Strategies and Attacks of Digital Militias in WhatsApp Political Groups

Daniel Kansaon[1*], **Philipe de Freitas Melo**[2,1], **Savvas Zannettou**[3], **Anja Feldmann**[4], **Fabricio Benevenuto**[1]

[1] Universidade Federal de Minas Gerais, Brazil
[2] Universidade Federal de Viçosa, Brazil
[3] TU Delft, Netherlands
[4] Max Planck Institute for Informatics, Germany
daniel.kansaon@dcc.ufmg.br, philipe.freitas@ufv.br, s.zannettou@tudelft.nl, anja@mpi-inf.mpg.de, fabricio@dcc.ufmg.br

Abstract

WhatsApp provides a fertile ground for the large-scale dissemination of information, particularly in countries like Brazil and India. Given its increasing popularity and use for political discussions, it is paramount to ensure that WhatsApp groups are adequately protected from attackers who aim to disrupt the activity of WhatsApp groups. Motivated by this, in this work, we characterize two types of attacks that may disrupt WhatsApp groups. We look into the *flooding attack*, where an attacker shares usually numerous duplicate messages within a short period, and the *hijacking attack*, where attackers aim to obtain complete control of the group. We collect a large dataset of 19M messages shared in 1.6K WhatsApp public political groups from Brazil and analyze them to identify and characterize flooding and hijacking attacks. Among other things, we find that approximately 7% of the groups receive flooding attacks, which are usually short-lived (usually less than four minutes), and groups can receive multiple flooding attacks, even within the same day. Also, we find that most flooding attacks are executed using stickers (62% of all flooding attacks) and that, in most cases, attackers use both flooding and hijacking attacks to obtain complete control of the WhatsApp groups. Our work aims to raise user awareness about such attacks on WhatsApp and emphasizes the need to develop effective moderation tools to assist group administrators in preventing or mitigating such attacks.

Introduction

The rise of messaging applications has significantly transformed how people communicate and interact. WhatsApp, in particular, has achieved widespread usage, especially in Brazil and India, where virtually every cell phone user has embraced the platform (Bianchi 2022). This tool has seamlessly integrated into people's lives, becoming indispensable in various daily activities, including event organization, entertainment, news consumption, and business searches. Its extensive presence has made it an essential and indispensable component of modern existence.

The immense popularity of WhatsApp, coupled with the nature of the communication it facilitates, has created a highly convoluted and fertile environment for the propagation of misinformation campaigns. A notable example of this phenomenon can be seen in India, where false reports of child abductions have led to numerous instances of mob violence and lynchings (Samuels 2020). Also, in Brazil, a fact-checking effort conducted during the 2018 Brazilian presidential election process revealed an alarming statistic: 88% of the most popular messages circulating on the platform were found to be false or misleading (Resende et al. 2019b). The environment created by messaging apps like WhatsApp is inherently complex. While these platforms offer a variety of tools to facilitate the rapid dissemination of messages, they also maintain a level of anonymity that conceals the authors of these messages.

Since Jair Bolsonaro's election in 2018, WhatsApp has evolved into a bustling media space for political militancy. The public space within the platform has emerged as a hub of communication and organization, enabling the seamless coordination of activists. These public groups connected hundreds of very active users dedicated to spreading information to participants and groups, creating a backbone for information propagation within WhatsApp (Melo et al. 2019b). Misinformation campaigns feed these groups of activists, who are moved by loyalty to the preferred candidate and tend to amplify the reach of the messages received, regardless of their truthfulness. At the same time, given the polarized nature of political discussions, groups of activists organize themselves into digital militias to fight with each other and to promote hostile interactions towards opponents. This no man's land created within WhatsApp by digital militias is nearly unexplored by the research community.

Motivated by this research gap, this work presents the first comprehensive study that analyzes the strategies and attacks employed by digital militias operating in public WhatsApp groups. By examining the dynamics and tactics used by these groups, this research sheds light on the complex landscape of online political engagement and the role played by these militias in dismantling and attacking the opposing side's groups. The first type of attack identified and evaluated in this study is the flooding attack, a commonly employed tactic to disrupt the activity of the opponent group. This attack involves overwhelming the group with a high volume of messages, often causing chaos. By inundating the group with an excessive number of messages, the attackers aim to disrupt the normal flow of conversation and prevent

[*] The work was partly done during an internship at MPI-INF.

the effective exchange of information among group members. The second type is the hijacking attack, which involves the unauthorized takeover of a WhatsApp group by a malicious user, who aims to disrupt and dismantle the group. The hijacker gains control over the group, often exploiting vulnerabilities in the group's administration and then taking destructive actions, such as removing all members or spreading harmful content. To understand these attacks, we conducted an extensive data collection of WhatsApp political public groups from Brazil. We gathered 19M messages shared in 1,645 public groups. Then, using a mixed-methods data-driven approach, we identify and characterize flooding and hijacking attacks on WhatsApp groups. We believe that combining both quantitative measurements and qualitative assessment of the content used in these attacks is a suitable approach for characterizing and understanding these attacks.

Main Findings: Our main findings are:

1. We find that flooding attacks are not a rare phenomenon on WhatsApp groups focusing on Brazilian politics; 7% of all groups in our dataset received flooding attacks.
2. Stickers can be misused by attackers to undertake flooding attacks (62% of all flooding attacks). Also, we find a large percentage of attacks that use offensive, repulsive, or sexually explicit stickers.
3. Flooding attacks are short-lived (98% of them are below 4 minutes) and groups may receive multiple flooding attacks, even within the same day.
4. We find that hijacking attacks are conducted with flooding attacks to take full control of the group and "silence" the group (e.g., by forcing people to leave the group).
5. We find evidence that groups change their name to obfuscate their political nature to avoid prosecution.

Implications. Our work and findings have many implications for various interested stakeholders. First, for end-users, our work raises awareness about these two types of attacks and how they are conducted on WhatsApp. For platform operators, our work and findings can shed light on the modus operandi of the attackers involved in flooding and hijacking attacks and assist them in adjusting their platform governance. For group administrators and third-party developers, our work can motivate them to implement and employ effective moderation tools (e.g., moderation bots) to identify disruptive behaviors in the group and remove the attackers before making drastic changes to the group.

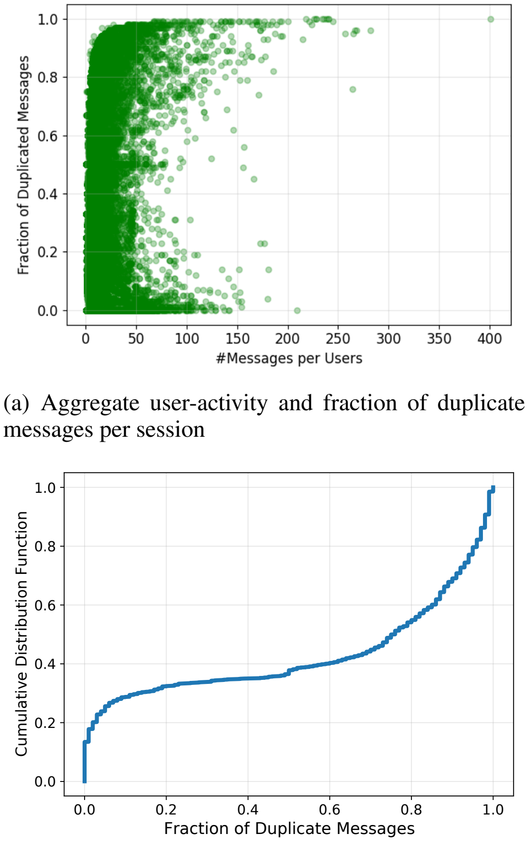
Data Collection

WhatsApp is Brazil's most popular messaging app, with millions of users sending messages daily in one-on-one chats, private groups, and public groups. WhatsApp messages are encrypted and conversations are confidential, except in public groups where anyone can join via an invitation link. Data collection on WhatsApp is, consequently, a difficult task. Previous works proposed overcoming this challenge by collecting public groups on WhatsApp, using social networks and online repositories to discover invitation links to public groups and join them, especially in political contexts (Kazemi et al. 2022; Melo et al. 2019a; Bursztyn and Birnbaum 2019). Then, all messages and group information are accessible. By implementing it on a large scale, it is possible to collect what is shared in thousands of groups.

