Characterizing community-changing users using text mining and graph machine learning on Twitter

Federico Albanese[1], Esteban Feuerstein[1], Leandro Lombardi[2] and Pablo Balenzuela[3]

[1]*Instituto de Ciencias de la Computación, CONICET- Universidad de Buenos Aires, Buenos Aires, Argentina*
[2]*Medallia, Buenos Aires, Argentina*
[3]*Instituto de Física de Buenos Aires (IFIBA), CONICET, Argentina*

Abstract

Even though the Internet and social media have increased the amount of news and information people can consume, most users are only exposed to content that reinforces their positions and isolates them from other ideological communities. This environment has real consequences with great impact on our lives like severe political polarization, easy spread of fake news, political extremism, hate groups and the lack of enriching debates, among others. Therefore, encouraging conversations between different groups of users and breaking the closed community is of importance for healthy societies. In this paper, we characterize and study users who change their community on Twitter using natural language processing techniques and graph machine learning algorithms. In particular, we collected 9 million Twitter messages from 1.5 million users and constructed retweet networks. We identified their communities and topics of discussion associated with them. With this data, we present a machine learning framework for social media users classification which detects users that swing from their closed community to another one. A feature importance analysis in three Twitter polarized political datasets showed that these users have low values of PageRank, suggesting that changes in community are driven because their messages have no resonance in their original communities.

Keywords

Social Media, text mining, graph learning, communities

1. Introduction

People with different political opinions and diverse backgrounds interact on social networks. However, this diversity does not translate to enriching debates between users with different profiles because they tend to cluster according to their beliefs, constituting homogeneous communities known as echo chambers [16]. Aruguete et al. focused on the interaction between users in political contexts and described how Twitter users frame political events by sharing content exclusively with like-minded users forming two well-defined communities [4]. A segregated partisan structure with extremely limited connection between communities of users with different political orientations on the retweet networks can be found in multiple papers, in different contexts and countries like, for instance, the 2010 U.S. congressional midterm elections [11], the 2011 Canadian Federal Election [13] or tweets about the death of Venezuelan President Hugo Chavez [20]. Well defined communities can also be found in different platforms [10, 25].

15TH ALBERTO MENDELZON INTERNATIONAL WORKSHOP ON FOUNDATIONS OF DATA MANAGEMENT, 2023
falbanese@dc.uba.ar (F. Albanese)

Previous works showed the dramatic consequences and negative effects of closed communities and echo chambers, which include the increase of negative discourse, hate speech and political extremism [19], confirmation bias (i.e. the users tendency to seek out and receive information that strengthens their preferred narrative) [25] and spreading of baseless rumors and fake news [12, 9].

In this paper, we propose a machine learning framework in order to characterize the users who break this logic and change who they interact with: the *community-changing users* (i.e. the Twitter users that first belonged to a well defined community and then start interacting mostly with different users swinging to another community). Analyzing users that switch their political community can offer valuable insights into the complex dynamics of electoral politics, as they may be the deciding factor in which party wins an election.

Three datasets were built and used in order to show that the methodology can be easily generalized to different scenarios. Namely, we examined three Twitter network datasets constructed with tweets from: 2017 Argentina parliamentary elections, 2019 Argentina presidential elections and 2020 tweets about Donald Trump. For each dataset, we analyzed two different time periods and identified the larger communities corresponding to the main political forces. Using graph topological information and detecting topics of discussion of the first network, we built and trained a model that classifies whether an individual will change his/her community and find relevant features of the community-changing users.

Our main contributions are the following:

1. We describe a generalized machine learning framework for social media users classification, in particular, to detect and characterize community-changing users. This framework includes natural language processing techniques and graph machine learning algorithms in order to describe the topics of interests and interactions of each individual.
2. We experimentally analyze the machine learning framework by performing a feature importance analysis. In particular we assert the importance of the low value of "PageRank"[23] measure for this specific task. An interpretation of this result is that a person changes their community because their message was not heard in their previous community.

