UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Philipe de Freitas Melo

ACTIVISM AND MISINFORMATION ON WHATSAPP: Measurement, Analysis, and Countermeasures

Belo Horizonte
2022

Philipe de Freitas Melo

ACTIVISM AND MISINFORMATION ON WHATSAPP: Measurement, Analysis, and Countermeasures

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Fabrício Benevenuto

Belo Horizonte
2022

Melo, Philipe de Freitas

M528a Activism and misinformation on whatsapp [recurso eletrônico] : measurement, analysis, and countermeasures / Philipe de Freitas Melo –2022.
1 recurso online (209 f. il, color.) : pdf.

Orientador: Fabrício Benevenuto de Souza.

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f.169-201
.
1. Computação – Teses. 2. WhatsApp (Aplicativo de mensagens) – Teses. 3. Desinformação Teses. 4. Redes sociais – Teses. 5. Disseminação de informação – Teses. 6. Fake News – Teses I. Souza, Fabrício Benevenuto de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*73(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INCIPIT VITA NOVA
7 DE SETEMBRO DE 1927

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Activism and Misinformation in WhatsApp: Measurement, Analysis, and Countermeasures

PHILIPE DE FREITAS MELO

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Fabricio Benevenuto
PROF. FABRÍCIO BENEVENUTO DE SOUZA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

PROF. HEITOR SOARES RAMOS FILHO
Departamento de Ciência da Computação - UFMG

PROF. ALTIGRAN SOARES DA SILVA
Instituto de Computação - UFAM

PROFA. RAQUEL DA CUNHA RECUERO
Centro de Letras e Comunicação - Universidade Federal de Pelotas

Belo Horizonte, 16 de dezembro de 2022.

To my lovely girlfriend, supportive parents, and awesome friends.

Acknowledgments

First of all, I would like to thank my parents, who support me in every step of my life and encourage me to go after my dreams. Also thanks to Marco, Dudu and Carolina, my brothers and sister, who accompanied me and helped me grow as a person. Without you, this journey would not be possible at all. Also to my six nephews Isabela, Gustavo, Max, Renata, Maria Eduarda and Ana Beatriz who help me to see the world even with the simplicity and lightness of children.

I would like to express my gratitude to Clara for her help with this work, not only for her emotional support, but also for the valuable discussions and even for her technical contributions, from which I could even say she is my co-author in all of this journey. I sincerely appreciate your patience and effective help during the many nights of work. Thanks for all attention and care I received, even when I could not return in an equal way. Your company made each challenge much easier than it could be.

Thanks, Luiz, my dear friend with whom I shared most of the good and bad moments in the university. Thanks to Lucas, Manu, Sâmara, Madeira and to all of my friends for the moments of joy. Without these, my life would not be so fun.

I would also like to thank my colleagues from the LOCUS lab, who are a great company not only when working, but also when having leisure time. In particular, I would like to thank Julio, who helped me to see the opportunities in life and to improve myself when I was struggling at UFMG.

To my advisor, Fabrício Benevenuto, thank you for the academic counseling, for all knowledge I acquired during the process, for every incentive and for believing in my potential as a researcher and as a professional. Moreover, thank you for being a friend over an advisor that I could fearless share my thought and ideas.

Finally, I would like to thank the Department of Computer Science and all its staff for attention and commitment to keep all program working.

To all who contributes in any way to my accomplishments throughout life, thank you very much.

“*You can’t make people listen. They have to come round in their own time, wondering what happened and why the world blew up around them.*”

Ray Bradbury, Fahrenheit 451

Resumo

A Internet tem conquistado um espaço cada vez maior em nossa vida. As mídias sociais, em especial as plataformas de mensagens instantâneas (PMIs) permeiam nossas relações e a forma com que nos comunicamos. Essas plataformas permitem usuários se comunicarem tanto de forma mais privada e pessoal, como também possuem grandes grupos de conversa público, no qual múltiplos usuários interagem entre si e compartilham interesses. Devido seu grande sucesso e popularidade, essas plataformas, infelizmente, também se tornaram um local fértil para campanhas de desinformação que espalham boatos, notícias falsas, discurso de ódio e linchamento online. Entretanto, a estrutura fechada de funcionamento dos serviços de mensagens, tais como WhatsApp e Telegram, dificultam a investigação do conteúdo que circula dentro destas plataformas.

Nesta tese, investigamos o ambiente do WhatsApp e a (des)informação que se espalha na plataforma. Essa tarefa apresenta um desafio único, pois o conteúdo que circula na rede do WhatsApp é protegido por uma camada de criptografia ponta a ponta e acontecem em canais privados ou de difícil acesso, enquanto esse conteúdo, geralmente anônimo, ainda consegue se espalhar de forma viral pela rede para milhares de usuários. Exploramos esse lado público do WhatsApp, fornecendo um conhecimento aprofundado sobre o ecossistema de mensagens instantâneas, revelando como as campanhas de desinformação abusam desse ciberespaço e também propondo contramedidas para enfrentar esse problema.

Primeiramente, desenvolvemos uma nova metodologia de coleta de dados no WhatsApp capaz de reunir dados em larga escala e de longo prazo de grupos públicos, na qual coletamos mais de 10 milhões de mensagens de grupos políticos no WhatsApp no Brasil. Esta metodologia abre um novo rumo nos estudos que investigam da plataforma do WhatsApp, revelando tendências e os conteúdos mais populares que são postados nele Em seguida, realizamos uma análise comparativa dos PMIs, proporcionando uma melhor compreensão desse universo de grupos públicos compartilhados nas redes sociais. Mostramos que existe uma enorme quantidade de grupos públicos do WhatsApp, destacando as principais características dos principais serviços de mensagens instantâneas.