In this work, we perform a similar large-scale data collection of messages shared in WhatsApp public groups in Brazil over almost three years, from March 2020 to December 2022. This period encompasses important political events, such as the COVID-19 pandemic and Brazilian elections (Reis et al. 2020). For our data collection, we focus on public WhatsApp groups related to Brazilian politics (i.e., political groups). To achieve this objective, our methodology was initiated by leveraging an initial set of keywords related to Brazilian politics, as originally formulated by (Resende et al. 2019b). Given the long-term duration of our data collection span, we subsequently expanded this keyword lexicon with newly emergent terms associated with political individuals and topics, which had arisen after the work conducted by Resende et al. (2019b). This enriched set of keywords is then employed to identify and retrieve invitation links to public WhatsApp groups disseminated across social media platforms such as Twitter and Facebook, as well as other online repositories listing public groups. Initially, we found 1,828 valid groups (i.e., unbroken URLs) from which we filtered 364 relevant groups to join. Importantly, this process was conducted iteratively, with periodic reiterations, as necessary to curate groups that remained contemporary and reflective of the evolving landscape of discourse and more recent political events, reaching a total of 1,645 public political WhatsApp groups collected. This iterative practice is essential in sustaining a prolonged and expansive data collection effort, given the inherently transient nature of communities within instant messaging platforms (Hoseini et al. 2020). In order to ensure the political nature of the groups included in our dataset, one of the authors meticulously undertook annotation. This process entailed a careful manual evaluation of each discovered group's title and description. Groups were joined only if they exhibited a clear association with the topic of politics.

Then, to extract data from groups, we use the approach proposed by (Garimella and Tyson 2018) and (Melo et al. 2019a) to find, process, and store the content shared in these groups. This methodology gets the WhatsApp database from the user's smartphone, directly accessing the messages received. Using the above methodology, we gathered data about all messages shared between March 2020 and December 2022. For each message, we extract (i) user ID, (ii) group ID, (iii) timestamp, (iv) a label on whether the message was forwarded, (v) text, (vi) media type (e.g., images, audio, and videos, sticker) and (vii) if available, the attached multimedia files (e.g., images, audio, and videos) downloaded through a *media_url* provided by WhatsApp.

Overall, we collected 19M messages shared in 1,645 WhatsApp groups from 189K users between March 2020 to December 2022. Our dataset is quite diverse and includes a lot of multimedia: 5.4M messages are text, 5.2M messages are video messages, 4.2M are image messages, 1.8M messages are stickers, 1.2M messages are links, 984K messages are audio, and we have 54K messages sharing documents. Note that on WhatsApp, all multimedia and files shared on the platform are assigned a unique identifier; hence, as-



(a) Aggregate user-activity and fraction of duplicate messages per session

(b) CDF of the fraction of duplicated messages in sessions with more than 60 messages per user.

Figure 1: Distribution of duplicate messages and user-activity in the 1-minute sessions.

sociating multimedia files sharing the same content is not straightforward. To overcome this challenge, we use the MD5 checksums of the files for audio files, videos, and documents and deduplicate the dataset based on the unique set of MD5 checksums. For images, we use Perceptual Hashing (Monga and Evans 2006), a technique that uses a fingerprinting algorithm to generate a hash for each image based on how the image looks. We treat two images to be the same if they have the same pHash.

Data Limitation. As with all studies focusing on messaging platforms like WhatsApp, we cannot assess how representative our dataset is, mainly because we do not have a vantage point to obtain a holistic or random sample of all WhatsApp public political groups from Brazil. Despite this, we believe that our data collection is extensive and for the purposes of our work, it allows us to characterize various types of attacks happening in political groups on WhatsApp.

Flooding Attack

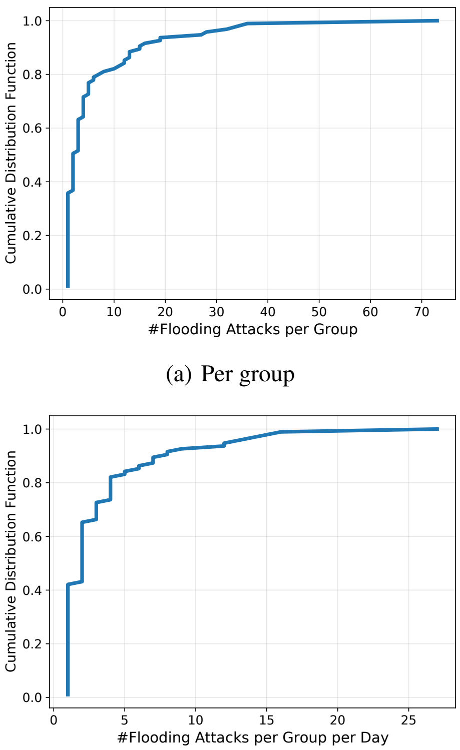
Flooding attacks are denial-of-service (DoS) attacks that aim to overwhelm a server and cause network disruption by creating network congestion (Yi et al. 2006). Flooding attacks are not only limited to computer networks, it is also a popular type of attack and can also affect other communication channels, like SMS (Hussain et al. 2021) and chat messages on online messaging platforms. On messaging platforms like WhatsApp, attackers can infiltrate a group and send a large volume of messages quickly, disrupting the group's regular operation and making it difficult for benign users to interact and chat in the WhatsApp group. Motivated by this, we aim to identify and characterize flooding attacks on WhatsApp. Furthermore, we aim to understand how these attacks are carried out and their impact on WhatsApp victim groups.

For our analysis of flooding attacks, we focus on the period between July 2022 and December 2022, which includes data from 1,267 groups and 12,167,529 messages. We focus on this period for many reasons; first, this is the most active period of our dataset, and second, our data collection included all activity related to sticker messages, which, as we will see later, are important for flooding attacks.

Identifying Flooding Attacks. The flooding attack occurs when an attacker sends a large volume of messages, usually containing identical content, to a WhatsApp group within a short period. To identify flooding attacks on WhatsApp groups, we devise the following methodology. First, we split the group's message-sharing activity into *sessions*, each comprising one minute of the group's activity. Then, for each session, we calculate: 1) the average number of messages per user; 2) the fraction of duplicated messages (i.e., sharing the same text, image, audio, video, sticker).

Figure 1(a) shows a scatter plot of these two metrics for each 1-minute session; we observe that the majority of the sessions have less than 60 messages per user, which is expected given that in most of the sessions, we expect to have benign conversations where many users share messages with relatively low frequency. Also, we observe that for a low average number of messages per user (less than 60), we have many sessions with a fraction of duplicated messages across the entire range of 0 and 1. On the other hand, when considering sessions with more than 60 messages per user, we observe that the fraction of duplicate messages is concentrated mainly on the limits. This is evident by looking at Figure 1(b), which shows the Cumulative Distribution Function (CDF) of the fraction of duplicated messages for all sessions with 60 messages per user or more. We observe that 60% of sessions with 60 messages per user have at least 60% of the messages shared within the session as duplicates (i.e., users sharing messages with identical content). Based on this session-based characterization, we assume that a group is under a flooding attack when there is an average number of messages of 60 messages or more and at least 60% of all the session messages are duplicates.

Using the above-mentioned methodology, we identify 893 flooding sessions in 95 WhatsApp groups (7,04% of all active groups from July 2022 to December 2022). Then, we aggregate the flooding sessions into flooding attacks. Since we create 1-minute sessions, flooding attacks may span multiple consecutive flooding sessions. Therefore, we combine all consecutive flooding sessions happening in the same group and treat them as part of the same flooding attack. Overall, we find 580 flooding attacks in 95 WhatsApp groups.