2. Data Collection

Twitter has several APIs available for developers. Among them is the Streaming API that allows the developer to download in real time a sample of tweets that are uploaded to the social network filtering it by language, terms, hashtags, etc. [21]. The data is composed of the tweet id, the text, the date and time of the tweet, the user id and username, among other features. In case of a retweet, it has also the information of the original tweet's user account.

For this research, we collected three datasets in two different periods of time: 2017 Argentina parliamentary elections (2017ARG), 2019 Argentina presidential elections (2019ARG) and 2020 United States tweets of Donald Trump (2020US). For the Argentinian dataset, the Streaming API was used during the week preceding the primary elections (from Aug 7^{th} to Aug 13^{th} 2017 and from Aug 5^{th} to Aug 12^{th} 2019) and the week before the general elections (from Oct 15^{th} to Oct 20^{th} 2017 and from Oct 20^{th} to Oct 27^{th} 2019). Keywords were chosen according to the four main political parties present in the elections. For the 2020US dataset, we used

"realDonaldTrump" (the official account of president Donald Trump) as keyword and the weeks from May 9^{th} to May 16^{th} and from June 10^{th} to June 16^{th} of 2020 as first and second time period respectively. Details can be found in the appendix. We have analyzed more than 9 million tweets and more than 1.5 million individuals in total.

Ethical Considerations and Data Availability: The datasets were constructed entirely with publicly available data as we do not collect any data form private accounts. For reproducibility, we also make publicly available the Ids.

https://github.com/fedealbanese/community-changing-users/

3. Methods

In this section, we will present the methodology employed to characterize the users. We describe how we calculate each feature and implement a supervised model that classifies users who changed their community over time. These models allow us to highlight which features are relevant characteristics of the users.

3.1. The retweet network

We represent the interaction among individuals in terms of a graph $G = (N, E)$, where users are nodes (N) and retweets between them are edges (E). Considering that a user can be retweeted multiple times by another user, this is well modeled by a directed and weighted graph. However, when a user n_1 retweets a tweet written by another user n_2, should the edge point form n_1 to n_2 or from n_2 to n_1? This definition has important implications. In the first scenario, the edges represent pointers to the "influencers" and important content generators. In the second scenario, the edges represent the flow of information through the network, going from the source to the user who spread the message. Indeed, there is no clear consensus in the scientific literature about which direction should be given to the edges: while some authors [3, 17, 31] use the first, others [27, 20, 11] prefer the second one. A priori, we cannot tell which direction is better for our purpose, so we decided to calculate the topological features in both scenarios. We named the directions of the edges RC (from Retweeter to content Creator) and CR (from content Creator to Retweeter).

In Fig. 1, we can visualize the retweet network for each time period and dataset. In the case of the US dataset, most of the users are concentrated in two groups, portraying the political polarization in that country. On the other hand, in the Argentinean dataset we can identify two large groups and also some smaller ones. The graph visualizations are produced with Force Atlas 2 layout using Gephi software [15].

3.2. Unsupervised Learning: Community Detection

In a given graph, a community is a set of nodes strongly connected among them and with little or no connection with nodes of other communities [32]. We detect the communities in the retweet network for each dataset using the Louvain method [6]. Given its stochasticity, we follow the solution proposed by Lancichinetti et al. [18] that runs the method several times (100 in our case). Then, only the nodes that were always consistently assigned to the same

