

Bots don't Vote, but They Surely Bother!

A Study of Anomalous Accounts in a National Referendum

Eduardo Graells-Garrido

Data Science Institute, Universidad del Desarrollo
Santiago, Chile
egraells@udd.cl

Ricardo Baeza-Yates

Institute for Experiential AI, Northeastern University
California, USA
rbaeza@acm.org

ABSTRACT

The Web contains several social media platforms for discussion, exchange of ideas, and content publishing. These platforms are used by people, but also by distributed agents known as bots. Although bots have existed for decades, with many of them being benevolent, their influence in propagating and generating deceptive information in the last years has increased. Here we present a characterization of the discussion on Twitter about the 2020 Chilean constitutional referendum. The characterization uses a profile-oriented analysis that enables the isolation of anomalous content using machine learning. As result, we obtain a characterization that matches national vote turnout, and we measure how anomalous accounts (some of which are automated bots) produce content and interact promoting (false) information.

CCS CONCEPTS

• Information systems → Social networks.

KEYWORDS

Social networks, bot detection, political polarization, stance classification.

1 INTRODUCTION

Social media platforms have acquired a crucial role in meaning-making processes within communities [8]. In the context of social changes and worldwide events, such processes have acquired more importance than ever. As technology evolves, the “social” has become more than just people: social platforms provide a myriad of services ranging from news, health, business, games, among others. The entities in these platforms are people, but also companies, political parties, and media sources of all sizes and credibility. Yet, not all accounts that pretend to be people are actual persons. Some of them are automated accounts. Although sometimes bots are benevolent [1], the last several years the focus of bots has been deceiving people by manipulating and amplifying social media content. This situation has promoted methods to detect and characterize bots [20], as well as to understand their role in social interactions mediated by these platforms [18].

In this paper we study the political discussion around the Chilean Constitutional Referendum, held in October 25th, 2020. This event was one of the consequences of the fiercest social outburst in the last decades [19]. It started on October 18, 2019, and it is considered an important event that has impacted Chile’s well-being, due to a “perfect storm” of situations, including the recent pandemic [13]. One of the main demands of the social movements involved was a referendum to draft a new constitution for the country, because

the current constitution was drafted during Pinochet’s dictatorship. Thus, the plebiscite enticed strong and polarizing discussions on social media, particularly in the micro-blogging platform Twitter. Being publicly accessible, the trending topics of Twitter are part of everyday conversations and media reports. Given how social media can shape people’s perception, and how this perception can be tied to voting turnout, here we aim to understand the role of bots in the discussion. Mainly, we focused on the volume of content published by bots, their potential synchronization, and their political leaning.

We applied an existing methodology for stance detection (not referenced for anonymity), which enabled us to classify Twitter accounts into in-favor or against a new constitution. Then, we applied an existing anomaly detection method, Isolation Forest [9, 10], to quantify how anomalous was each account with respect to their behavior in the platform. We interpreted the global patterns of anomalous behavior, and then established a criteria to define a bot. As result, we observed that the stance classification produced results aligned with the election turnout; that the fraction of bots is small (0.66%) but their impact is much larger; and that, in terms of interaction and information diffusion, there are bot communities in both sides of the political spectrum, yet the larger communities were right-leaning, against the drafting of a new constitution.

2 DATA

We connected to the Twitter Streaming API using a system designed to crawl Chilean tweets. The query parameters were keywords related to mainstream political discussion in Chile, including keywords related to the two stances of the referendum: to approve (*Apruebo* in Spanish) the drafting of a new constitution, or to reject it (*Rechazo* in Spanish). We studied the period between August 1st, 2020, and October 25th, 2020. In total, we obtained 2.3M tweets from 251K users (see Figure 1) after a cleaning process. This represents about 10% of all Twitter users in Chile¹ and about 1.3% of the Chilean population at that time. Of those tweets, 32% were retweets, 6% were quotes of other tweets, and 9% were replies to other tweets.

3 PREDICTING STANCE

Given the size of the discussion under analysis, manually labeling the user profiles into the two stances {*apruebo*, *rechazo*} is expensive and impractical. In view of this difficulty, we predicted stances using a classifier trained on a labeled subset of the data set. This subset is labeled automatically from a list of seed patterns and keywords for each stance, as they are an effective mechanism to predict the community a user belongs to [3].