(a) Per group

(b) Per group per day

Figure 2: CDF of the number of flooding attacks per group: a)for the entire period of our dataset; b) per group per day. We focus on the groups that received at least one flooding attack during our dataset.

Characterizing Flooding Attacks. Having identified a set of flooding attacks, we aim to characterize these attacks, focusing on understanding how these attacks are executed in WhatsApp groups. We start our characterization by looking into the groups that are the recipients of the flooding attacks. Figure 2 shows the CDF of the number of flooding attacks received per group (Figure 2(a)), as well as the CDF of the number of flooding attacks per group per day (Figure 2(b)). Almost half of the groups receive only one flooding attack throughout our dataset. At the same time, we observe that many groups receive flooding attacks; e.g., 20% of the groups receive more than seven flooding attacks throughout our dataset (see Figure 2(a)). More worrying is the finding that groups receive multiple flooding attacks within the same day. We find that 57.89% of the groups receive more than one flooding attack within the same day (see Figure 2(b)), highlighting the prevalence and gravity of these attacks, especially in political groups. To better illustrate this phenomenon, we present a case study of a single WhatsApp group that received multiple flooding attacks throughout our dataset in Figure 3. This specific group received in total four flooding attacks, with the first three increasing in intensity, as observed by the increasing number of duplicate messages shared during the attacks. Overall, the finding that WhatsApp groups are the recipients of multiple flooding attacks, and sometimes within the same day, indicates that the group's administrators can not prevent or moderate these attacks effectively. This is likely due to the absence of effective moderation tools that can assist group administrators in tackling attackers that share many duplicate messages within a short period of time (Melo et al. 2019b).

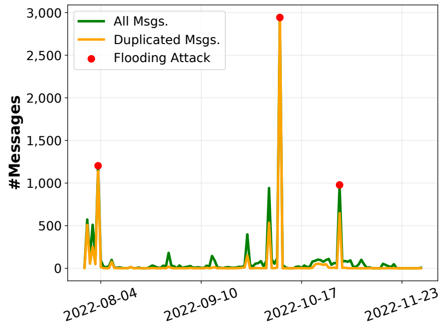


Figure 3: Group targeted by multiple flooding attacks.

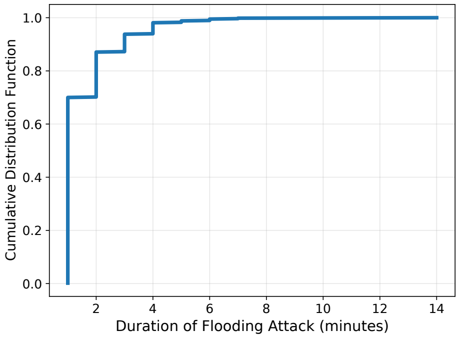


Figure 4: CDF of the duration of flooding attacks.

Next, we look into the duration of flooding attacks. Given that flooding attacks may consist of multiple 1-minute flooding sessions, we calculate each attack's duration by summing all the consecutive 1-minute flooding sessions. Figure 4 shows the CDF of the duration (in minutes) of the flooding attacks observed in our dataset. We observe that, in general, flooding attacks are short-lived; 70% of the flooding attacks have a duration of up to one minute, and 98.1% of the flooding attacks have a duration of up to four minutes.

Given that flooding attacks are short-lived, we then turn our attention to looking into the type of messages that are disseminated during the flooding attacks. We expect that attackers are sending media or types of messages that can send en-masse in a short period. To shed some light on the modus operandi of the attackers, for each flooding attack, we identify the types of messages that are disseminated during the flooding attack. Figure 5 shows the prevalence of the flooding attacks across the various message types. Most attacks are carried out using exclusively stickers (54.13%), text (22.06%), or a combination of both text and stickers (5.86%). Stickers are small images and can even be ani-

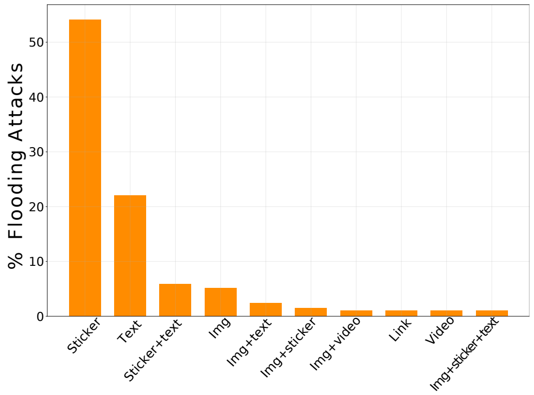


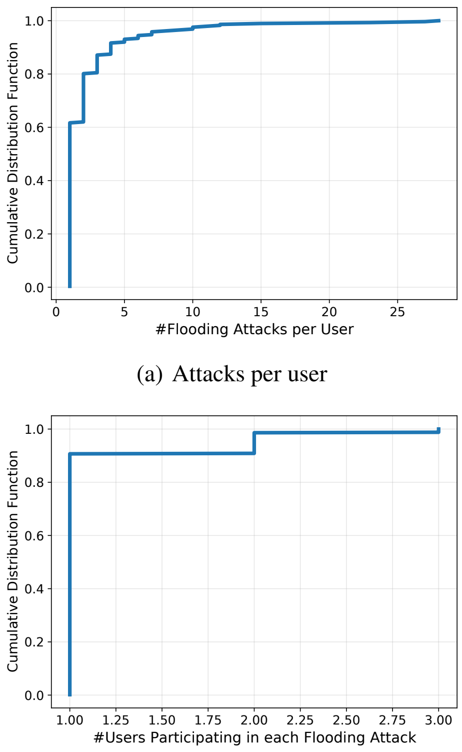
Figure 5: Percentage of flooding attacks for each different type of message in our dataset.

mated. They serve multiple purposes, like sharing emojis and memes that can be easily and quickly shared in a group. Given that stickers are also customizable, and group participants can create new stickers, attackers may create some offensive stickers and then disseminate them in the group to undertake a more severe attack.

Characterizing stickers used in flooding. Thus far, we have observed that stickers are one of the most popular message types for undertaking flooding attacks. To better understand flooding attacks, here, we aim to characterize the content of the stickers by undertaking a qualitative analysis using thematic coding analysis (Braun and Clarke 2006). To do this, we extracted a random sample of 100 stickers used during flooding attacks and constructed a codebook after two researchers went through the stickers, created initial codes, discussed them together, and refined them iteratively until no more changes were made, reaching as result ten codes:

- **Political Agenda**: Promotion of political views or ridicule/criticism of the opposite political side.
- **Meme:** Non-political memes and animated content.
- **Porn:** Explicit sexual content, nudity, and pornography.
- **Abstract:** Random stickers and miscellaneous topics.
- **Disgusting:** Repulsive and disgusting content, usually involving bodily waste.
- **Religious Attack:** Employing religious symbols and irony to mock or satirize religion.
- **Violence:** Content that promotes violence.
- **LGBTQ+ Attack:** Content attacking LGBTQ+ people.
- **Antisemitism:** Promoting nazism or attacking Jewish.
- **Racism:** Content promoting racist views.

Having constructed the codebook, two researchers independently coded another sample of 100 messages; we find a Cohen's Kappa coefficient of 0.936, indicating high agreement between them, hence the rest of the stickers are coded from a single annotator. Overall, we coded all 1,667 stickers used in flooding attacks. We find that almost 60% have a Political Agenda, which shows that the attacks also aim to provoke the opposite side and sometimes promote their ideologies. Memes (17.87%) are frequently employed to inundate the group. Abstract (8.62%) stickers are readily accessible, with some being standard stickers commonly found on many phones, indicating that users choose random stickers to inundate the group. More worrying is the fact that we find a substantial percentage of stickers containing harmful content like Porn (12.34%), Violence (1.12%), and Disgusting (2.79%), including repulsive content and offensive imagery. The attacker's goal extends beyond merely targeting the group; it also involves creating discomfort by disseminating offensive and repugnant content. Other instances of hate content include Religious Attacks (1.18%), LGBTQ+ Attacks (0.87%), Antisemitism (0.81%), and Racism (0.50%). The substantial degree of harmful stickers used in flooding attacks highlights the gravity and potential impact of such attacks on WhatsApp users.