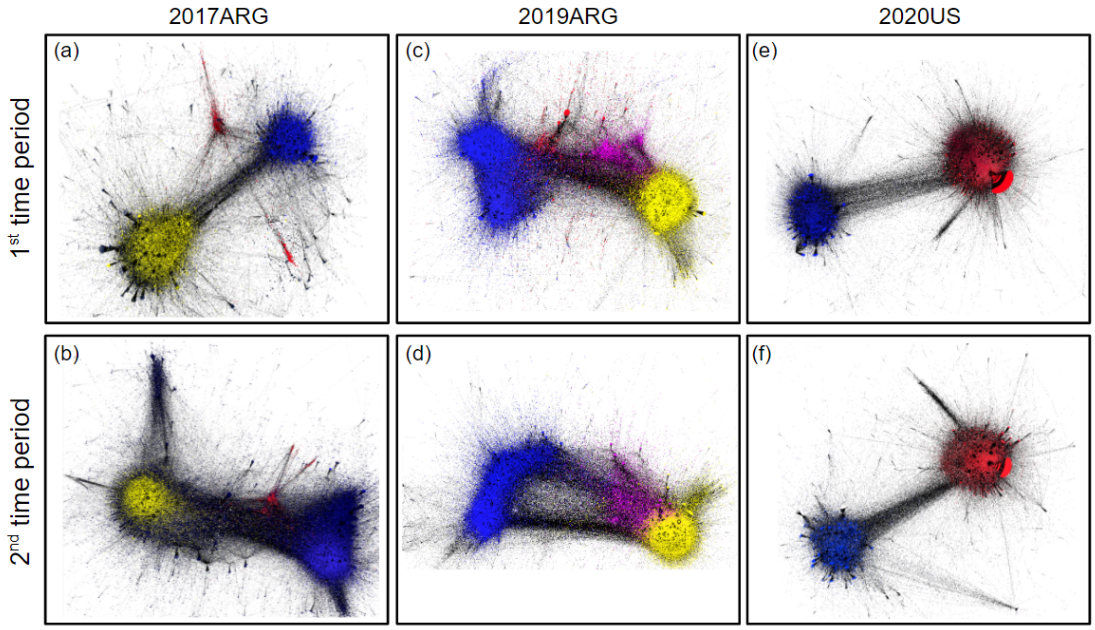

Figure 1: Retweet networks: (a) 2017 Argentina primary election; (b) 2017 Argentina general election; (c) 2019 Argentina primary election; (d) 2019 Argentina general; (e) first time period of the 2020US and (f) second time period of the 2020US. Each node is a Twitter user (colored depending on its community) and each edge (directed and weighted) represents the retweets between two given users (in black).

community in all iterations were considered in this work, in order to minimize the possibility of an incorrect labeling. We also only consider the users that received or made more than 5 retweets at each time period.

Despite the fact that the algorithm found several communities, we just considered the 4 largest ones for the Argentinean datasets and the 2 larger ones for the US dataset since these contain more than 90% of the users. We examine the text of the tweets and the users with the highest degree of each of the biggest communities and found that each one had a clear political orientation corresponding to the four biggest political parties in the election (beeing "Cambiemos", "Unidad Ciudadana", "Partido Justicialista" and "1 Pais" for 2017ARG and "Frente de Todos", "Juntos por el Cambio", "Consenso Federal" and "Frente de Izquierda-Unidad" for 2019ARG). Regarding the 2020US dataset, the 2 biggest communities corresponded to Republicans and Democrats accounts. The United States has a bipartisan political system which can be seen in Fig. 1, where only two big clusters concentrate almost all of the users and interactions. In contrast, the Argentinean datasets have two principal communities and some minor communities as well. This network topology with highly connected and polarized clusters had been reported in previous works [4, 11, 29].

3.3. Graph Features

Given that the analyzed datasets comprise two snapshots of the retweet network separate in time, we need to fully characterize the users in the early networks in order to properly identify those users that change their community. With this goal, we computed the following metrics for each user in the network: Degree, Indegree, Outdegree, PageRank, betweenness centrality, clustering coefficient and cluster affiliation (the detected community). As we mentioned earlier, it's important to note that the direction of the edges of the network drastically affects the value of these metrics. Consequently, we calculated them with both interpretations. All these metrics were used as features in the machine learning classification task and feature importance analysis.

3.4. Natural Language Processing Features

The features described above are based on user interaction and arise from the topology of the retweet network. We also characterized the topics of discussion during the first period of each data set by analyzing the texts of the tweets.

The features described above are based on the interaction of users and arise from the topology of the retweet network. We also characterize the topics of discussion during the first period of each dataset analyzing the texts of the tweets. Similarly to previous works [1, 24], first the tweets were described as vectors through the Term Frequency - Inverse Document Frequency (tf-idf) representation [26] and we used 3-grams and a modified stop-words dictionary that not only contained articles, prepositions, pronouns and some verbs but also the names of the politicians, parties and words like "election". Then, we performed Non-Negative Matrix Factorization (NMF) [30] to cluster our corpus of texts in topics. Finally, users were also characterized by a vector where each cell corresponds to one of the topics and its value to the percentage of tweets the user tweeted with that topic.