¹<https://dataportal.com/reports/digital-2020-chile>, p. 38.

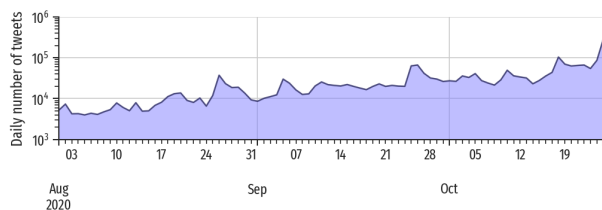


Figure 1: Weekly volume of content in the data set.

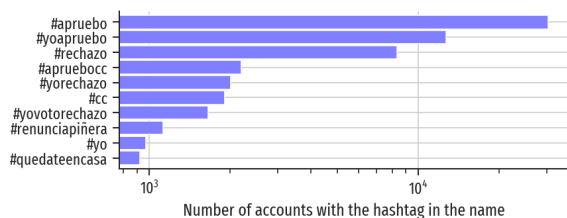


Figure 2: Frequency of top hashtags found in full names within profiles.

To identify seeds, we explored the data set to seek for terms that could be mapped to the *apruebo* and *rechazo* stances. This included hashtags (*#apruebo*, *#yoapruebo* –I approve–, *#votoapruebo* –I vote approve–; and their counterparts). We noticed that a relevant fraction of users self-reported their stances in the full name section of their profiles by including hashtags (see Figure 2). The seed terms are not necessarily frequent, but they are discriminating, *i.e.*, it is likely that someone in its corresponding category would use the term, and not from the other. The list is built iteratively in the sense of running the first steps from this section up to the classification step, and then exploring the usage of discriminating terms by accounts in each group to look for other potential seeds. Additionally, when we observed an account that was remarkably associated with a stance but was classified as the other, we manually labeled that account. In total, we manually labeled just 8 *apruebo* accounts, and 50 *rechazo* accounts.

Next, we propagated the user labels from the previous step to the rest of the data set. We used the XGBoost classifier that trains decision trees using gradient boosting [4]. The input feature matrix is the concatenation of several matrices:

- An account-term matrix, that encodes the number of times each account has used each term.
- A profile-term matrix, analogous to the previous one, but this time for the terms contained in the full name and biographical self-description of each user.
- A profile-domain matrix, mapping to each user’s home page its main domain (*e.g.*, *twitter.com*) and their main top level domain (*e.g.*, *.com*).
- Three adjacency matrices based on the interactions in the discussion: retweets, replies, and quotes.

- A user-stance interaction matrix for each type of interaction, where each cell contains the number of times the corresponding user has interacted with other users that were pre-labeled with a stance.

We removed terms that were used for labeling from the feature matrix, as they perfectly separate users from both groups and our goal is to classify users who do not use these terms in their content. Then, we trained the classifier using the set of labeled users.

Then, we predicted the stance of the rest of the data set. For a given account u , the classifier outputs a value $p_a(u)$ for each stance a that lies in $[0, 1]$, corresponding to the fraction of decision trees that vote for the corresponding stance. We applied a small threshold ($p_a(u) \geq 0.55$) to consider predictions with at least a small confidence by the classifier. Those accounts who cannot be classified were marked as *undisclosed*.

As result, we predicted 81.20% of accounts in *apruebo*, 17.34% of accounts in *rechazo*, and 1.46% as *undisclosed*. This matches well the referendum results, where 78.28% voted *apruebo*, whereas 21.72% voted *rechazo*. The distribution is similar, hinting that Twitter is a powerful signal when analyzing national-level events, even when the sample is not representative of the overall country demographics. This was a result obtained even before the culmination of the study, as one week before the election we shared a preliminary prediction using this method with journalists.

To characterize each stance, we estimated which terms and features were most associated with each stance using the log-odds ratio (see Figure 3). As expected, features related to homophily are important, such as retweeting and quoting pre-labeled users. With respect to content, we found that in *apruebo*, words like *dignidad* (dignity) and *pueblo* (people) exhibit its political left-leaning, whereas the Chilean flag emoji, mentions to *libertad* (liberty) and *patriotas* (patriots) exhibit its right-leaning nationalism.