Também exploramos o comportamento dos grupos políticos públicos do WhatsApp, investigando as mensagens publicadas e alguns eventos capturados por nosso conjunto de dados, incluindo a desinformação compartilhada nesses períodos. Em relação a esses grupos, exploramos ainda como a desinformação se espalha por meio da rede WhatsApp, avaliando como algumas funcionalidades do sistema impactam no fluxo de disseminação da informação. Nossos resultados revelam que esses recursos ajudam a desacelerar a disseminação do conteúdo, entretanto não evitam que ele se torne viral. Por fim, apresentamos duas medidas práticas e viáveis para se combater desinformação no WhatsApp. Propomos uma arquitetura compatível com a criptografia do WhatsApp e que não viole a privacidade dos usuários. Essa abordagem depende do uso de conteúdo que foi previamente checado por agencias especializadas e poderia ser facilmente implementado no aplicativo do WhatsApp para ajudar os usuários a identificar desinformação no aplicativo. Em segundo lugar, desenvolvemos e colocamos no ar um sistema real - WhatsApp Monitor - que processa os conteúdos mais populares compartilhados todos os dias nos grupos públicos do WhatsApp, ranqueando e exibindo as imagens, vídeos, áudios e imagens mais enviadas em uma interface web que já é utilizada por muitos jornalistas e pesquisadores no combate a desinformação.

Palavras-chave: WhatsApp, Desinformação, Redes Sociais, Disseminação de informação, Fake News.

Abstract

The Internet has gained an increasing space in our lives. Social media, especially the Instant Messaging Platforms (IMPs), permeate our relationships and the way we communicate. These platforms allow users to communicate both in a more private and personal conversation, as well as engaging in large public chat groups, in which multiple users interact with each other and share interests. Due to their great success and popularity, these platforms have unfortunately also become a fertile environment for misinformation campaigns that spread rumors, false news, hate speech, and online lynching. However, the closed operating structure of messaging services, such as WhatsApp and Telegram, makes it difficult to investigate the content that circulates within these platforms.

In this thesis, we investigate the WhatsApp ecosystem and the (mis)information that spreads on the platform. This task presents a unique challenge, as the content circulating on the WhatsApp network is protected by an end-to-end encryption layer and takes place in private or hard-to-access channels, while, at the same time, this generally anonymous content still manages to spread virally over the network to thousands of users. We focus on this public side of WhatsApp, providing an in-depth understanding of the instant messaging ecosystem, revealing how misinformation campaigns abuse this cyberspace, and also proposing countermeasures to address this issue.

First, we developed a new WhatsApp data collection methodology capable of gathering large-scale and long-term data from public groups, in which we collect more than 10 million messages from political groups on WhatsApp in Brazil for more than 3 years. This methodology opens up a new path for studies investigating the WhatsApp platform, revealing trends and the most popular content posted on it. Then, we performed a comparative analysis of IMPs, providing a better understanding of this universe of public groups shared on social networks. We show that there is a substantial amount of public WhatsApp groups and shed light on the main characteristics of the most used instant messaging services.

We also explore the behavior of WhatsApp's public political groups, investigating their published messages and some events captured by our dataset, including the misinformation shared during these periods. In relation to these groups, we also explore how misinformation spreads through the WhatsApp network, evaluating how some features of the system impact the flow of information dissemination. Our results reveal that these features help to slow down the spread of content, but do not prevent it from going viral. Finally, we present two practical and viable measures to combat misinformation on

WhatsApp. We propose an architecture compatible with WhatsApp E2E encryption and that does not violate any users' privacy. This approach relies on using content that has been previously checked out by specialized agencies and could be easily deployed into the WhatsApp app to help users identify misinformation in the chat. Second, we developed and launched a real system – *WhatsApp Monitor* – that processes the most popular content shared every day in WhatsApp public groups, ranking and displaying the most uploaded images, videos, audios and images in a web interface that it is already used by many journalists and researchers to fight misinformation.

Keywords: WhatsApp, Misinformation, Social networks, Information dissemination, Fake News.

List of Figures

List of Tables

List of Abbreviations

API Application Programming Interface

E2EE End-to-end Encryption

CDF Cumulative Distribution Function

IMP Instant Messaging Platform

OSN Online Social Network

PII Personally Identifiable Information

SEI Suscetible-Exposed-Infected Model

SIR Suscetible-Infected-Recovery Model

Wapp WhatsApp

Contents

Chapter 1

Introduction

With the Internet increasingly present in our daily lives, online platforms determine the way we communicate and keep in touch with other people. While some traditional Online Social Networks (OSN) such as Facebook, Twitter and Instagram still very popular around all the globe, the Instant Messaging Platforms (IMP) have also been gaining space and the preference of users The most popular of them, the WhatsApp, was released in 2009 and acquired by Facebook in February 2014, and today has a widespread adoption that achieved more than 2 billion monthly active users in 2020 (WHATSAPP, 2020b).

These IMPs such as WhatsApp are designed for mobile devices and require a smartphone, but they are also accessible from desktop computers[1]. These applications allow users to send text messages, make voice and video calls, and share media files of images, audios, videos, documents, and other content[2]. Furthermore, many of them present an end-to-end encryption to provide more safety and privacy for their users[3], which is a communication encryption that only the final users are able to read the exchanged messages. Moreover, in WhatsApp, one has the possibility to create group chats, which is a key aspect that makes this platform an effective tool to chat with not only one, but multiple other people at once.

