EPJ .org
● ● ● ● ● ●

O EPJ Data Science

a SpringerOpen Journal

REGULAR ARTICLE   **Open Access**

Check for updates

Evolution of the political opinion landscape during electoral periods

Tomás Mussi Reyero[1], Mariano G. Beiró[1,2]* , J. Ignacio Alvarez-Hamelin[1,2] , Laura Hernández[3] and Dimitris Kotzinos[4]

*Correspondence: mbeiro@fi.uba.ar
[1]Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, C1063ACV Buenos Aires, Argentina
[2]CONICET – Universidad de Buenos Aires. INTECIN, Paseo Colón 850, C1063ACV Buenos Aires, Argentina
Full list of author information is available at the end of the article

Abstract

We present a study of the evolution of the political landscape during the 2015 and 2019 presidential elections in Argentina, based on data obtained from the micro-blogging platform Twitter. We build a semantic network based on the hashtags used by all the users following at least one of the main candidates. With this network we can detect the topics that are discussed in the society. At a difference with most studies of opinion on social media, we do not choose the topics a priori, they emerge from the community structure of the semantic network instead. We assign to each user a dynamical topic vector which measures the evolution of her/his opinion in this space and allows us to monitor the similarities and differences among groups of supporters of different candidates. Our results show that the method is able to detect the dynamics of formation of opinion on different topics and, in particular, it can capture the reshaping of the political opinion landscape which has led to the inversion of result between the two rounds of 2015 election.

Keywords: Social media; Elections; Opinion modelling; Twitter data

1 Introduction

Our understanding of how opinion is formed and evolves in society has benefited since the last decade from the rapidly increasing amount of data diffused by on-line social networks. In this way, large scale, cross-cultural studies that were difficult to perform based on standard off-line surveys become feasible.

In spite of the different biases that are known to affect studies based on on-line social networks, in terms of age, gender, residence location, social status, etc., the enormous amount of information they convey remains useful in particular to detect trends in the evolution of social opinion, at least restricted to the users of these platforms whose amount increases continuously. Moreover nowadays traditional broadcasting media, like radio or television, diffuse information or opinions selected from on-line social networks thus coupling this large but biased set of users with the general population.

The micro-blogging platform Twitter has been widely used in order to study the evolution of social opinion on different topics [1, 2], as well as the properties of the social interaction networks that result from the different functionalities offered by the platform

🐎 Springer

(mentions, retweets, follower-followee) [3, 4]. The intrinsic properties of the platform, like the small size of posts (tweets limited to 280 characters), or its simplicity of usage, (a message can easily be re-transmitted, a mentioned user can be alerted, etc.) make it an excellent tool to study situations where the time scale for the opinion's evolution is short as in electoral processes [1, 5–7] or in social protests [8–10].

The possibility to predict opinion evolution using Twitter, which is of particular interest during an electoral campaign, has been seriously challenged [11–13]. Nevertheless, this kind of studies remain interesting because of their explanatory power. For instance, it was possible to unveil the spontaneous character of the emergence of off-line demonstrations during the Spanish social movement 15M by correlating the intensity of posts at a given location with the different gatherings observed off-line [8]. Also it was possible to identify the origin of a denigrating campaign against one of the potential candidates to the last French presidential election by studying the community structure of a network of retweets [1].

Twitter users share text messages, images and videos, and they may associate their posts to a concept, by the means of *hashtags* (words beginning by the character "#"), in order to install a given concept to be discussed on the public ground.

Among a vast literature devoted to studies of social phenomena based on Twitter data, one can identify those studies that concentrate on the structure of the different networks that can be defined (follower-followee, retweets, mentions and answers), and those which try to infer the opinion dynamics, based on text mining and analysis. Both aspects, structure and content, are nevertheless entangled, as users are mainly exposed to the content produced by those other users that they have decided to follow or by those selected by the algorithms of the platform [3, 4]. This situation has been shown to lead to the phenomenon known as *echo chambers* or *bubbles*, meaning that in fact, in the worldwide open field of Twitter (as well as that of other internet platforms) users are mostly and sometimes uniquely exposed to the same information, frequently the one that comforts their own opinion, thus limiting the possibilities of a real discussion [14–16].

An interesting way to study the structure of opinions in Twitter exploits the hashtags chosen by the users, assuming that this choice reveals a concept that the user wishes to address. In a recent work [17], topics are defined by determining the community structure in a weighted network of hashtags, where two hashtags are connected if they appear together in the same tweet. Assuming that the coexistence of hashtags is semantically meaningful, the community structure of such network can reveal the general topics under discussion. In this way, the users may be characterized by a topic vector, with a dimension equal to the number of communities detected and where each component informs about the interest of the user on the different topics. The authors show that the similarity among users connected by a follower-followee relationship or by a mention relationship is higher on average than the similarity among a sample of random users.

In this work we extend the ideas developed in [17] to a dynamical study of the rapidly evolving opinion landscape that takes place in a society during an electoral campaign. Specifically we understand by *evolution of opinion* the evolution of the preferences of the social actors concerning the topics discussed in the platform, and we are particularly interested in detecting whether specific groups or users synchronize around some specific topics at some given times.

Table 1 Basic statistics for each electoral process. (∗) AAUs: active Argentinian users

	2015 elections	2019 elections
Number of AAUs [∗]	218k	586k
Number of tweets by AAUs captured	50M	280M
Time frame	1 Jul 2015/31 Mar 2016	1 Jan 2019/10 Dec 2019
Primaries	9 Aug 2015	11 Aug 2019
First round	25 Oct 2015	27 Oct 2019
Second round (ballottage)	22 Nov 2015	–

The proposed method allows us to recover the dynamics of the political tendency without introducing questions to the population, which are known to be subject to different bias (of formulation, false declarations, etc.) and *without imposing a priori*, neither an ontology nor the number of topics to be inspected. Our method just extracts the information coded in the data with the only assumption that two hashtags used in the same tweet are semantically related. Our results show that, in spite of the limitations of studies of opinion using Twitter described above, this method is able to capture the opinion evolution of the users with a high enough time-scale resolution so as to detect, for example, the reconfiguration of the political landscape taking place in the short period between the first and second round of the election, which, in one of the cases presented here, overturned the score of the first round of the election.

2 Methods and data set

2.1 Data capture

This study is based on data captured during the two recent Argentinian presidential campaigns, in 2015 and in 2019. The periods of data capture extend from 1 July 2015 to 31 March 2016, and from 1 January 2019 to 10 December 2019, the main elections being held on 25 October 2015 and 27 October 2019 (see Additional file 1 for a detailed description of the electoral processes).