4 QUANTIFYING ANOMALOUS BEHAVIOR

We want to know to which extent the discussion was influenced by anomalous accounts. First, we downloaded the largest data set of bots available [5], and found that only two accounts from our data set were on it, both wrongly marked as bots: one account was from a legitimate media platform and the other was from a right-wing politician. Since we do not have known bot labels for users, and bot classifiers, such as Botometer,² tend to rely on the full recent content published by accounts (in contrast to our content-based crawling approach), and may not work well in other languages than English [17], we implemented a method to quantify the anomaly (or lack thereof) of every account in our study.

We based our work in the Isolation Forest (IF) [9, 10] model. IF is an unsupervised anomaly detection that quantifies the distance of a given observation to the rest of the data. The model has two assumptions regarding anomalies: first, they are few; second, they are very different to normal observations. As such, anomalies can be succinctly described with respect to the rest of the data set. The model does so by building an ensemble of trees, where each tree learns a description of a sample of the data based on binary partitions of its feature space. An anomaly score for an observation is derived from the average path length in all trees. It is equivalent

²<https://botometer.osome.iu.edu/>.

Bots don't Vote, but They Surely Bother!

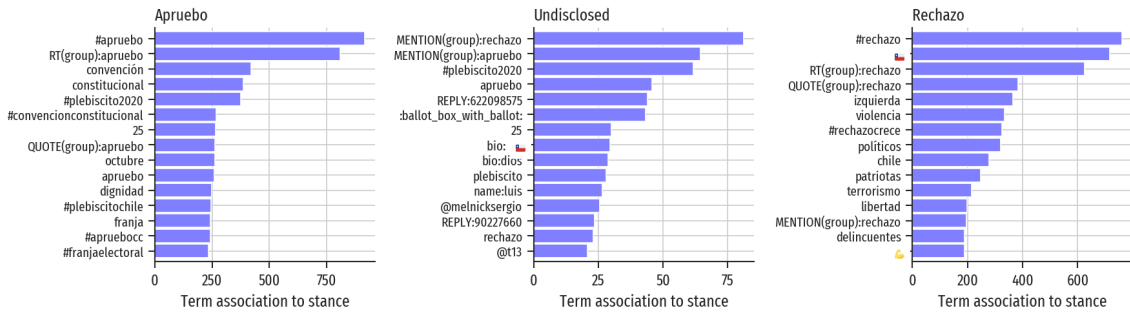


Figure 3: Top terms and features associated with each stance.

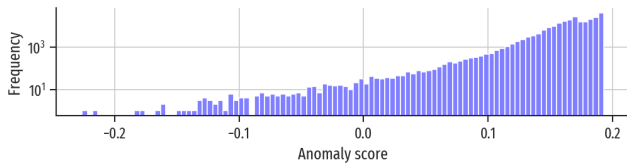


Figure 4: Distribution of anomaly scores.

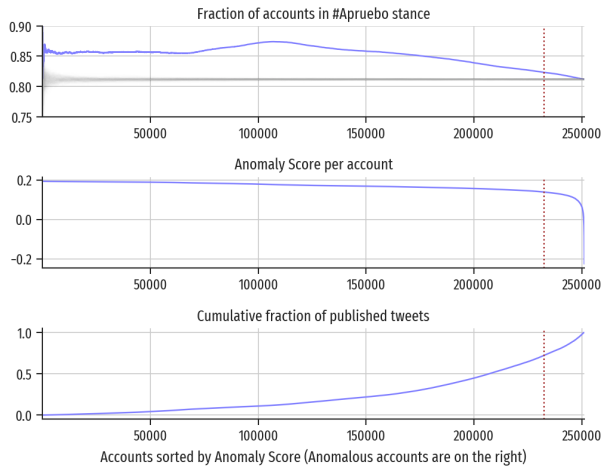


Figure 5: Top: fraction of accounts in *apruebo* (blue line) with respect to the number of accounts over the anomaly threshold (red dotted line), compared with a null model (gray lines). Center: Anomaly scores per account. Bottom: Cumulative fraction of published tweets.

to the number of splittings required to isolate the observation. The greater the score, the less anomalous an observation is.