Summing all those features with the speed of a synchronous interaction, easiness of access via just Internet access, and the associated low cost compared to Short Message Service (SMS) available to mobile phone users (CHURCH; OLIVEIRA, 2013), these applications have drastically changed the format on how users interact and communicate on Web. Initially, online communications were dominated by a one-to-one paradigm, which is characterized by private and immediate exchange of messages between two equal partners of communications (CAMERON; WEBSTER, 2005; SEUFERT et al., 2016). With the emergence of OSN, some messaging systems became more popular (such as Facebook, Twitter, MySpace), adding a one-to-many layer to the interactions on Web (SEUFERT et al., 2015), where users publish messages and broadcast them, so everyone can read, or at least a restricted list of friends or followers. Recently, because of the democratiza-

[1] <https://faq.whatsapp.com/general/download-and-installation/about-supported-operating-systems/>

[2] <https://www.whatsapp.com/features/>

[3] <https://faq.whatsapp.com/general/security-and-privacy/end-to-end-encryption>

tion of access to smartphones and Internet, IMPs have supplemented this context with also a group communication (MONTAG et al., 2015; SEUFERT et al., 2015), allowing conversations of a fixed group of users that can equally participate in the chat.

Through WhatsApp group communication, users can talk with each other instantaneously from almost everywhere (as they have Internet access) and about anything, sharing common interests in a fixed ecosystem. They also can create and invite others to their public groups by an invitation link. This group communication in WhatsApp is a widespread practice that covers highly connected regions such as Europe and North America, and also including a significant number of users in developing regions, such as Brazil and India (GARIMELLA; TYSON, 2018). As WhatsApp became a powerful tool that provides its users an accessible messaging service, it attracted the attention of many people and organization wanting to explore its full potential. Indeed, the universe of WhatsApp public groups is a valuable source not only to study and understand the behavior of online users, but it is also of big interest in the economy of business models. Actually, WhatsApp itself released in 2018 a standalone app targeted at companies, called WhatsApp Business[4], to allow companies to interact with customers via WhatsApp client (PEREZ, 2018). Despite that, differently from other OSNs, the encrypted and closed nature of the WhatsApp makes it a hard task to investigate what happens inside those groups, and even finding trending topics on there could represent a challenge. On the other hand, WhatsApp and others IMPs turned into an object of considerable concern because of people exploiting this environment to spread rumors, fake news, hate speech, and lynching campaigns that bring a dreadful impact to the entire society.

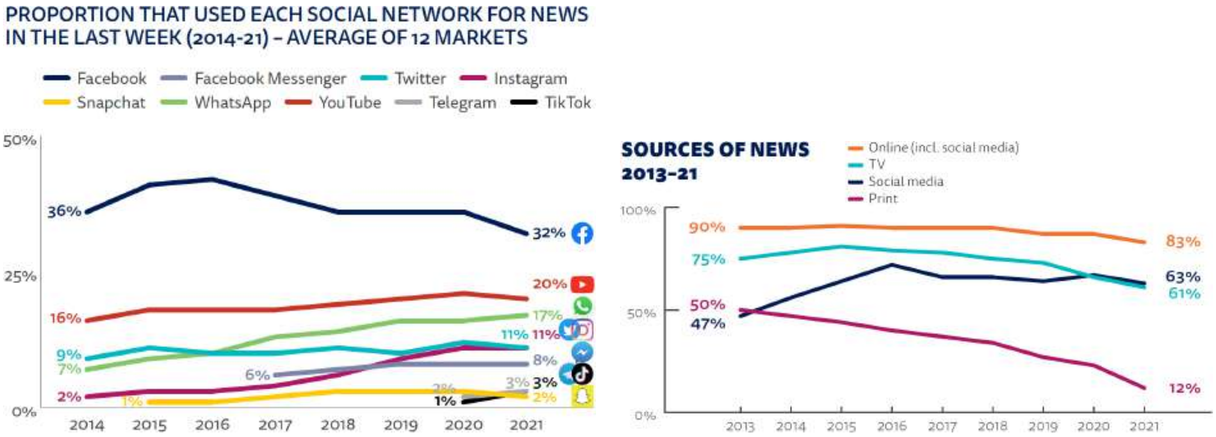
1.1 Motivation

Globally, more than a third of the world's population actively use online social networks, including messaging platforms and other digital platforms (GALLAGHER, 2017). These platforms have significantly changed the way users interact with each other, communicate and inform themselves. A extended survey performed by NEWMAN et al., 2019 on Reuters with users from around all the world observed an expressive growth in the proportion of people that used social networks for news consumption. Additionally, in Brazil, 63% of users said that they consume news through social medias – surpassing even the audience of news from TV (61%) for the first time in historical records – and 58% of them also share news via this media.

However, social medias, unfortunately, have become key environments for fake news

[4] <https://www.whatsapp.com/business/>

Figure 1.1: Results from Reuters of the source people use to consume news.



(a) Use of social media for news.

(b) Sources of news consumption in Brazil.

Source: Reuters Institute Digital News Report 2021 (NEWMAN et al., 2021). .

dissemination. Despite the numerous benefits that these systems bring to our society, they have been reportedly abused by misinformation, especially for political purposes (BESSI; FERRARA, 2016; RIBEIRO et al., 2019). Election after election, we observed new and different forms of abuse and complex strategies of opinion manipulation through the spread of misinformation. The 2016 presidential elections in the USA are still remembered for a 'misinformation war' that happened mostly through Twitter and Facebook. The notorious case involved an attempt to influence through targeting advertising (RIBEIRO et al., 2019) .

