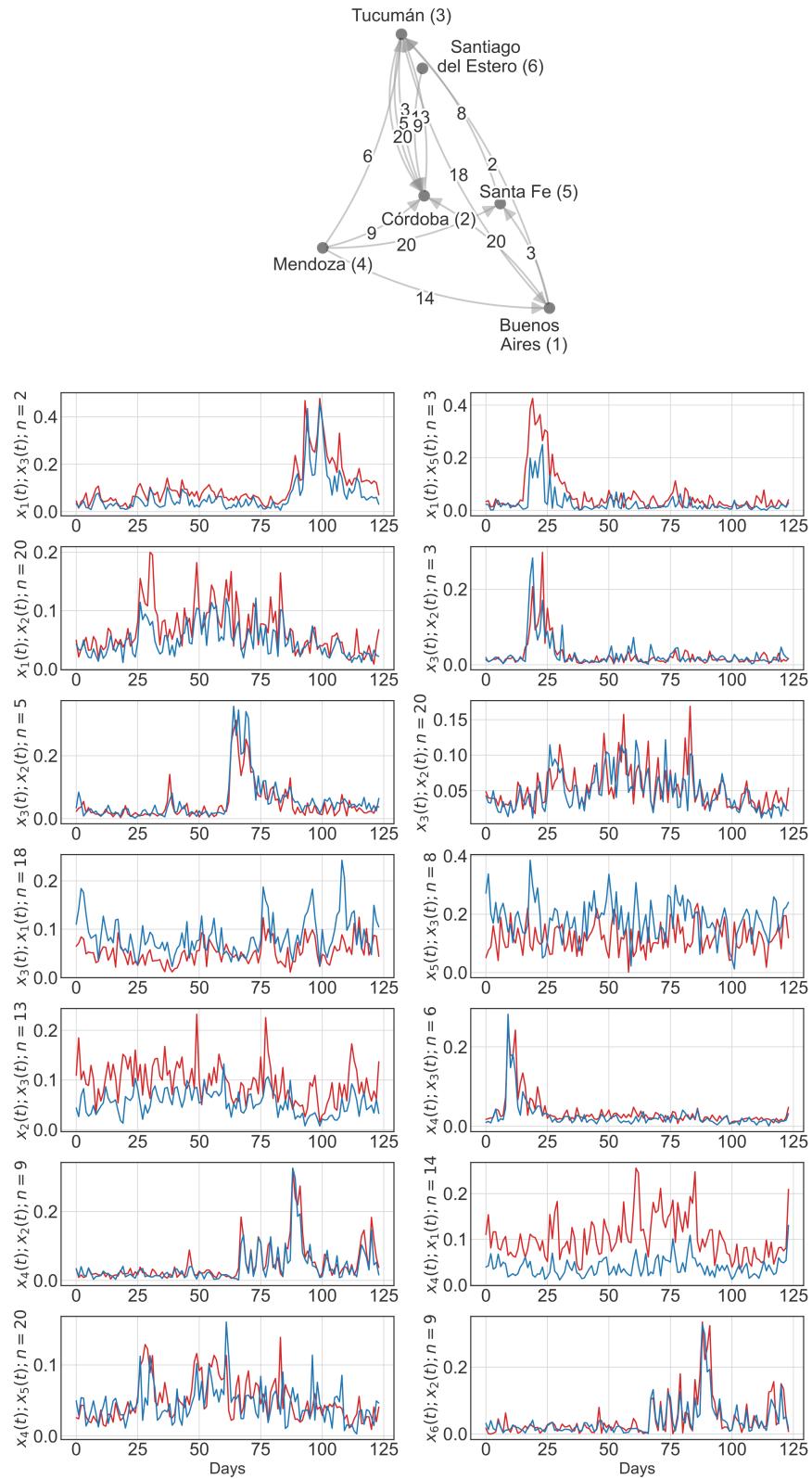


FIG. 15. **Top panel:** Network constructed by selecting the 14 links with highest TE value. **Bottom panels** display the two time series of each link.