To define how to estimate the distance between accounts, we built a feature matrix with the following elements per account:

- The number of active days in the data set, *i.e.*, the number of days with published content by each account.
- Relative amount of content: the number of published tweets, retweets, quotes, and replies, log-transformed and divided by the amount of active days.

- A daily rhythm, consisting of the total amount of content published divided by the number of active days.
- Ratio of friends (followees) over followers.
- The number of digits in the account username.
- A flag regarding the use of the default profile image.
- Whether the account is on a connected component of interactions, and which component.
- Account age in days.
- Relative global behavior: the log-transformed number of globally published tweets, friends, and followers, divided by the account age.

After applying the model, we sorted the accounts with respect to their anomaly scores (see Figure 4 for the distribution). In data sets of similar size, it has been determined that around 7.5% of accounts are bots [12]. We looked at the distribution of anomaly scores and the cumulative fraction of published tweets (see Figure 5, center), and we observed that anomaly scores present relevant values in a smaller proportion of accounts, and that anomalous accounts tend to publish more than normal ones. It is known that the distribution of published tweets follows a power-law and that few accounts generate most of the content. Those accounts could also be anomalous, but not necessarily bots.

To understand whether there is a relationship between anomaly and political position, we compared the fraction of accounts in *apruebo* at every incremental subset of accounts sorted by anomaly score. We compared this distribution with a null model where the political stance was permuted at random (see Figure 5, top). In the null model, the anomaly score is not correlated with the fraction of accounts in *apruebo*, which is what we expected. However, in the observed distribution, there is a complex relationship between being anomalous and the fraction of accounts in *rechazo*. This result hints that most anomalous activity is associated with right-wing politics. Bots need a significant investment, so this is not surprising either.

5 DISCRIMINATING BOTS

The anomaly score points accounts that could be bots, but additional criteria is needed to identify them. One element of this criteria is to consider the age of accounts: old accounts may be anomalous with respect to their behavior, such as a high frequency of publications, but this may be a natural behavior of the population. Hence, we separated accounts into five groups, with the first group defined as

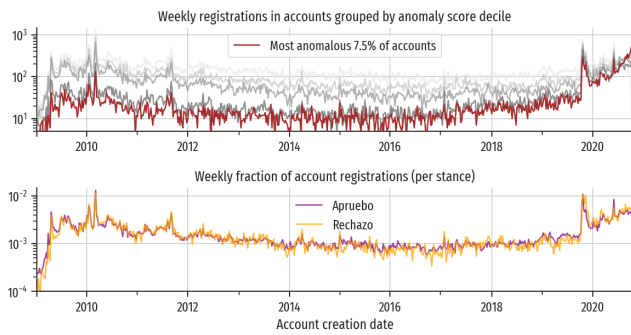


Figure 6: Weekly registrations with respect to anomaly score (top), and stance (bottom, normalized).

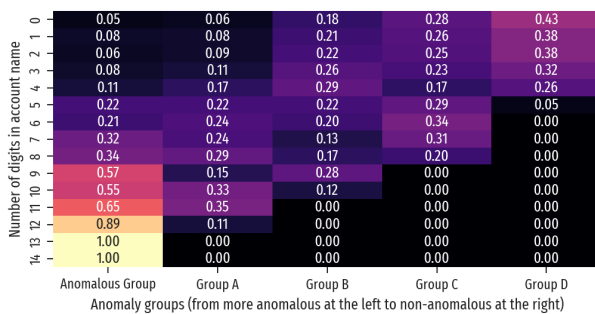


Figure 7: Distribution of accounts according to their anomaly score per number of digits in the username.

