Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups

Victor S. Bursztyn
v-bursztyn@u.northwestern.edu
Department of Computer Science
Northwestern University
Evanston, Illinois, USA

Larry Birnbaum
l-birnbaum@northwestern.edu
Department of Computer Science
Northwestern University
Evanston, Illinois, USA

ABSTRACT

There is growing concern about the use of social platforms to push political narratives during elections. One very recent case is Brazil's, where WhatsApp is now widely perceived as a key enabler of the far-right's rise to power. In this paper, we perform a large-scale analysis of partisan WhatsApp groups to shed light on how both right- and left-wing users leveraged the platform in the 2018 Brazilian presidential election. Across its two rounds, we collected more than 2.8M messages from over 45k users in 232 public groups (175 right-wing vs. 57 left-wing). After describing how we obtained a sample that is many times larger than previous works, we compare right- and left-wing users on their social network metrics, regional distribution, content-sharing habits, and most characteristic news sources.

CCS CONCEPTS

• **Networks** → **Online social networks**.

KEYWORDS

chat applications, WhatsApp, elections, partisanship, data collection, social network analysis

ACM Reference Format:
Victor S. Bursztyn and Larry Birnbaum. 2018. Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

On October 28th 2018, amid strong polarization and conspiracy theories that flooded social media, Brazilians elected far-right candidate Jair Bolsonaro their next president. With over 120M users in Brazil – the second largest market in the world – the role that WhatsApp played in this electoral process has emerged as a major focus of attention and controversy due to its alleged importance in a large number of successful political campaigns.

Indeed, WhatsApp in Brazil connects an audience comparable in scale to television's to content created and distributed with almost no barriers or filters other than user curation. Although the promise of inexpensive one-to-one mobile communication may have sparked this popularity – WhatsApp's 1.5B users are largely in developing countries – group chats arguably made it catch fire. The app allows users to create groups where up to 256 users can share text and multimedia messages, transforming these groups into highly active social spaces. Users can also create public invites to their groups and share them as URLs across the web, transforming these groups into small public forums.

However, the fact that all messages circulate with end-to-end encryption hinders transparency and even law-enforcement, making WhatsApp a fertile ground for bad actors. A recent study on the Brazilian electorate found that false stories that circulated massively through WhatsApp's network were more far-reaching than initially assumed, as it revealed that 90% of Bolsonaro's supporters think they are truthful [2]. But measuring effects on the surface only to speculate about such a complex network isn't enough to protect future elections from the use of WhatsApp as a political weapon. Instead, it's necessary to learn how to measure partisan activity at scale.

In practice, this is challenging because it requires finding a large number of partisan WhatsApp groups and following the digital rallies that happen in them. Next, we need to find meaningful ways to analyze such rallies and characterize each partisan group. There are many analyses that can be made to compare right- and left-wing users in WhatsApp: How their social networks are structured, how much do they represent the larger population of voters, and what types of content are consumed and shared by them.

Addressing the challenges involved in analyzing partisan activity in WhatsApp, we make the following contributions:

- We introduce a new data collection method capable of growing a sample as much as necessary and towards specific directions in the political spectrum;
- Our code, released to the research community upon publication, can mine invites to other public WhatsApp groups from the groups already joined;
- We analyze the first large-scale dataset of partisan activity in WhatsApp using a variety of methods and standpoints, contributing with real measurements about right- *vs.* left-wing users in the platform.

2 RELATED WORK

Several works indicate a recent rise of interest in analyzing public WhatsApp groups:

Rosenfeld *et al.* [8] provided an initial study of WhatsApp messages with a particular focus on predictive analysis. Although it comprised 4M messages, the dataset spanned only 100 users and didn't include the actual contents of those messages – only a handful of meta-data. They noted that a small majority of messages originated from group messaging and used it to distinguish WhatsApp from plain texting.

Garimella and Tyson [3] collected the first large-scale WhatsApp dataset. To do so, they looked for invites to public WhatsApp groups in websites that aggregated and organized information about such groups as well as by searching for the string "chat.whatsapp.com" in Google. After automating the process of joining groups using Selenium over WhatsApp's web client, their work was able to collect 454,000 messages from 45,794 users in a six month period, encompassing a total of 178 groups about a wide variety of themes.