3.5. Feature importance analysis

Given that our objective was to characterize users who change their community and start interacting with users from other clusters, we implemented a machine learning model which classifies users and then performed a feature importance analysis. The instances of the model were the Twitter users who were active during both time periods [7] and belonged to one of the biggest communities in both time periods networks. Consequently, the number of users considered at this stage was reduced. Individuals were characterized by a feature vector with components corresponding to the mentioned topological metrics and others corresponding to the percentage of tweets in each one of the topics. The information used to construct these feature vectors was gathered only from the first time period, to avoid data leakage. The target was a binary vector that takes the value 1 if the user changed communities between the first and the second time periods and 0 otherwise. The summary of the datasets is shown in Table 1.

We apply the gradient boosting technique XGBoost [8], which uses an ensemble of predictive models and has proven to be efficient in a wide variety of supervised scenarios outperforming previous models [22]. We use a $67/33$ random split between train and test. In order to do

Table 1
Summary of the datasets used in the experiments.

	2017ARG	**2019ARG**	**2020US**
$\#Individuals$	21134	26118	116854
$\#Communities$	4	4	2
$\#TextFeatures$	9	7	6
$\#GraphFeatures$	10	10	10
$Trainingsetsize$	14159	17499	78292
$Testsetsize$	6975	8619	38562

Table 2
Summary of the results of the XGBoost models (ROC AUC).

	2017ARG	**2019ARG**	**2020US**
XGB (text)	0.7339	0.6683	0.6839
XGB (graph)	0.7664	0.7995	0.7425
XGB (text + graph)	**0.7925**	**0.8019**	**0.7614**

hyper-parameter tuning of the XGBoost models, we use the randomized search method [5] over the training dataset with 3-fold cross-validation.

Finally, we performed random permutation of the features values among users in order to understand which of them are the most important in the performance of our model (using the so-called Permutation Feature Importance algorithm [2]). In this way, we could identify the most important characteristics that separates the users that do change their community from those that do not change who they interact with.

4. Results

We trained three different gradient boosting models for each dataset: the first one was trained only with the features obtained via text mining (how many tweets of the selected topics the user talks about); a second one was trained just with features obtained through complex network analysis (degree, PageRank, betweenness centrality, clustering coefficient and cluster affiliation); and the last one was trained with all the data. In this way, we could compare the importance of natural language processing and the complex network analysis for the task of classifying community-changing users.

In Table 2 we can see the area under the ROC (receiver operating characteristic) curve [28] of the different models for each dataset. The best performance is obtained in all cases by the machine learning model built with all the features of the users, which is able to more efficiently classify the users who changed their community. This result is expected, since an assembly of models manages to have sufficient depth and robustness to understand the network information, the topics of the tweets and the graph characteristics of the users. Also, the model trained with graph features outperformed the model with only text features in all three cases.

We performed random permutation of the features values among users for the model trained

Table 3
Average $PageRank_{CR}$ of the users who changed their community and the users who did not change their community.

	2017ARG	**2019ARG**	**2020US**
changed	1.32e-5	3.81e-6	3.16e-6
did not change	1.55e-5	4.43e-6	3.47e-6

with all features (text+graph). We found that the most important feature in all cases corresponds to the node's connectivity: $PageRank_{CR}$, where the edges point from the tweet source (the content creator) to the user who retweeted. The feature importance coefficients of the $PageRank_{CR}$ are 1635 (2017ARG), 2836 (2019ARG) and 843 (US2020). All other features display even lower coefficients. In particular, the other $PageRank_{RC}$ (corresponding to the other direction of the edges) had importance feature coefficients of 717, 1202 and 527 for each dataset respectively (a reduction greater than 40%). This means that there is a clear privileged direction of edges for the task of detecting the users who changed their community.

