(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures

Gustavo Resende[1], Philipe Melo[1], Hugo Sousa[1], Johnnatan Messias[2]
Marisa Vasconcelos[3], Jussara M. Almeida[1], Fabrício Benevenuto[1]
[1]Universidade Federal de Minas Gerais, Computer Science Department, Brazil
[2]Max Planck Institute for Software Systems (MPI-SWS), Germany
[3]IBM Research
{gustavo.jota,philipe,hugosousa,jussara,fabricio}@dcc.ufmg.br
johnme@mpi-sws.org,marisaav@br.ibm.com

ABSTRACT

WhatsApp has revolutionized the way people communicate and interact. It is not only cheaper than the traditional Short Message Service (SMS) communication but it also brings a new form of mobile communication: the group chats. Such groups are great forums for collective discussions on a variety of topics. In particular, in events of great social mobilization, such as strikes and electoral campaigns, WhatsApp group chats are very attractive as they facilitate information exchange among interested people. Yet, recent events have raised concerns about the spreading of misinformation in WhatsApp. In this work, we analyze information dissemination within WhatsApp, focusing on publicly accessible political-oriented groups, collecting all shared messages during major social events in Brazil: a national truck drivers' strike and the Brazilian presidential campaign. We analyze the types of content shared within such groups as well as the network structures that emerge from user interactions within and cross-groups. We then deepen our analysis by identifying the presence of misinformation among the shared images using labels provided by journalists and by a proposed automatic procedure based on Google searches. We identify the most important sources of the fake images and analyze how they propagate across WhatsApp groups and from/to other Web platforms.

CCS CONCEPTS

• **Networks** → **Online social networks**; • **Mathematics of computing** → *Network flows*; • **Information systems** → *Mobile information processing systems*; *Chat*.

KEYWORDS

WhatsApp groups, misinformation, information dissemination, social network structure, fake images

1 INTRODUCTION

WhatsApp is a world-wide popular messaging app with more than 1.5 billion active users [2] which is currently the main messaging app in many countries, including India, Brazil, and Germany. Nearly everyone with a smartphone uses WhatsApp in Brazil (about 120 million active users [15]) to keep in touch with friends and family, do business, as well as read the news.

There are key features in WhatsApp that make this app unique. First, any communication within the app is end-to-end encrypted, meaning that messages, photos, videos, voice messages, documents, status updates, and calls are only seen by those involved in the communication. Second, WhatsApp allows users to easily create and organize chat groups. These groups, which are limited to 256 members, are by default private, as group administrators decide who can join them. However, a group manager may choose to share the link to join it in websites or social networks. In such a case, anyone with access to the link can join the group, which becomes, from a practical perspective, publicly accessible. Finally, WhatsApp provides features for viral spreading, allowing users to broadcast an initial message to 256 contacts or groups or forward content to 20 contacts or groups[1].

Recent events have raised serious concerns that WhatsApp can become a fertile ground for groups interested in disseminating misinformation, especially as part of articulated political campaigns. In 2018, unfounded allegations disseminated over WhatsApp have fueled mob lynching in India that killed more than 20 people in a two-month window [9]. The 2018 Brazilian elections experienced an information war organized within WhatsApp where false rumors, manipulated photos, decontextualized videos, and audio hoaxes have become campaign ammunition and went viral on the platform with no way to monitor their full reach or origin [15].

This paper provides a large scale investigation of information dissemination within *WhatsApp groups*. We focus on *political-oriented* publicly accessible groups as we expect greater user engagement in topics of stronger social impact. We also offer a first look into *misinformation* dissemination within *WhatsApp*. More specifically, we tackle the following research questions.

RQ1: What kind of content is shared in WhatsApp *publicly accessible* groups? Are there fake news or misinformation in these messages? **RQ2:** What is the interplay between WhatsApp and other Web platforms (i.e. social networks such as Twitter, forums, and websites) in the dissemination of political content and, in particular, misinformation?

[1]The message forwarding was limited to 5 groups in India and 20 in the rest of the world along the period this work was developed. Currently, the limit has been updated to 5 worldwide.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.

To answer these questions, we first identify publicly accessible groups related to Brazilian politics in WhatsApp, by searching the Web and other social networks such as Twitter and Facebook for invitation links to WhatsApp groups. These groups are suitable for activism and political engagement, making them a potential target of misinformation campaigns that might attempt to maximize the audience of a story with misinformation by sharing it with people that are engaged in supporting political candidates. We joined those groups and gathered the content shared within them for time periods corresponding to two major social mobilization events in Brazil: (i) a national truck drivers' strike[2] (May 21^{st} to June 2^{nd}, 2018); and (ii) the first round of the 2018 Brazilian general elections campaign (August 16^{th} to October 7^{th}, 2018), with 141 and 364 groups monitored, respectively.

We start our investigation by first analyzing the content shared and the user interactions within the monitored WhatsApp groups to understand how users disseminate information in such environments. Our results show that images are often the most shared type of media, and they usually carry satires, news, and activism-related content. We also show that WhatsApp has a network configuration similar to many other online social networks (e.g., Twitter or Facebook) which connects thousands of users. Thus, it has the potential to make any information become viral. Moreover, we analyze how the content dissemination crosses the boundary between WhatsApp and other Web platforms.

Further, we explore the presence of misinformation campaigns in the monitored groups. We identify misinformation in image content by relying on two sources: (i) a Brazilian fact-checking agency; and (ii) a proposed automatic procedure that exploits the results of Google searches to identify images that appear in well-known fact-checking websites. Our results show a considerable number of images checked as containing misinformation, which was largely disseminated in the monitored groups. We assess the occurrence of such images in Web domains and Twitter accounts then analyze the network of misinformation propagation. Our analyses reveal that a few groups are the most responsible for disseminating images with misinformation. Moreover, by comparing the timestamps when an image first appeared on WhatsApp (as captured by our data) and on other Web applications, we find that WhatsApp was the primary source of 30% of the identified images containing misinformation.

The remainder of the paper is organized as follows. Next section discusses related work. Then, Section 3 details our methodology and summarizes the data gathered. Section 4 characterizes the content, the network structure of the groups we monitored and how information propagates in such groups. Section 5 presents our main results on the dissemination of misinformation in those groups. Finally, Section 6 concludes the paper.

2 RELATED WORK

Fake news have encountered a suitable means for fast, cheap, and easy dissemination in social media systems. Indeed, these platforms have been a main vehicle for public opinion manipulation and fake news dissemination [19, 21, 25]. Studies during the 2016 U.S. presidential election campaign observed a strong correlation of the number of visits to fake news websites (i.e., sites that deliberately publish hoaxes and misinformation) and aggregate voting patterns at state and county levels [7].