Caetano *et al.* [1] provided an initial study of political discussions in Brazilian WhatsApp groups and, at the same time, offered valuable measurements for a non-electoral setting. Their dataset comprised 273,468 messages from 6,967 users in a one month period, encompassing a total of 81 groups. From these, only eight groups were selected and further analyzed – four being political and four non-political. More recently, Resende *et al.* [7] described a system aimed at helping journalists to report on WhatsApp activity during the electoral process. Their data comprised 210,609 messages from 6,314 users in a one month period, encompassing a total of 127 public groups dedicated to political themes. A fraction of this sample, however, may be considered politically neutral or mixed: there were 26 "debate groups" and 30 "news sharing groups", so partisan groups may be limited to 71 groups.

In [6], Resende *et al.* provide a more up-to-date view on their collected dataset: it comprises 789,914 messages from 18,725 users, but it's still limited to the first round of the Brazilian election and filled with heterogeneous groups.

3 METHODOLOGY

Our data collection started on September 1, 2018, and was concluded for the purpose of this analysis on November 1, 2018, thus spanning exactly two months. It included meaningful events that preceded the electoral race (e.g., Brazil's Supreme Court barring former President Lula from the race on September 11) as well as the first and second rounds of the election (on October 8 and 28, respectively).

All data was collected using a dedicated smartphone with 64GB of storage, allowing for a maximum of 1GB of data per average day in our study. WhatsApp's daily backups were observed so that our portfolio of groups would never exceed this safe margin. For our setting, in practice, it allowed the tracking of 700-800 public groups – the total fluctuated as groups were closed or added.

In this section we describe our data collection methodology in two parts. First, we discuss the basic steps also implemented by previous works [1, 3, 6, 7]. Second, we describe our improvements, which substantially enhance data collection, now made available at a public repository. As with previous works, we emphasize that the resulting data collection complies to WhatsApp's privacy policy.

Data Collection - Base Method

Public WhatsApp groups are characterized by the ability to join them through a public URL created by their owners, called "an URL invite". At the same time, all WhatsApp groups are limited to 256 users. Considering this, Caetano *et al.* [1, Figure 2] summarizes the base method in three steps: **1)** Searching the web for invites to public WhatsApp groups; **2)** Trying to join them; and **3)** Extracting data from them.

For step **#1**, a typical solution is to look for invites to public WhatsApp groups in a set of publicly accessible sources. These sources include: (i) websites aimed at organizing information about public WhatsApp groups; (ii) the web, in general, by searching for the string that is the prefix of URL invites ("chat.whatsapp.com") in Google; and (iii) the social web, by performing the same search in Facebook, Twitter, YouTube, and Instagram. Furthermore, within these sources, it's often possible to filter groups dedicated to political discussion: (i) by selecting categories most likely to host such groups; (ii) and (iii), by compiling a list of keywords referring to the political right and to the political left (e.g., candidate names, vice-presidents, parties), and combining this list with the string "chat.whatsapp.com". In our implementation, this process resulted in about 100 valid URL invites.[1]

For step **#2**, a typical solution is to use the Selenium-based Python script published by Garimella and Tyson [3] to automate the joining process. Their code: (i) receives a list of URL

[1]The full list of keywords we used is available at https://github.com/vbursztyn/whatsapp-data-collection/blob/master/keywords_invites.csv

invites, (ii) opens a browser window, (iii) loads WhatsApp's web client, and (iv) simply tries to join each group sequentially. These attempts aren't necessarily successful because of the group size limitation. However, since the joining process is now automated, new attempts can be scheduled until a spot appears.

For step **#3**, the solution varies. It ranges from simply exporting group activity using WhatsApp's chat export feature to obtaining access to WhatsApp's local DB [3] or using a third-party API to scrape WhatsApp's web client [6, 7].

Data Collection - Enhancements

We added step **#4** to Caetano *et al.* [1, Figure 2]: **4)** Mining new URL invites sent to the groups already joined.

Based on the automated joining script by Garimella and Tyson [3], our code: (i) opens a browser window, (ii) loads WhatsApp's web client, (iii) inserts the string "chat.whatsapp.com" into the search bar, (iv) waits for the results, (v) scrolls an arbitrary amount of times, (vi) mines all invites that are loaded in the browser, (vii) retrieves each group's information, and finally (viii) saves all information in a table that can be *a.* manually managed, and *b.* passed on to the automated joining script. Step **#4** is particularly powerful as it creates a loop between joining groups and extracting more invites from their messages. This loop can be used to grow the sample as much as necessary and towards specific directions – for instance, by mining more invites from left-wing groups. In practice, we are able to explore the interconnected nature of partisan WhatsApp groups.