When comparing the $PageRank_{CR}$ (PR) averages of these users with the users that did not change their community, we observed that the latter had higher values in all cases (Table 3). We applied the Kolmogorov-Smirnov test [14] to the PR distributions of each set and found that these differences were statistically significant in all cases ($p < 0.001$). The $Pagerank$ measures how relevant or important a user is in the retweet network based on the retweets of their messages and the importance of the users who retweeted. The direction of $PageRank_{CR}$ represents the information flow in a network, starting from the tweet creator and then spreading through the network. The fact that the community-changing users had statistically lower $PageRank_{CR}$ values means that these users were less relevant to the tweeter conversation and their messages did not spread in their original community. A possible interpretation of these results is that a user changes community when they do not have strong affinities with their community and their messages have no response.

The fact that the $PageRank_{CR}$ is the most important feature is also consistent with the model trained with network features getting a better AUC than the model trained with the texts of the tweets in the three datasets.

5. Conclusion

In this paper we presented a machine learning framework approach in order to identify and characterize users who changed their community for another one. The framework includes natural language processing techniques to detect their topics of interest and graph machine learning algorithms in order to describe how an individual interacts with other users. The framework was applied to three different datasets with similar results, showing that the methodology can be easily generalized.

We found that the users who changed communities had statistically lower values of $PageRank_{CR}$. This graph feature was also the most important indicator of the classification task in all three datasets according to the feature importance analysis. In particular, our results also show that

there is a clearly privileged direction on the network for this task, with the edges going from the content creator to the retweeter. A possible interpretation for these last two results is that users change who they interact with when they do not have strong affinities with other users, their messages have no response and are not being “heard" by their community.

References

[1] Albanese, F., Pinto, S., Semeshenko, V., Balenzuela, P.: Analyzing mass media influence using natural language processing and time series analysis. Journal of Physics: Complexity **1**(2), 025005 (2020)

[2] Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)

[3] Amati, G., Angelini, S., Capri, F., Gambosi, G., Rossi, G., Vocca, P.: On the retweet decay of the evolutionary retweet graph. In: International Conference on Smart Objects and Technologies for Social Good. pp. 243–253. Springer (2016)

[4] Aruguete, N., Calvo, E.: Time to# protest: Selective exposure, cascading activation, and framing in social media. Journal of communication **68**(3), 480–502 (2018)

[5] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of machine learning research **13**(2) (2012)

[6] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)

[7] Cazabet, R., Rossetti, G.: Challenges in community discovery on temporal networks. In: Temporal Network Theory, pp. 181–197. Springer (2019)

[8] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)

[9] Choi, D., Chun, S., Oh, H., Han, J., et al.: Rumor propagation is amplified by echo chambers in social media. Scientific reports **10**(1), 1–10 (2020)

[10] Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the National Academy of Sciences **118**(9) (2021)

[11] Conover, M.D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on twitter. In: Fifth international AAAI conference on weblogs and social media (2011)

[12] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences **113**(3), 554–559 (2016)

[13] Gruzd, A., Roy, J.: Investigating political polarization on twitter: A canadian perspective. Policy & internet **6**(1), 28–45 (2014)

[14] Hodges, J.L.: The significance probability of the smirnov two-sample test. Arkiv för Matematik **3**(5), 469–486 (1958)