(a) Attacks per user

(b) Users per attack

Figure 6: CDF with flooding attacks per user.

Characterizing text used in flooding. Text messages are also popular for flooding groups; 31.3% of flooding attacks used text messages. Here, we aim to characterize the text content shared during flooding attacks. To do this, we categorize 20% of all flooding text messages. Initially, two human evaluators examined 30% of these messages and established the five codes:

- **Overload Attack Message**: A long message with many characters designed to slow down the phone.
- **Political**: Messages strengthen their own political stance.
- **Meaningless**: Laughter or random characters with no specific meaning.
- **Accusation/Attack**: Messages to provoke or incite political reactions.

- **Word/Phrase**: Words or phrases lacking a particular discernible purpose.

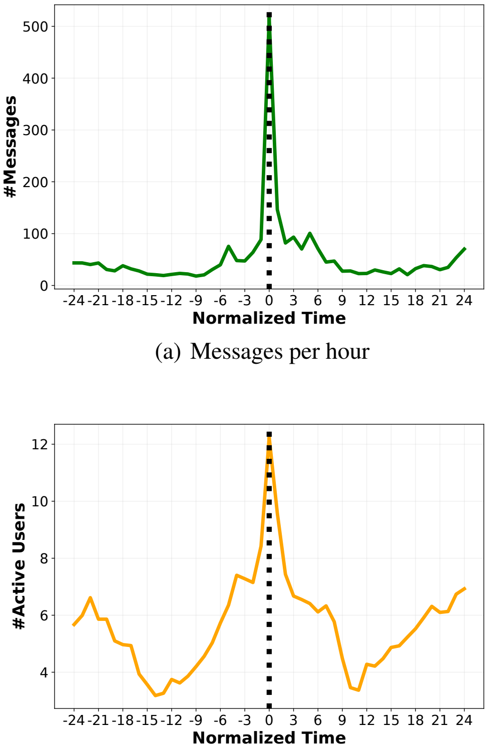
Subsequently, 30% of all text messages used in flooding attacks were labeled. Within this set of messages, 54.88% were identified as Overload Attack. This shows the malicious intent of the attacker, who not only floods the group but also attempts to slow down users' phones. Furthermore, an additional 25.61% Meaningless messages, characterized by random characters seemingly typed on the keyboard without any discernible meaning. Another 14.64% contained political Accusations/Attacks designed to provoke the opposing political group. Finally, 4.88% consisted of random Words/Phrases lacking a specific discernible meaning.

To conclude our characterization of the flooding attacks, we look into the users who participate in flooding attacks (i.e., attackers). To identify attackers, we extract the most active users in each flooding attack and we treat the user as an attacker if the number of messages they shared during the attack period is 20% or more of the entire session activity. Figure 6 shows the CDF of the number of flooding attacks per user, as well as how many users are participating in the same flooding attack. We find that 38.3% of the users that participated in flooding attacks participated in more than one attack throughout our dataset (see Figure 6(a)). Also, we find that most of the flooding attacks are executed by a single attacker (90%, see Figure 6(b)).

Impact of Flooding Attacks. Having characterized flooding attacks, here, we aim to analyze the impact of these attacks on WhatsApp groups. To achieve this, we compare the activity within the group before and after each flooding attack. Given that we already showed that flooding attacks are short-lived, we focus on 24 hours before and after the attack. Then, we calculate the hourly number of messages and active users and present the results in Figure 7. We focus here on cases where we only have one flooding attack within 48 hours (corresponds to 21.03% of all flooding attacks), as subsequent flooding attacks will affect the WhatsApp group activity. We observe a substantial increase in both metrics (number of messages and active users) during the flooding attack. This result is expected for the number of messages, given that the attack itself is based on the creation of a large number of messages. On the other hand, the increase in the number of active users is likely due to benign users enquiring about the attack. After the attack, we observe that the group is restored to its regular activity three hours following the flooding attack; on aggregate, we have a similar number of messages and active users before and after the flooding attacks. Overall, these results highlight that flooding attacks can potentially disrupt the groups' activity, however, their impact appears limited to only a few hours.

To conclude our analysis of the impact of the flooding attacks, we perform a small-scale qualitative analysis of text messages sent by benign users to shed some light on the impact on WhatsApp users. In this direction, we labeled 20% of all text messages sent by benign during the flooding attack, corresponding to 243 messages. Initially, 30% of messages were labeled to find these five codes:

- **Complaining/Cursing**: Users express dissatisfaction or use profanity to describe flooding attacks.



(a) Messages per hour

(b) Users per hour

Figure 7: Number of messages and number of active users before and after flooding attacks. We normalize the time and we focus on 24 hours before and after each attack (time 0 corresponds to the flooding attack).

- **Requesting Moderation**: Users request the administrator to take action and eliminate the attacks.
- **News**: Political news or new media forwarded.
- **Interaction**: Users engage with one another.
- **Miscellaneous**: Laughing, meaningless content or message that we could not identify.

Next, all other messages were labeled by two human evaluators, reaching a kappa coefficient agreement of 0.70 (Cohen 1960). Among all the messages, 31.28% specifically pertain to Complaining/Cursing about the ongoing attack, while 13.17% of messages were out of a lack of Moderation and requested the administrator's intervention to remove the attacker. For instance, some examples we observed during flooding attacks from benign users are: *"Where is the admin?"*, *"Hey admin ban the attacker"*, *"What is it?"*, *"my cell phone is crashing"*(translated messages from Portuguese). Furthermore, 25.93% of the messages involve users interacting with each other or responding, while 16.05% consist of News content shared during the attack. Lastly, 13.58% constitute Miscellaneous, which could not be clearly identified or categorized.

Takeaways. The key takeaways from our analysis of flooding attacks on WhatsApp groups are as follows:

- Flooding attacks are not a rare phenomenon on WhatsApp when considering Brazilian groups related to politics. We find that 7.04% of all monitored WhatsApp groups experienced at least one flooding attack throughout our dataset.
- Flooding attacks are short-lived (70% of the flooding attacks have a duration of up to one minute), and they are usually undertaken by a lone wolf (i.e., 90% of all flooding attacks consist of a single attacker).
- We find that WhatsApp groups are the recipients of multiple flooding attacks (even within the same day), which indicates that the moderators or group administrators cannot proactively prevent flooding attacks.
- Many flooding attacks are made using the large dissemination of stickers; 62.57% of all flooding attacks in our dataset use stickers. In addition, based on our manual annotations, we find flooding attacks that use offensive stickers including pornography, violence, and provocative messages.
- By analyzing the impact of flooding attacks, we find that the irregularities in the group activity due to the attacks last for only a few hours (usually three hours), which indicates that both the attack and its impact are short-lived.

Group Hijacking Attack

In the previous section, we investigated flooding attacks on WhatsApp groups and observed that WhatsApp groups tend to return to regular operation/activity after the flooding attacks. Here, we investigate a more severe attack, the Group Hijacking Attack, where an attacker aims to obtain complete control of the group and potentially make drastic or catastrophic changes. The attack is initiated when an attacker obtains administrator rights in the group either via unauthorized means (e.g., by compromising the account of an administrator or the group creator) or by enticing other administrators to promote the attacker to an administrator, or by identifying groups where the creator left the group. Having obtained administrator rights, an attacker can make important changes in the group, such as removing group participants, repurposing the group by changing group metadata, or even archiving/deleting/privatizing the group.