Last, in our work, we inspected each public WhatsApp group to assess whether its cover photo, group title or group description would explicitly support a specific candidate. 232 groups had *all the three* elements explicitly supporting a candidate, thus being deemed partisan. Among these, 175 groups were clearly right-wing, as they supported far-right candidate Jair Bolsonaro, whereas 57 groups were classified as left-wing. These 57 groups either supported Fernando Haddad, the left-wing runner-up, or Ciro Gomes and Marina Silva, who were center-to-left candidates that didn't make it to the second round and then declared support for Haddad. Therefore, as the BBC [5] did in their analysis of WhatsApp in India, we aggregate candidates to the left of Bolsonaro to represent the political left although the two partisan groups aren't equidistant from the center.

Our code is available at https://github.com/vbursztyn/whatsapp-data-collection

WhatsApp's Privacy Policy

Like previous projects [1, 3, 6, 7], our work is based on public WhatsApp groups, which are accessible to any user with valid URL invites. These invites, especially in the case of partisan groups during the Brazilian election, were widely disseminated by group owners. This work is similarly compliant with WhatsApp's privacy policy as it states that all users must be aware that their data will be shared with other members once they participate of a group, public or not. Unlike previous works, no third-party tools were used for data extraction: our data collection used WhatsApp's chat export feature, meaning that all data we have accessed, processed, and analyzed were selected and sent by email through the WhatsApp app. Many other systems rely on the same chat export feature, which is limited to the 10,000 most recent messages if multimedia files are included or to the 40,000 most recent messages otherwise.

4 PARTISANSHIP: GENERAL CHARACTERISTICS

Social Network Metrics

In this analysis we evaluate structural properties of the networks formed by right- and left-wing users in the 232 partisan groups identified. To do so, we construct networks where nodes represent active users (i.e., users who sent at least one message to at least one group) and edges represent pairs of users co-participating in a same group. Since we have three times as many right-wing groups (175 *vs.* 57), networks will have different sizes. Indeed, as Table 1 shows, the right-wing network has 39,035 nodes (users) and 8.4M edges, while the left-wing has 6,242 nodes and 0.9M edges.

However, a more detailed analysis tells a different story. The difference in the number of connected components isn't proportional to the difference in sizes, which is confirmed by the extent of their largest connected components (LCCs): 95% of nodes in the right-wing network belong to a single connected component, while this value is down to only 80% of nodes in the left-wing network. Additionally, despite the difference in sizes, right-wing users have a smaller average path length (APL): 3.03 *vs.* 3.13 for left-wing users. In other words, **we find that right-wing users are more tightly connected in WhatsApp**.

It's also worth noting that our results for clearly partisan groups show a substantially smaller APL compared to Resende *et al.*'s [6, Table 4] results for "political groups", in general: 3.03 & 3.13 (ours) *vs.* 3.95 (theirs). This happens despite the fact that our networks have 4+ times as many nodes (45,277 *vs.* 10,860). Therefore, **we find that partisan groups are more tightly connected when compared to political groups, in general**.

Representation of Real Population of Voters

In this analysis we evaluate a possible link between partisan activity in WhatsApp and the real population of voters, as we conjecture that the regional distribution of users in our sample should reflect the regional distribution of voters in a constituency.

Table 1: User network metrics.

	Right-Wing	Left-Wing
# of groups	175	57
# of nodes	39,035	6,242
# of edges	8,423,514	872,957
# of components	1,830	1,249
Largest connected component	95.31%	80.01%
Average path length	3.03	3.13

Table 2: Content-sharing habits.

	Right-Wing	Left-Wing
# of messages (%)	2,392,851 (100%)	429,835 (100%)
# of multimedia messages (%)	1,113,821 (46.55%)	129,328 (30.09%)
# of messages with URLs (%)	279,196 (11.67%)	50,608 (11.77%)
# from YouTube (%)	157,208 (56.31%)	22,378 (44,22%)
# from WhatsApp (%)	42,414 (15.19%)	9,902 (19.57%)
# from Facebook (%)	30,172 (10.81%)	10,127 (20.01%)
# from Twitter (%)	5,602 (2.01%)	3,111 (6.15%)
# from Instagram (%)	6,586 (2.36%)	663 (1.31%)