One of the platforms that play a major role in this scenario is the WhatsApp. With 2 billions users (WHATSAPP, 2020b), it is the third most used social media in the world (Fig. 1.1(a)), just behind YouTube and Facebook itself (TANKOVSKA, 2021). Regarding messaging services, it is the most used IMP and currently one of the main forms of communication and information diffusion in many countries, including India, Brazil and Germany (NEWMAN et al., 2019). Nearly everyone who owns a smartphone uses WhatsApp in Brazil – about 120 million active users (MAGENTA; GRAGNANI; SOUZA, 2018), which is more than half of the Brazilian population – to keep in touch with friends and family, do business, as well as keep up with the news.

WhatsApp, as one might expect, is more focused on private conversations. Having said that, it is clear that WhatsApp, nevertheless, is used very differently across countries. The presence of large groups chats is what provides to WhatsApp application such variety of purposes and uses. NEWMAN et al., 2019 found that majority of WhatsApp users in Brazil (58%) use groups to interact with people they don't know (along with other countries like Turkey (65%), Spain (40%) and Malaysia (60%)). By contrast, only a minority of users in Australia (27%) and the UK (12%) seem to use WhatsApp in this manner. Because of that, while many people use WhatsApp for direct, safe, and personal conversation, another large portion of users use the service for mass communication, connecting to new people and sharing common topics of interest through its network.

With such popularity and big audience, WhatsApp gained a lot of attention in Brazil, being used as a political weapon (RICARD; MEDEIROS, 2020) and spreading false news. Actually, in Brazil, around 18% of the users are part of political and news groups (KALOGEROPOULOS, 2019). WhatsApp has been pointed as playing a essential role in 2018 Brazilian Elections and has emerged as an ideal tool for mobilizing political support and spreading fake news (NEMER, 2018). During Brazilian elections, then, WhatsApp was actively abused to send out misinformation campaigns, with wide use of manipulated images and memes containing all kinds of political attacks. A recent study showed that 88% of the most popular images shared in the last month before the Brazilian elections were fake or misleading (TARDAGUILA; BENEVENUTO; ORTELLADO, 2018). In India, specialists also raised worries about WhatsApp misinformation (GARIMELLA; ECKLES, 2020; BADRINATHAN, 2020). Fake rumours spread through WhatsApp were directly responsible for multiple cases of lynching and social unrest (ARUN, 2019) and associated vigilante violence(BANAJI et al., 2019), leading to requests by India government to WhatsApp to cut down potentially harmful content going viral(RAJ, 2021; BENGALI, 2019). Those examples not only show the danger of misinformation, but how it effectively disseminates through the network of WhatsApp.

During COVID-19 pandemic, in which people are more isolated and highly dependent on technologies to communicate to each other. In this scenario, WhatsApp is also an important resource for the users to keep in contact with family and friends, to meet with colleagues in work, companies also uses WhatsApp Business to reach theirs customers, and, unsurprisingly, to inform themselves, even about health measures. In an extended survey, authors from NEWMAN et al., 2020 found almost a quarter (24%) of users used WhatsApp to find, discuss, or share news about COVID-19 and around 18% joined a support or discussion group with people they did not know on either Facebook or WhatsApp to talk about COVID-19(NEWMAN, 2020). In this context, misinformation is very concerning and requires even more evidence-based infodemic interventions by health and specialized agencies, with greater engagement in pandemic misinformation management efforts (VIJAYKUMAR et al., 2021). Particularly in Brazil, misinformation was extremely abused, undermining strategies to contain the COVID-19 pandemic and bringing risks that imperil the whole population (SOARES et al., 2021b; RICARD; MEDEIROS, 2020).

Recent studies have shown the harmful potential of fake news in society (REIS et al., 2019, 2019; RIBEIRO et al., 2017). Still, little was explored towards the understanding and, principally, tackling misinformation on WhatsApp. As mentioned above, WhatsApp has some peculiar characteristics that transform it into a unique problem. The easy access, the public group chat format, and the end-to-end encryption change way (mis)information are shared there and requires specific measures to face this problem.

Moreover, WhatsApp shows a very ephemeral structure. Unlike Facebook or Twitter, where content remains stored for years, available for simple reference to whoever has access to it, WhatsApp has a much more temporary operation, even including an option to send messages that self-destruct after a short period of time[5]. There, conversations get lost after groups are broken up or if users wipe their cell phones. Even new users in a group chat do not have access to past content, they are able to see only what was sent after joining that group. Therefore, a study that gather a large-scale data from WhatsApp, especially during a period when the platform proved itself to be so important to explain the events of our country, also serves as a historical record of what actually happened in this environment parallel to our real life. Gathering and analyzing this kind of data is an attempt to avoid that such valuable information get vanished forever in times of the digital era.

Thus, the motivation in this thesis lies in exploring this new ecosystem of IMPs, particularly the WhatsApp, investigating in depth their characteristics and functioning to provide society with a better comprehension of these platforms, what happens in this environment, and useful mechanisms for combating the hazardous content of misinformation that circulates in there.

1.2 Thesis Statement

In this thesis, the main aim is to investigate WhatsApp environment and the (mis)information that spreads over the platform. This task presents a unique challenge, as the content circulating on WhatsApp network is encapsulated in group chats and conversations protected by an end-to-end encryption layer. Yet, this content still manages to spread virally to the point of reaching a large part of the population and causing serious effects on society. Even though WhatsApp is a service that claims for safety, privacy, and encryption, it also provides tools for messages to disseminate in a bulk, reaching multiple people, and becoming very popular among the users and acquiring also a public aspect. In other words, the platform is designed for personal and private communication, but at the same time, WhatsApp also provides tools that allow the content to go viral through its network.

Here, we will explore this public aspect of WhatsApp, using the public chat groups that can be accessed by everyone. This thesis will supply a deep understanding about the information messaging ecosystem created through WhatsApp, providing tools to collect and analyze data from this social media, unveiling how misinformation campaigns abuse

[5] <https://blog.whatsapp.com/introducing-disappearing-messages-on-whatsapp>

from this cyberspace and proposing countermeasures to tackle this problem within such platforms.

1.3 Research Goals

The hypothesis explored by this thesis is that the data extracted from Instant Messaging Platforms such as WhatsApp may be valuable to understand the online environment and how it is related to real-world events. Furthermore, that there are features and characteristics in those IMPs that facilitate misinformation to disseminate through their networks, but there are proper means and actions that could be taken to combat this problem, creating a better and healthier space to people communicate online.

Starting from these points, the general objective of this thesis can be divided into the following specific goals (RGs):

RG1 – Collect and build a large-scale and long-term dataset of WhatsApp that can be explored to understand real-world events.

Due to the closed nature of WhatsApp, the task of retrieving data from this social media is an open challenge itself. First, the data itself on WhatsApp is very ephemeral, that is, it has a temporary and instantaneous character. It is not possible, for instance, to have access to the retroactive data of a given group on WhatsApp. A member only has access to data relative to the period of time in which they were part of that chat. In addition, conversations are stored on users' devices for a short period of time. Old media is difficult to recover and WhatsApp itself offers a functionality that auto-destructs a message after seven days, becoming impossible to recover it if it was not stored otherwise. The public sphere of WhatsApp groups is also constantly evolving. New groups are created, and old groups are deactivated, creating a flow of change in which users migrate without remaining static in this digital landscape. Summing this with the fact that most chats are private, and does not exist any official API for gathering this kind of data, there is an undeniable importance evaluated to the existing efforts in the direction of building a WhatsApp dataset and there is still much space to be explored in the development of a framework to collect WhatsApp data. For example, how to find the groups, how to select groups to collect, how to access, to store, to organize and to display data. All those steps are not well-defined in the literature. Thus, the results obtained in the thesis concerning each of them have a lot to aggregate for many researches. In top of that, a cautious attention should be paid to ethical research precepts, following strict rules to respect and conceal sensitive information from other users.

In addition, since the difficulties of the platform, the field lacks sets of data to support better understanding of the events taking place within WhatsApp groups. By developing a methodology of collection in large-scale and long-term of WhatsApp public groups, this study provides valuable information related to some events (i.e., the 2018 Brazilian presidential elections and COVID-19 pandemic) that played a big role in the recent years of our society and how they unfolded through the lenses of WhatsApp.

RG2 – Understanding how the structure of the WhatsApp impacts on the dissemination of information within its network. Once one is able to access and analyze the data from WhatsApp, it is natural that the next step is understanding the characteristics of the platform. On WhatsApp, probably one of its points of greatest interest is how fake news and misinformation spread through the network and how these messages manage to reach so many people in a theoretically private and encrypted space. A tool that definitely helps this process along with group-based communication is the ability to forward messages to friends or groups. To try to get around this problem, Facebook – the WhatsApp owner – implemented limitations on the number of contacts a person can *Forward* a message, reducing the maximum to 5 other contacts[6] and allowing that very shared messages to be forwarded only to a single contact[7]. Notwithstanding, there is no clue on how this or any other WhatsApp feature can impact or help to prevent the dissemination of misinformation. In this direction, this thesis aims to investigate the viral aspect of WhatsApp, showing that it has potential to spread messages to the entire network and evaluating how its structure and tools affect in the dissemination. Furthermore, we want to understand the contrast in which a private, personal, and encrypted application permits such different goals of public, massive, and viral uses of WhatsApp.

RG3 – Finding measures that can be taken to combat misinformation circulating in WhatsApp. After collecting data and observing how fake news spread there, one may ask what are the means to combat misinformation in this environment, or even if there is any due to all complications investigated until this point. Therefore, the third goal of this work will focus on the fight against fake news and misinformation, with the objective of proposing a realistic methodology that can be adopted by instant messaging services to prevent the dissemination of unwanted content. Furthermore, it assesses other measures taken in the direction to aid the hard task of fact-checking, verifying that if it is possible to offer techniques that seek to diminish the consequences of misinformation shared on Instant Messaging Platforms.

[6] <https://blog.whatsapp.com/more-changes-to-forwarding>

[7] <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>

1.4 Chapters Organization

This thesis is organized as follows: Chapter 2 presents the background of related work, Chapter 3 enters the instant messaging platforms universe, explaining what they are and their mains differences between each others, in Chapter 4 we go further on WhatsApp details and how it works. Chapter 5 presents a detailed explanation of our collection methodology and the overview of data collected. Chapter 6 makes a comparative analysis of IMPs, understanding the ecosystem of public group chats of those platforms shared on the Web. Chapter 7 performs an extensive analysis on information dissemination of WhatsApp content, in Chapter 8 we propose two distinct and viable countermeasures for tackling misinformation on WhatsApp, and, finally, Chapter 9 show the final discussion, with the conclusions achieved and plans for next steps of this work.

Chapter 2

Background and Related Work

In this chapter, we present a summary of background information and related work that is fundamental to the understanding of this thesis. We discuss previous efforts related to social networks, more particularly, about messaging platforms, especially about WhatsApp and also regarding misinformation present on these digital platforms.

There is a rich body of previous work focusing on analyzing online social networks, other emerging social networks, and Web communities. Specifically, these works focus on exploring various aspects of the mainstream social networks such as Twitter (KWAK et al., 2010; CHA et al., 2010; MISLOVE et al., 2011), YouTube (CHA et al., 2007; FIGUEIREDO; BENEVENUTO; ALMEIDA, 2011), Reddit (GILBERT, 2013; SINGER et al., 2014; BAUMGARTNER et al., 2020a), Flickr (CHA et al., 2010; MISLOVE et al., 2008), and Facebook (VISWANATH et al., 2009; LIU et al., 2011; SPEICHER et al., 2018; RIBEIRO et al., 2019; RIBEIRO; BENEVENUTO; ZAGHENI, 2020). Those studies vary from measuring digital platforms to investigating their users and content posted there. They give us a wide comprehension on how those platforms operates, deep insights related to the networks structure, dissemination of information within the platform, and how many important events take place in the digital world of social networks. This help us to understand better this environment while also shed light in some issues and challenges that this brings to us such as hate speech (SILVA et al., 2016; DAVIDSON et al., 2017; MONDAL; SILVA; BENEVENUTO, 2017) and misinformation (LORENZ-SPREEN et al., 2020; SHU et al., 2017; ALLCOTT; GENTZKOW, 2017; SPEICHER et al., 2018) present in many of these social medias.

Furthermore, many works have also been focusing on analyzing and measuring emerging fringe social networks such as 4chan (HINE et al., 2016; BERNSTEIN et al., 2011), an anonymous image board in a forum structure. Many of these alternative platforms to conventional networks claim to be a safer space for discussion and freedom of speech. Some of them appear especially as a response to users who have been suspended on other social networks for violating their terms of service, such as Gab (ZANNETTOU et al., 2018a; LIMA et al., 2018) and Parler (ALIAPOULIOS et al., 2021), alt-right microblogs very similar to Twitter that have many users banned from there and migrate to these alternative spaces, in which a large portion of posts feature far-right, anti-Semitism, and conspiracy theories such as the QAnon. There is also Mastodon (RAMAN et al., 2019), another decentralized microblog, BitChute (TRUJILLO et al., 2020) – a video site with alt-right content usually banned from YouTube, and Voat (RIBEIRO et al., 2021). Understanding these platforms and the exchanges between them are the focus of many efforts. While mainstream social medias start to moderate content and ban their users for abuse of hate speech and fake news, those alternative platform emerge as parallel communities where users banned from traditional ones join and replicate the toxic content in a space generally free of any moderation. JOHNSON et al., 2019 analyzed this global ecology of online hate and its adaptive dynamics. They show that these communities have developed strongly interconnected and resilient hate groups that are able to jump between platforms and migrate their content when a single platform (e.g., Facebook) applies content moderation policies, including pointing out Telegram and WhatsApp as their destination. groups. Likewise, RIBEIRO et al., 2021 showed that when online communities face restrictive moderation measures, their users can migrate to other platforms with more flexible policies. Their results suggest that users post more often on such alternative platforms and tend to become even more toxic on them. Moreover, motivated by the overwhelmingly large number of social networks available, previous works also focus on analyzing multiple social networks and measuring the interplay between them (MISLOVE et al., 2007; ZANNETTOU et al., 2017, 2018b; CHANDRASEKHARAN et al., 2017).

Numerous research studies explored the particular issue of misinformation that emerged in the ecosystem of the OSNs (BESSI; FERRARA, 2016; LAZER et al., 2018). Specially in the political context, many have investigated how social networks are exploited in the organization of political movements, such as the role of social networks during the Arab Spring in the Middle East (HOWARD; HUSSAIN, 2011) and the different narratives built on Twitter about the White Helmets in Syria (STARBIRD et al., 2018). More recently, some researches focus on investigating attempts to manipulate political discourse with the use of social bots and even state-sponsored trolls (ALLCOTT; GENTZKOW, 2017; FERRARA, 2017; ZANNETTOU et al., 2019a). RIBEIRO et al., 2019 evaluated the use of the Facebook advertising platform to carry out political campaigns that exploit targeted marketing as a means of disseminating false advertisements. There are also a number of recent efforts that investigate the role of political campaigns through social networks, primarily focused on the 2016 US presidential elections, exploring the creation of *bots* on Twitter and its role in engaging users in political discourse (Freitas et al., 2015; BESSI; FERRARA, 2016). Other works evaluated the use of the Facebook advertising platform to carry out political campaigns that exploit targeted marketing as a means of disseminating false advertisements or on divisible themes, inciting polarization and hatred (MONDAL; SILVA; BENEVENUTO, 2017) in social networks.

In addition, there are efforts that aim at exploring strategies for automatic detection of fake news (REIS et al., 2019; SHU et al., 2017). In this direction, there is a vast group of efforts that propose solutions to mitigate the fake news problem based on various artificial intelligence techniques such as classic supervised approaches (CONROY; RUBIN; CHEN, 2015), active learning (BHATTACHARJEE; TALUKDER; BALANTRAPU, 2017) and deep learning (WANG et al., 2018).

However, there is a new universe of instant messaging platforms that has been emerging and gaining a substantial popularity among users around the world, but which has still been little explored by researchers. With the emergence of such platforms as some of the most popular social media that people use to communicate to each other, a new environment with a new range of new peculiarities gained prominence. Unfortunately, they have been identified as playing a central role in disseminating misinformation and fake news campaigns in some events around the globe. Next, this scenario of IMPs will be further discussed, with their specific characteristics and previous work that have studied how they are being used in recent years.

2.1 Instant Messaging Platforms

Many of the previous works related to this area focus on analyzing and measuring different messaging platforms in different contexts by looking at the networks composed by the public group chats present in each of them.