[15] Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout

algorithm for handy network visualization designed for the gephi software. PloS one **9**(6), e98679 (2014)
[16] Jamieson, K.H., Cappella, J.N.: Echo chamber: Rush Limbaugh and the conservative media establishment. Oxford University Press (2008)
[17] Kogan, M., Palen, L., Anderson, K.M.: Think local, retweet global: Retweeting by the geographically-vulnerable during hurricane sandy. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. pp. 981–993 (2015)
[18] Lancichinetti, A., Fortunato, S.: Consensus clustering in complex networks. Scientific reports **2**(1), 1–7 (2012)
[19] Lima, L., Reis, J.C., Melo, P., Murai, F., Araujo, L., Vikatos, P., Benevenuto, F.: Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 515–522. IEEE (2018)
[20] Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M.: Measuring political polarization: Twitter shows the two sides of venezuela. Chaos: An Interdisciplinary Journal of Nonlinear Science **25**(3), 033114 (2015)
[21] Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In: Seventh international AAAI conference on weblogs and social media (2013)
[22] Nielsen, D.: Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU (2016)
[23] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
[24] Pinto, S., Albanese, F., Dorso, C.O., Balenzuela, P.: Quantifying time-dependent media agenda and public opinion by topic modeling. Physica A: Statistical Mechanics and its Applications **524**, 614–624 (2019)
[25] Quattrociocchi, W., Scala, A., Sunstein, C.R.: Echo chambers on facebook. Available at SSRN 2795110 (2016)
[26] Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)
[27] Rath, B., Gao, W., Ma, J., Srivastava, J.: From retweet to believability: Utilizing trust to identify rumor spreaders on twitter. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 179–186 (2017)
[28] Rice, M.E., Harris, G.T.: Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. Law and human behavior **29**(5), 615–620 (2005)
[29] Stewart, L.G., Arif, A., Starbird, K.: Examining trolls and polarization with a retweet network. In: Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web. vol. 70 (2018)
[30] Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 267–273 (2003)
[31] Yang, M.C., Lee, J.T., Lee, S.W., Rim, H.C.: Finding interesting posts in twitter based on

retweet graph analysis. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 1073–1074 (2012)
[32] Yang, Z., Algesheimer, R., Tessone, C.J.: A comparative analysis of community detection algorithms on artificial networks. Scientific reports **6**(1), 1–18 (2016)

References

[1] Albanese, F., Pinto, S., Semeshenko, V., Balenzuela, P.: Analyzing mass media influence using natural language processing and time series analysis. Journal of Physics: Complexity **1**(2), 025005 (2020)
[2] Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)
[3] Amati, G., Angelini, S., Capri, F., Gambosi, G., Rossi, G., Vocca, P.: On the retweet decay of the evolutionary retweet graph. In: International Conference on Smart Objects and Technologies for Social Good. pp. 243–253. Springer (2016)
[4] Aruguete, N., Calvo, E.: Time to# protest: Selective exposure, cascading activation, and framing in social media. Journal of communication **68**(3), 480–502 (2018)
[5] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of machine learning research **13**(2) (2012)
[6] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
[7] Cazabet, R., Rossetti, G.: Challenges in community discovery on temporal networks. In: Temporal Network Theory, pp. 181–197. Springer (2019)
[8] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
[9] Choi, D., Chun, S., Oh, H., Han, J., et al.: Rumor propagation is amplified by echo chambers in social media. Scientific reports **10**(1), 1–10 (2020)
[10] Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the National Academy of Sciences **118**(9) (2021)
[11] Conover, M.D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on twitter. In: Fifth international AAAI conference on weblogs and social media (2011)
[12] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences **113**(3), 554–559 (2016)
[13] Gruzd, A., Roy, J.: Investigating political polarization on twitter: A canadian perspective. Policy & internet **6**(1), 28–45 (2014)
[14] Hodges, J.L.: The significance probability of the smirnov two-sample test. Arkiv för Matematik **3**(5), 469–486 (1958)

[15] Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one **9**(6), e98679 (2014)
[16] Jamieson, K.H., Cappella, J.N.: Echo chamber: Rush Limbaugh and the conservative media establishment. Oxford University Press (2008)
[17] Kogan, M., Palen, L., Anderson, K.M.: Think local, retweet global: Retweeting by the geographically-vulnerable during hurricane sandy. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. pp. 981–993 (2015)
[18] Lancichinetti, A., Fortunato, S.: Consensus clustering in complex networks. Scientific reports **2**(1), 1–7 (2012)
[19] Lima, L., Reis, J.C., Melo, P., Murai, F., Araujo, L., Vikatos, P., Benevenuto, F.: Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 515–522. IEEE (2018)
[20] Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M.: Measuring political polarization: Twitter shows the two sides of venezuela. Chaos: An Interdisciplinary Journal of Nonlinear Science **25**(3), 033114 (2015)
[21] Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In: Seventh international AAAI conference on weblogs and social media (2013)
[22] Nielsen, D.: Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU (2016)
[23] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
[24] Pinto, S., Albanese, F., Dorso, C.O., Balenzuela, P.: Quantifying time-dependent media agenda and public opinion by topic modeling. Physica A: Statistical Mechanics and its Applications **524**, 614–624 (2019)
[25] Quattrociocchi, W., Scala, A., Sunstein, C.R.: Echo chambers on facebook. Available at SSRN 2795110 (2016)
[26] Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)
[27] Rath, B., Gao, W., Ma, J., Srivastava, J.: From retweet to believability: Utilizing trust to identify rumor spreaders on twitter. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 179–186 (2017)
[28] Rice, M.E., Harris, G.T.: Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. Law and human behavior **29**(5), 615–620 (2005)
[29] Stewart, L.G., Arif, A., Starbird, K.: Examining trolls and polarization with a retweet network. In: Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web. vol. 70 (2018)
[30] Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 267–273 (2003)