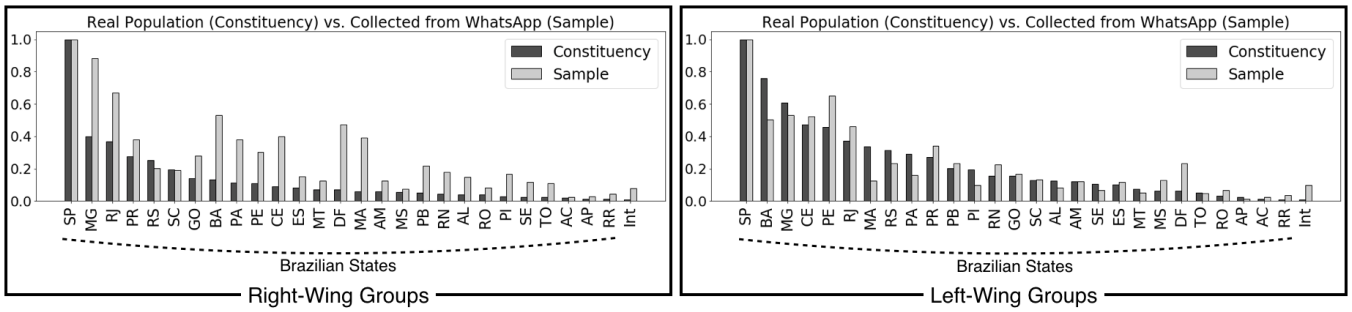

Figure 1: Right- and left-wing users organized by state and normalized.

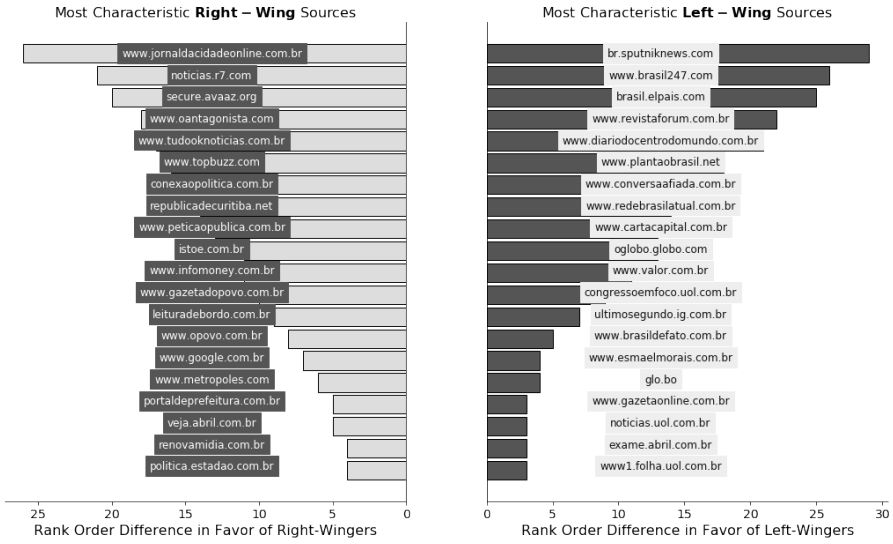

Figure 2: Largest rank order differences indicate news sources that are most characteristic of each partisan group.

First, in our sample, we obtain two regional distributions by processing the state area codes extracted from the distinct phone numbers found in each partisan group. Next, we compare these distributions with the final election results for each candidate in each state [9]. We normalize the populations from each state by the ones from São Paulo ("SP"), as it's where both Bolsonaro and Haddad garnered the most votes (15.3M and 7.2M, respectively). Note in Figure 1 that "SP" is represented by the highest bars, normalized to 1.0. All other bars would decay similarly if the same ratio held. This way, states where the "Sample" bar exceeds the "Constituency" bar are overrepresented in our sample (e.g. "MG" in right-wing groups), while states where the opposite happens are underrepresented (e.g., "BA" in left-wing groups).

It's worth noting that **two regions are extraordinarily overrepresented in both partisan groups**: the one comprised of Brazil's administrative district (**"DF"**), which revolves around political activity, and the one comprised of

voters who live abroad (**"Int"**). Although a possible explanation could be that these two regions were disproportionately engaged in political activism, a deeper analysis of the international numbers is warranted as they could include ghost accounts created through third-party services.

Also, among right-wing users, most regions are overrepresented. It should be noted that these distortions could have several origins: from differences in regional populations of WhatsApp users to possible differences in the sharing of invites between groups, which would cause our data collection method to mine more groups from more interconnected regions. The body of knowledge on Twitter mining suggests there could be a number of biases in this population [4], thus some analyses should be made with caution – especially if describing the actual constituencies.