Social bots are one of the most common types of manipulation attack, emulating real users, posting content and interacting with real users and other bots [5, 16]. They were used on Twitter during the 2016 U.S. presidential campaign to manipulate discussion [1] reaching about 20% of all posts about presidential elections. Another misinformation campaign was observed during the 2017 French presidential election, in which bots' posts with unauthentic documents about a candidate quickly spread on Twitter two days before the final voting [4]. Facebook was also a target of misinformation spread aiming at influencing American voters during the 2016 presidential campaign. Using Facebook Ads platform, groups linked to the Russian Intelligence Research Agency (IRA) bought about 3,000 ads linked to 470 user accounts targeting voters from the swing states [12, 21]. Since then, Facebook has performed measures to mitigate fake news dissemination such as removing fake accounts related to political movements and working directly with fact-checking websites [13]. However, fake news is not only disseminated exclusively by social bots or through ads but also by real users. A recent study analyzed over 126,000 cascades of fact-checked news stories on Twitter, finding that fake news was 70% more likely to be retweeted than true stories, and humans are more likely to spread fake news than bots [25].

More recently, WhatsApp has also become a powerful tool to influence people during political campaigns, especially in countries in South America, Africa, and Southeast Asia. In these locations, WhatsApp groups constitute the majority of online communication, enabling politicians to reach a larger share of voters, even those residing in areas with precarious Internet connectivity. Other features that make communication via WhatsApp attractive for this kind of marketing are: it is cheaper, messages are often not contextualized and it can be used to target small groups with specific messages [20, 23]. This was observed in Brazil, where family groups were responsible for 51% of the dissemination of fake news on WhatsApp during the period of the 2018 presidential elections [10]. Because of the end-to-end encryption, it is hard to track the dissemination of misinformation in such platforms. Yet, recent efforts have gathered and analyzed data from WhatsApp chat groups [8, 20], focusing on textual interactions in these groups. Our present effort provides a deeper understanding of the content exchanged in WhatsApp, unveiling, among other findings, the spread of misinformation campaigns through images in the platform.

3 METHODOLOGY

In this section, we present how we gathered data of WhatsApp groups and the methods used to process and analyze it.

3.1 Data Collection

As a first step of our data collection, we had to identify a considerable number of publicly accessible groups. To that end, we used the URL pattern `"chat.whatsapp.com"`, which is commonly used in invitations to join WhatsApp groups, as a search query and submitted it to Google, Twitter, and Facebook search engines. We restricted our search space to groups related to Brazilian politics,

[2]This was a massive event, with strong political connotation, that affected the whole country, paralyzing many economic and social sectors of the nation.

by including in each search query a word from a dictionary related to the 2018 Brazilian elections[3]. This dictionary contains the name of politicians, political parties, as well as words associated with political extremism. Finally, we performed a manual inspection of the collected group names to filter out those unrelated to politics. In total, we found 3,444 distinct links for publicly accessible groups, out of which only 1,828 were valid (i.e., unbroken).

As a second step, we selected a number of valid groups to monitor. This monitoring involves joining each group using a cell phone. Thus, the number of groups (see Section 3.3) monitored was constrained by the available devices and their resources (memory). We joined each selected group using our available cell phones and a tool developed by Garimella *et al.* [8]. We then periodically downloaded all data shared in each group and stored them in a database. Specifically, stored data can be grouped into: images, videos, audio messages, external links, and text messages. From each message, we extracted its group name (i.e., the group the message was posted), a group ID, a user ID, and timestamp. That is, we mapped telephone numbers and user names onto unique user identifiers[4], discarding the original information afterwards. For the media messages, we also downloaded their respective files and used their filenames as a reference to the message.

As we will show in Section 3.3, images are the most frequent type of *media* content in our collected data, as well as an important source of misinformation. Thus, we delved further into the images shared on the monitored groups and developed a tool to collect Web pages in which those images have appeared. The tool exploits the capability of searching for images provided by Google search, where a user can submit an image as query, and obtain as result webpages that include matching images along with their post dates. As will be discussed in Section 4, this information allows us to analyze temporal sharing patterns such as the time interval between first appearance of an image on WhatsApp and on the Web.

In order to explore the content veracity of the images shared in the groups, we extended the tool to automatically identify whether each image had been previously checked as *fake* by a number of fact-checking websites. We further elaborate on this process in Section 5.2.

3.2 Data Limitation

To our knowledge, this work is the first effort that aims to explore the political debate in WhatsApp. Also, it proposes a methodology to infer which identified publicly accessible groups are related to politics. Unfortunately, we are not aware of an approach that would allow us to assess the representativeness of our data as even the total number of groups available in the country is *not* of public knowledge. We emphasize, though, that all sensitive information (i.e., user names and phone numbers) were *not* stored in our dataset.

3.3 Events Captured in our Dataset

Our data collection focuses on the time period of two major social mobilization events in Brazil: (i) a national truck drivers' strike (May 21^{st} to June 2^{nd} 2018); and (ii) the first round of the 2018 Brazilian presidential elections campaign (August 16^{th} to October 7^{th} 2018). The truck drivers' strike was a stoppage of autonomous truck drivers all over Brazil. The strikers began a mobilization through their leaders in social networks, expressing against frequent adjustments and without minimum predictability in fuel prices, especially diesel, made by the state-owned company Petrobras[5]. The shutdown and blockades of highways in 24 states and in the Federal District caused the unavailability of food and medicine around the country, shortages and high gasoline prices, with long queues to fuel. We analyze the data collected for each aforementioned period as a separate dataset, contrasting our findings across periods.

Table 1: Overview of our datasets.

	Truck Drivers' Strike	Election Campaign
#Groups	141	364
#Total Users	5,272	18,725
#Total Messages	121,781	789,914
#Text Messages	95,424	591,162
#Images	11,610	110,954
#Videos	9,752	73,310
#Audios	4,995	14,488
#URLs	11,728	92,654

Table 1 provides an overview of our two datasets, showing the total number of messages shared as well as the number of messages per type of content (text[6], image, video and audios). Note that most shared messages are indeed textual content, but image is the most frequent type of *media* content in both datasets, reaching roughly 10% and 15% of all content shared in the monitored groups during the strike and election period, respectively. Note also the large number of links to websites (last row) present in the text messages.

We shared a sample of our dataset to a set of journalists which led to wide press coverage. Particularly, a BBC story[7] manually analyzed the popular content from the truck drivers' strike data and suggested that it was organized in a decentralized way through WhatsApp groups. Audios were often used to call truck drivers to join and remain in the strike, while videos and images were used to disseminate scenes of the strike and its consequences in different cities. Similarly, another BBC story[8] analyzed popular content collected during a week of the electoral period, reporting presence of misinformation in different kinds of content, but mainly in images. They reported audios describing conspiracy theories, manipulated photos, fake polls, attacks to the traditional media and to famous personalities, as well as the instigation of hate, especially against LGBT and feminists.