[31] Yang, M.C., Lee, J.T., Lee, S.W., Rim, H.C.: Finding interesting posts in twitter based on retweet graph analysis. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 1073–1074 (2012)
[32] Yang, Z., Algesheimer, R., Tessone, C.J.: A comparative analysis of community detection algorithms on artificial networks. Scientific reports **6**(1), 1–18 (2016)

A. Appendix

In this section we specified the keywords used for collecting the tweets.

2017 Argentina parliamentary elections:

The tweets were restricted to be in Spanish and the following terms were chosen as keywords for tweeter: the candidates for Senate of the main four parties: their name and official user on Twitter (i.e., “SergioMassa”, “Massa”, “RandazzoF”, “Randazzo”, “estebanbullrich”, “Bullrich”, “CFKArgentina”, “CFK” and “Kirchner”). The name of the official accounts of the first candidates for deputies of the parties (i.e., “felipe_sola”, “BuccaBali”, “gracielaocana” and “fvallejoss”). The name of the official accounts of political parties on Twitter (i.e., “1PaisUnido”, “1Pais”, “FJCumplir”, “Frente Justicialista”, “cambiemos”, “UniCiudadanaAR” and “Unidad Ciudadana”). The President of Argentina and the governor of the province of Buenos Aires at the time of elections (i.e., “mauriciomacri”, “Macri” and “mariuvidal”).

2019 Argentina presidential election:

The tweets were restricted to be in Spanish and the following terms were chosen as keywords for tweeter: “Elisacarrio”, “OfeFernandez_”, “PatoBullrich”, “macri”, “macrismo”, “mauriciomacri”, “pichetto”, “MiguelPichetto”, “JuntosPorElCambio”, “alferdez”, “CFKArgentina”, “CFK”, “kirchner”, “kirchnerismo”, “FrenteTodos”, “FrenteDeTodos”, “Lavagna”, “RLavagna”, “Urtubey”, “UrtubeyJM”, “ConsensoFederal”, “2030ConsensoFederal”, “DelCaño”, “NicolasdelCano”, “DelPla”, “RominaDelPla”, “FitUnidad”, “FdeIzquierda”, “Fte_Izquierda”, “Castañeira”, “ManuelaC22”, “Mulhall”, “NuevoMas”, “Espert”, “jlespert”, “FrenteDespertar”, “Centurion”, “juanjomalvinas”, “Hotton”, “CynthiaHotton”, “Biondini”, “Venturino”, “FrentePatriota”, “RomeroFeris”, “PartidoAutonomistaNacional”, “Vidal”, “mariuvidal”, “Kicillof”, “Kicillofok”, “Bucca”, “BuccaBali”, “chipicastillo”, “Larreta”, “horaciorlarreta”, “Lammens”, “MatiasLammens”, “Tombolini”, “matiastombolini”, “Solano”, “Solanopo”, “Lousteau”, “GugaLusto”, “Recalde”, “marianorecalde”, “RAMIROMARRA”, “Maxiferraro”, “fernandosolanas”, “MarcoLavagna”, “myriambregman”, “cristianritondo”, “Massa”, “SergioMassa”, “GracielaCamano”, “nestorpitrola”.

2020 tweets of Donald Trump:

The following term was used as a keyword for the twitter API: “realDonaldTrump”. In addition, the tweets were restricted to be in English.