Content-Sharing Habits

In this analysis we evaluate the 2.39M messages sent in right-wing groups and the 0.43M messages from left-wing groups to outline the content landscape in each partisan group, as seen in Table 2.

Interestingly, right-wing users send multimedia messages at a substantially higher rate: 46.55% *vs.* 30.09%. Despite the small sample from Caetano *et al.* [1, Figure 5], we highlight that these numbers are much higher than the 20% they had found for political groups in a non-electoral setting. Considering their baseline, **use of multimedia messages by right-wing users more than doubled compared to what was seen one year before**.

Previous works [1, 6, 7] found that roughly 10% of all text messages contain URLs. Among these, they further found that YouTube tops the list of popular domains followed by Facebook and WhatsApp. Based on a substantially larger sample, our results seem to confirm theirs.

Our results suggest that **the electoral process could be a strong driver for the use of multimedia messages in partisan groups**, especially among right-wing users, whereas the same effect isn't seen in the total amount of URLs shared. However, similarly to what was noted in the United States, YouTube appears to play a role in information diffusion especially for the political right – **56.31% of right-wing URLs are YouTube videos**.

News Consumption

In this analysis we evaluate how right- and left-wing users consume news by calculating their most characteristic news sources. To do so, we count the most frequent news sources among right-wing messages and compile a rank with their top 30 sources ($Rank^{RW}$), doing the same for the left ($Rank^{LW}$).

Consequently, for a given source α, consider that $Rank^{RW}_{index}(\alpha)$ returns the rank order of α among *right-wing* messages (or 30 if α isn't in the rank, referring to the last position of a top 30). Likewise, $Rank^{LW}_{index}(\alpha)$ returns the rank order of α among *left-wing* messages (adopting 30 as fallback).

We calculate a score for α among right-wing users by calculating the difference in rank orders, as follows:

$$Score^{RW}_{\alpha} = Rank^{LW}_{index}(\alpha) - Rank^{RW}_{index}(\alpha)$$

Resulting in the most characteristic news sources shown in Figure 2, which matches domain knowledge at the same time that it uncovers lesser-known sources.

5 CONCLUSION & FUTURE WORK

In this paper, we performed the first large-scale analysis of partisan WhatsApp groups in the context of Brazil 2018 presidential election. The methodology we disclosed allowed a sample that is, at the same time, more specialized and substantially larger than described in previous works. We were able to analyze how right- and left-wing users organized a myriad of small, constant rallies in WhatsApp, finding a number of distinct characteristics within right-wing groups – right-wing users are more prevalent, tightly connected, geographically distributed, and shared more multimedia messages and YouTube videos. Future work will target more specific behaviors such as expression of distrust or promotion of certain types of information across the political spectrum.

REFERENCES

[1] Josemar Alves Caetano, Jaqueline Faria de Oliveira, Hélder Seixas Lima, Humberto T Marques-Neto, Gabriel Magno, Wagner Meira Jr, and Virgílio AF Almeida. 2018. Analyzing and Characterizing Political Discussions in WhatsApp Public Groups. *arXiv preprint arXiv:1804.00397* (2018).

[2] Folha de São Paulo. 2018. 90% of Bolsonaro's Supporters Believe in Fake News, Study Says. https://folha.com/0futoq3n

[3] Kiran Garimella and Gareth Tyson. 2018. WhatsApp, Doc? A First Look at WhatsApp Public Group Data. *In Proceedings of the ICWSM* (2018).

[4] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM* 11, 5th (2011), 25.

[5] BBC Audiences Research. 2018. Duty, Identity, Credibility: "Fake News" and the Ordinary Citizen in India. https://downloads.bbc.co.uk/mediacentre/duty-identity-credibility.pdf

[6] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proceedings of the 2019 World Wide Web Conference.*

[7] Gustavo Resende, Johnnatan Messias, Márcio Silva, Jussara Almeida, Marisa Vasconcelos, and Fabrício Benevenuto. 2018. A System for Monitoring Public Political Groups in WhatsApp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web.* ACM, 387–390.

[8] Avi Rosenfeld, Sigal Sina, David Sarne, Or Avidov, and Sarit Kraus. 2016. WhatsApp Usage Patterns and Prediction Models. *ICWSM/IUSSP Workshop on Social Media and Demographic Research* (2016).

[9] Tribunal Superior Eleitoral. 2018. 2018 Electoral Results. http://divulga.tse.jus.br/oficial/index.html