4 MESSAGE CONTENT AND DISSEMINATION

In this section, we characterize the content of the messages, the network structure of the groups as well as the temporal patterns of message sharing in our datasets.

[3]https://goo.gl/PdwAfV

[4]Throughout the paper we refer to such identifiers as users. Yet, they are indeed unique telephone numbers, as we are not able to identify multiple devices of the same user.

[5]http://www.petrobras.com.br/en/

[6]Only messages that are entirely composed of textual content are counted as text messages.

[7]https://www.bbc.com/portuguese/brasil-44325458

[8]https://www.bbc.com/portuguese/brasil-45666742

4.1 Media Content and URLs

As our first analysis, we focus on the content shared, notably media content (audios, videos, and images). We also briefly discuss the presence of URLs in textual messages. The discussion is based on Table 1 as well as on Figure 1, which shows the complementary cumulative distribution of the number of messages of different media types shared per day, across all monitored groups.

4.1.1 URLs and Webpage domains. As shown in Table 1, a total of 11,728 URLs were shared (as part of text messages) in the monitored groups during the truck drivers' strike period. During the election campaign, this number reached 92,654 URLs. They correspond to 8% and 9% of the total amount of messages we gathered for each period, respectively. The nature of these URLs varies from links to news websites, blogs, entertainment, and other social networks to even links to other WhatsApp groups. Also, 69% of all URLs shared during the truck drivers' strike period are unique. This fraction drops to 45% during the election campaign, indicating less diversity and more repetition of the links during that period.

4.1.2 Audios. Our datasets contain 4,995 and 14,488 audio messages during the truck drivers' strike and election campaign periods, respectively, which correspond to 4% and 2% of all messages gathered during each respective period. As shown in Figure 1, audio is the least frequent media type shared in the monitored groups. For example, during the truck drivers' strike, up to 450 audios were shared daily, in 60% of the days (see Figure 1(a)). During the election campaign period, up to 500 audios were shared in the same fraction of the monitored days. Note however that, this number reached a peak two days before the election day, with 1,002 audio messages shared on all monitored groups.

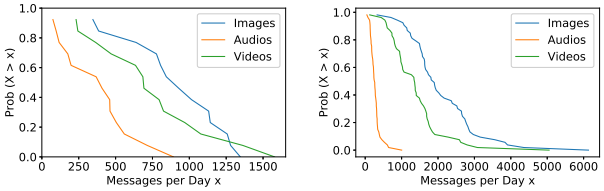
4.1.3 Videos. In total, 9,752 videos were shared during the truck drivers' strike, which corresponds to 8% of all messages shared during the period. Also, according to Figure 1, up to 800 videos were shared daily in 60% of the days, reaching up to 1,578 videos on a single day. During the election campaign period, 73,310 videos (9% of all messages) were shared in total, with 1,300 videos being shared daily in 60% of the days (up to 5,052).

4.1.4 Images. As mentioned, images represent the most popular media content shared on the monitored groups, with a total of 11,610 images (almost 10% of all messages) and 110,954 images (15%) shared during the truck drivers' strike and election campaign periods, respectively. According to Figure 1, this higher frequency happens on a daily basis. Up to 1,000 and 2,000 images were shared in 60% of the days during the truck drivers' strike and during the election campaign, respectively. Even more, in about 5% of the days, the number of images shared on the groups on a single day exceeded 1,350 and 6,100 in the two periods. For the election campaign (the latter), these days are in the week before the election day.

Given the popularity of images, we next deepen our analysis of the dissemination of type of content in the monitored groups.

4.2 Characterizing WhatsApp Images

A WhatsApp group is usually meant to be a space for discussions about a specific subject such as politics, education, games. However, the content shared itself may diverge from the group subject given the will of their participants. For example, as in other Web systems, WhatsApp groups are susceptible to spam activity (e.g. advertisements or inappropriate content). To understand the kinds of images shared on our selected groups, we first categorize the images by performing content labeling and analyze the distribution of images across categories. We also discuss the appearance of the same images on other websites and social networks.



(a) Truck Drivers' Strike (b) Election Campaign

Figure 1: Numbers of daily messages with media content shared on all groups.

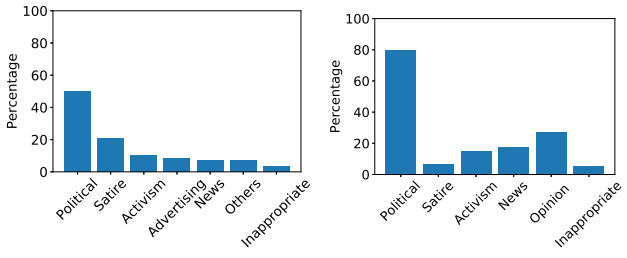
Table 2: Image categories used to label the sampled images.

Category	Description
Political	Information about a candidate or party
News	News information with a quote
Advertising	Advertisement of product, service or company
Satire	Humorous content regarding current events
Inappropriate	Illicit products, violence, hate speech or pornographic content
Activism	Popular movements and protests
Opinion	Expression of a personal opinion or comment
Others	Image does not fit in any other category

4.2.1 Content Labeling. We asked three volunteers to label a sample of the most shared images during each period. The sample from the truck drivers' strike period contains a selection of the top-20 most shared images on each day, with a total 220 images. For the election campaign period, the sample contains the top-100 most shared images, considering the whole monitored period. In order to identify duplicates of the same image, we used the *Perceptual Hashing (pHash)* algorithm [17] to calculate a fingerprint for each image. We were then able to group images having the same hash-values based on human eye perception as duplicates.

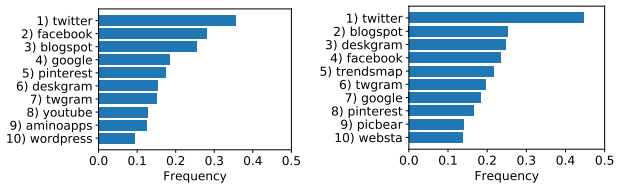
A taxonomy guideline document with instructions was given to the volunteers with the following directions: (i) observe an image and, read the text on it, if available; (ii) if there is a text, check the existence of any citation to a website or other source; (iii) check if the following content types are present in this post: *Political Content; News; Advertising; Opinion; Satire; Activism*; (iv) identify possible inappropriate, offensive or even illegal content by checking for the presence of *Dissemination of Hate; Violence; or Promotion of Illicit Products as Inappropriate Content*; (v) you can classify a post with more than one category (e.g., News and Political Content) or none of them; and finally (vi) if you cannot fit the image in any of the listed categories or are unable to establish its category, label the image as *Others*. Table 2 lists the categories used to label the sampled images.