The capture is based on the *active* followers (we define a user as active in Twitter if he/she posted at least one tweet during the first month of capture) of the candidates for president or for deputy-president of each of the main political parties participating in these elections. We filter those users whose profile location is set to some city/province in Argentina, in order to focus on Twitter users that are residents in the country, and we capture and process all their tweets in the period. Table 1 gives a summary of the basic statistics for each dataset.

A detailed description is provided in Additional file 1 regarding the number of tweets captured daily and their classification as original tweets, simple retweets, retweets with comment and replies, and how the supporters are classified into political parties, along with an analysis of the geographical and gender distribution of our user base.

2.2 Definition of topics and user's description vectors

Hashtags are keywords created and chosen by the users, which can be interpreted as representing the engagement of users with events, ideas or different discussion subjects. If two hashtags highly co-occur (i.e., they frequently appear together in the same tweet) it is a reasonable hypothesis to assume a semantic association between them. Following the ideas developed in [17], we build a complex weighted network based on hashtags' co-occurrence. Then, the topics of discussion arise as communities measured on this net-

work, which we detect using the OSLOM algorithm [18]. OSLOM is a local community detection method that builds the communities guided by their statistical significance, which can help distinguish topics of different sizes. This might be relevant to capture the opinion of small political groups, for example. Besides, it can deal with weighted graphs as in this case, and detect overlapping communities (which might here occur when some hashtag is used ambiguously by different users). In Additional file 1 we provide a comparison against the Infomap community detection algorithm for the purpose of this study.

As the topics automatically *emerge* from the community detection algorithm, which is completely agnostic regarding their meaning and does not pre-determine their number, in this work there is no selected ontology. It is worthwhile noticing that the knowledge of the topics' meaning is not required by the method presented here. However, the inspection of the hashtags composing each topic reveals its meaning to a reader understanding the language of the tweets and the social issues of the studied system. As we will show later, this knowledge facilitates the analysis and shows the coherence of the obtained results with the chronology of the political and social events, but is not a pre-requisite of the method.

We describe the interests of each user i by means of a user description vector $\boldsymbol{d_i}$ of dimension N_T, the number of topics (communities) found, which informs about the topic preferences of user i.

This description vector is computed in the following way:

1. We build a user-topic matrix, U, where each element, u_{ij}, gives the absolute number of times that user i has used a hashtag that belongs to the community identified as topic j.

2. We compute the global topic vector $\boldsymbol{T} = \sum_{i}^{N} \boldsymbol{u_i}$, of dimension N_T where $\boldsymbol{u_i}$ is the i-th row vector in the user-topic matrix, and N the size of the population. This vector gives, in each component, the total number of usages of a topic in the population.

3. We define the vector $\boldsymbol{v_i}$ which gives the difference between the frequency of usage of each topic by user i and its global frequency of usage in the population.

$$\boldsymbol{v_i} = \frac{\boldsymbol{u_i}}{\|\boldsymbol{u_i}\|_1} - \frac{\boldsymbol{T}}{\|\boldsymbol{T}\|_1}. \tag{1}$$

Here the norm $\| \cdot \|_1$ must be understood as the sum over all the components in the space of dimension N_T. The vectors of Eq. (1) thus inform about whether user i has addressed each of the identified topics more or less than on average.
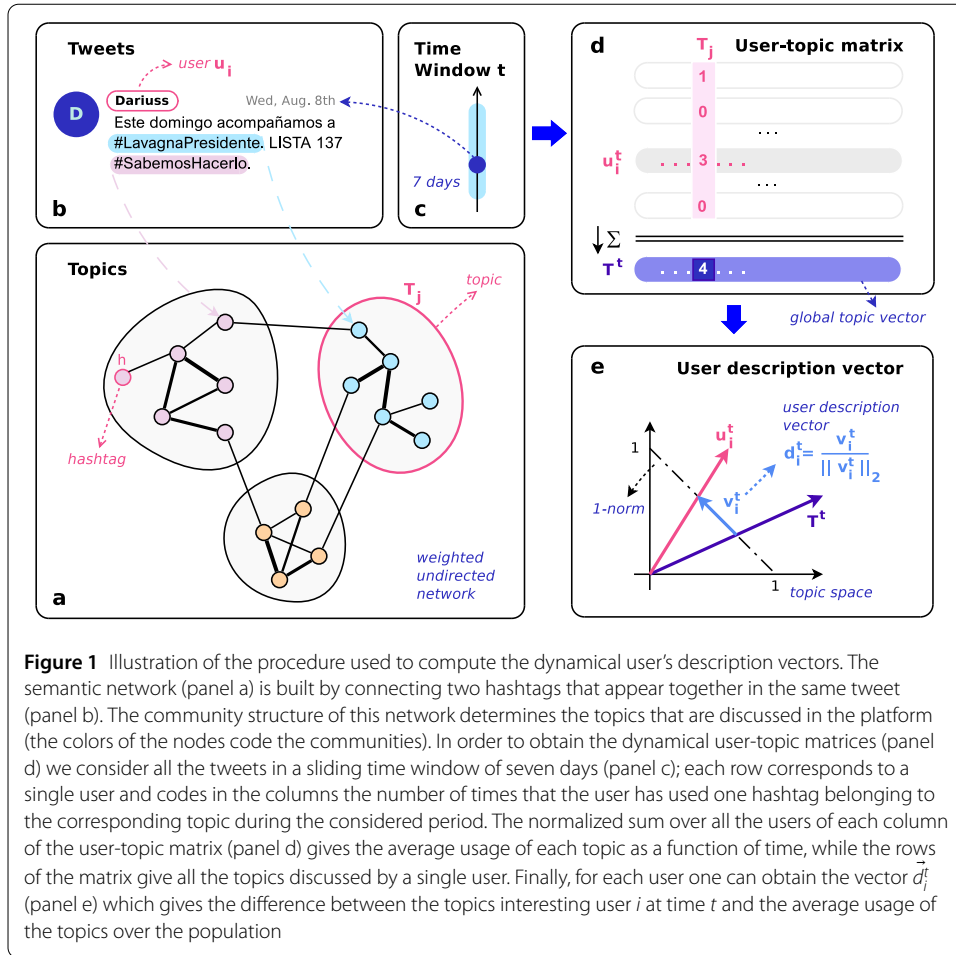
4. As we are only interested in the orientation of the description vectors, they are normalized as:

$$\boldsymbol{d_i} = \frac{\boldsymbol{v_i}}{\|\boldsymbol{v_i}\|_2}, \tag{2}$$

where $\|\boldsymbol{v_i}\|_2$ is the standard euclidean norm in the topic hyperspace of dimension N_T.