The group hijacking attack is analogous to attacks aiming to compromise accounts on social media platforms (Egele et al. 2013) to either repurpose it (Elmas, Overdorf, and Aberer 2023), steal personal information, or share misleading information. Here, we focus on understanding and characterizing this phenomenon through the lens of political groups from Brazil on WhatsApp. In political groups, hijacking attacks can be used to disrupt the discussions of supporters from the other party or attack them by sharing provoking content within their group. Overall, there is a pressing need to understand this phenomenon, as it has the potential to increase online political polarization in the WhatsApp ecosystem. Here, our analysis for hijacking attacks focuses on the entire collected dataset.

To identify potential group hijacking attacks, we focus on groups that had at least one change in the group's name throughout the period of our dataset. We find that 563 groups (34.28% of all groups) had at least one name change; not all name changes pertain to attacks. To identify potential attacks, we then manually annotate all the 563 groups and their respective name changes to identify whether the name change is suspicious (e.g., the group's name before and after the change substantially differ semantically) by two evaluators with a 0.7 kappa coefficient. We find a total of 33 groups with substantial semantic differences in the two group names, which corresponds to 5.86% of all groups with at least one name change in our dataset.

Categories	Original Name	Changed Name
Conflicting	*Bolsonaro's slogan*	*Lula 2022*
	Evangelic & Lula	*100% Bolsonaro*
	Brazil Patriot	*Antifa Action*
	Bolsonaro the myth	*Arthur King*
	Bolsonaro Reigns	*Left Reigns*
	Bolsonaro's Slogan	*Brazil is Lula*
	Lula vs Bolsonaro the Myth	*Lula big vs Bolsonaro*
	*Haddah is a sh***	*Antifascist*
Offensive	*Bolsonaro 2022*	*Gay Group*
	Alliance for Brazil	*Golden Shower*
	Captain gallows	*Devil Gallows*
	Just patriots	*Bozo the shit*
	22	*Prostitution Group*
	Bolsonaro 2022	*Bolsonaro's Brothel*
Explicit	*Entourage PT*	*157 by: lagxzada*
	Alliance for Brazil	*Hacked*
	Anything goes!	*Archived by Apocalypse*
	Itu antifascism	**bot.py**
Context Switch	*We with Bolsonaro*	*Grandma's recipe*
	Antifascist Alliance	*Banana Cake*
	Beloved Brazil	*Birthday Uncle Dudu*
	Strategic Right	*Cake Recipes*
	Brazil-CE	*Yoga group*
	Bolsonaro President	*play the Siri*
	United right	*Grandma's recipe*

Table 1: Groups with suspicious name changes (translated names from Portuguese).

Characterizing Groups with name changes. Based on our manual annotations, we categorize these 33 groups into four high-level categories (see Table 1[1]).

- **Conflicting:** These are groups that, after the group name change, a contradiction emerges in the political ideology when comparing the period before and after the change in the group's name. We find four cases of groups in our dataset with conflicting political leanings. For instance, a group named "Bolsonaro reigns" was renamed to "Left Reigns," indicating the substantial shift in the group's ideology, likely because of a group hijacking attack.
- **Offensive:** Groups where the group's name became more offensive or toxic after the name change. We find in total five such cases of groups in our dataset. For instance, we find a group that was changed from "Bolsonaro 2022" to "Gay Group" and a group changed from "Patriots and the Captain" to "bordel bozo" (Bolsonaro's brothel).
- **Explicit:** These are groups where the name change indicates that there was an attack on the group. We find four such cases; some examples include a group initially named "Alliance for Brazil" and then "Hacked" and a group renamed "Archived by apocalipse."

[1] Due to the space limitations, we include a subset of the groups classified as Context Switch.

- **Context Switch:** Refers to groups where the name change indicates a substantial shift in the group's context and topic of discussion. We found 15 groups with context switching. For instance, a group called "We with Bolsonaro" was renamed to "Grandma's recipe," a topic that has nothing to do with political discussions.

Identifying and Annotating Hijacking Attacks. Just because some WhatsApp groups have suspicious changes in their metadata does not necessarily mean they are the victims of hijacking attacks. Therefore, to detect hijacking attacks, it is paramount to analyze and understand what happened in the WhatsApp groups after the name changes and what messages were shared (if any) after the name change. To do this, we performed a manual annotation on the 33 WhatsApp groups that had suspicious name changes based on our previous annotations. In particular, for each group, we plot and evaluate the message activity (i.e., number of messages shared per day, before and after the name change) and manually read messages before and after the name change. Our analysis yields several interesting observations. First, we find that 15 out of the 33 groups that had suspicious name changes pertain to a group hijacking attack. For 15 of the groups, we observe that after the hijacking attack, the activity (in terms of messages shared) of the group becomes zero and we find messages that indicate that indeed an attack occurred. For instance, Fig. 8(g) shows an example of a group that received a hijacking attack, and shortly thereafter the group is destroyed and has no further activity. In addition, the attacker provokes the participants who complain and say that the group was overthrown, in addition to celebrating the success of the action (see yellow box in Fig. 8(g)). The rest of the groups (all part of the Context Switch category), we observe that are not the victims of group hijacking attacks, but rather, the group had a name change in an attempt to obfuscate the political nature of the group. Below, we present some examples of hijacking attacks and examples of context switching.

Characterizing Hijacking Attacks. Figure 8 shows 9 out of the 15 groups that were the recipients of hijacking attacks (others omitted due to space). For these suspicious names, we have read all messages on the days before and after the group name change. The messages sent on the attack day help us understand what happens in the group. Looking at Figure 8(c), 8(g), 8(e) and 8(h), we realize that the attacker sometimes interacts with the participants, threatening them *"make me admin or they will be hacked"*, sending messages that the group was taken down *"it's over for you, this group is now ours"*, or even celebrating successful group destruction, *"It was a pleasure to take down this group with you 2x on the same day"*. In other cases, the attacker does not interact, but messages from participants indicate that an attack is taking place. Figure 8(a) shows an example where some users ask the admin's support to moderate and remove the attacker because the group is under attack. In Figure 8(b), the last message was sent by a user informing the group was attacked and the participants needed to leave.

Upon careful observation, we have identified a pattern in hijack attacks: most groups had an attack shortly after we joined the group, typically within a maximum of 10-20 days from its initial creation, 11 of 15 hijacking groups. Moreover, 9 out of the 15 hijacked groups ceased their activities within a maximum of 5 days. In Figure 8(g), we see an example of a group that suffered an attack on the same day of its creation, and the group was destroyed. This pattern suggests that attackers are specifically targeting recently formed groups. The public group is shared on social networks, and many users join, even malicious users, who take advantage of the recently created group to ask for help with group moderation and gain admin privileges. Each group contains at least one admin responsible for moderating the group, but others can also be added to help organize the group. The attacker can gain the confidence of the moderator by assuming the role of an admin within the group, once a malicious user obtains admin privileges, the group becomes vulnerable to destruction. In some cases, the group has vulnerabilities and, by default, allows users to join with administrator privileges. Moreover, the attacker can intimidate group members and foster a hostile environment. In Figure 8(c), we can observe how the attackers orchestrate a flooding attack toward the admin's private chat. This can coerce the admin into either leaving the group or adding the attackers as additional admins. In Figure 8(g), the attackers also flooded the group and sent a message stating that the group's original creator had been removed and that the group now belonged to the attackers. In two other groups, the attacks did not occur immediately after their creation, but within days of joining the groups. This suggests a scenario where the group admin reshared the invite link to attract new users, and attackers discovered a new group to infiltrate.

Out of the 15 groups that were targeted in the attacks, one group managed to avoid complete destruction because the admin promptly took action by renaming the group and removing the attacker's presence. This case shows that quick administrator action can help mitigate the damage caused by an attack. In practice, we realized that the attacked groups did not have an active and engaged administrator, which facilitates the action of the attackers. In Figure 8(d) and 8(i), we show how users are complaining about admin action: *"where are the admins?"*. In Figure 8(a), the user goes so far as to say that the administrator is sleeping.