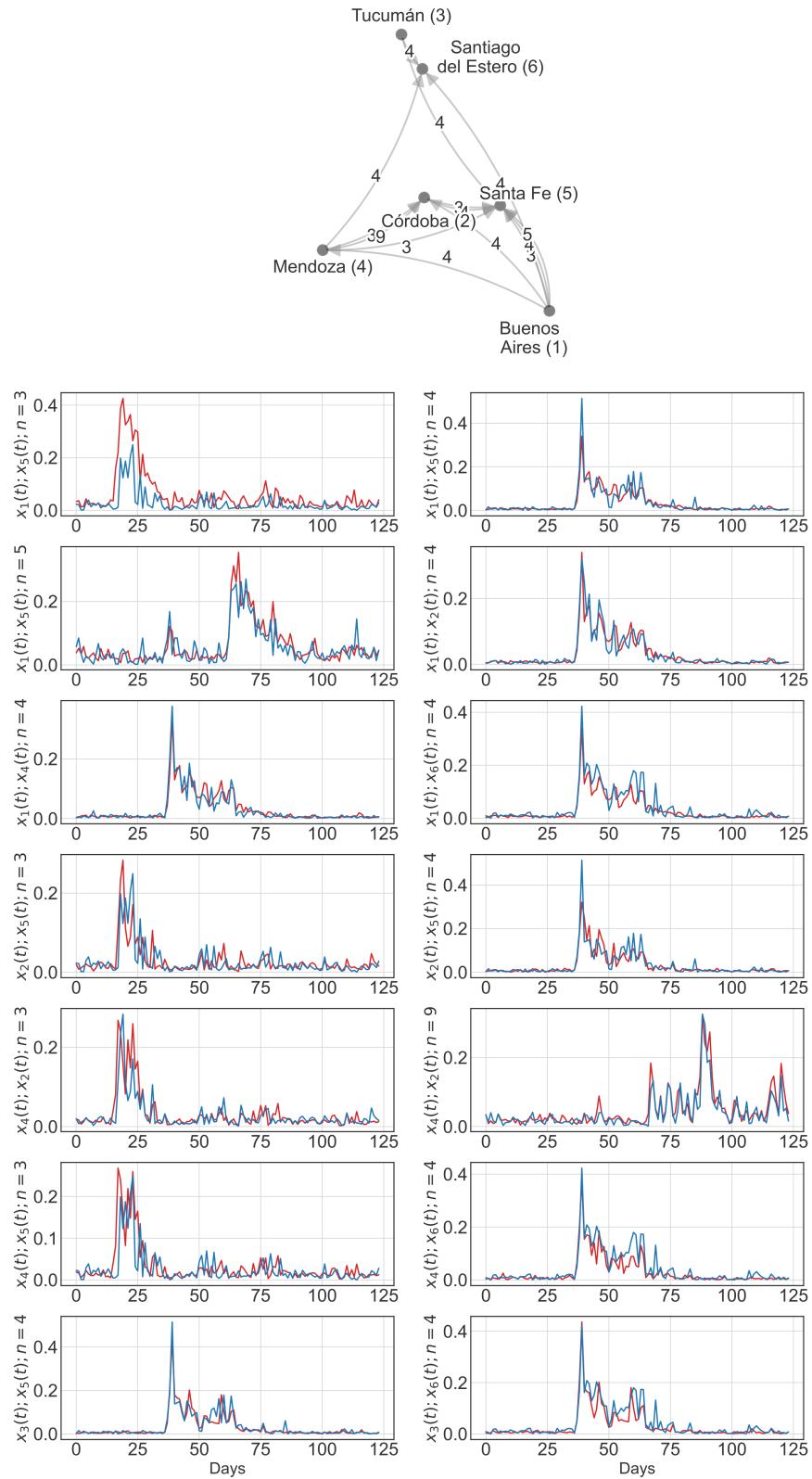


FIG. 16. **Top panel:** Network constructed by selecting the 14 links with highest *GC* value. **Bottom panels** display the two time series of each link.