Dynamical measurements   In order to track the evolution of the users' interests we apply the aforementioned procedure to sliding time windows of 7 days, thus producing a series of user-topic matrices U_t, one for each day. We shall call $\boldsymbol{d_i^t}$ the description vector for user i at discrete time t.

The full procedure is illustrated in Fig. 1.

Figure 1 Illustration of the procedure used to compute the dynamical user's description vectors. The semantic network (panel a) is built by connecting two hashtags that appear together in the same tweet (panel b). The community structure of this network determines the topics that are discussed in the platform (the colors of the nodes code the communities). In order to obtain the dynamical user-topic matrices (panel d) we consider all the tweets in a sliding time window of seven days (panel c); each row corresponds to a single user and codes in the columns the number of times that the user has used one hashtag belonging to the corresponding topic during the considered period. The normalized sum over all the users of each column of the user-topic matrix (panel d) gives the average usage of each topic as a function of time, while the rows of the matrix give all the topics discussed by a single user. Finally, for each user one can obtain the vector $\vec{d}_i^t$ (panel e) which gives the difference between the topics interesting user i at time t and the average usage of the topics over the population

2.3 Measuring the similarity between groups of users

We define the similarity between a pair of users i and j as the cosine similarity between the corresponding description vectors.

By construction, the cosine similarity detects vectors with similar directions in the hyperspace of dimension N_T. A high (low) value of the cosine similarity between two users reveals that there are common topics that they address with a frequency higher (lower) than the average.

As the description vectors are normalized, the similarity reduces to the inner product:

$$s(i,j) = \langle \boldsymbol{d_i}, \boldsymbol{d_j} \rangle. \tag{3}$$

We also define the average description vector of a group of users G, of cardinal $|G|$:

$$\boldsymbol{D_G} = \frac{\sum_{i \in G} \boldsymbol{d_i}}{|G|}. \tag{4}$$

Now we can introduce two indices measuring *collective similarities*:

- The *cohesion* of a group of users, *intra-group similarity* or *self-similarity*, $s(G, G)$, defined as the average similarity between all its users, and computed in the following

way:

$$s(G, G) = \frac{\sum_{i,j \in G} s(i,j)}{|G|^2} = \frac{\sum_{i \in G} \langle \boldsymbol{d_i}, \boldsymbol{D_G} \rangle}{|G|} = \langle \boldsymbol{D_G}, \boldsymbol{D_G} \rangle = \|\boldsymbol{D_G}\|^2. \tag{5}$$

- The *cross-group similarity* is the average similarity between members of different groups G_1 and G_2, namely $s(G_1, G_2)$:

$$s(G_1, G_2) = \frac{\sum_{i \in G_1, j \in G_2} s(i,j)}{|G_1| \cdot |G_2|} = \frac{\sum_{i \in G_1} \langle \boldsymbol{d_i}, \boldsymbol{D_{G_2}} \rangle}{|G_1|} = \langle \boldsymbol{D_{G_1}}, \boldsymbol{D_{G_2}} \rangle. \tag{6}$$
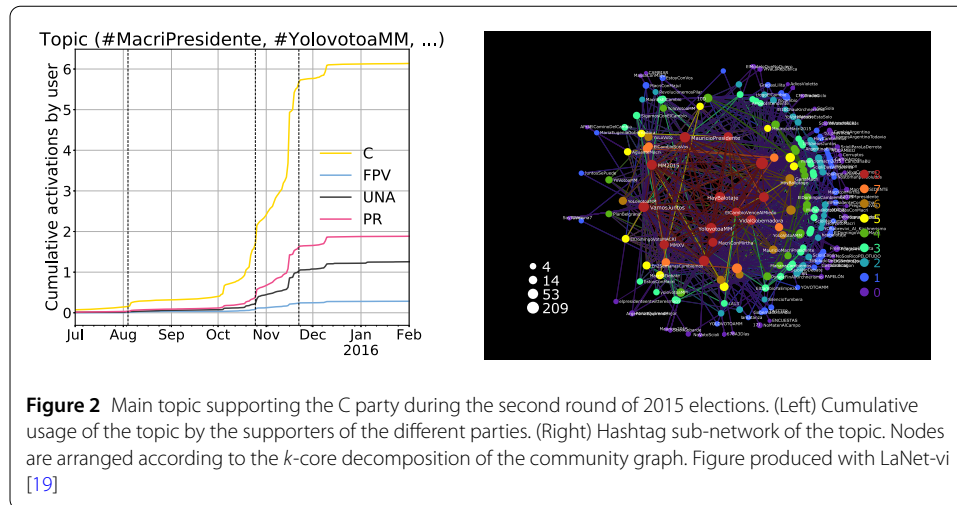
3 Results

The results presented in this section are based on tweets collected as described in Sect. 2, for the two last presidential elections in Argentina. Table 2 defines the acronyms of the political parties intervening in each election, along with a rough characterization of their position in the political spectrum and their performance. Details of the retrieval and cleaning methods of the data-set can be found in Additional file 1.

As explained in Sect. 2.2, we built the semantic network with the assumption that hashtags used in the same tweet carry some semantic similarity. The community structure of such network reveals the topics that are discussed in the society, and the description vectors allow us to characterize the interests of each user, in the topic space.

In Fig. 2 we show as an example the structure of a topic (right panel) composed of hashtags supporting the *Cambiemos* party (C), one of the two major parties in the second round of 2015 election. As the topics emerge from the community analysis without any a priori information introduced into the system (they are arbitrarily labelled by a number), it is the inspection of the hashtags composing the community that informs about the subject to which the topic is related. On the left panel we show the cumulative number of supporters of each political party referring to that topic. This shows that, although it is not always true that people choose a hashtag only to support the idea it conveys (notice that members of other parties, including the strongest opponent, FPV, also use the considered topic), on

Table 2 Acronyms and characteristics of the main contenders of 2015 (top) and 2019 (bottom) presidential elections