Finally, among the hijacked groups, we found that 4 also had a flooding attack on the day the group was renamed. This suggests a strategy employed by the attackers: hijacking and flooding attacks are used to gain control and disrupt the group. By flooding the group, the attackers create chaos and confusion, facilitating their hijacking actions. Flooding attacks are not just random messages; sometimes they are selected to evoke fear or intimidation among the group's users. In Figure 8(d), the attackers flood the group with pornographic messages and stickers.

Characterizing Context Switching. Looking for cases that had significant name changes but were not identified as an attack, we find context-switching groups. We noticed that 11 out of 15 cases occurred on specific days, coinciding with either the second round of the Brazilian presidential election or the final days of 2022 when Bolsonaro exited the Brazilian government. During that period, the Superior Electoral

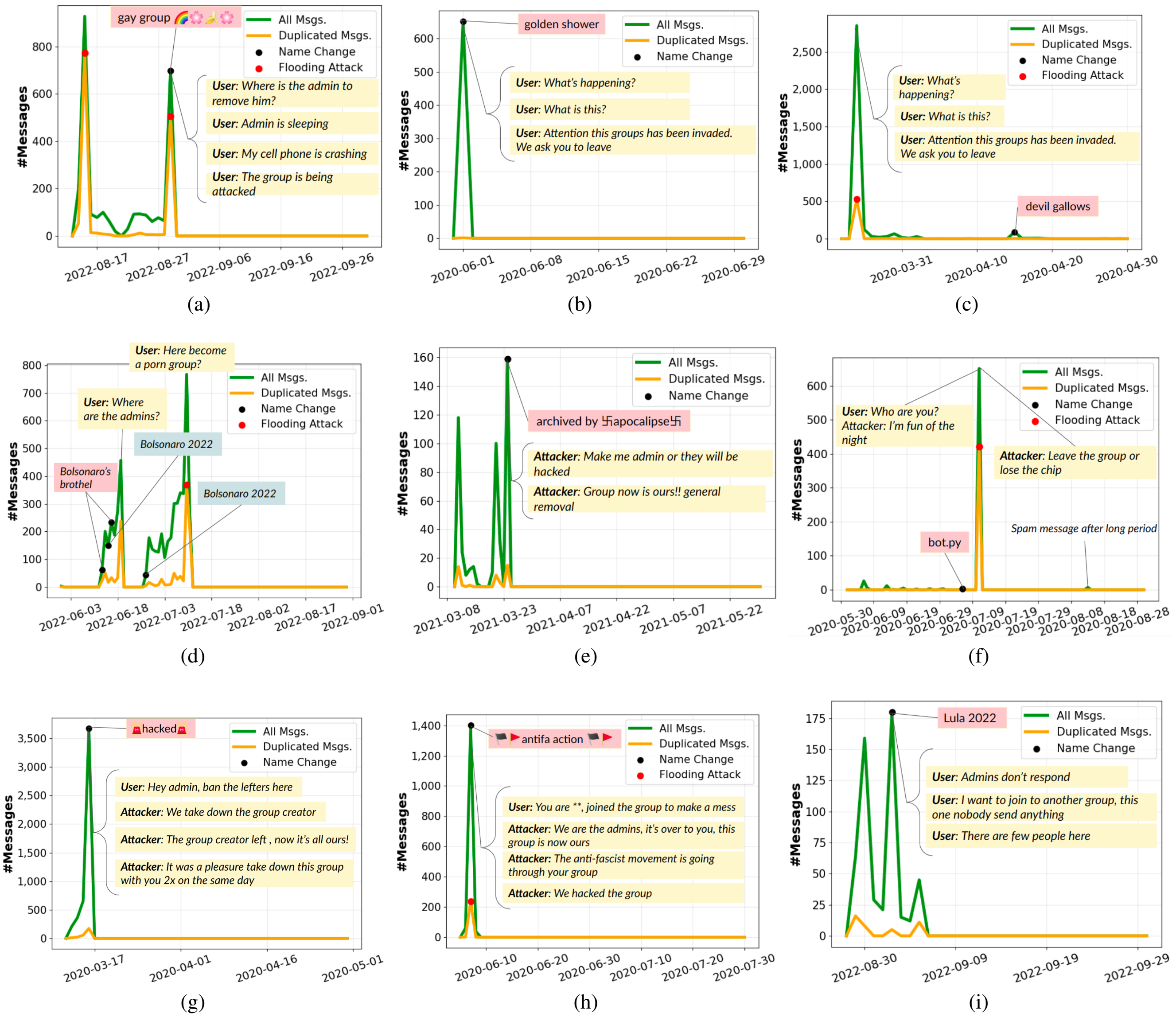


Figure 8: Examples of hijacking attacks. Yellow rectangles show some examples of messages from users, red and blue rectangles correspond to name changes, black circles to name changes, and red circles to flooding attacks. Figures (a), (b), (c), and (d) show examples of Offensive name changes, (e), (f), and (g) show examples of Explicit name changes, and (h) and (i) show examples of conflicting name changes.

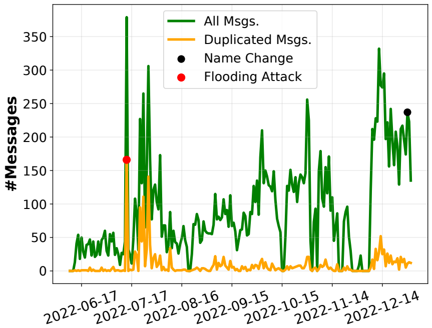


Figure 9: Switch Context Example.

Court (TSE) issued numerous court orders to shut down several WhatsApp and Telegram groups used to coordinate protests alleging election fraud (Mello and Galf 2022), which resulted in the invasion of the Brazilian Congress, Supreme Court, and presidential offices (Nicas and Spigariol 2023). These groups argue that there may be prosecution and therefore they need to camouflage to avoid censorship by the next government. Upon examining the activity within these groups, it becomes evident that despite the name change, they sustained a consistent level of message exchange and active user participation (see Figure 9). The change in group names serves as camouflage, offering participants a sense of confidence and security to express their thoughts and opinions freely. By adopting new names, these groups seek to preserve a level of anonymity and protect themselves from potential sanctions (Mello and Galf 2022).

Takeaways. The main takeaways from this section are:

- The hijacking attack targets recently formed groups, 11 out of the 15 hijacked groups being compromised shortly after we joined the group, either within a few days or 10-20 days of the group's creation.
- The hijacking attack goes further than flooding attacks, taking control of the group and disrupting the overall interaction. In 9 of 15 hijacking attacks, the group activity stops within a maximum of 5 days.
- 11 of 15 of the groups classified as context switch did this on presidential election day or the last day of Bolsonaro's presidency. These groups are from Bolsonaro supporters, and by adopting new names, aim to maintain a certain level of anonymity and shield themselves from possible sanctions (Mello and Galf 2022).
- Among the hijacked groups, 4 also had a flooding attack on the same day as the group renaming, indicating a connected strategy to disrupt and gain control over groups.

Related Work

WhatsApp has gained significant attention due to the spread of misinformation on its network (Resende et al. 2019b). In India, the spreading of rumors caused a series of violent lynchings around the country (Arun 2019), while in Brazil, there is evidence that fake news campaigns circulating on WhatsApp have interfered in the results of presidential elections (Resende et al. 2019b; Bursztyn and Birnbaum 2019; Machado et al. 2019). During the COVID-19 pandemic, WhatsApp was also pointed as an important vector for spreading fake news about health (Javed et al. 2020; Vijaykumar et al. 2021). Other issues regarding misinformation content disseminated through WhatsApp were reported in different locations such as in India (Kazemi et al. 2022), in Indonesia (Kwanda and Lin 2020), in Pakistan (Javed et al. 2020), U.K. (Vijaykumar et al. 2021), Ghana (Moreno, Garrison, and Bhat 2017), Nigeria (Cheeseman et al. 2020), Spain (Elías and Catalan-Matamoros 2020).