After each of the three volunteers annotated each image according to the categories in Table 2, we measured the inter-annotator agreement in terms of the *Fleiss's* κ [6]. We assumed that consensus



(a) Truck Drivers' Strike (b) Election Campaign

Figure 2: Distributions of image categories.



(a) Truck Drivers' Strike (b) Election Campaign

Figure 3: Most popular domains for images shared on WhatsApp publicly accessible groups.

was reached if the null hypothesis of negative or no agreement $\kappa = 0$ can be rejected. Since the same image may fit more than one category, we applied the test individually for each category, averaging the κ scores obtained. The test result for the *Others* category indicated a poor agreement for the sample collected during the election campaign period. This result was then disregarded. Overall, we obtained moderate agreement among the annotators for both periods, with average κ equal to 0.6 and 0.42 for the truck drivers' strike and election campaign samples, respectively. This result is reasonable given that some categories are very broad and distinctions are somewhat blurred. In the following, we assume that an image belongs to a category if at least two of the annotators agreed upon that category.

Figure 2 shows the distributions of the image categories in each sample. In both periods, most images are related to politics (50% during the truck drivers' strike and 80% during the election campaign). The large fractions of satire and activism during the truck drivers' strike are also worth noting: we observed a lot of memes about the movement and messages supporting the truck drivers. A small fraction of images shared during that period contains advertisement. In contrast, images with opinions from personalities were much more popular during the election campaign, corresponding to 20% of all images in our sample. News and images with activism were also frequently shared during that period, but no image in our sample contain explicit advertisements.

4.2.2 WhatsApp Images on other Websites. We now analyze the extent to which the images shared on our monitored WhatsApp groups have also appeared on other websites including social networks and blogs. We do so by searching for the observed images using the Google Images search engine, as discussed in Section 3.1.

Figure 3 shows the most popular domains returned by Google Images for the images shared in each monitored period. Notice that online social networks like Twitter, Facebook, Aminoapps, and Pinterest are among the most frequent domains where the images were also posted, for both periods. Similarly, image apps like Deskgram and Twgram as well as Blogspot are popular domains, especially the latter, suggesting that a large fraction of image content shared on WhatsApp groups may indeed have blogs as possible sources. TrendsMap, a website that shows visualizations of the trends on Twitter, was popular for images shared during the election campaign period, while more than 10% of the images shared during the truck drivers' strike period also appeared on YouTube, as thumbnails of videos.

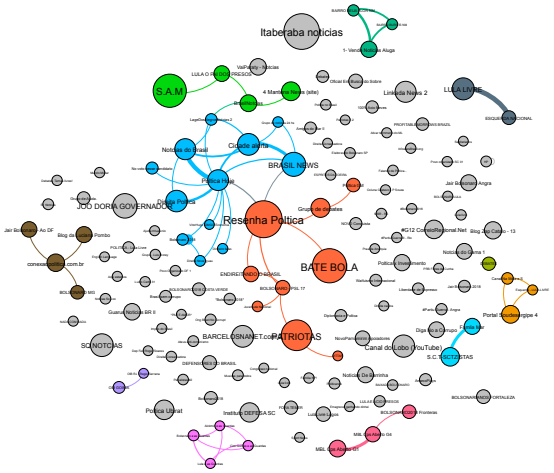
Table 3: Sharing of images on monitored WhatsApp groups.

	Truck Drivers' Strike			Election Campaign		
	Mean	Std. Dev	Max	Mean	Std. Dev	Max
#Images per Group	91	143.250	1,011	345	580.531	4,320
#Users per Group	17	21.745	110	34	41.347	211
#Images per User	5	10.831	197	10	32.322	1,612
#Shares for Image (total)	1	1.978	58	1	4.512	125

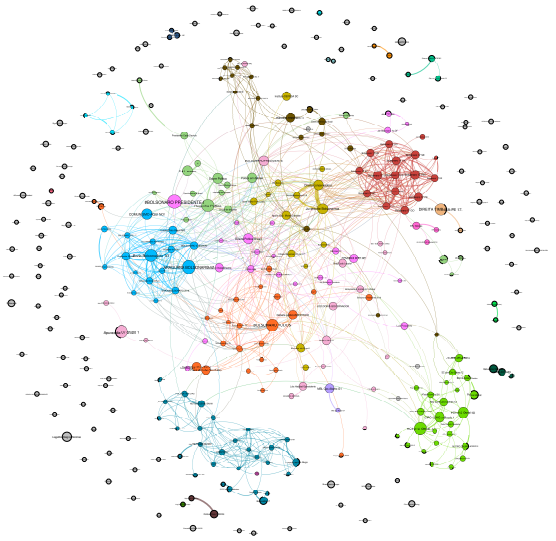
4.3 Network Structure

In this section, we analyze sharing patterns of images on the selected WhatsApp groups by studying the structure of the networks that emerge from the participation of users in different groups. To better understand this network structure, we first discuss some key measures related to the sharing of images within each group. For each period, Table 3 presents averages, standard deviations, and maximum values of the numbers of images shared by each user and within each group as well as number of users sharing images in each group and the total number of times each image was shared (across all groups). Overall, all averages are larger for the election campaign period. Although some differences may be (partially) credited to a longer monitoring period, we note that the intensity of image sharing per user was indeed higher during the campaign, with an average twice as larger and a peak eight times as larger than in the strike period. Similarly, we do observe a significant increase in the number of users sharing images in the groups. Note however that most images are shared only a few times (once, on average), as only a few images are widely shared. Interestingly, we found that the groups with the largest number of images shared during the strike (*"Resenha Política"* with 1,011 images) and during the election campaign (*"#BOLSONARO PRESIDENTE"* with 4,320 images) are indeed the groups with the largest numbers of users sharing this type of content (110 and 211, respectively).

We modeled the interactions across groups by means of two network models, one at the group level and one at the user level. That is, we built a *group network* where each node represents one monitored group and edges are added connecting groups that have at least one member in common sharing image content. Figure 4 shows the group networks built for each monitored period. The size of each node represents the number of users who shared image content in the group. Although many groups are somewhat isolated or weakly connected to the rest, we do note the presence of several clusters of groups which are strongly interconnected by sharing many members in common. This may facilitate the flow of information across group boundaries.



(a) Truck Drivers' Strike Network



(b) Election Campaign Network

Figure 4: Group networks (nodes are groups and edges connect groups with users in common).

We also modeled the relationship between users by building a *user network* where each node is a user and an edge is added between two nodes if the corresponding users have shared image content in at least one group in common. Node size represents the number of groups in which the user shared images. The user networks are naturally larger and harder to visualize. For illustration purposes, Figure 5 shows a subgraph of the network built for the election campaign, with 5,700 nodes. The network structure of the groups is evidenced by the clusters formed. We note a large number of users blending together connecting to each other inside those groups. Most users indeed form a single cluster, connecting mostly to other members of the same community. On the other hand, there are also a few users who serve as bridges between two or more groups linked by multiple users at the same time. Furthermore, a few users work as big central hubs, connecting multiple groups simultaneously. Lastly, some groups have a lot of users in common, causing these groups to be strongly inter-connected, making it even difficult to distinguish them.