Y	C	Party Name	Acron.	Political/econ. orientation	PASO	1st Rnd	2nd Rnd
2015	🔵	Frente para la Victoria (in power)	FPV	Center-left, econ. interventionist	38.67%	37.08%	48.66%
	🟨	Cambiemos (challenger)	C	Center-right, econ. liberal	30.12%	34.15%	51.34%
	⬛	Unidos por una Nueva Alternativa	UNA	Economically interventionist	20.57%	21.39%	–
	🟥	Progresistas	PR	Socially and econ. liberal	3.47%	2.51%	–
2019	🟨	Juntos por el Cambio (in power)	JPC	Center-right, econ. liberal	31.80%	38.67%	
	🔵	Frente de Todos (challenger)	FDT	Center-left, econ. interventionist	47.79%	48.24%	
	🟦	Consenso Federal	CF	Center, econ. liberal	8.15%	6.14%	
	🟥	Frente de Izquierda	FI	Leftist party	2.83%	2.16%	
	⬛	Frente Despertar	FD	Right party	2.16%	1.47%	

Figure 2 Main topic supporting the C party during the second round of 2015 elections. (Left) Cumulative usage of the topic by the supporters of the different parties. (Right) Hashtag sub-network of the topic. Nodes are arranged according to the *k*-core decomposition of the community graph. Figure produced with LaNet-vi [19]

average, our method does correctly capture the expected preferences of the users. A similar description of a topic in support of the largest opponent party, FPV, can be found in Additional file 1.

Argentinian law imposes to the citizens the obligation to vote. As a consequence, not only high participation rates are observed but also political discussions occupy a significant part of the public attention. In both elections the political opinion was highly polarized, with two main antagonist parties dominating the political spectrum. Other two or three smaller parties may, on certain occasions, play the role of a pivot for the determination of the final result; therefore understanding the evolution of the opinion of their supporters is a crucial issue. We will see that this was the case of 2015 election, where a second round was necessary to determine the winner. A second round only takes place if no party obtains either *(i)* more than 45% of the votes or; *(ii)* more than 40% and a 10% difference with the second most voted party.

Unlike in 2015 election, no second round took place in 2019. In fact that year, the primary elections played the role of a first round. The electoral rule establishes that the primary elections, called PASO, *Primarias Abiertas, Simultaneas y Obligatorias* meaning "open, simultaneous, and compulsory primary elections", take place simultaneously and all the competing parties are bound to present at least one candidate. Due to the particular political configuration of that moment, no party took the risk to divide its votes into several candidates, and therefore they presented one single candidate each. Under such circumstances, these primaries were seen as rehearsal of the general election.

Both situations are excellent case studies for this work because they present a short time window with a fast dynamics of political opinion, and a wide variety of subjects of discussion (with or without political character) that capture the attention of Twitter users. We will show that our method is able to detect in the data the reconfiguration of the opinion landscape, as different groups of users favours more or less some discussion topics, for two different elections corresponding to different political situations.

3.1 2019 elections

The two main parties intervening in this election were *Juntos por el Cambio* (JPC), whose candidate was the incumbent, and the challenger (and previous ruler) *Frente de Todos* (FDT).
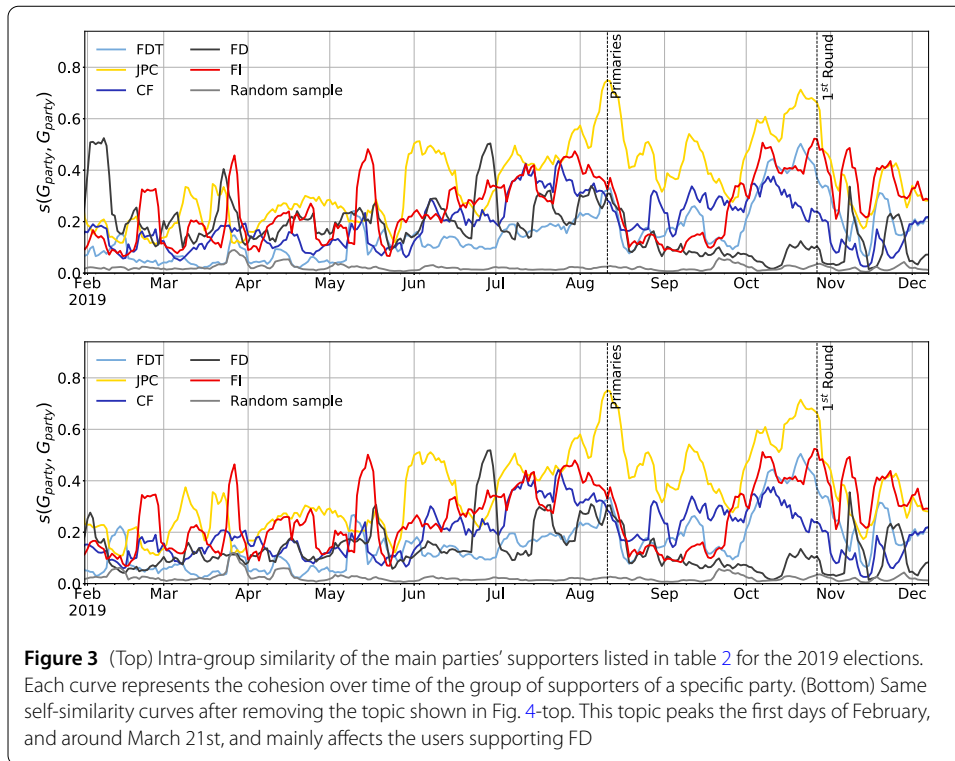
Figure 3 (Top) Intra-group similarity of the main parties' supporters listed in table 2 for the 2019 elections. Each curve represents the cohesion over time of the group of supporters of a specific party. (Bottom) Same self-similarity curves after removing the topic shown in Fig. 4-top. This topic peaks the first days of February, and around March 21st, and mainly affects the users supporting FD

Figure 3 shows the self-similarities for the parties participating to the 2019 election as a function of time, together with the self-similarity among a randomized sample containing the same number of users from each party, which serves as a baseline. The ruling party shows, in general, a higher self-similarity than the others which increases as the PASO approaches, reflecting the strong cohesion of its supporters. Interestingly, peaks of strong self-similarity are also observed in the curves corresponding to the two minority parties at both extremes of the political spectrum, FI and FD, at different points in time. This is the signature of a coherent reaction among the supporters of one party, probably to some event in real life, while for all the other users that event does not trigger a particular reaction. It is plausible to assume that this happens when the event is in resonance with the political traditions of one of these parties.

Let us recall that the description vectors of the users contain a large diversity of topics, many of which do not have a political character. When the public discussion is dominated by one of these topics (for instance a football championship) the differences among the supporters of different parties may be partially and temporarily washed out. This effect is enhanced far from the electoral dates, where we can observe that all the parties fluctuate around the same value of self-similarity, except for the isolated peaks already mentioned.