Furthermore, not only text messages are abused in this environment (Resende et al. 2019a; Caetano et al. 2019), but also multimedia content can contain misinformation, such as audios (Maros et al. 2020) and images (Reis et al. 2020; Garimella and Eckles 2020). These questions regarding misinformation on WhatsApp are also perceived by the users, who point to WhatsApp as one of the platforms they worry most about false information (Newman et al. 2021). Because of that, understanding how users make use of WhatsApp and what happens inside the large public chats under WhatsApp encrypted networks has become a relevant issue.

One of WhatsApp's key characteristics is the chat groups. Seufert et al. (2016) explored user interactions, revealing the vital role of group chats for communication within the app, including usage patterns and topics discussed. O'Hara et al. (2014) studied WhatsApp user habits. Through interviews, they uncovered how users prefer WhatsApp for maintaining close ties, not just in one-on-one interactions but also in group chats, which facilitate a sense of belonging, identity, and collective experience. Blabst and Diefenbach (2017) explored how people utilize WhatsApp for daily communication, highlighting user perceptions of specific features such as group chats, last seen, and read receipts in terms of communication quality and well-being. Rosenfeld et al. (2016) found that two-thirds of user conversations are direct one-to-one messages, while the remaining occur within groups.

Public group chats are the focus of many studies investigating WhatsApp, focusing on how it was abused for misinformation campaigns. Garimella and Tyson (2018) developed a methodology to find and collect WhatsApp data from public groups. They find 2K groups on Twitter and join a set of them to characterize WhatsApp users in India. Resende et al. (2019b) examined over 400 public political groups in Brazil, focusing on image-based misinformation dissemination. Melo et al. (2019a) develop a system that gathers and processes data from 1.1K public groups in Brazil and India, allowing users to explore the most shared content through a Web interface. In subsequent work, Melo et al. (2019b) investigates the impact of message forwarding limits on the spread of messages in WhatsApp public groups. (Bursztyn and Birnbaum 2019; Machado et al. 2019; Resende et al. 2019a; Caetano et al. 2019) employed online repositories of public WhatsApp groups to engage in political discussions and analyze misinformation circulating within these groups.

Some studies, like Abu-Salma et al. (2017), examine security and privacy features in Telegram. They discover that despite users feeling secure, the app provides limited security benefits, with many using the less secure default chat mode. Espinoza et al. (2017) provided a privacy analysis of E2EE features in LINE, identifying several vulnerabilities and challenges in app design. (Rösler, Mainka, and Schwenk 2018) analyze encryption challenges in group communication on WhatsApp, Signal, and Threema, proposing a group chat security model for instant messaging platforms tailored to group dynamics. (Schrittwieser et al. 2012) investigated earlier versions of WhatsApp, discovering significant security vulnerabilities that enabled attackers to hijack accounts, spoof sender IDs, or enumerate subscribers.

The opacity of WhatsApp makes it difficult to understand and monitor its internal dynamics. Moreover, there is insufficient empirical research on how political groups are targeted by activists and harmful users within this environment. Public groups tend to be highly connected, forming hierarchies of information, and reinforcing echo chambers. However, invitation links are easily found and shared on social media, which allows all kinds of users to gain access. This lack of knowledge conceals interactions within the platform and inhibits the development of effective strategies to moderate and reduce harm. This information gap highlights the need for further research to create a safer environment.

Ethics & Broader Perspective

In this study, we collected data from public political groups on WhatsApp. We acknowledge that these groups may contain personal opinions and sensitive information of the participants. Therefore, we took measures to protect the privacy and anonymity of the group members. All sensitive information such as usernames and phone numbers were not stored in our dataset (we only store hashes). We believe that our work's benefits outweigh the potential harms that may arise. Our work aims to shed light on attacks that may harm WhatsApp groups, which is important as it helps raise user awareness about these attacks, as well as encourages platforms to enhance their governance and moderation tools in an attempt to either prevent or mitigate such attacks.

Discussion & Conclusion

In this work, we explored two kinds of attacks that can disrupt the activity of WhatsApp groups, particularly flooding attacks that aim to disseminate many messages within a short period and hijacking attacks that aim to take control of the group and drastically change its purpose. We collected a large-scale dataset of 1.6K WhatsApp groups related to Brazilian politics between March 2020 and December 2022, including 19 million messages. Then, we devise a methodology to identify and characterize flooding attacks and investigate hijacking attacks by focusing on WhatsApp groups with suspicious name changes. Among other things, our analysis shows that flooding attacks are not rare when considering political groups in Brazil. It is likely a way for people from one party to attack people from another party, aiming to disrupt their conversations. We find that flooding attacks are usually short-lived, and most flooding attacks are made by one attacker. Also, we find that WhatsApp groups are the recipients of multiple flooding attacks, even within the same day, which likely highlights the lack of effective tools that assist the group moderators and administrators who aim to maintain the group's harmony. In regard to hijacking attacks, we find a handful of such attacks in our dataset, and we find that in most cases, the attacker's goal is to close or remove the group. Finally, we find that WhatsApp groups can be the recipients of both flooding and hijacking; the attackers first flood the group and then hijack the group entirely. Overall, our study is a significant leap towards demystifying the dark side of WhatsApp groups, particularly hostile intergroup interactions across the political spectrum. Our work has important implications for many stakeholders, including users and platforms like WhatsApp and Telegram. Below, we discuss how our study benefits these parties.

Raising Awareness. Our work highlights the prevalence and gravity of these attacks, identifying that they are not rare in Brazilian political WhatsApp groups. Additionally, our qualitative analysis highlights that a significant portion of these attacks contains harmful content, such as hateful or offensive stickers, as well as overly long messages to disrupt WhatsApp's regular operation on users' phones. Taken together, these findings show that these attacks are an important problem, and our work has the potential to raise awareness about these attacks among both WhatsApp users and operators of messaging platforms. For WhatsApp users, raising awareness of these attacks is important as it allows them to be more prepared when the attack is underway and try to protect themselves. For messaging platforms, our work and findings can be used to raise awareness about how these attacks are performed on their platforms, which is vital for designing effective moderation tools.

Need for timely moderation. Our work highlights the necessity for moderation tools to quickly and effectively moderate such attacks. Messaging platform operators can utilize our framework and data analysis pipeline to develop tools for detecting and preventing such attacks from impacting other WhatsApp users. For instance, platforms like WhatsApp can introduce throttling mechanisms to prevent users from sharing the same message in a short period or impose message length restrictions to prevent attackers from sending long messages that could slow down other users' phones. Also, our work emphasizes the need to have tools to detect harmful content (e.g., disgusting or hateful content), which we have observed to constitute an essential aspect of these attacks. For instance, platforms may introduce classifiers to detect and prevent the addition of hateful stickers by attackers. Overall, research is needed on deploying these moderation tools in end-to-end encrypted environments like those found in modern messaging platforms such as WhatsApp.

Acknowledgements

This work was supported by grants from CAPES, CNPq, FAPEMIG, and FAPESP.

References

Abu-Salma, R.; Krol, K.; Parkin, S.; Koh, V.; Kwan, K.; Mahboob, J.; Traboulsi, Z.; and Sasse, M. A. 2017. The Security Blanket of the Chat World: An Analytic Evaluation and a User Study of Telegram. In *Proc. the EuroUSEC*.

Arun, C. 2019. On WhatsApp, Rumours, and Lynchings. *Economic & Political Weekly*, 54(6): 30–35.

Bianchi, T. 2022. WhatsApp in Brazil - Statistics & Facts. https://bit.ly/3J2kbcG. Accessed: 2023-05-15.

Blabst, N.; and Diefenbach, S. 2017. Whatsapp and Wellbeing: A Study on Whatsapp usage, communication quality and stress. In *HCI*.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Bursztyn, V. S.; and Birnbaum, L. 2019. Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups. In *ASONAM*, 484–488.