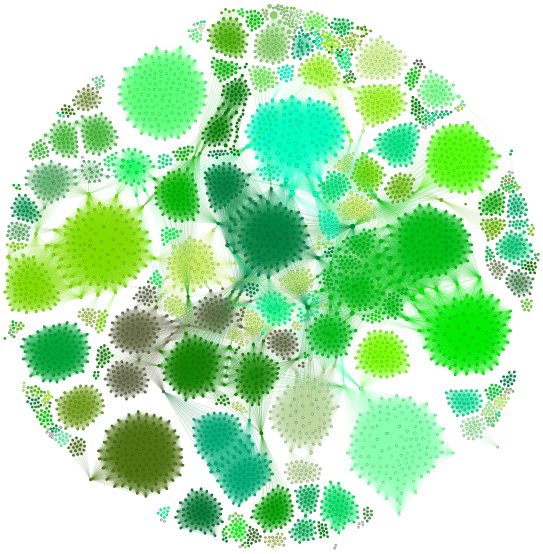


Figure 5: User network (nodes are users and edges connect users with group in common. Subgraph of election period).

Table 4: Network metrics for WhatsApp graphs.

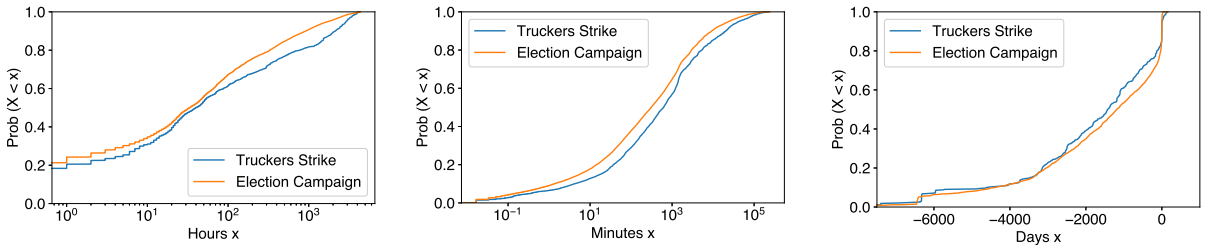
	#Nodes	#Edges	Avg. Degree	Diameter	APL*	Density	LCC**
Group Network Truck Drivers' Strike	136	55	0.809	8	2.975	0.006	25
Group Network Election Campaign	333	842	5.057	8	3.459	0.015	206
User Network Election Campaign	10,860	492,217	90.91	9	3.952	0.008	8,934

*Average Path Length. **Largest Connected Component.

To better understand the properties of these graphs, Table 4 shows various network metrics computed for the group and user networks. It presents numbers of nodes and edges, average node degree, network diameter, average path length (APL), network density, and the size of the largest connected component (LCC).

Focusing on the group networks, we see that the graph for the election campaign is more complex and more densely connected, with more clusters (i.e., communities) of groups and edges emerging between them, and a larger fraction of nodes belonging to the largest connected component (62%). Despite such differences, the average path length between the groups is only slightly larger (3.46, against 2.97 in the strike period). Also, although network density (ratio of number of edges in the graph to the maximum number of edges possible) is low for both periods (under 2%), it is higher during the election campaign. We observe similar properties in the user network with a small average shortest path length (3.95) and higher largest connected component (82% of the users).

Therefore, as illustrated in Figures 4 and 5 and in Table 4, WhatsApp is more than just a mobile network that provides end-to-end encrypted communication between two users. It exhibits network properties very similar to many other social networks such as Twitter or Facebook, connecting thousands of users and having the potential to make a piece of information become viral.



(a) Lifetimes (b) Burst times (c) Interval between first share on WhatsApp and Web

Figure 6: Distributions of temporal properties characterizing propagation dynamics of images.

4.4 Propagation Dynamics

We analyze the propagation dynamics of images within the monitored WhatsApp groups by means of two metrics, namely lifetime and burst time. The lifetime of an image is the time interval between the first and the last time the image was shared, as captured by our datasets. Burst time is the time interval between consecutive shares of the same image, irrespective of the group. In our analysis, we first identify the set of images shared during each monitored period and then compute their lifetimes and burst times considering an extended monitored period from April 23^{rd} to October 22^{nd} 2018. We also analyze the propagation dynamics of images as they cross the boundaries of WhatsApp, appearing elsewhere on the Web.

4.4.1 Lifetimes. As shown in Figure 6(a), the distributions of lifetimes are mostly similar[9]: around 20% of the images have lifetimes under 1 hour, and 40% of them have lifetimes under 20 hours. Yet, many images last quite longer on the system: more than 30% of the images have lifetimes exceeding 100 hours.

4.4.2 Burst Times. Figure 6(b) shows the distributions of burst times. Once again, both distributions are mostly similar, although there is a tendency of images being reshared within shorter time intervals during the election campaign. For example, in 40% of the cases, images were reshared within up to 120 and 100 minutes during the strike and election campaign periods, respectively. Yet, some images were reshared very sporadically: in around 20% of the cases, burst times exceed 5 and 2 days during the same periods.

4.4.3 Propagation to and from the Web. We also analyze the propagation of images across the boundaries between WhatsApp and the Web. Specifically, we analyze the difference between the time an image was first shared on a monitored group and the time when it was indexed by Google. The latter is taken as an estimate of the time it first appeared on the Web. A positive difference suggests that the image was first shared in one of the monitored groups and then published on the Web. A negative difference may suggest the image was first posted on the Web[10]. Figure 6(c) shows the cumulative distribution of such time differences for images shared during the two monitored periods. Note that, in both cases, the time differences are indeed negative for most images (80%). Yet, 14% of the images were posted on the same day on the Web and on the WhatsApp groups, and 6% were first shared on WhatsApp.

5 MISINFORMATION ON WHATSAPP

In this section, we look at the presence of misinformation in the images shared on WhatsApp groups. First, we discuss two techniques used to identify misinformation in the images in our datasets. We analyze their characteristics and propagation, and compare these images with the rest of our WhatsApp data.

5.1 Labeling with a Fact-Checking Agency

We created a list of the most shared images during the election campaign period and gave them to one of the most important fact-checking agencies in Brazil, *Lupa* [11]. They checked the veracity of each of these images following a methodology similar to other fact-checking agencies around the world (e.g., the American *Politifact*[12] and the Argentinian *Chequeado*[13]). They first analyzed where these images contained factual information as opposed to opinions since it is not possible to check the latter. Out of a total of 61 images, 47 were marked as factual. Out of these factual images, they found that 22 had already been checked by other fact-checking agencies: 17 images had been checked as containing misinformation, and only 5 images had been checked as true. These results show an expressive number of images with misinformation in WhatsApp during the 2018 Brazilian elections. In terms of percentages, 36.2% of the images with factual information where checked as containing misinformation, whereas 53.2% of them include misleading and inconclusive content (not supported by public information), and only 10.6% were verified as true. Examples of images checked as misinformation are shown in Figure 7.