In order to further investigate what are the topics behind the observed isolated peaks, there are two approaches. On the one hand, if one has some insight about the important events or discussions in the society at the date of the peak, it is possible to proceed to a careful inspection of the dominant topics at the date of the peak. This can be done using the platform we have created to analyse the evolution of the different topics [20]. Let us consider, as an example, the two nearby sharp peaks observed in Fig. 3 by the end of March 2019 which correspond to the curves of FD and FI (black and red respectively).
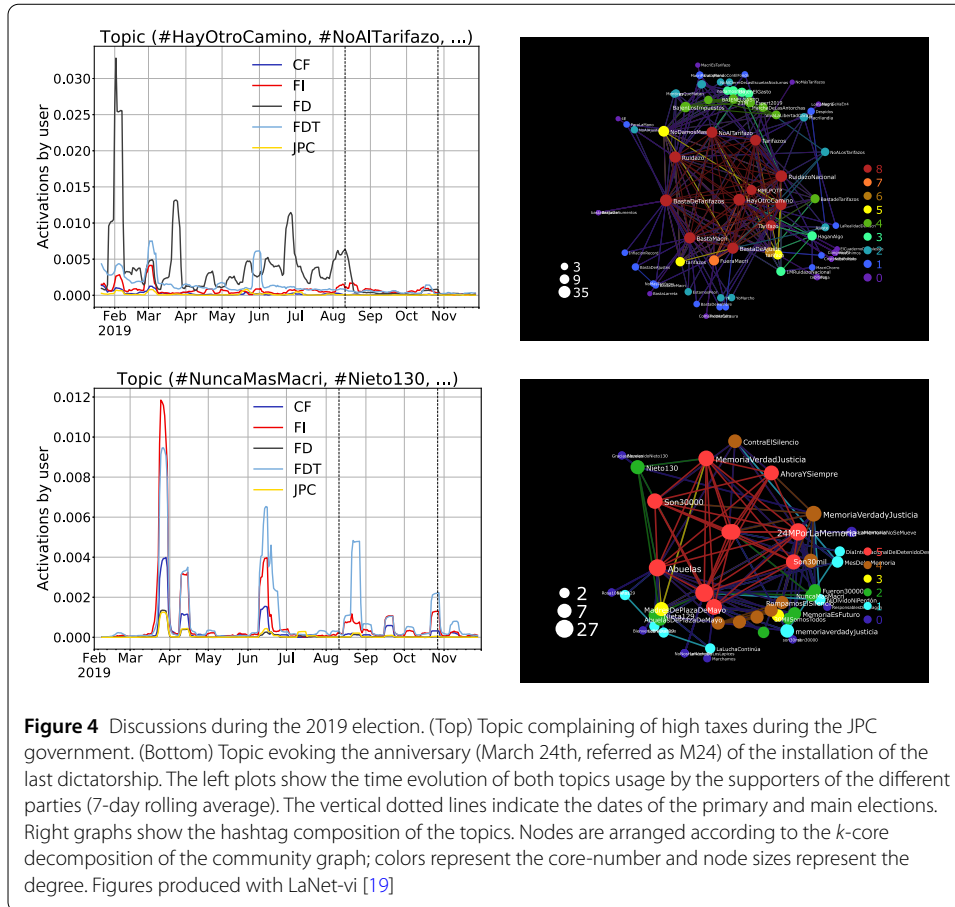
The peak of the FD self-similarity curve (in black) in the top panel of Fig. 3 occurs in March 21st 2019, coincident in time with an important demonstration against tax rises which took place in front of the National Congress. A reasonable conjecture is to hypothesize that this event in real world could have triggered discussions on-line that synchronize some particular users. The accuracy of such guess can be checked by removing the usages of hashtags in that topic from the user-topic vectors, recomputing the similarities and observing the modifications (if any) of all the curves. In this case the peak of FD (see the bottom panel of Fig. 3) disappears, showing that it was the tax topic which was responsible of that peak. Moreover, with this method we detect another peak in the FD curve, at the beginning of February, also disappearing, which reveals that this group was also active about the topic at that early date. We also observe that this topic only synchronizes the activity of FD followers as no significant changes are seen on the rest of the self-similarity landscape (compare both panels of Fig. 3). The tax rise is a subject that usually interests right and liberal parties, however we observe that JPC supporters (mainly liberals) did not synchronize around this topic, which can be due to the fact that their party, being in power, was responsible for the tax rise. This can be checked by an inspection of the topics that interested JPC supporters at that time using the platform [20] where it is possible to confirm that the smaller peak of the JPC curve (yellow) observed in Fig. 3, does not correspond to the tax topic.

Concerning the neighbouring peak in the FI curve (in red in Fig. 3), it is situated around March 24th. This date corresponds to the remembrance of the victims of the last dictatorship in Argentina (1976-1983), a subject of main concern for the leftist parties (FI and FDT). The bottom panel of Fig. 4 depicts the temporal behaviour of the corresponding topic. It can be seen that FI and FDT are the most concerned with the topic during that day, while the more conservative JPC and FD (right wing) are scarcely active in the topic. Interestingly this topic is later reactivated periodically, but mainly due to the activity of FDT, which is a major, composite party including an important left wing. Again, by removing the usages of hashtags in this topic, we can verify that the peak in the self-similarity was effectively due to it (we provide this comparison in Additional file 1).

So, the inspection of the topics shows that the three peaks observed very near the end of March 2019 in the self-similarity curves of FD, FI, and FDT supporters correspond to different discussions. In this way the evolution of the self similarity captures the synchronization of some groups around important topics discussed on the platform. Other significant peaks have been equally identified and the interested reader can find a more detailed description in Additional file 1, as well as a dynamical inspection of the topics in the platform for topic visualisation and analysis that we have created [20].

On the other hand, without any insight on the determinant events occurring in the studied society, no guess can be made. Nevertheless, it is worthwhile noticing that our method allows for an automatic and agnostic detection of the topics around which the groups synchronize, leading to the peaks in the self similarity. It is enough to systematically remove one topic at the time and monitor the height of the self-similarity curve, as has been described above for the selected topics responsible of the peaks in March (Fig. 3). When the right topic has been hit, the chosen peak, and eventually few others are strongly modified, thus signaling the discussion that has synchronized the corresponding group of users.