Caetano, J. A.; Magno, G.; Gonçalves, M.; Almeida, J.; Marques-Neto, H. T.; and Almeida, V. 2019. Characterizing Attention Cascades in WhatsApp Groups. In *WebSci*, 27–36.

Cheeseman, N.; Fisher, J.; Hassan, I.; and Hitchen, J. 2020. Social Media Disruption: Nigeria's WhatsApp Politics. *Journal of Democracy*, 31(3): 145–159.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Egele, M.; Stringhini, G.; Kruegel, C.; and Vigna, G. 2013. Compa: Detecting compromised accounts on social networks. In *NDSS*.

Elmas, T.; Overdorf, R.; and Aberer, K. 2023. Misleading repurposing on twitter. In *Proc. the ICWSM*, volume 17, 209–220.

Elías, C.; and Catalan-Matamoros, D. 2020. Coronavirus in Spain: Fear of 'Official' Fake News Boosts WhatsApp and Alternative Sources. *Media and Communication*, 8(2): 462.

Espinoza, A. M.; Tolley, W. J.; Crandall, J. R.; Crete-Nishihata, M.; and Hilts, A. 2017. Alice and bob, who the FOCI are they?: Analysis of end-to-end encryption in the LINE messaging application. In *7th USENIX FOCI*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles. Accessed: 2024-04-04.

Garimella, K.; and Eckles, D. 2020. Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review*.

Garimella, K.; and Tyson, G. 2018. WhatApp Doc? A First Look at WhatsApp Public Group Data. In *ICWSM*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hoseini, M.; Melo, P.; Junior, M.; Benevenuto, F.; Chandrasekaran, B.; Feldmann, A.; and Zannettou, S. 2020. Demystifying the Messaging Platforms' Ecosystem Through the Lens of Twitter. In *IMC*.

Hussain, B.; Du, Q.; Sun, B.; and Han, Z. 2021. Deep Learning-Based DDoS-Attack Detection for Cyber–Physical System Over 5G Network. *IEEE Transactions on Industrial Informatics*.

Javed, R. T.; Shuja, M. E.; Usama, M.; Qadir, J.; Iqbal, W.; Tyson, G.; Castro, I.; and Garimella, K. 2020. A First Look at COVID-19 Messages on WhatsApp in Pakistan. In *ASONAM*, 118–125.

Kazemi, A.; Garimella, K.; Shahi, G. K.; Gaffney, D.; and Hale, S. A. 2022. Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp. *(HKS) Misinformation Review*, 3(1).

Kwanda, F. A.; and Lin, T. T. C. 2020. Fake news practices in Indonesian newsrooms during and after the Palu earthquake: a hierarchy-of-influences approach. *iCS*, 23(6): 849–866.

Machado, C.; Kira, B.; Narayanan, V.; Kollanyi, B.; and Howard, P. 2019. A Study of Misinformation in WhatsApp Groups with a Focus on the Brazilian Presidential Elections. In *WWW*, 1013–1019.

Maros, A.; Almeida, J.; Benevenuto, F.; and Vasconcelos, M. 2020. Analyzing the Use of Audio Messages in WhatsApp Groups. In *WWW*, 3005–3011.

Mello, P. C.; and Galf, R. 2022. TSE dá ordens em série para derrubar grupos golpistas que se multiplicam nas plataformas.

Melo, P.; Messias, J.; Resende, G.; Garimella, K.; Almeida, J.; and Benevenuto, F. 2019a. WhatsApp Monitor: A Fact-Checking System for WhatsApp. In *ICWSM*, volume 13, 676–677.

Melo, P.; Vieira, C. C.; Garimella, K.; de Melo, P. O. V.; and Benevenuto, F. 2019b. Can WhatsApp Counter Misinformation by Limiting Message Forwarding? In *CNA*, 372–384.

Monga, V.; and Evans, B. L. 2006. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *Transactions on Image Processing*, 15(11): 3452–3465.

Moreno, A.; Garrison, P.; and Bhat, K. 2017. Whatsapp for Monitoring and Response During Critical Events: Aggie in the Ghana 2016 Election. In *ISCRAM*, 645–655.

Newman, N.; Fletcher, R.; Kalogeropoulos, A.; and Nielsen, R. K. 2021. *Reuters Institute Digital News Report 2021*. Reuters Institute for the Study of Journalism.

Nicas, J.; and Spigariol, A. 2023. Bolsonaro Supporters Lay Siege to Brazil's Capital. https://nyti.ms/49m15sB. Accessed: 2023-05-15.

O'Hara, K. P.; Massimi, M.; Harper, R.; Rubens, S.; and Morris, J. 2014. Everyday Dwelling with WhatsApp. In *CSCW*, 1131–1143.

Reis, J. C. S.; Melo, P.; Garimella, K.; Almeida, J. M.; Eckles, D.; and Benevenuto, F. 2020. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and India Elections. *ICWSM*, 14(1): 903–908.

Resende, G.; Melo, P.; C. S. Reis, J.; Vasconcelos, M.; Almeida, J. M.; and Benevenuto, F. 2019a. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In *WebSci*, 225–234.

Resende, G.; Melo, P.; Sousa, H.; Messias, J.; Vasconcelos, M.; Almeida, J.; and Benevenuto, F. 2019b. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *WWW*, 818–828.

Rosenfeld, A.; Sina, S.; Sarne, D.; Avidov, O.; and Kraus, S. 2016. WhatsApp usage patterns and prediction models. In *ICWSM*.

Rösler, P.; Mainka, C.; and Schwenk, J. 2018. More is Less: On the End-to-End Security of Group Chats in Signal, WhatsApp, and Threema. In *EuroS&P*, 415–429.

Samuels, E. 2020. How misinformation on WhatsApp led to a mob killing in India. https://wapo.st/3U84dUQ. Accessed: 2023-05-15.

Schrittwieser, S.; Frühwirt, P.; Kieseberg, P.; Leithner, M.; Mulazzani, M.; Huber, M.; and Weippl, E. 2012. Guess Who's Texting You? Evaluating the Security of Smartphone Messaging. In *NDSS*, 415–429. Internet Society.

Seufert, M.; Hoßfeld, T.; Schwind, A.; Burger, V.; and Tran-Gia, P. 2016. Group-based Communication in WhatsApp. In *IFIP Networking Conf. and Workshops*, NETWORKING.

Vijaykumar, S.; Jin, Y.; Rogerson, D.; Lu, X.; Sharma, S.; Maughan, A.; Fadel, B.; de Oliveira Costa, M. S.; Pagliari, C.; and Morris, D. 2021. How shades of truth and age affect responses to COVID-19 (Mis)information: randomized survey experiment among WhatsApp users in UK and Brazil. *Humanities and Social Sciences Communications*, 8(1).

Yi, P.; fei Hou, Y.; Zhong, Y.; Zhang, S.; and Dai, Z. 2006. Flooding attack and defence in ad hoc networks. *JSEE*, 17(2): 410–416.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes!
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the claims in the abstract and introduction accurately reflect the paper's contribution and scope.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we state in the Introduction why our mixed-methods approach is suitable and appropriate for understanding and characterizing these two attacks on WhatsApp.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, because as described in Section , we do not have access to representative samples from WhatsApp so we can not make any claims about the data used and its representativeness.
   (e) Did you describe the limitations of your work? Yes, the limitations of our work are mainly related to data collection (see Section).
   (f) Did you discuss any potential negative societal impacts of your work? No, because we do not foresee any potential negative societal impact from this work.
   (g) Did you discuss any potential misuse of your work? No, because we do not foresee any potential misuse of this work. Our research aims to raise awareness and inform the public and messaging platform operators about the existence and modus operandi of these attacks on WhatsApp.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we describe measures we take to prevent or mitigate potential negative outcomes of our research in Section , which includes a discussion about how we dealt with sensitive information.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read the ethics review guidelines and ensured that our paper conforms to them.

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA