Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Don't Break the Chain: Measuring Message Forwarding on WhatsApp

Philipe de Freitas Melo[1, 2], **Mohamad Hoseini**[3], **Savvas Zannettou**[4], **Fabricio Benevenuto**[1]
[1] Universidade Federal de Minas Gerais, Brazil
[2] Universidade Federal de Viçosa, Brazil
[3] Max Planck Institute for Informatics, Germany
[4] TU Delft, Netherlands
philipe.freitas@ufv.br, mhoseini@mpi-inf.mpg.de, s.zannettou@tudelft.nl, fabricio@dcc.ufmg.br

Abstract

WhatsApp has evolved into a popular communication tool, facilitating the exchange of billions of multimedia messages globally. With its large public groups and forwarding features, the platform has enabled messages to go viral, rapidly disseminating across the WhatsApp network. This brought WhatsApp to a central position of spreading misinformation campaigns, prompting the company to implement measures to counter bulk message dissemination, such as limiting forwards. Despite these measures, there remains a gap in our understanding of how forwarded messages function within this ecosystem and the effectiveness of the restrictions in containing the spread of viral content. In this study, we analyze 10 million messages from 1,101 public WhatsApp groups dedicated to political discussion in Brazil, focusing on forwarded content. We investigate the structure of message forwarding, assess the reach of *Forwarded Many Times* labeling mechanism, and evaluate the platform's ability to detect duplicated media. Our findings reveal that forwarded messages constitute a substantial portion of the content shared in public WhatsApp groups. Moreover, we discover that measures implemented by WhatsApp to restrict the dissemination can be easily circumvented, allowing users to intentionally bypass the architecture of the system and share media beyond the imposed limits. Notably, we identify that 59% of duplicated content flagged as FMT by WhatsApp does not receive the corresponding flag and find evidences of misinformation circulating virally within them. This provides valuable insights into the dynamics of forwarded messages on WhatsApp and highlights the need for more effective strategies to combat the spread of viral content within the platform.

Introduction

Digital platforms have changed the way people interact globally, heavily impacting how people interact and communicate. In this context, messaging applications, particularly WhatsApp with over 2 billion users globally (WhatsApp 2020b), have become an essential part of daily life for smartphone users due to its simplicity, security, and popularity. WhatsApp is usually cheaper and more attractive than Short Message Services (SMS) as it requests only a phone number and Internet access while offering various communication features, including multimedia messaging, voice/video calls, large group chats and ease to forward content (Resende et al. 2019). Its popularity remarkably observed in countries like India and Brazil (Melo et al. 2019a), where almost everyone with a smartphone uses it.

On top of that, all messages are encrypted and hence anonymous when forwarded beyond the groups and chat boundaries. Therefore, tracing the origin of a message that has spread across the entire network is challenging. This leads us to an issue regarding the significant concern of misinformation dissemination through forwarded messages on WhatsApp (Jakesch et al. 2021). The enclosed nature of the platform combined with the ease of transferring multimedia and sharing information with public groups makes WhatsApp unique, in which an anonymous encrypted message can be viral, reaching multiple users in a short period of time (Melo et al. 2019b). Forwarding features of WhatsApp play an important role in this scenario, as it allows virality under encryption, creating a suitable environment for misinformation. Once provoked to deal with misinformation, WhatsApp has restricted the number of messages users can share. According to WhatsApp, such limitations were able to reduce the number of forwards globally by 25% (WhatsApp 2020a). Even though this helped to slow down spreading, it does not constrain virality within WhatsApp (Melo et al. 2019b). In fact, many users forward helpful information, as well as entertainment content and things they find meaningful; however, there is a significant increase in the amount of forwarding, which overwhelms users and contributes to the dissemination of misinformation.

More recently, in order to prevent the massive virality of messages, WhatsApp also introduced users to the concept of messages that have been *"Forwarded Many Times"* . These messages are labeled to indicate they did not originate from close contact. More precisely, a message receives this label when it was forwarded through a chain of sharing reaching far away from its original sender. WhatsApp states this kind of message is less personal compared to typical messages, and the system limits them to only be forwarded up to one chat at a time (WhatsApp 2020a). Although these changes have as objective to counter the wide dissemination of messages within the WhatsApp network, there are no efforts to investigate the influence of this functionality on message exchange on WhatsApp or in any other instant messaging platform under encryption. We do not know whether WhatsApp

is effective in identifying viral messages using this restriction strategy, or if labeling messages as viral is enough to reduce their reach. Since flagging the messages is the only indication users have to distinguish popular messages from those of more personal nature, the comprehension of this tool is essential to understand the information propagation in such a particular and closed environment.

In this study, we want to investigate the impact of this forwarding on the content that circulates on WhatsApp networks. We are particularly interested in evaluating the importance of the process of flagging viral messages made by WhatsApp. Through our study, we aim to answer the following research questions:

- **RQ1:** How does the forwarding work on WhatsApp? What kind of messages are forwarded on the platform and how WhatsApp flags them?
- **RQ2:** What is the effectiveness of this approach of flagging and restrict forwarded content in detecting and block the spreading of viral content?

To achieve that, we proposed an extended characterization of messages frequently forwarded on and how users make use of the forward tool within WhatsApp using a large set of almost 10M real-world message data from over 1100 public WhatsApp groups related to political discussions in Brazil. We analyzed the volume of forwarded content by different media types and also compute similarity between multimedia messages to count their popularity. After that, we evaluate the spreading of those media within the public chat groups monitored compared to the capacity of WhatsApp strategy to identify and flag duplicated instances of popular content circulating within its network.

Main Findings: Our results shed light on the viral and public nature that WhatsApp messages can assume. We find that the forward feature play an important role in the dynamic of WhatsApp network, and most of the content posted in many of the public groups monitored actually originated from an external source functioning as large repositories for dumping messages. Furthermore, we discovered that, although WhatsApp is able to detect instances of viral multimedia messages shared and flag them as "Forwarded Many Times" (FTM), a large part of them lack proper labeling and circulate throughout the platform without any warning, including misinformation messages. This suggests that WhatsApp's actions to avoid content to further spread cannot prevent users from widely sharing multimedia content. The ease with which we discover a user bypass the forwarding limitations evidences how this encrypted platform can be abused, enabling misinformation to go viral without any alert and without tracking who are the original authors. Finally, we show that, by using techniques of merging content by similarity, it would be possible to track and flag more viral content and that it could be useful to counter viralization of misinformation within this platform.

Viral Forwarding Under Encryption

During the 2018 Brazilian election, those engaged in WhatsApp groups observed the platform's remarkable capacity for viral spread and its societal impact (Resende et al. 2019). The misuse of WhatsApp, particularly within public groups, has led to real-world consequences in India, such as mob lynchings. (Mukherjee 2020). WhatsApp's forwarding feature emerged as an effective tool for disseminating content (Melo et al. 2019b), guaranteeing the survival of content dropped on other platforms (Reis et al. 2020a) and even providing it in an anonymous and untraceable way to its creators due to the end-to-end encryption (Kazemi et al. 2022).

WhatsApp recognized its role as a platform for mass communication. In response to the significant viral sharing observed, it promptly implemented self-regulatory measures to mitigate this phenomenon, such as limiting simultaneous forwarding to five recipients in 2019[1] and identifying widely shared content (WhatsApp 2020a), imposing more restricted rules to it during COVID-19 pandemic in 2020 (WhatsApp 2020a). While WhatsApp claims to keep the app more personal and private (WhatsApp 2020a), the flow of messages can easily get viral within its well-connected network (Melo et al. 2019b), which may confuse the user about the authenticity behind a message. The dual role of WhatsApp raises questions about users' intentions regarding message forwarding. (Malka, Ariel, and Avidar 2015).

The ambiguity of WhatsApp's use is not unique to this platform. Experts have argued that social network platforms should also be treated as media companies (Napoli and Caplan 2017). They defend that if social networks would not be considered purely as technology ones, but also as media companies, they could be held more accountable for the content they display to users. (Malka, Ariel, and Avidar 2015) studied how Israeli citizens use *WhatsApp* during wartime, where they observed that the platform works as a multifunctional channel of communication to get the news as well as interpersonal chats to be in contact with those on the battlefield. During the Ukraine war, a similar duality between mass communication and private chat is highlighted on Telegram, another instant messaging application. Both Russians and Ukrainians made use of the app to get updates and news about the war even public figures, while many experts raised concerns about privacy within the platform (Allyn 2022). Despite the differences in intention of messages between public and private, the information presented to users remains largely the same on WhatsApp. The interface provides minimal clues, such as the "Forward" and "Forwarded Many Times" label, to help users discern the content's origin. Therefore, understanding how people utilize Forwarding and how WhatsApp manages this process is crucial for comprehending challenges around the platform.

The decision to forward messages on WhatsApp reflects the receiver's choice to share received content with others (Bakare, Abdurrahaman, and Owusu 2022). This process, akin to word-of-mouth in digital spaces, involves stages such as receiving a message, deciding to open or read it, understanding the content, and choosing whether to forward it (Phelps et al. 2004). Various factors can influence this decision, including sender appeal, emotional response, entertainment value, usefulness, ease of forwarding, and social identity (Bakare, Abdurrahaman, and Owusu 2022; Yu and Kamarulzama 2016; Karimiyazdi and Mokhber 2015).

[1] https://blog.whatsapp.com/more-changes-to-forwarding

Figure 1: A forwarding chain that shows when a message receives the label "Forwarded Many Times" .

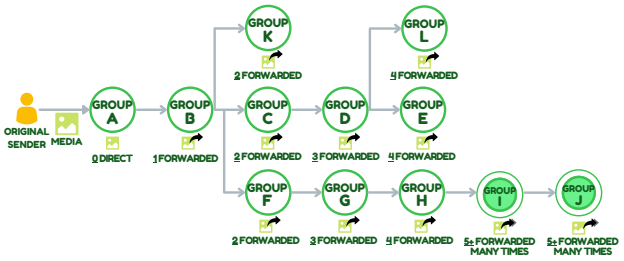


Figure 2: Flowchart with a scenario of a forked thread of a spreading message chain on WhatsApp.

Furthermore, when forwarding a viral message, users implicitly endorse its content, thereby enhancing the message credibility (Harvey, Stewart, and Ewing 2011). According to (Berger 2016), contagious content evokes strong emotions, both positive and negative, which can prompt users to share a message.

WhatsApp Forwarding Mechanism

When a user sends a message in any chat on WhatsApp, it can be easily shared by others using the Forward button. Forwarded messages are then marked with a "Forwarded" label and a symbol, indicating that the message originated from someone else. However, there are limits and exceptions in the forwarding process. Users can forward a message to up to five chats simultaneously, with a maximum of one group chat included[2]. Additionally, WhatsApp introduced the concept of messages that have been marked as *"Forwarded Many Times"* . Messages labeled as such present a different visualization, showing they did not originate from a close contact. These messages have been forwarded through a chain of five or more forwards away from their original sender (Fig. 1). When a message is forwarded many times, the system understands that this could be a potential viral or harmful content, and it restricts the further forwarding of such messages to just one chat at a time (WhatsApp 2020a).

For a message to potentially go viral on WhatsApp, it needs to traverse a path of at least five hops from the author before reaching the final recipient. However, identifying the exact number of hops in this process is not always straightforward. There are scenarios in which users can forward a message to multiple chats in parallel (up to five), with each parallel path receiving its own counter. This means forwards made in these parallel paths may not trigger the "Forwarded Many Times" labeling, as shown in Fig. 2.

However, the counter for "Forwarded Many Times" is

[2]https://faq.whatsapp.com/887468535575482/

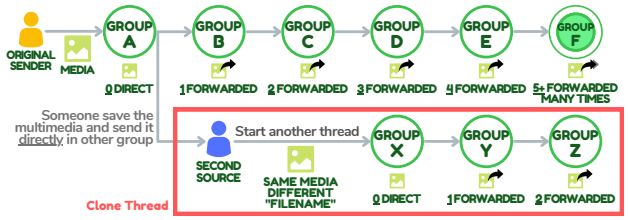


Figure 3: Flowchart with a scenario of a spreading message with a second user breaking the chain and starting a new cloned thread with the same media file.

created only for chains of messages that are repeatedly forwarded without interruption. If a user downloads a message and sends it directly to someone else without using the forward button, it resets the counter kept by WhatsApp, interrupting the viral tracking and creating a new cloned thread for that message, as shown in Figure 3. Since downloading a multimedia message and sending it directly from the phone's gallery is usually effortless, this method can effectively bypass WhatsApp restrictions and viral labeling. Moreover, due to WhatsApp encrypted nature, new messages derived from the clone will be considered entirely different, as WhatsApp cannot detect that the new message has the same content. Then, the network of shared messages on WhatsApp may build a complex flow, in which even though there are multiple appearances of the same piece of media, only a single one could be flagged as FMT.

WhatsApp introduced symbols and labels to help users distinguish between personal and widely shared messages. The FMT label is the main feature used by WhatsApp to identify potentially viral content. However, despite the straightforward flagging system, certain instances may require a deeper understanding during actual app usage. Some nuances can be obscure to users, especially concerning the FMT label, which may be misleading. In the following sections, we delve into a comprehensive data collection of messages sent on WhatsApp to gain a better understanding of how content operates within this ecosystem.

Data Collection

Despite the billions of messages exchanged daily through WhatsApp, accessing this content is challenging due to the private and encrypted nature of the platform. Additionally, companies do not share large-scale representative data. However, numerous public WhatsApp groups are shared on the Web using an invitation URL. Hence, an effective way to investigate WhatsApp data is by joining these public groups and collecting their messages to develop a robust dataset (Garimella and Tyson 2018; Melo et al. 2019a).

In this work, we deploy this methodology to collect large-scale data from public WhatsApp groups dedicated to political discussions in Brazil, particularly during presidential elections between June and October 2022. We initiated our data collection by obtaining a set of invitation URLs to public WhatsApp groups from various online sources, including search engines and social networks such as Twitter and

Facebook. These were identified using a curated list of keywords related to the Brazilian political context, which was initially developed by (Resende et al. 2019). To ensure the relevance of our dataset to the 2022 political landscape, we expanded this keyword list to include new terms related to the 2022 period. These additional terms encompassed the names of new candidates, political parties, and emerging topics within both the right-wing and left-wing political spectrums. Our data collection process is an ongoing effort to capture the dynamics of political discourse over time. As such, we periodically repeated the entire procedure to continually identify emerging WhatsApp groups that engage with recent events. This step is essential given the ephemeral ecosystem of groups on instant messaging platforms (Hoseini et al. 2020).

To guarantee that all the groups included in our dataset were genuinely political in nature, one of the authors manually annotated each discovered group. This labeling process involved inspecting the group's title and description to determine if it was related to Brazilian politics. We finally manually selected and joined only those groups with a clear association for inclusion in our dataset. After completing this meticulous process, our dataset consisted of a total of 1,101 public WhatsApp groups. These groups are primarily chat groups managed by individuals affiliated with political parties, local community leaders, and engaged users participating in political discussions.

After that, we extract the database from the smartphone with the WhatsApp account configured, which includes all groups monitored and gather the data about all messages, with the following information: (i) user ID, (ii) ID of the group in which the message was posted, (iii) timestamp, (iv) whether the messages were forwarded labeled or not, (v) text of the message, (vi) media type of the message (e.g., images, audio, and videos) and, when available, (vii) actual attached multimedia files (e.g., images, audio, and videos) downloaded through a (viii) filename in a unique media_url provided by the WhatsApp message.

For processing WhatsApp data, each message is stored isolated, and we do not have any prior metric about merged content on chats. Therefore, to track aggregated metadata about distinct media being shared and forwarded, we must also perform a post-processing step of merging multimedia messages by their content similarity to find near-duplicates within our dataset and measure their popularity. For audio, videos, and documents, we combine identical files using the MD5 checksum of each media file, as it allows us to detect copies of the same file and then track shares of the same content. Conversely, images are easier to manipulate than other multimedia formats and require better strategies to detect and merge duplicated content on WhatsApp (Melo et al. 2019a). To relate all messages representing the same content, we calculated a perceptual hash.

Using Perceptual Hash to Compare Images

For merging two pieces of information containing an image as a unique item, our approach uses the perceptual hash algorithm pHash (Zauner 2010) to calculate a fingerprint for every image. Unlike cryptographic hashing algorithms such as MD5 and SHA, perceptual hashing considers the visual attributes of the file. Therefore, small divergences in the content will result in slight differences between hashes, making it possible to compare them and see similar content.

Using visual hashing for detecting similar images is a well-known method used by researchers and in industry, especially for digital forensics and cybercrime studies on Web (Hao et al. 2021). Its structure and easiness of computing allow the hashes to be used to explore the immense universe of online images and find abusive content on the Web. However, this technique has limitations, particularly in cases in which images are manipulated to have different hashes, evading detection (Struppek et al. 2022). This can be used to fool perceptual-based systems and even real-world search engines (Hao et al. 2021). Furthermore, in the context of misinformation imagery spread online, the false information of a fabricated image can emerge exactly from just minor alterations made to the source. By utilizing a hashing method capable of effectively generalizing image characteristics, we can relate the original image and its edited similar copies. However, we can still discriminate between legitimate and fake content by using exactly equal hashes as a threshold to determine duplicated images to avoid the risk of combining different content as a single one.

Next, we further evaluate the impact of the hashing algorithm regarding the time to data collection. For those experiments, we used a total of 905K image files, 375K video files, and 80K audio files from an extended dataset from WhatsApp that were processed using different hash algorithms. In Table 1, there is a summary of the experiments with different hash methods. There, besides the pHash algorithm, we compare it to other popular visual hashing methods such as Average Hash (aHash), Differential Hash (dHash), Wavelet image hash (wHash), and Facebook PDQ[3]. While cryptographic hashing reduced the total number of unique image files by 4%, perceptual hashes matched up to 29% of the files containing distinct content. Moreover, it does not take much more time to process compared to the checksum method.

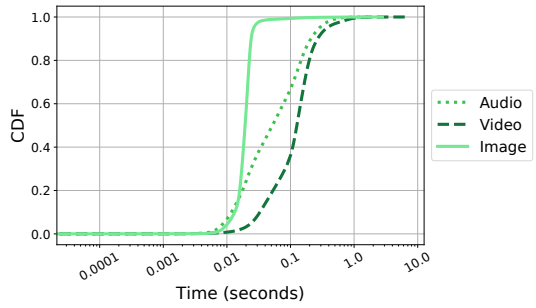
Figure 4(a) shows the time required to process the checksum hash for each media format. Images are the fastest media to process the hash, taking mostly between 0.01 and 0.03 seconds. Even though there are more than twice as many image files as video files, we observed that the time required to process these videos was practically three times greater than the total time of the images (5.2 hours to complete all images and 16 hours to all videos).

Figure 4(b) shows the time required by different hash algorithms to process the image data of WhatsApp; we observe that the checksum is the fastest method to compute the hash. However, it is not a perceptual hash. The Facebook algorithm PDQ is the method that required more computation time to complete the task. We can note the impact of these time differences on processing the entire dataset. The total time required to process all 900k images using PDQ was almost ten days, while pHash took about 6 hours to run all images. For these reasons, pHash was selected to process

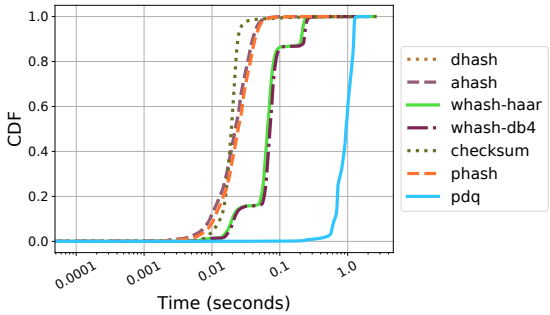
[3]https://github.com/facebook/ThreatExchange/blob/master/hashing/hashing.pdf

	checksum (MD5)	pHash	aHash	dHash	wHash (haar)	wHash (db4)	PDQ
Unique Hashes	867,857	714,114	646,533	741,533	652,472	729,840	783,205
Matching Content	4%	21%	29%	18%	28%	19%	14%
Total Time Spent (min)	316	395	359	361	1,202	1,308	14,044

Table 1: Comparison of different visual hashing methods for processing image data (based on 906,671 files).



(a) Multimedia Checksum



(b) Images Hashes

Figure 4: Time needed to process the checksum and perceptual hashes for multimedia types within WhatsApp data.

image data on WhatsApp due to its effectiveness in detecting duplicates of the content shared in public groups. Also, pHash is faster than more complex hash algorithms such as wHash and PDQ, while it is also as fast as some simple hash methods such as average hash. Although measuring the distance between two pHashes is possible, this study merges only images with identical hashes. This inflexible threshold is necessary to distinguish problematic images (e.g., original and false images) made by making small image manipulations, slightly changing the content to mislead the user. Therefore, it is wanted that the original and slightly manipulated images are stored distinctly.

Dataset Overview

Table 2 provides an overview of the final collection. Our dataset comprises 9.78M messages, with half being multimedia (i.e., images, videos, audio, documents, and stickers). In sum, more than 2.5M (26.13%) messages were forwarded (received "Forward" label by WhatsApp), and 491K (5.02%) received the "Forwarded Many Times" flag. As we do not have a simple, unique hash for text comparison as other media types, we do not calculate unique text messages among our data. We also examined the frequency of appearances for each unique media item within our dataset, as illustrated in Figure 5. As expected, for all types of multimedia, over half of distinct media items were shared more than once on WhatsApp. This indicates that users are frequently inclined to reproduce messages within the platform, even under imposed restrictions. There's a pressing need to examine how content is shared in this enclosed environment of an instant messaging service.

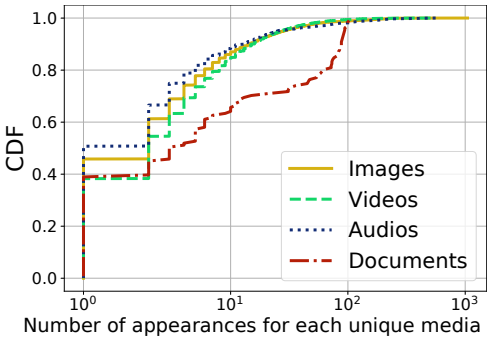


Figure 5: Distribution of the duplicated appearances for each unique multimedia in our dataset per media type.

While documents are the least common media type in our dataset, their distribution reveals that this media is generally duplicated more frequently. On the other hand, documents exhibit a maximum of 100 appearances for a unique piece of media, while other media formats in our dataset contain instances of messages with hundreds or even over a thousand occurrences for images. This observation highlights the substantial presence of multimedia content on WhatsApp groups, often with multiple appearances of the same media. This pattern underscores an important behavior that a significant portion of the discourse within these public WhatsApp groups consists of replicated content rather than novel contributions from their members.

Ethical Consideration and Limitations

Our data collection covers a large dataset from public groups on WhatsApp, but we are aware that most of the conver-

	#Messages	#Forwarded	#Forwarded Many Times	#Unique Media
Text	4,750,816	794,842	54,106	-
Video	2,180,920	1,125,999	299,050	795,901
Image	1,361,021	540,647	105,371	636,925
Sticker	1,034,283	1,335	0	90,890
Audio	441,320	84,552	29,918	333,057
Document	15,673	8,666	2,643	5,639
Total	**9,785,951**	**2,557,026**	**491,438**	**1,862,412**

Table 2: Overview of our political dataset from WhatsApp. We divide the messages according to their media type.

sations occur in private channels, and the data collection may not be representative of the entire WhatsApp network, and our results refer to the content that circulates on the public layer of the platform. Nevertheless, evidence suggests that public groups form the key backbone of misinformation campaigns on WhatsApp, especially in political contexts (Resende et al. 2019; Melo et al. 2019b; Jakesch et al. 2021). Moreover, Brazil has a substantial user base on WhatsApp, and the similarities between the Brazilian group ecosystem and those of other major user bases such as India and Indonesia (Melo et al. 2019b) make this project valuable and applicable to various contexts. This research sheds light on the opaqueness in which WhatsApp is exploited for mass communication, thereby contributing to greater transparency for a platform that holds a pivotal role in our society but remains relatively closed under encrypted architecture.

Our methodology follows a strategy of joining only publicly accessible WhatsApp groups shared on the Web for general discussion. Our focus is exclusively in the propagation of information through the WhatsApp network. Therefore, we do not keep any sensitive data from users nor groups, using only the shared content from messages. All users and messages are anonymized by hashes to make comparisons and analysis of propagation. Furthermore, we do not make available any data that could put at risk the privacy of the groups or their members.

Measuring Forwarding Dynamics

In this section, we analyze the forwarding dynamics on WhatsApp and assess the effectiveness of WhatsApp measures for flagging and limiting forwarded messages.

Content Flagged as Forwarded by WhatsApp

We begin by examining the prevalence of forwarded content within the public chat groups for political discussion that we monitored. WhatsApp labels messages helpfully as "*Forwarded*" and "*Forwarded Many Times*," allowing us to quantify the extent of forwarded content in these groups. As depicted in Figure 6, our dataset reveals that a significant proportion of messages in these groups received the "Forwarded" label. In fact, we can observe that in most groups, over 20% of the content is forwarded, and in the top 11% of groups, more than half of the messages are forwarded. This suggests that a substantial portion of the content in these groups did not originate within the group itself but

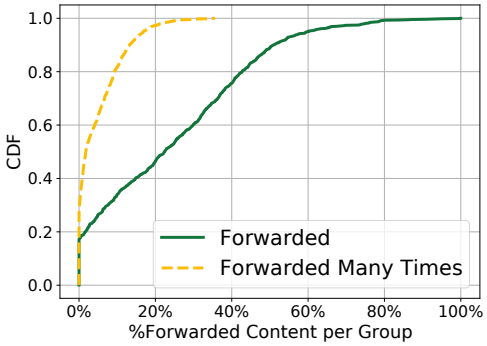


Figure 6: Portion of forwarded content in each group.

was shared from external sources. Remarkably, there are instances of groups where more than 80% of the messages are forwarded, suggesting that these groups primarily serve as repositories for dumping externally sourced content.

This highlights the critical role of message sharing within the communication ecosystem of WhatsApp. Much of the content within the monitored groups did not originate there, demonstrating that users frequently employ the forwarding feature to redistribute messages among different chats. This behavior facilitates the quick dissemination and virality of content within this enclosed ecosystem despite the encryption applied to the messages.

Next, we delve into the ability of WhatsApp to flag forwarded messages based on media type. Figure 7 illustrates the proportion of messages that were directly sent by users, labeled as "*Forwarded*", and labeled as *""Forwarded Many Times" "* for various media formats. While text messages and stickers have a negligible percentage of Forwarded content, other media types exhibit a higher prevalence, with documents leading with 75.6%, followed by videos (65%), images (54.5%), and audio (18.2%).

We observe an expected result that multimedia content is more prone to be forwarded by users than plain text. Given the nature of the application, which is mainly designed for text-based chat, creating an original text message is straightforward. In contrast, producing audio, images, videos, or documents is more time-consuming, which may explain why users often opt to forward these formats instead of composing new multimedia content. Users, therefore, more frequently share content using the forwarding features

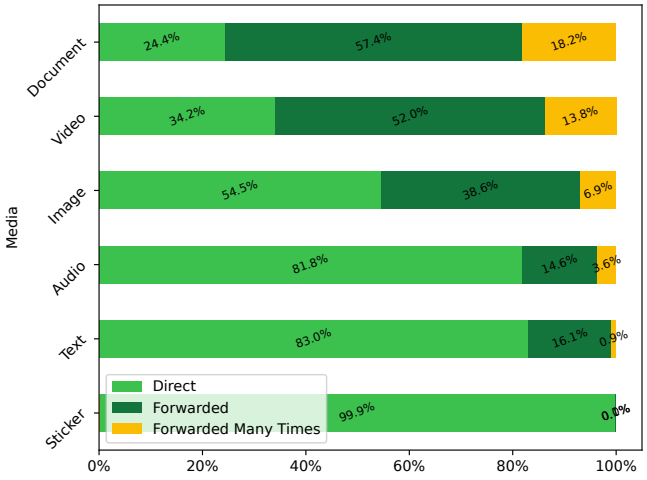


Figure 7: Portion of forwarded/viral messages sent per type.

offered by WhatsApp than create their own new messages, especially for more complex medium formats. However, by doing that, those media may rapidly spread through the network, flowing beyond individual groups' boundaries and making it more challenging to track the origin of that message. These findings reinforce the public and viral nature that this platform can assume, challenging the conventional perception of instant messaging services as exclusive and private platforms.

Cloned Content

Even though the "Forwarded" and "FMT" labels from WhatsApp help to understand content dissemination in the platform, they may not capture all the paths in which messages circulate through the network. These labels rely on an individual identifier assigned to each forwarded message, which allows WhatsApp to track the forwarding chain and calculate the FMT flag. However, this approach has limitations. Users can circumvent it by saving and sending content directly or sharing the same content obtained from external sources on WhatsApp. In these situations, the forwarding chain is broken, and consequently, the FMT label cannot track the further spreading.

This challenge is particularly pronounced when tracking duplicated content circulating within instant messaging platforms. As each message is treated separately, identical messages may come from isolated sources, and WhatsApp will not even notice duplicity, creating what we define as two distinct **clone threads** for the same message.

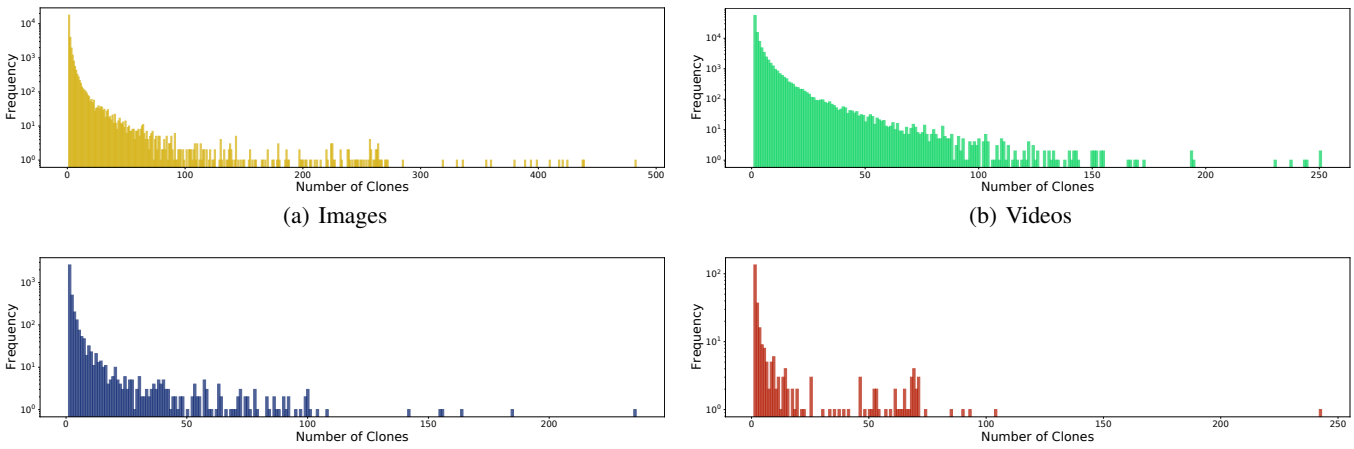
To address these limitations, we developed a methodology that utilizes hashing algorithms to compare multimedia content, as described in Section , enabling us to detect near-duplicates of media shared on WhatsApp. This approach enables us to better quantify the spread of a single piece of content, even when labels of WhatsApp are absent. We compare these media hashes to the individual identifiers of WhatsApp, revealing the number of clone threads for each unique piece of media.

Figure 8 presents a histogram with the number of clones detected for each distinct media marked as "Forwarded Many Times" in our dataset, categorized by message format. A single forwarding chain for a media indicates that WhatsApp successfully consolidated all occurrences and encapsulated them under a single identifier. Alternatively, the existence of multiple clones suggests that the popularity of the content is fragmented among various distinct identifiers, presenting a challenge for WhatsApp to identify distinct occurrences as identical content. While some media exhibit only one identifier, a significant number display numerous clones, with instances having more than 200 clones. This prevalence of clones suggests that users are effortlessly bypassing WhatsApp forwarding mechanism by reproducing multimedia content through alternative means rather than using the built-in Forward function. Therefore, relying solely on FMT labeling may not adequately measure the popularity of a message.

WhatsApp employs the FMT metric to diminish the reach of viral content within its network, restricting the number of times a user can forward such content, arguing that this limitation is "*a way to help keep conversations on WhatsApp intimate and personal´´* and this also helps *´´slow down the spread of rumors, viral messages, and fake news*"[4]. However, users can copy a message and send it directly to recipients rather than using the built-in Forwarding function, easily bypassing any restrictions imposed by WhatsApp. Consequently, they can share viral content without constraints. In conclusion, while WhatsApp FMT labeling is a valuable tool, it cannot prevent users from widely sharing multimedia content. The ease with which users can bypass the forwarding counting system by creating multiple clones of the same content sheds light on how the platform can be exploited, facilitating the viral spread of harmful content such as misinformation, hate speech, or conspiracy theories without adequate warning to WhatsApp users.

Discrepancy between forwards and duplicates "Forwarded Many Times" is the label used by WhatsApp to define potentially viral content. If one single occurrence of a message is marked as such, we would expect that all other appearances of the same messages should receive the same label. Given that a message gets a FMT label after five steps away from its source, we would expect most of the appearances of a popular message to be flagged as such. However, our analysis highlights a significant discrepancy between the total appearances and the instances flagged as "Forwarded Many Times" by WhatsApp. We evaluate the proportion of multimedia messages lacking this label, quantifying how users bypass the WhatsApp forwarding detection system and contribute to information spreading. For instance, we observed numerous instances of media with over 200 appearances in our dataset, but with less than 20% of them labeled as FMT by WhatsApp. Only 40 occurrences of these media are defined as potentially viral for WhatsApp, while the remaining 160 instances of that media went unnoticed by the WhatsApp labeling mechanism and, consequently, also for the users. Thus, there is a clear gap between the number of times a media is shared on WhatsApp and the capacity of

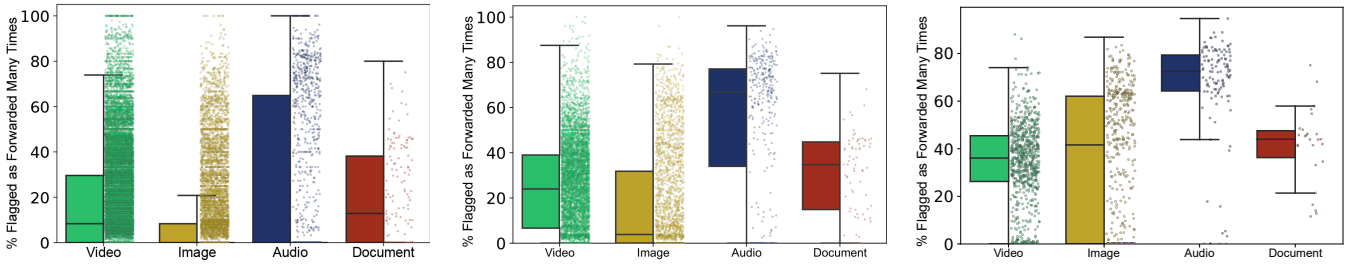
[4]https://faq.whatsapp.com/1053543185312573



(a) Images (b) Videos

(c) Audios (d) Documents

Figure 8: Histograms of number of clones, duplicated media within messages on WhatsApp per type.



(a) Appearances $\geq$ 10 (b) Appearances $\geq$ 25 (c) Appearances $\geq$ 100

Figure 9: Distribution of the portion of occurrences flagged as "Forwarded Many Times" by WhatsApp per media type.

the system to find and track them.

To better understand the implications of the ability of WhatsApp to flag data, we investigate the distribution of flagged "Forwarded Many Times" content relative to the total appearances of each media type in the boxplot of Figure 9. Media with more shares tend to have a higher percentage of instances flagged by WhatsApp. For instance, considering sets of media with at least 10, 25, and 100 appearances, the percentage of unflagged instances decreases as media become more widely shared.

However, even in the case of media with over a many appearances in our dataset (Figure 9(c)), none of them have all 100% of their instances flagged. Approximately 14% of these media lack any flagged instances, WhatsApp did not recognize not even one as FMT. Moreover, only 31% have more than half of their appearances labeled as "Forwarded Many Times" . This suggests that, from the users' perspective, it can be challenging to distinguish whether content originated from a personal contact or if it is a viral message, as the presence of a label would suggest.

It is also interesting to note some differences in flagged data between media types. Most audio data have more than 50% of their occurrences tagged as FMT, while for video and images, there are much fewer instances tagged as such. In particular, for images, we have a considerable amount of media, and none of the shares were labeled. This suggests that images and videos are harder to track their viral dissemination and that it is easier for WhatsApp to keep the forwarding chain for other media formats like audio and documents, given the percentage of instances they actually flagged.

Finally, we measured the gap between the total of appearances of each duplicated media and those instances labeled as "Forwarded Many Times" by WhatsApp.To accomplish that, we count all distinct multimedia messages that received at least one "Forwarded Many Times" label, which gives us a total of 136,924 distinct media. Then, summing the number of appearances for all those media, we estimate that they were shared 1,037,113 times within our data. Finally, we calculate how many of these instances did not receive a "Forwarded Many Times" flag by WhatsApp. We have found that **612,855 (59%)** occurrences were not labeled by WhatsApp. This analysis reveals the extent to which WhatsApp labeling system may fail to capture the full view of media sharing activity on the platform, highlighting potential gaps in monitoring and tracking forwarded content.

Content Analysis

After examining forwarding aspects of the messages, we perform an in-depth analysis of two case studies on our content to provide further insights into dynamics of WhatsApp: (i) five representative images from our dataset and (ii) some piece of texts explicitly instructing users to bypass WhatsApp forwarding mechanism.

Top Images Our first approach is selecting the five images with the most viral shares within our data. To identify the top images, we first filtered all images that received at least one "Forwarded Many Times" tag from WhatsApp, then ranked them based on the total number of occurrences and the proportion marked as FMT. With that rank, we manually investigate the top five regarding total numbers of shares, viral shares, whether the image represents misinformation or not, and instructions to readers to share with their contacts. As expected, all of them are related to the political context of Brazilian elections. Figure 3 summarizes our findings.

To identify if the image is fake, we rely on specialized fact-checking agencies in Brazil (Reis et al. 2020b; Resende et al. 2019). We manually search each image on search engines (i.e. Google and Bing) to find if any fact-checking agencies have checked this information, and only if we find the exact same image checked, we assume the content is fake. Notably, in three out of the five most viral images within our dataset, we have found they are fake, which evidences misinformation is among the most popular content we investigate. Additionally, in two of these fake images, there are clear instructions to users to widely spread its content over WhatsApp, which shows to be an effective way to share a message, even with restriction limits imposed. It is important to highlight that WhatsApp was actually able to identify most of the share instances of these images as viral (i.e. they were tagged as FMT), however, about 30% of their occurrences lack such label and circulated freely through the platform without any warning.

The other two images also demonstrate interesting findings regarding forwarding on WhatsApp. Both of them are explicit chain messages, which are texts that attempts to convince the recipient to make a number of copies and pass them on to numerous other recipients. This suggests the popularity of such chains within WhatsApp. For example, the last image is just a screenshot of plain text with sensationalist content: "*If you send it to just 20 contacts in one minute... the whole of Brazil will unmask this criminal. DO NOT break this chain. The unwary need to be enlightened before it is too late*".

Another interesting finding about observing the images is that the image with more occurrences, with 1051 shares over the dataset, has only one single message tagged as viral by WhatsApp. This means that for 99.9% of its occurrences, it would not be possible to identify that it is viral in any other way in WhatsApp than through our similarity methodology.

Investigate Users Intention Next, we investigate whether the potential for circumventing WhatsApp's forwarding mechanism is something inherent of the system architecture design or arises from intentional or malicious user behavior. Our next analysis performs an examination of user intentions to verify if there are indications that users actively seek to evade this system, or, alternatively, could the cloning, downloading, and resharing of content within WhatsApp we present in this work be a natural consequence of information dissemination online?

Many fake WhatsApp messages directly request users to forward them to others (Mohan, Nagadeepa, and Bharathi 2020), highlighting the significance of forwarding behavior also in the context of misinformation. Since asking for sharing is a common feature users turn to reach a broad audience, as we also observe in images, we investigate how they do this in text messages. To do this, we compiled a list of keywords related to sharing requests, such as forwarding, share, copy and paste, and some variations of these words in Portuguese. Next, we search and filter text messages containing such words, merging similar messages using the Jaccard Index to facilitate the evaluating process. After that, an author manually goes through these messages to investigate the intention of sharing them by the users.

Analyzing the messages, we find evidence of deliberate content copying-pasting and resharing to bypass the FMT label of WhatsApp. Below, we exemplify some of our findings of pieces of chain text messages and how many times each one appeared within our dataset, supporting the idea that users consciously circumvent the WhatsApp forwarding mechanism to avoid restrictions.