Cross-similarity curves reveal other aspects of the evolution of the opinion during the electoral process. The brown curve in the top panel of Fig. 5 shows the cross-similarities

Figure 4 Discussions during the 2019 election. (Top) Topic complaining of high taxes during the JPC government. (Bottom) Topic evoking the anniversary (March 24th, referred as M24) of the installation of the last dictatorship. The left plots show the time evolution of both topics usage by the supporters of the different parties (7-day rolling average). The vertical dotted lines indicate the dates of the primary and main elections. Right graphs show the hashtag composition of the topics. Nodes are arranged according to the *k*-core decomposition of the community graph; colors represent the core-number and node sizes represent the degree. Figures produced with LaNet-vi [19]

between the two main antagonistic parties in 2019, JPC and FDT, compared to the respective self similarities (i.e., the same curves shown in Fig. 3), plotted as a reference. As expected, we observe that the higher the self-similarities of the groups, the lower the cross-similarity among them. The cross-similarity strongly decreases in the vicinity of the primary and the main elections.

The inspection of the dynamical behaviour of the cross-similarities between supporters of the minority parties with the two major ones is particularly interesting, as they show to what extent the subjects of interest of the minority parties shift to those of the winner near the final round of the election, revealing their contribution to the final outcome. The remaining panels in Fig. 5 show a clear difference of behaviour among smaller parties. In the second panel, while FD (Right party) has, most of the time and in particular near the electoral periods, a positive cross-similarity with the JPC in power, it holds almost always a negative cross-similarity with the challenger (FDT). Interestingly by mid-november 2019 this tendency is reverted, and a (small) positive similarity with FDT is observed at the same point where a strong self-similarity appears for the supporters of FD party. An inspection of the active topics of the moment reveals that this activity is related with Latin-american international politics, which mobilizes both parties but on opposite sides of the opinion spectrum. This shows that although we do not perform any sentiment analysis here, the attention of the groups about a given subject is correctly captured independently of the opinion of the users on that subject.
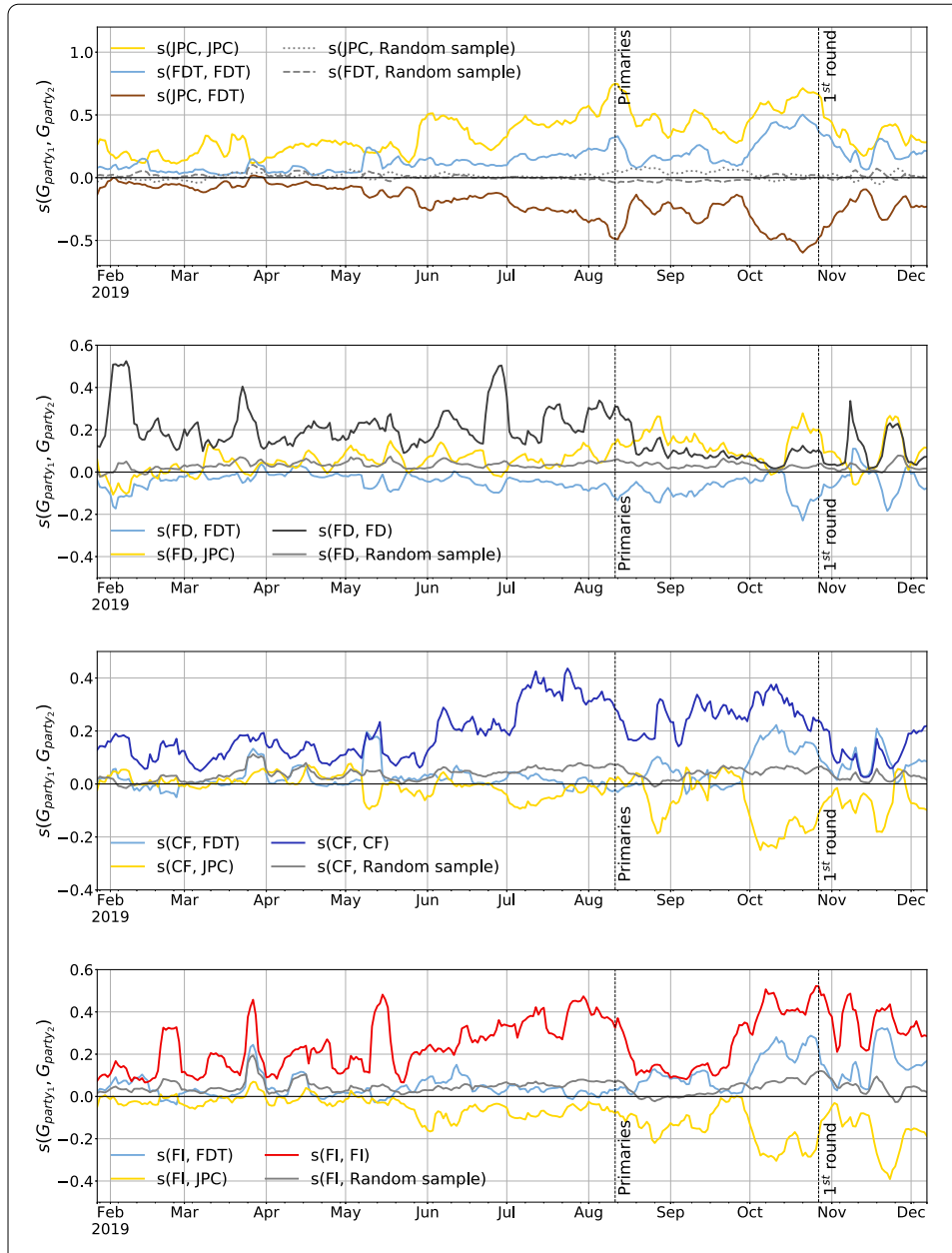
Figure 5 Cross-group similarity between supporters of different parties in 2019. In the top panel, the brown curve represents the cross-similarity between the supporters of the two main competing parties, JPC and FDT. The self-similarity of each party is also shown for the sake of comparison (yellow and light blue curves). The 3 remaining panels show the cross-similarity between the supporters of each of the minority parties (FD, CF and FI, respectively) and those of the two major ones in 2019, JPC (in yellow) and FDT (in light blue). In each of these panels the self-similarity of the corresponding minority party (FD, CF and FI) is reproduced as a reference (black, blue and red curves, respectively)

On the contrary, the supporters of CF (dissident liberal party) show a complete different behaviour. Both cross-similarities with the two major parties are quite low and fluctuating far from electoral periods. This tendency remains before the primary elections showing the strong support of their own candidate. However, before the first round a strong positive similarity with FDT develops (after some fluctuating period), this sudden change of

preferences of the CF supporters signals that during this period, their interests are in line with those of the challenger FDT, whom they supported from the first round.