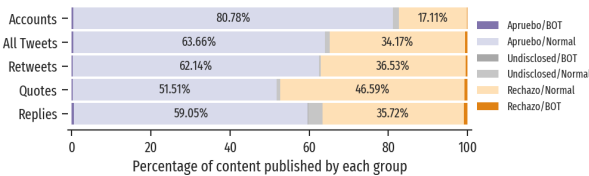


Figure 8: Distribution of content with respect to predicted stance and assigned bot status.

the 7.5% most anomalous, and the other four groups defined as the subsequent accounts in evenly-spaced ranges of anomaly score (see Figure 6). We observe that the most anomalous group tends to be on the lower-bound of registrations per week until one week after the beginning of the study (August 8th, 2020). After that, it became the most active group in terms of registrations. In comparison, the distribution of registrations with respect to stance does not present differences between stances, implying that registration date may help us to point to bots regardless of their political position.

Another important feature is the number of digits in a username, as a high number of digits may indicate an account with a randomly generated username. Indeed, the model found that less anomalous accounts tend to have less than four digits in their names (a feasible explanation of this limit is that some accounts have a year in their

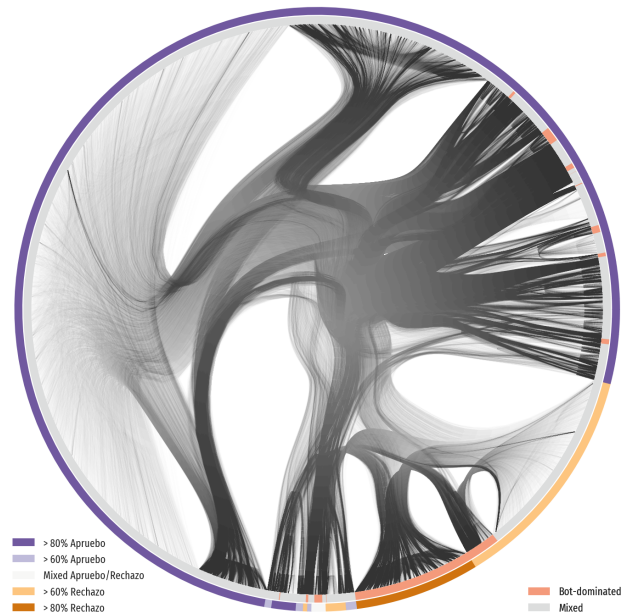


Figure 9: Network of retweets. Communities are represented by the two rings on the outside, one colored according to political stance (outer), and the other colored according to the presence of bots (inner). Edges are lines between nodes, where the origin of the edge (the retweeting account) is colored in light gray, and the destination of the edge (the retweeted account) is colored in dark gray.

username), whereas highly anomalous accounts have up to 14 digits (see the distribution in Figure 7).

Then, we assigned bot status to an account that lied in the anomaly group (7.5% more anomalous), that has been registered from August 8th onward, and that has more than four digits in the username. Only 0.66% of accounts are labeled as bots with this criteria. Next, we compared the distribution of accounts and the published content taking into account both, predicted stance and bot status (see Figure 8). We observed that *rechazo* publishes a greater amount of content than expected given its number of accounts, however, the activity of bot accounts does not seem suspicious in terms of content volume. This hints that bot activity in this study could be related to bursts of coordinated action, for instance, to establish a trending topic, rather than continuous generation of content. It suggests, as well, that bot detection is tied to political activity, and thus, a generic bot detector may not perform well without considering the political context.

Finally, to explore the potential coordination between bots, we estimated a hierarchical community structure using a Stochastic Block Model [16] (see Figure 9), under the assumption that, if there is coordinated action between bots, then those accounts should belong to the same community. We studied the largest connected component of the retweet network, with 143K accounts and 527K weighted edges. In its 118 detected communities, we standardized the fraction of bots within each, and then we classified each into

one of three groups: it has more bots than expected (more than one standard deviation of bot presence, 17 communities), and mixed (the rest, 89 communities). We find it interesting that there are no communities without bot accounts, probably indicating that there are false positives in our criteria. There are bot-dominated communities in both political stances, however, the *rechazo* ones are larger. These communities tend to show a high popularity, with inter- and intra-community retweets. This is a sign of coordinated action, although the effects of these actions have yet to be determined.