- *Please, copy and paste to be able to send it to 5 people, as it was limited to 1. (...)* (412 times)
- *Copy, paste, and forward as much as possible. If you only forward, soon you can only forward one at a time.* (322 times)
- *Copy and paste if you're forwarding to avoid the double arrow that makes it difficult to send in batches of five.* (170 times)
- *(...) and the maneuver of sending one message to one group at a time began! LET'S GO, BRAZIL, GO GET THEM! ATTENTION! In this case, just copy and paste the message once, and it will go back to sending to up to 5 groups normally.* (111 times)
- *(...) if you copy and paste and send it to 5 groups, this text will not be blocked* for excessive sending* (84 times)
- *copy and paste to forward without WhatsApp restrictions* (25)
- *ATTENTION!* WhatsApp *will remove the forward option* in the latest update because they want to STOP the spread of videos from the protests. *TO CIRCUMVENT THIS CENSORSHIP* and continue sharing as much as possible all videos and content about the protests, share this step-by-step guide to *how to forward to up to 50 contacts* using WhatsApp Web with all your contacts. To use it on the computer, install this extension: [link]. Access WhatsApp Web and go to the *settings* menu and enable the option: *Increase the native sharing limit from 5 contacts to 50 contacts* Now, when you use the forward option, *you can select up to 50 contacts at once*!"* (7 times)

Those messages not only ask users to share them, but they also point to the WhatsApp imposed limits in using the Forward button and instruct how other users can bypass the barrier. One message even presents a supposed software that allows a user to expand the forward limits. This leads us to believe that even though part of the forwarding can happen from organic behavior due to the WhatsApp design choices, there are those who exploit the system to intentionally make their messages viral. While organic behavior may not always bring harm to users, this can be particularly weaponized by misinformation campaigns to threaten demo-

Description	Total Shares	FMT Shares	Fake?	Ask users to share?
A hashtag campaign from a right-wing candidate to be shared over WhatsApp.	741	506	No	**Yes**
Fake photo accusing a candidate of being a friend of a criminal who attacked Bolsonaro.	581	420	**Yes**	**Yes**
Fake photo accusing a candidate of not using national flags.	585	415	**Yes**	No
Fake photo accusing a vice presidential candidate of being part of a criminal organization.	546	382	**Yes**	**Yes**
A WhatsApp chain message asking users to share it with as many people as possible.	526	348	No	**Yes**

Table 3: Description of top five most viral images with "Forwarded Many Times" label within our data collection.

cratic processes in countries where WhatsApp has a huge reach among the population.

It becomes evident that user intentions play a significant role in the sharing ecosystem of WhatsApp. We have unveiled instances where users consciously seek to evade WhatsApp's forwarding metrics by distributing messages that encourage the copy-paste-share strategy. These findings have direct implications for WhatsApp ongoing efforts to curb the spread of misinformation and viral content within its platform. This also demonstrates the challenges faced not only by WhatsApp, but also by other messaging platforms in maintaining content integrity.

Take-aways

We analyzed a large data set of almost 10M messages and 1,101 WhatsApp public groups, and our main takeaways are:

- Forwarding is a key component of many group-based communications on WhatsApp, as 11% of groups in our dataset have more than 50% of their messages being flagged as forwarded, and, most of the multimedia content appeared more than once within our collection.
- Although WhatsApp has a metric to count and track forwarding chains to combat message virality, there are ways in which users can bypass this metric and create multiple cloned content within the platform, making even harder the task of detecting and blocking the spread of viral messages within public groups.
- As a result, we find that 59% of all appearances from viral multimedia messages did not receive FMT flag, which represents a total of 612,855 instances of duplicated messages that were unflagged.
- We find evidence that WhatsApp users intentionally aim to bypass WhatsApp mechanisms by nudging other users to share information in a way that will not get detected.
- Also, we find that among the most viral images in our dataset, there is a substantial percentage of them that are sharing misinformation.

These findings reveal that the strategy of WhatsApp to mitigate the spread of potentially viral harmful content within its network, based solely on forwarding tracking, fails in detecting over half of the instances of popular media captured in our dataset. WhatsApp affirms that a message earns the FMT status after five hops of sharing, with the intent of alerting users. It implies that this measure is sufficient to counter the proliferation of viral content on its network. However, our study challenges this assumption, suggesting that multimedia content on WhatsApp can still achieve extensive dissemination across the network, and the forwarding metric employed by WhatsApp is not consistently effective in detecting all instances. Given that users may expect that label to support their decision whether the content is viral or not, especially in an environment of political discussion where misinformation is a known issue, the absence of this flag can mislead users into believing the authenticity of the content, potentially intensifying the spread of misinformation within the platform.

Related Work

WhatsApp, a widely used instant messaging platform, has become a focal point for the dissemination of misinformation, posing significant challenges in various global contexts. Notably, during the Indian general elections, WhatsApp played a pivotal role in the rapid spread of rumors and fake news (Jakesch et al. 2021). In Brazil, the impact of misinformation campaigns on WhatsApp was also pronounced during presidential elections, raising concerns about their influence on electoral outcomes (Resende et al. 2019; Machado et al. 2019). Similar issues related to misinformation on WhatsApp have been reported globally, including in Indonesia (Kwanda and Lin 2020), Pakistan (Javed et al. 2020), Nigeria (Cheeseman et al. 2020), and Spain (Elías and Catalan-Matamoros 2020). Additionally, the COVID-19 pandemic witnessed WhatsApp as a prominent vector for the propagation of false health-related information (Vijaykumar et al. 2021).

Misinformation concerns on WhatsApp are not confined to textual content alone; multimedia elements such as audio messages (Maros et al. 2020) and images (Reis et al. 2020b) also serve as conduits for misleading information. Users themselves expressed concerns regarding the prevalence of false information on WhatsApp (Newman et al. 2021). Misinformation, along with hate speech, conspiracy theories and extremism, are also prominent problems in the context of other instant messaging platforms (Hoseini et al. 2020). On Telegram, evidence was also found of the abuse of the platform to spread political disinformation (Júnior et al. 2022), and be exploited by terrorist organizations(Prucha 2016; Yayla and Speckhard 2017). Discord has been used for organizing extremist rallies, e.g., the "Unite the Right" in Charlottesville in 2017 (Roose 2017) and for disseminating potentially harmful and sensitive material (Cox 2018). Problems with misinformation were also highlighted on the Japanese LINE (Funke 2018) and on WeChat in China (Lu et al. 2020; Guo and Zhang 2020).

The group chat feature is central to these dynamics on

WhatsApp (Seufert et al. 2016), and a significant portion of conversations occurs within these groups (Rosenfeld et al. 2016). Researchers have developed methodologies to collect WhatsApp data from public groups (Garimella and Tyson 2018), scrutinized political misinformation in these environments (Resende et al. 2019; Bursztyn and Birnbaum 2019), and created systems for data exploration (Melo et al. 2019a). Some studies have probed the impact of message forwarding limits on message dissemination within public groups (Melo et al. 2019b). Beyond misinformation campaigns, research has investigated message cascades on WhatsApp (Caetano et al. 2019) and the privacy and security aspects of instant messaging platforms (Abu-Salma et al. 2017; Rösler, Mainka, and Schwenk 2018). Despite these research efforts, a crucial gap persists in understanding the dynamics of message forwarding within the closed network structure of WhatsApp. Our work addresses this gap by systematically analyzing the forwarding of messages in WhatsApp, shedding light on the mechanics and challenges of message propagation within this unique digital ecosystem.

Conclusion

In this study, we have delved into the intricate dynamics of message forwarding within WhatsApp, uncovering how this feature transforms an encrypted and private platform into a medium for mass communication also. The ability to share messages rapidly within WhatsApp complex network can lead to unintended viralization, where users may find their content spreading across the platform without their consent or even knowledge. At the same time, users can receive messages without knowing their true origin, blurring the lines between personal conversations and viral content, creating a thriving environment for the spread of misinformation.