Figure 7(a) is an edited image of the Brazilian former president Dilma Rousseff, who was impeached in 2016 [22], alongside Fidel Castro, former president of Cuba. At the time this picture of Castro was taken, Dilma was 11 years old. Thus, the image is clearly fake. It was the most popular image in the analyzed period. Figure 7(b) is an edited image of the former Brazilian president Lula, imprisoned for corruption at the time of monitoring [3, 24], meeting the aggressor responsible for stabbing the then presidential candidate Jair Bolsonaro during a campaign rally [18]. The intention of the image was to associate Lula with the attack against Bolsonaro.

[9] We note that the somewhat shorter lifetimes for images shared during the election campaign may be a side effect of the interruption of monitoring.
[10] Our analysis is constrained by the view of WhatsApp provided by our datasets.
[11] https://piaui.folha.uol.com.br/lupa
[12] https://www.politifact.com
[13] https://chequeado.com

(a) Fabricated image of former president Dilma Rousseff next to Fidel Castro.

(b) Fabricated image of Bolsonaro's aggressor next to former president Lula.

Figure 7: Images checked as containing misinformation by both fact-checking methodologies.

5.2 An Automatic Methodology for Finding Misinformation

Recently, Facebook has announced partnerships with many third-party fact-checking organizations, through which Facebook demote or reduce the visibility of links rated as false [14]. This kind of partnership neglects misinformation in images, as fact-checkers only provide rates to links containing stories with misinformation. Next, we provide a strategy to connect the false stories found in images shared with external links that appear on the Web, providing a simple way for Facebook to demote links containing images with misinformation identified on WhatsApp.

First, we identified the main fact-checking agencies in Brazil[14]. We then automatized the process of searching each image shared on the WhatsApp groups on the Web by using the Google Image search. Given the search results for an image, we checked whether any of the returned pages belong to one of the fact-checking domains. If so, we parsed the fact-checking page and automatically labeled the image as fake or true depending on how the image was tagged on the fact-checking page.

We applied this methodology to all images in both datasets. We found only 2 images with misinformation in the truck drivers' strike dataset and 70 images containing misinformation in the elections dataset. Thus, we restricted our focus to the election campaign period. We compared the 70 images with misinformation identified by the automatic process with the 17 images checked as fake by Lupa (see previous section), obtaining an overlap of only 2 images. Thus we built a single dataset of 85 images with misinformation identified by official fact-checking agencies.

In the following, we analyze the images in this dataset, focusing on other websites where they also appear. We also compare properties of these images with those of the other images shared during the election campaign period. To distinguish between them we refer to the former as misinformation and to the later as unchecked, since the veracity of their content was not necessarily checked. We cannot guarantee the absence of misinformation in the unchecked images, given that such an assertion is restricted by the availability of checked facts. Yet, we expect that we were able to catch most images containing misinformation in our dataset, especially those with greater impact on users, as they most probably were identified by the fact checkers.

[14]Fact-checking agencies: *Boatos.org*: https://www.boatos.org; *e-Farsas*: http://www.e-farsas.com; *Comprova*: https://projetocomprova.com.br; *Lupa*: https://piaui.folha.uol.com.br/lupa; *Globo G1*: http://g1.globo.com/fato-ou-fake; and *Aos Fatos*: https://aosfatos.org.

Table 5: Overview of images shared during election campaign period: misinformation versus unchecked content.

	Misinformation	Unchecked
#Groups in which images were shared	157	351
#Users who shared images	624	10,339
#Unique images	85	69,590
#Total shares of images	1,168	109,791

5.3 Images with misinformation on WhatsApp and on the Web

Table 5 presents a comparison of the images with misinformation and with unchecked content shared during the election campaign, showing the numbers of distinct images, users who shared those images, groups in which those images were shared and total number of shares. Note that, even though the number of distinct images with misinformation is small (85), these images summed up 1,168 shares posted by 624 different users in 157 different groups. Despite representing less than 1% of all images shared, these images appeared in 44% of the monitored groups in the period of the election campaign, effectively reaching a large user population. Also, note that nearly 5.7% of all users shared images with misinformation.

5.3.1 Network of propagation in WhatsApp groups. We analyzed the propagation of images with misinformation on the WhatsApp groups by building a network model representing the groups in which the images with misinformation *first appeared*. Specifically, we built a directed graph where each node represents a group and a directed edge from node A to node B was added if the same image with misinformation was first shared in group A and then appeared in group B. To build this graph we considered only groups in which at least 2 distinct images with misinformation were shared during the period. The weight of an edge is defined as the number of images containing misinformation that were first shared in a group and then co-occurred in the other. The size of a node represents the number of images with misinformation posted on that particular group while the color represents the sum of the outgoing edges, that is, the total number of images that were "first seen" in that group and then spread to the rest of the network.

Figure 8 shows the network of propagation of images with misinformation in the monitored WhatsApp groups during the election campaign. Note that some nodes are darker (larger out-degree) than others, suggesting they are the main "seeds" of the images with misinformation in the graph. It is worth noting that the group in which the largest number of images with misinformation first appeared (largest node) is indeed the group with the largest number of users and largest number of images shared in general. Yet, we note that some large nodes have very light colors (e.g. *"ARAGUANA BOLSONARO 1"* and *"BOLSONAROPRESIDENTE"*), meaning that although many images containing misinformation were shared in them, they acted more as receptors than seeds, since their out-degrees are small. These results seem to suggest that fewer groups are responsible for the spreading of a large fraction the images with misinformation in WhatsApp.