Finally the bottom panel in Fig. 5 shows the cross-similarity between the FI party and the two main competitors. Although the leftist party has a very small influence in the country, the interest of this curve is to reveal that indeed its cross-similarity with FDT is clearly positive (and negative with JPC). This is a clear evidence of the existence of a strong leftist component in FDT, as mentioned above, which is completely absent in JPC.

3.2 2015 elections

The discussion of the results concerning this previous election, is interesting not only as a test of the validity of our method in a different political context, but also because unlike in 2019, this election required two rounds to establish the winner. Even more interesting, the rank of the two first qualified parties was overturned in the second round. We will show that our method is able to identify the details of the evolution of the opinion of the supporters of the smaller parties towards that of the final winner, contributing to overturn the result of the first round.

The main political parties intervening in this election, which are formally different from those of 2019 election, although there are important overlaps, are listed in Table 2.

The dynamics of the self-similarities, as well as that of the cross-similarity between the two largest parties, follow a similar pattern as those of 2019 elections and are detailed in Additional file 1.

The most interesting feature of this electoral period is revealed by the cross-similarities between each one of the two smaller parties against the two leaders. Figure 6 shows the cross-similarity of the two small parties, PR (top) and UNA (bottom) with the two leaders, FPV (blue curve) and C (yellow curve), along with the self-similarity of the corresponding
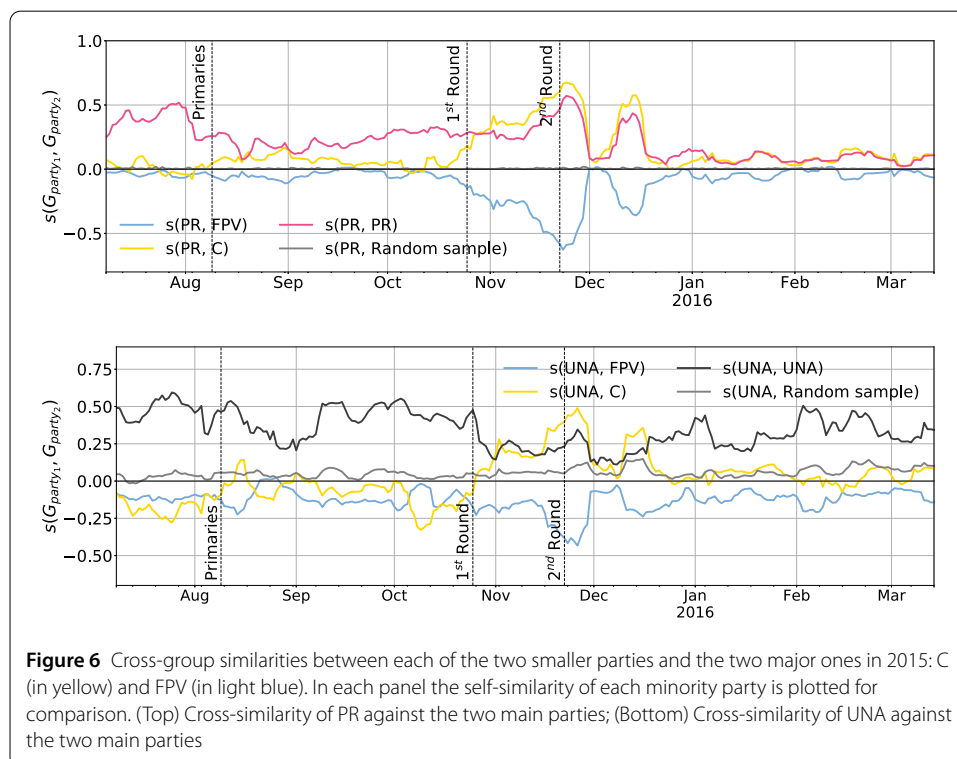

Figure 6 Cross-group similarities between each of the two smaller parties and the two major ones in 2015: C (in yellow) and FPV (in light blue). In each panel the self-similarity of each minority party is plotted for comparison. (Top) Cross-similarity of PR against the two main parties; (Bottom) Cross-similarity of UNA against the two main parties

small party, for comparison. This analysis shows a small positive (negative) cross-similarity of the PR supporters with the C (FPV) party, compared to their own self-similarity. This reveals than out of the electoral period, the interests of the PR supporters have little overlap with either of the dominant parties. However, as the first round approaches, the behaviour of the PR changes and they clearly show a community of interests with the supporters of C.

A similar behaviour, though more enhanced, happens with the UNA supporters. With a negative cross-similarity against both major parties before the elections, it is between the two rounds that the cross-similarity with the C party suddenly strongly increases, showing a clear change on the topics chosen by the UNA supporters. This change in the alignment of the UNA between the two rounds turned out to be decisive to 2015 election results and the victory of C party [21], and in Additional file 1 we show that this change in similarity was in effect due to the topic in support for C.

4 Summary and conclusions

We have performed a study of the dynamics of opinion during an electoral process, based on data obtained from the micro-blogging platform Twitter. While this subject has been explored in several works [1, 5–7, 22–25], here we apply a different, user-centered, perspective of the discussions that are taking place in the platform. Most previous works on the subject define a set of keywords, hashtags or mentioned users (e.g., political candidates) to be tracked, and thus they obtain a dataset of tweets which are inherently political. Instead, by defining a set of seed users and capturing all the content that their followers generate, we have information about the evolution of the users' opinion on different topics, and we are not only restricted to a subset of their tweets.

Following Ref. [17], the topics to study are not set a priori, but emerge from the community structure of a semantic network. This network is built with the assumption that two hashtags used in the same tweet carry some semantic relationship. The disclosed communities provide a representation of the topics under discussion in the society in a multidimensional topic space. In this work we add the temporal dimension to the topic vectors, and therefore we are able to study the dynamics of the opinion of party supporters during the electoral campaign with great detail.

As discussed in the introduction, the known biases of the population using Twitter to foster political discussions, which lead among others, to an over representation of an urban male population [26, 27] hampers the possibility of prediction of electoral results. Instead, we show here how we can follow the evolution of political opinion through the different stages of an electoral period. The case of the 2015 elections in Argentina shows that our method captures the details of the reshaping dynamics of the opinion that took place between the two rounds of the election where the initial result was overturn.

Although we cannot expect to predict the outcome of an election, one could still attempt to detect massive opinion changes on *real time*. In this respect, it is worthwhile recalling some technical details in order to understand the possibilities and the limitations of the method developed here. In this work, the topics are determined by the community analysis of an *aggregated* semantic network, meaning that it has been built using the tweets collected during the whole electoral period. Tracking the semantic network in real time has the drawback of starting with a small network, with new hashtags entering as time evolves, which could hamper the correct initial determination of the topics by lack of data.