WhatsApp has implemented features like the "Forwarded" and "Forwarded Many Times" labels to help users identify shared messages and restrict the further spread of popular/viral content. However, our analysis reveals that these labels alone may not be sufficient, potentially confusing users and blurring the distinction between private and public conversations. Notably, 59% of popular media did not receive a proper "Forwarded Many Times" label and circulated in the groups we monitored without any warning about its virality. Furthermore, we discovered that three out of five of the most viral images are, in fact, fake content. Also, we find evidences of content creators intentionally asking users to share their messages and instructing how to bypass the restrictions. These findings suggest that misinformation campaigns are actively exploiting those issues to reach a broader audience within the network.

Moreover, end-to-end encryption architecture increases the difficulty in tracking the origins of viral content, allowing many messages to circulate freely, evading detection and moderation from the platform. Our findings underscore the complexity of combating misinformation and abusive content within the messaging platform ecosystem without violate users privacy. Even the platform's built-in measures can be easily circumvented by user actions, revealing the need for a more comprehensive approach to addressing these issues. The debate between ensuring user privacy through encryption and implementing tools to counter viral misinformation remains a crucial one for policymakers and platform operators. To address these challenges, our study proposes a solution using perceptual hashes to detect duplicate multimedia messages shared in public groups. By employing similarity methods, messaging apps could identify and block potential harmful messages before they gain further traction within the platform. Our approach demonstrates that there are means to enhance content moderation within this enclosed environment without violating users privacy.

In conclusion, the trade-offs between security, encryption, and countering emerging issues like viral misinformation need to be carefully considered. Safeguarding the WhatsApp environment against malicious users exploiting vulnerabilities is imperative. By doing so, we can create a safer digital space that minimizes the spread of viral misinformation and harmful content, preserving the platform's integrity.

Acknowledgments

This work was supported by a grant 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 *EuroUSEC '17*. Internet Society.

Allyn, B. 2022. Telegram is the app of choice in the war in Ukraine despite experts' privacy concerns. Online. *NPR – National Public Radio*. https://www.npr.org/2022/03/14/1086483703/telegram-ukraine-war-russia.

Bakare, A. S.; Abdurrahaman, D. T.; and Owusu, A. 2022. Forwarding of Messages Via WhatsApp: The Mediating Role of Emotional Evocativeness. *Howard Journal of Communications*.

Berger, J. 2016. *Contagious: Why things catch on*. Simon and Schuster.

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. ACM.

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

Cox, J. 2018. The Gaming Site Discord Is the New Front of Revenge Porn. *The Daily Beast*. [Online].

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.

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

Funke, D. 2018. How misinformation spreads on Line – one of the most popular messaging apps in Southeast Asia. *Poynter*. https://www.poynter.org/fact-checking/2018/how-misinformation-spreads-on-line-\%c2\%97-one-of-the-most-popular-messaging-apps-in-southeast-asia/. [Online].

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.

Guo, L.; and Zhang, Y. 2020. Information flow within and across online media platforms: An agenda-setting analysis of rumor diffusion on news websites, Weibo, and WeChat in China. *Journalism Studies*, 21(15): 2176–2195.

Hao, Q.; Luo, L.; Jan, S. T.; and Wang, G. 2021. It's Not What It Looks Like: Manipulating Perceptual Hashing Based Applications. In *CCS '21*, 69—85. ACM.

Harvey, C. G.; Stewart, D. B.; and Ewing, M. T. 2011. Forward or delete: What drives peer-to-peer message propagation across social networks? *Journal of Consumer Behaviour*, 10(6): 365–372.

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'20*.

Jakesch, M.; Garimella, K.; Eckles, D.; and Naaman, M. 2021. Trend Alert: A Cross-Platform Organization Manipulated Twitter Trends in the Indian General Election. *CSCW'21*, 5.

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.

Júnior, M.; Melo, P.; Kansaon, D.; Mafra, V.; Sá, K.; and Benevenuto, F. 2022. Telegram Monitor: Monitoring Brazilian Political Groups and Channels on Telegram. In *Hypertext'22*, 228–231.

Karimiyazdi, R.; and Mokhber, M. 2015. Improving viral marketing campaign via mobile instant messaging (MIM) applications. *Journal of Advanced Review on Scientific Research*, 10(1): 20–33.

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.

Lu, Z.; Jiang, Y.; Lu, C.; Naaman, M.; and Wigdor, D. 2020. *The Government's Dividend: Complex Perceptions of Social Media Misinformation in China*, 1–12. ACM.

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 *The Web Conference*, 1013–1019. ACM.

Malka, V.; Ariel, Y.; and Avidar, R. 2015. Fighting, worrying and sharing: Operation 'Protective Edge'as the first WhatsApp War. *Media, War & Conflict*, 8(3): 329–344.

Maros, A.; Almeida, J.; Benevenuto, F.; and Vasconcelos, M. 2020. Analyzing the Use of Audio Messages in WhatsApp Groups. In *The Web Conf. 2020*, WWW'20. .

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

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 *Complex Networks*, 372–384.

Mohan, R.; Nagadeepa, C.; and Bharathi, N. 2020. Follow and/or forward: Impact of e-WoM on WhatsApp health messages. *Science, Technology and Development*, 9(1): 60–64.

Mukherjee, R. 2020. Mobile witnessing on WhatsApp: Vigilante virality and the anatomy of mob lynching. *South Asian Popular Culture*, 18(1): 79–101.

Napoli, P.; and Caplan, R. 2017. Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, 22(5).

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

Phelps, J. E.; Lewis, R.; Mobilio, L.; Perry, D.; and Raman, N. 2004. Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research*, 44(4): 333–348.

Prucha, N. 2016. IS and the Jihadist Information Highway–Projecting Influence and Religious Identity via Telegram. *Perspectives on Terrorism*, 10(6).

Reis, J. C.; Melo, P.; Garimella, K.; and Benevenuto, F. 2020a. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *(HKS) Misinformation Review*.

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

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

Roose, K. 2017. This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville. *New York Times*. [Online].

Rosenfeld, A.; Sina, S.; Sarne, D.; Avidov, O.; and Kraus, S. 2016. WhatsApp usage patterns and prediction models. In *ICWSM/IUSSP Workshop on Social Media and Demographic Research*.

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.

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.

Struppek, L.; Hintersdorf, D.; Neider, D.; and Kersting, K. 2022. Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 58—-69. ACM.

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).

WhatsApp. 2020a. Keeping WhatsApp Personal and Private. Online. *Facebook Newsroom*. https://about.fb.com/news/2020/04/whatsapp-message-forward-limit/.

WhatsApp. 2020b. Two Billion Users – Connecting the World Privately. Online. *WhatsApp Blog*. https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately.

Yayla, A. S.; and Speckhard, A. 2017. Telegram: The mighty application that ISIS loves. *International Center for the Study of Violent Extremism*.

Yu, C. W.; and Kamarulzama, Y. 2016. Viral Marketing via the New Media: The Case of Communication Behaviour in WhatsApp. *IBAICM 2016*, 3: 82–102.

Zauner, C. 2010. *Implementation and benchmarking of perceptual image hash functions*. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg, AT.

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 forwarding mechanism of 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 mechanisms to bypass WhatsApp restrictions on forwarding.
   (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