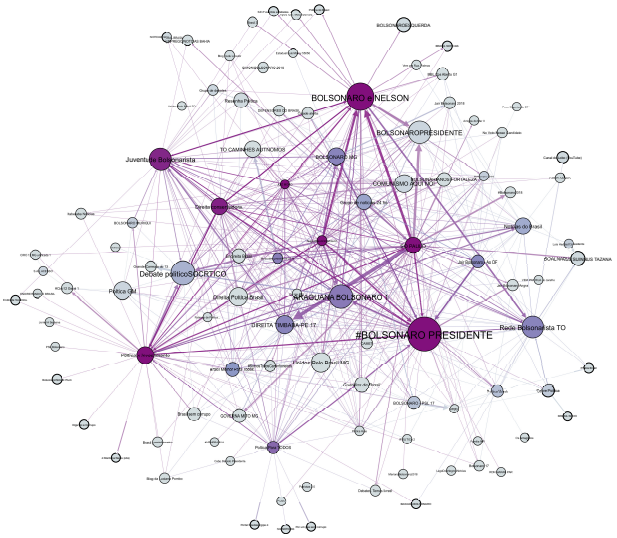


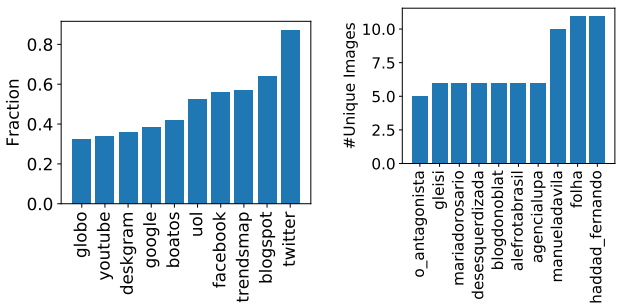
Figure 8: Network of misinformation propagation on WhatsApp groups (election campaign period).

5.3.2 Presence on the Web. We also analyzed the presence of images with misinformation on the Web. Figure 9(a) shows that these images frequently appeared in other social networks and blogs, notably Twitter, as also observed for images in general (Figure 3). Figure 9(b) shows the Twitter accounts that most shared these images, presenting the number of different images with misinformation tweeted by each account. Among them, there are some official journalistic accounts (*folha* and *agencialupa*) and official accounts of the presidential candidate Fernando Haddad and his vice, Manuela d'Avila. We note these profiles posted images containing misinformation with the purpose of repudiating them, acting as fact-checking accounts. Yet, some other accounts acted as misinformation broadcasters by spreading it further through the network.

5.3.3 Propagation Dynamics. Recall that, in Section 4.4 we analyzed image propagation dynamics by characterizing their lifetimes and burst times within WhatsApp groups as well as the time interval between their first appearance in WhatsApp and on the Web. We here revisit this analysis comparing the same metrics for images with misinformation and images with unchecked content, focusing on the election campaign period. The results are shown in Figure 10. For each metric, we compared the two distributions using the Kolmogorov-Smirnov test [11] with 95% confidence level, with the null hypothesis that two samples have the same distribution.

We found no statistical difference between the distributions of lifetimes of images with misinformation and images with unchecked content (p-value of 0.78). For both types of content, around 70% of the images remain in the system for up to 100 hours. In contrast, the distributions of burst times are statistically different (p-value of 2e-30). Burst times tend to be shorter for images with misinformation, suggesting a faster propagation of this type of content. For example, in 60% of the cases, an image with misinformation is reshared within 100 minutes. The fraction of such burst times reduces to 40% for images with unchecked content.

Similarly, the distributions of the time interval between first appearance on WhatsApp and on the Web are also clearly different (p-value of 2.4e-47). The vast majority (95%) of images with



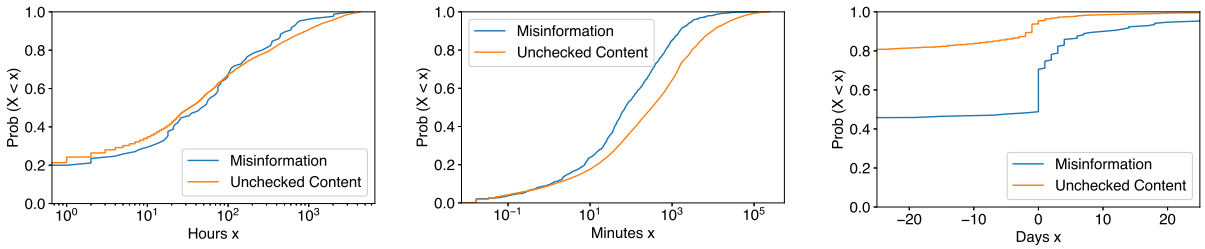
(a) Most popular domains. (b) Most popular Twitter accounts.

Figure 9: Images with misinformation on the Web.

unchecked content were first posted on the Web (negative intervals). Only 3% of them were shared first on the monitored groups (positive intervals) whereas 2% appeared on both Web and WhatsApp on the same day. In contrast, only 45% of the images with misinformation were shared first on the Web, 20% of them were shared on both platforms on the same day, and 35% were shared first on the WhatsApp group. These results seem to suggest that WhatsApp acted as a source of images with misinformation during the election campaign period.

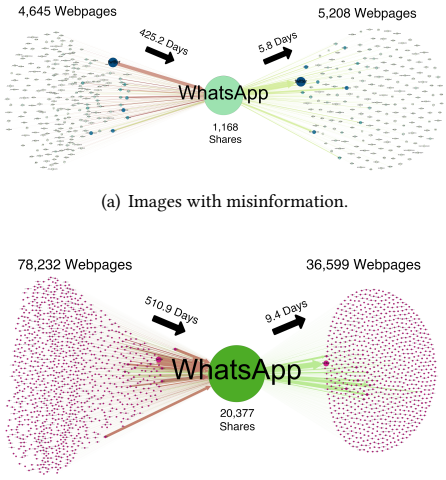
To further investigate the sharing of image content on the monitored WhatsApp groups and on the Web, we propose a visualization by means of a directed network, as shown in Figure 11. The network contains a central node representing WhatsApp (i.e., the monitored groups); the other nodes represent Web domains in which the images shared on WhatsApp also appeared. A directed edge from a node/domain to the central node implies that an image first appeared on that domain and later it was shared on WhatsApp. A directed edge from the central node to a node/domain implies the opposite. Thus, to improve readability, we plot nodes representing domains in which the images appeared before being shared on WhatsApp to the left of the central node, and nodes representing domains in which the images appeared after being shared on WhatsApp to its right. The size of a node representing a domain captures the number of webpages in that domain in which images shared on WhatsApp appeared. The color of an edge represents the average time difference between the first appearance of an image on WhatsApp and on the specific domain, considering all images posted on that domain (green is faster than red). We emphasize that this representation captures the temporal ordering of the first appearance of an image within WhatsApp and on the Web, as captured by our dataset. Although it may provide hints about the propagation of image across the boundaries between WhatsApp and the Web, we cannot claim they map exactly the actual information flow.

Figure 11 shows the network representations for images with misinformation and images with unchecked content. In addition to the network itself, each figure shows, for each group of domains, the total numbers of pages containing shared images as well as the average time interval between the first appearance of an image on the Web and on WhatsApp. We note that images that were first published on the Web take much longer to reach the WhatsApp groups (more than a year) than the other way around (only a few days) for both types of images. The average time interval is 73



(a) Lifetime (b) Burst Time (c) Interval between first share on WhatsApp and Web

Figure 10: Temporal properties of propagation of images with misinformation versus images with unchecked content.



(a) Images with misinformation.

(b) Images with unchecked content.

Figure 11: Network representation of images shared on WhatsApp and on the Web.

times longer for images with misinformation and 54 times longer for images with unchecked content. Also, in general, images with misinformation cross the boundaries between WhatsApp and the Web much more quickly: 425 days from the Web to WhatsApp and less than 6 days from WhatsApp to the Web, on average (as opposed to 511 and 9 days, respectively, for images with unchecked content). Moreover, the numbers of domains (and webpages) on both sides of the central node are much more balanced for images with misinformation. This suggests, once again, that images with misinformation are much more often spread from the WhatsApp groups to the rest of the Web than images with unchecked content.

6 CONCLUDING REMARKS

This work presents an analysis of messages shared on publicly accessible WhatsApp groups related to politics during two major events in Brazil. We found that images are the most popular type of media content shared on this platform during both periods analyzed. Moreover, by manually labeling these images, we found the frequent presence of satire, activism, and personal opinions, and much of this content came from other social networks. We also analyzed the temporal patterns of the propagation of these images within WhatsApp groups, finding that a large fraction of images remained being shared on the platform for quite some time (more than 4 days), and often within short time intervals (a couple of hours). We also found that most images shared on WhatsApp were actually posted first elsewhere on the Web. As a complement, we characterized the network structure of the monitored WhatsApp groups, showing how they connect with each other and offering insights into how information may propagate across them and to/from the Web.

We also proposed a methodology to automatically identify images with misinformation, and used it to investigate the sharing of this type of content in the monitored groups. We characterized the propagation dynamics of these images, contrasting it with the patterns observed for the other (unchecked) image contents. We found that images with misinformation tend to be reshared within shorter time intervals and are much more often shared first on WhatsApp and then on the Web. This observation suggests that WhatsApp may have been a relevant source of images with misinformation to the Web during the analyzed period.

As a final contribution, we also designed and deployed the **WhatsApp Monitor**[15], a Web-based system to help the top Brazilian official fact-checking agencies and journalists. Our system displays the most popular content shared in the monitored publicly accessible groups on a daily basis. This allows journalists to get an idea about critical content that is worth being fact-checked. We emphasize that sensitive information such as user names and phone numbers are discarded and thus are not shown by our system.

To our knowledge, this is the first effort to build a system of its kind. Our system has already been used by more than a hundred journalists with an editorial line and by three fact-checking agencies which explicitly mentioned our system as a data source. We hope this system can be useful during future election campaigns and other major events in Brazil and other countries.

More broadly, we expect this study to drive follow-up investigations covering other types of content as well as delving further into the interplay between WhatsApp groups and the Web as channels for information propagation.

ACKNOWLEDGMENTS

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-1, as well as grants from CNPq, CAPES, and Fapemig.

[15] http://www.whatsapp-monitor.dcc.ufmg.br

REFERENCES

[1] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21, 11 (2016).

[2] Josh Constine. 2018. WhatsApp hits 1.5 billion monthly users. $19B? Not so bad. (Jan 2018). https://techcrunch.com/2018/01/31/whatsapp-hits-1-5-billion-monthly-users-19b-not-so-bad/ "[Online; posted on 31-Jan-2018]".

[3] Shasta Darlington. 2018. As 'Lula' Sits in Brazil Jail, Party Nominates Him for President. *N. Y. Times* (Aug 2018). https://www.nytimes.com/2018/08/05/world/americas/lula-brazil-election-luiz-inacio-lula-da-silva.html

[4] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22, 8 (2017).

[5] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59, 7 (June 2016), 96–104.

[6] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

[7] Adam Fourney, Miklos Z. Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proc. of the CIKM*.

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

[9] Vindu Goel, Suhasini Raj, and Priyadarshini Ravichandran. 2018. How WhatsApp Leads Mobs to Murder in India. (July 2018). https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html "[Online; posted on 18-Jul-2018]".

[10] Juliana Gragnani. 2018. Pesquisa inédita identifica grupos de família como principal vetor de notícias falsas no WhatsApp. (April 2018). https://www.bbc.com/portuguese/brasil-43797257 [Online; posted on 20-April-2018].

[11] Frank J. Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[12] Young Mie Kim, J. Hsu, D. Neiman, C. Kou, L. Bankston, S. Kim, R. Heinrich, R. Baragwanath, and G. Raskutti. 2018. The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook. *Political Communication* 0, 0 (2018), 1–27.

[13] Mallory Locklear. 2018. Researchers say Facebook's anti-fake news efforts might be working. (Sep 2018). https://www.engadget.com/2018/09/14/facebook-fake-news-efforts-working/ [Online; posted on 14-Sep-2018].

[14] Tessa Lyons. 2018. Hard Questions: How Is Facebook's Fact-Checking Program Working? https://newsroom.fb.com/news/2018/06/hard-questions-fact-checking. (2018). "[Online; posted on 14-Jun-2018]".

[15] Matheus Magenta, Juliana Gragnani, and Felipe Souza. 2018. How WhatsApp is being abused in Brazil's elections. (October 2018). https://www.bbc.com/news/technology-45956557 "[Online; posted on 24-Oct-2018]".

[16] Johnnatan Messias, Lucas Schmidt, Ricardo Rabelo, and Fabrício Benevenuto. 2013. You followed my bot! Transforming robots into influential users in Twitter. *First Monday* 18, 7 (July 2013).

[17] Vishal Monga and Brian L. Evans. 2006. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE Transactions on Image Processing* 15, 11 (2006), 3452–3465.

[18] Dom Phillips. 2018. Jair Bolsonaro: Brazil presidential frontrunner stabbed at campaign rally. *the Guardian* (Sep 2018). https://www.theguardian.com/world/2018/sep/06/brazil-jair-bolsonaro-far-right-presidential-candidate-stabbed

[19] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019).

[20] 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 (WebMedia '18)*. ACM, New York, NY, USA, 387–390.

[21] Filipe N. Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabrício Benevenuto, Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. 2019. On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*'19)*. Atlanta, USA.

[22] Simon Romero. 2016. Dilma Rousseff Is Ousted as Brazil's President in Impeachment Vote. *N. Y. Times* (Aug 2016). https://www.nytimes.com/2016/09/01/world/americas/brazil-dilma-rousseff-impeached-removed-president.html

[23] Tactical Tech. 2018. WhatsApp: The Widespread Use of WhatsApp in Political Campaigning in the Global South. (2018). https://ourdataourselves.tacticaltech.org/posts/whatsapp/ "[Online; posted on 03-Jul-2018]".

[24] David Treece. 2018. Brazil's ex-president Lula imprisoned to keep him out of the election. (Jun 2018). https://www.theguardian.com/world/2018/jun/08/brazils-ex-president-lula-imprisoned-to-keep-him-out-of-the-election-letters

[25] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.