A compromise situation would be to start by a semantic network aggregated during some initial period, in order to set the terms of the public discussion. From that starting point, one could then incorporate the new hashtags to the existing semantic network, following a sliding time window. In this way, the topics could be recalculated and the analysis of the similarities could be performed *almost* in real time with just a small lag. It is expected that the more hashtags enter the semantic network the more accurate will be the opinion landscape mapping. This assumption lays on the implicit hypothesis that the system tends to a steady (or meta-stable) state compatible with the aggregated network. However, this may not always be the case. It is easy to imagine that some event, like the present Covid19 pandemic, introduces at some point in time an important, short-time scale, modification of the semantic network. In such cases, integrating new hashtags as described above, may capture extreme patterns that could be caused by the appearance of a rare event triggering the modification of the structure of the semantic network, almost in real time. This work is in progress.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-021-00285-8.

Additional file 1. Supplementary information (PDF 1.5 MB)

Availability of data and materials
The tweet id's of the tweets produced by the active argentinian users (AAUs) during both captures are available under request, as well as the daily user-topic matrices (with anonymized user id's).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JIAH, MGB and LH conceived the research project. TMR and MGB wrote the code for extracting the hashtag network, computing the matrices and the evolving similarities, and produced the figures. MGB and LH wrote the manuscript. JIAH, MGB, LH and DK analysed and discussed the results. All authors contributed in proofreading and discussing the manuscript. All authors read and approved the final manuscript.

Author details
[1]Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, C1063ACV Buenos Aires, Argentina. [2]CONICET – Universidad de Buenos Aires. INTECIN, Paseo Colón 850, C1063ACV Buenos Aires, Argentina. [3]Laboratoire de Physique Théorique et Modélisation, UMR-8089 CNRS, CY Cergy Paris Université, 2 Av. Adolphe Chauvin, 95302 Cedex, France. [4]ETIS UMR-8051 CY Cergy Paris Université, ENSEA, CNRS, 2 Av. Adolphe Chauvin, 95302 Cedex, France.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Gaumont N, Panahi M, Chavalarias D (2018) Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 french presidential election. PLoS ONE 13(9)
2. Boutet A, Kim H, Yoneki E (2013) What's in Twitter, I know what parties are popular and who you are supporting now! Soc Netw Anal Min 3(4):1379–1391

3. Himelboim I, Smith M, Shneiderman B (2013) Tweeting apart: applying network analysis to detect selective exposure clusters in Twitter. Commun Meth Meas 7(3–4):195–223

4. Barberá P (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Polit Anal 23(1):76–91

5. Ahmed S, Jaidka K, Cho J (2016) The 2014 Indian elections on Twitter: a comparison of campaign strategies of political parties. Telemat Inform 33(4):1071–1087

6. Caldarelli G, Chessa A, Pammolli F, Pompa G, Puliga M, Riccaboni M, Riotta G (2014) A multi-level geographical study of italian political elections from twitter data. PLoS ONE 9(5)

7. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Fourth international AAAI conference on weblogs and social, Media

8. Borge-Holthoefer J, Rivero A, García I, Cauhé E, Ferrer A, Ferrer D, Francos D, Iniguez D, Pérez MP, Ruiz G et al (2011) Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. PLoS ONE 6(8)

9. Alvarez R, Garcia D, Moreno Y, Schweitzer F (2015) Sentiment cascades in the 15m movement. EPJ Data Sci 4(1):6

10. Howard PN, Duffy A, Freelon D, Hussain MM, Mari W, Maziad M (2011) Opening closed regimes: what was the role of social media during the arab spring? Available at SSRN 2595096

11. Murthy D (2015) Twitter and elections: are tweets, predictive, reactive, or a form of buzz? Inf Commun Soc 18(7):816–831

12. Gayo-Avello D (2012) No, you cannot predict elections with Twitter. IEEE Internet Comput 16(6):91–94

13. Chung JE, Mustafaraj E (2011) Can collective sentiment expressed on Twitter predict political elections? In: Twenty-fifth AAAI conference on artificial intelligence

14. Kelly J, Francois C (2018) A vision of division. MIT Technol Rev 121(5):22–27

15. Nikolov D, Oliveira DF, Flammini A, Menczer F (2015) Measuring online social bubbles. PeerJ 1:38

16. Eady G, Nagler J, Guess A, Zilinsky J, Tucker JA (2019) How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. Sage Open 9(1):2158244019832705

17. Cardoso FM, Meloni S, Santanche A, Moreno Y (2019) Topical alignment in online social systems. Front Phys 7:58

18. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS ONE 6(4)

19. Alvarez-Hamelin JI, Alain Barrat AV, Dall'Asta L, Beiró MG LaNet-vi software. https://lanet-vi.fi.uba.ar/index.php

20. Guerrero F, Schapira M, De Rosa R, Alvarez-Hamelin JI, Beiró MG Argentinian elections 2019 online platform. http://elecciones2019.fi.uba.ar/

21. La Nación Massa D "Hay Una Enorme Mayoría de Los Que Nos Votaron Que Va a Votar a Macri", 2015-11-18. (accessed on 2021-04-20). https://www.lanacion.com.ar/politica/massa-hay-una-enorme-mayoria-de-los-que-nos-votaron-que-van-a-votar-a-macri-nid1846618/

22. Varol O, Ferrara E, Ogan CL, Menczer F, Flammini A (2014) Evolution of online user behavior during a social upheaval. In: Proceedings of the 2014 ACM conference on web science, pp 81–90

23. Ma Z, Sun A, Cong G (2013) On predicting the popularity of newly emerging hashtags in t witter. J Am Soc Inf Sci Technol 64(7):1399–1410

24. Zhang X, Chen X, Chen Y, Wang S, Li Z, Xia J (2015) Event detection and popularity prediction in microblogging. Neurocomputing 149:1469–1480

25. Bovet A, Morone F, Makse HA (2018) Validation of Twitter opinion trends with national polling aggregates: hillary clinton vs Donald trump. Sci Rep 8(1):1–16

26. Vaccari C, Valeriani A, Barberá P, Bonneau R, Jost JT, Nagler J, Tucker J (2013) Social media and political communication. A survey of Twitter users during the 2013 Italian general election. Riv. Ital. Sci. Polit. 43(3):381–410

27. Barberá P, Rivero G (2015) Understanding the political representativeness of Twitter users. Soc Sci Comput Rev 33(6):712–729