6 CONCLUSIONS

In this work we performed a detailed analysis of Twitter discussion in an historical national event in Chile, from the lens of anomalous activity, including bots. We found that, under strict criteria, the number of bots in the discussion is small, and that in terms of produced content, bot accounts do not differ from regular accounts. The difference lies in the network behavior: there are bot-dominated communities in the information diffusion network, and these communities have a high in-degree. This suggests for future work that, although they may not be an army of bots, these small squads in coordination with regular accounts, may influence what is being discussed. This influence has a clear political objective, as *rechazo* (right-leaning) bots form large communities in comparison with *apruebo* bots, which seem to be scattered around larger communities of regular accounts. Since the inner workings of trending topics in Twitter is unknown, this evidence provides information that helps people, journalists and politicians to put the digital discussion into perspective. In particular, bots amplify political polarization and makes more difficult to distinguish the reality with the perception of it.

Future work could focus on strengthening the pipeline of analysis, in particular the bot criteria. On the one hand, previous reports indicate a larger fraction of bots, thus, we may be providing a lower bound of this quantity. On the other hand, our criteria did not include the community detection step, which may help to identify false positives, or the difference between well-behaved and deceiving bots. We will also study how bots behave in the exit referendum of the constitutional process. That is, approving or not the new constitution.

In terms of representativeness, we acknowledge that Twitter is a biased sample of the population [2]. The similarity between our stance prediction and the election results also adds evidence to this aspect. Although the representativeness of such insights is yet to be determined, we propose for future work to disentangle national from local representativeness of results.

ACKNOWLEDGMENTS

This project uses the graph-tool [15], scikit-learn [14], XGBoost [4], numpy [6], pandas [11], and matplotlib [7] libraries. We thank Paula Vasquez-Henrriquez for her comments.

REFERENCES

- [1] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. 2021. People Are Strange When You're a Stranger: Impact and Influence of Bots on Social Networks. *Proceedings of the International AAAI Conference on Web and Social Media* 6, 1 (Aug. 2021), 10–17. <https://ojs.aaai.org/index.php/ICWSM/article/view/14236>
- [2] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (2018), 54–61.
- [3] John Bryden, Sebastian Funk, and Vincent AA Jansen. 2013. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science* 2, 1 (2013), 1–9.
- [4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, USA, 785–794.
- [5] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, Minnan Luo, and Minnan Luo. 2021. TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (Virtual Event, Queensland, Australia) (CIKM '21)*. ACM, New York, NY, USA, 4485–4494. <https://doi.org/10.1145/3459637.3482019>
- [6] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with NumPy. *Nature* 585, 7825 (2020), 357–362.
- [7] John D Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering* 9, 03 (2007), 90–95.
- [8] Anastasia Kavada. 2020. Creating the collective: social media, the Occupy Movement and its constitution as a collective actor. In *Protest Technologies and Media Revolutions*. Emerald Publishing Limited, Bingley, UK.
- [9] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *IEEE International Conference on Data Mining*. IEEE, Pisa, Italy, 413–422.
- [10] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 1 (2012), 1–39.
- [11] Wes McKinney et al. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* 14, 9 (2011), 1–9.
- [12] Marcelo Mendoza, Maurizio Tesconi, and Stefano Cresci. 2020. Bots in social and interaction networks: detection and impact estimation. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–32.
- [13] Mauricio Morales Quiroga. 2021. Chile's perfect storm: social upheaval, COVID-19 and the constitutional referendum. *Contemporary Social Science* 16, 5 (2021), 556–572.
- [14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [15] Tiago P Peixoto. 2014. The graph-tool python library.
- [16] Tiago P Peixoto. 2014. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X* 4, 1 (2014), 011047.
- [17] Adrian Rauchfleisch and Jonas Kaiser. 2020. The false positive problem of automatic bot detection in social science research. *PLoS one* 15, 10 (2020), e0241045.
- [18] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.
- [19] Nicolás M Somma, Matías Bargsted, Rodolfo Disi Pavlic, and Rodrigo M Medel. 2021. No water in the oasis: the Chilean Spring of 2019–2020. *Social Movement Studies* 20, 4 (2021), 495–502.
- [20] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *International AAAI Conference on Web and Social Media*. AAAI, Montreal, Canada, 280–289. <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587>