More particularly, in Telegram, previous work focuses on collecting data from Telegram and studying emerging research problems. BAUMGARTNER et al., 2020b, for instance, collected and made it publicly available a large-scale dataset of 27K Telegram groups and 317M messages. ANGLANO; CANONICO; GUAZZONE, 2017 and SATRYA; DAELY; NUGROHO, 2016 investigated the artifacts generated by the Telegram application, while ABU-SALMA et al., 2017 performed a user study to understand user perceptions related to Telegram's security. NASERI; ZAMANI, 2019 focused on the spread of news on Telegram by collecting data from five official Telegram channels (i.e., Telegram channels that are used by news outlets). NOBARI; RESHADATMAND; NESHATI, 2017 collected data from 2.6K Telegram groups and channels and performed a structural analysis of the content posted within those groups/channels. AKBARI; GABDULHAKOV, 2019 investigated the ban of the Telegram platform by Russia and Iran after Telegram refused to provide access to encrypted data posted among users of the platform.

A large body of work in this scenario examines the use of Telegram in the Mean East. NIKKAH; MILLER; YOUNG, 2018 studied the use of Telegram by Iranian immigrants with a focus on understanding how Telegram groups are moderated. HASHEMI; CHAHOOKI, 2019 conducted a large-scale analysis on 900K Iranian channels and 300K Iranian groups, aiming to distinguish groups into the ones that are high-quality (e.g., business-related) and low quality (e.g., dating groups). ASNAFI et al., 2017 analyzed the use of the Telegram platform in Iranian libraries. Finally, previous work focuses on studying how Telegram is exploited by terrorist organizations like ISIS (PRUCHA, 2016; YAYLA; SPECKHARD, 2017; SHEHABAT; MITEW; ALZOUBI, 2017). Such organizations exploit the Telegram platform for their communication purposes to spread propaganda, and possibly recruit new members.

There are also some previous research on Discord that, though it started with a focus on providing a messaging platform for the online gaming community, is nowadays a very popular IMP, used by the general public for various purposes. Discord is a fast-growing messaging platform that especially attracts the young population. By analyzing this platform, some studies could shed light on the use or abuse of messaging platforms by the young demographic. Discord has been used for organizing extremist rallies, e.g., the “Unite the Right” rally in Charlottesville in 2017 (ROOSE, 2017), and for disseminating potentially harmful and sensitive material, e.g., revenge porn (COX, 2018). HAMRICK et al., 2018 studied pump and dump schemes on the market of cryptocurrencies by analyzing data obtained from Discord. LACHER; BIEHL, 2018 examined the use of Discord for teaching purposes. They provide details on how Discord can be used by instructors as a class collaboration tool, and they discuss issues related to moderation and accountability. JIANG et al., 2019 studied the moderation challenges that exist on Discord, and in particular on voice-based channels. Similarly, KIENE; HILL, 2020 focused on the moderation of Discord platform and in the use of bots for moderating content posted on Discord servers.

WeChat is a Chinese social media platform developed by Tencent. First released in 2011, it became the world’s largest standalone mobile app in 2018 (MILLWARD, 2014) with over 1 billion monthly active users (JAO, 2018). WeChat provides instant messaging functionalities such as text messaging, voice and video calls, chat groups, but also works as a social networks in which users have their profile, friends, and post content on this timeline. Because of its wide range of functions, this described as China’s “app for everything” and a “super app”. Actually, this platforms is well known for the mass surveillance of huge volume of users data as Chinese government tracks the activity from users within the platforms (MCDONELL, 2019; COCKERELL, 2019), monitoring even group chats (DOU, 2017).

Another important IMPs in Asia is LINE app. Released in 2011, it is still the most used app in Japan, Taiwan, and Thailand (MILLWARD, 2017). With 220 million monthly active users, its popularization is principally due to it being the first of instant messaging to introduce stickers back in 2013 (RUSSELL, 2013). Stickers are larger-scale emoticon images used in text chats. They are popular among some, specially in Asia because they help convey emotion, and are more visual than blocks of text, which WANG, 2016 shows that it is an important feature of text chats in LINE as it provides more intimacy in the mobile communication environment.

OHASHI; KATO; HJORTH, 2017 explored LINE usage to maintain familial ties in Japan, where they found users in both individual and group chats keep very personal and affective relationships through the app, but they observed that LINE is also an integral part of the social media landscape in which they utilize social media applications to gather and share information about changes in social and political climate. Their study indicates that mobile media started to take on new types of both personal–private and political–public textures in the lives of their users. The social media influence performed by LINE was studied even for smoking habits. KULSOLKOOKIET et al., 2018 conducted an experiment with multimedia anti-smoking campaigns (text, pictures, and video) circulating on LINE instant messaging application to investigate how it influences in changes of smoking behavior participants, showing that this content on LINE favored people to cut smoking practices. As other more popular IMP, security and privacy through encryption became a standard concern in LINE, ESPINOZA et al., 2017 provided an analysis of E2EE features in LINE, identifying several vulnerabilities and challenges in app design, discussing, then, research directions to better bridge vendors, researchers, and end-users around security issues.

Moreover, LINE and WeChat are the focus for many scholars to study post-truth politics and fake news in Pacific-Asia (YEE, 2017). CHANDRA et al., 2017 concerned about group feature that can connect us with numerous other groups in spreading of fake news. They investigated higher education students participating of many groups to show college students' activities and the intention of awareness to receive fake news in social media. FUNKE, 2018 analyzed how misinformation spreads on LINE as it is becoming the major app for newsgathering, especially for the younger generation in Japan and Southeast Asia. He shows the growth of many groups that disseminate fake stories on LINE. Some malicious groups attract users and when they reach enough followers, they change the name and profile picture and start to send misinformation and often selling products. LINE misinformation is of particular concern in Taiwan. The country is facing a flood of misleading or factually incorrect news items that seek to influence the public on its most popular IMP (WYTZE, 2017).

In China, social media environments also become the dominant source of information and news and, consequently, suffer from challenges related to misinformation. There, WeChat is the mains actor in this scenario, being target of criticism for spreading fake news (LI, 2017). Some studies, for instance, have found that WeChat plays a central role in the distribution of misinformation among the Chinese immigrant community in America, including fake news, to politically motivate and organize them for conservative causes, especially against affirmative action and ethnic data disaggregation (ZHANG, 2018; CHEN; LIANG; CAI, 2018).

LU et al., 2020 studied and trustworthiness of online information perceived by WeChat users. They interviewed 44 Chinese WeChat users to understand how individuals perceive misinformation and how it impacts their news consumption practices. They work claim that the complex aspects of Chinese regulations over media impacts on varied opinions expressed by users about the credibility of online information sources. Furthermore, although most participants said that their opinions are not influenced, many admitted that they find difficulties to distinguish credibility of the content received, including sponsored efforts by the government or companies. GUO; ZHANG, 2020 examined the diffusion of day-to-day online rumors on WeChat and other platforms in China. They observed that, despite some of the strictest media censorship rules in the world, online rumors continue to permeate China's Internet. The government-controlled news websites are supposed to fight against online rumors, but these websites have also been found to carry unverified information They found a large portion of misinformation on WeChat and in many different topics ranging from food and personal safety, health, fraud, and highly amount of political motivated rumors.

As a result, China have implemented numerous anti-rumor regulations. In 2016, they criminalized manufacturing or spreading online rumors that undermine economic and social order. In 2017, a law called Provisions for the Administration of Internet News Information Services demanded that internet news providers reprint information published by government-acknowledged news organizations without distorting or falsifying news information, another government regulation requires microblogging service providers to establish an anti-rumor mechanism, which includes publicizing and refuting rumors when they arise. That is, social media platforms are prohibited from publishing their own independent articles or disseminating information without attribution, while Chinese authorities have also directly investigated social media companies for facilitating rumor spreading (REPNIKOVA, 2018).

With this, we can observe that even with severe rules and government regulation, misinformation continues to be an issue in instant messaging platforms around the world. We need to give special attention to this problem and find ways to bypass these challenges.

2.2 WhatsApp

Even when compared to all other mainstream OSN on the Web, WhatsApp is still one of the most used social media in the world, with 2 billion users, only behind

Facebook itself in numbers of users (NEWMAN et al., 2019). It is the most popular IMP (TANKOVSKA, 2021), it is extremely popular and the main online communication tool for people in many places. In some countries, the application's penetration is even more impressive: the WhatsApp application is used by more than half of the Brazilian population, with about 140 million users in Brazil (CECI, 2022). In India, around 390 million users are active in the app per month (JAIN, 2021).

Notwithstanding, relatively little attention has been given by researches to WhatsApp and others IMPs compared to OSNs. Facebook company, up to this date, does not provide any API to formal access to WhatsApp data[1], and keep all data proprietary. This, summed with the closed nature of the system, with private chats and E2EE encryption, make it harder for one investigating what is going on within the messages of the platform.

Because of those limitations, the majority of earlier studies in this field have taken a qualitative approach using interviews, diary studies, surveys, and focus groups (CHURCH; OLIVEIRA, 2013). Those works are extremely useful to provide a better understanding of the social aspects of WhatsApp usage. They help in knowing who are the WhatsApp users, the reasons behind they use the application, the relationship between the user with the app, and the implications of the communication with users' contacts through the use of this technological tool. However, these earlier works may lack in deep comprehension of the big picture of the entire WhatsApp ecosystem, the structure, and flow of information behind the complex network of friends created on WhatsApp platform, or the substantial volume of content that circulate in this messaging app every day.

Albeit WhatsApp has been released back in 2009, only recently researchers have taken more quantitative and systematic methodologies, investigating the WhatsApp network in a larger scale. With the development of tools that make WhatsApp data collection possible, especially those that involve joining public WhatsApp groups massively (GARIMELLA; TYSON, 2018). Research on WhatsApp groups, then, has shifted from working only ***with*** users, through surveys and interviews to accessing their personal experiences, to working ***as*** users themselves, by joining hundreds to thousands of public WhatsApp groups and be also part of the instant messaging ecosystem (CHANG, 2020).

Next, we will go deeper into the works that analyzed WhatsApp, evaluating the current state of the art in the area and the main remaining issues to be addressed by researches.

CHURCH; OLIVEIRA, 2013 start by exploring how and why people have adopted and appropriated IMPs applications, specifically WhatsApp, in their daily lives. They and compared how messaging practices differ between using WhatsApp and traditional SMS, discovering that users believed WhatsApp offers benefits such as cost, sense of community and more immediacy compared to SMS. O'HARA et al., 2014 focused on the

[1] WhatsApp Business has some tools that support third-party companies to manage their accounts and customers, but none are available to researches purposes.

habits of WhatsApp users. With their interviews they revealed an intimacy of relationships on WhatsApp and that users favor WhatsApp to maintain intimate relationships with strong ties not only for one-to-one relationships with friends but also in group chats, as they enable a sense of belonging, identity and experience of the collective. BLABST; DIEFENBACH, 2017 also explored how people use WhatsApp to communicate in their daily lives. They investigate how specific WhatsApp features (i.e., single and group chats, last seen and read receipts) are perceived by the users as communication quality and well-being. While they discovered that the high number of single chats on WhatsApp is positively correlated with perceived communication profundity but also with perceived stress and waste of time, the high number of group chats might support a less interactive usage of just scrolling and possibly not even reading, which is also pointed as a possible waste of time. SEUFERT et al., 2015 focus on how group-based communication occurs within WhatsApp by conducting interviews and a survey. Later, SEUFERT et al., 2016 investigated the group-based communication in WhatsApp based on the analysis of messaging logs. They evaluated a series of characteristics of WhatsApp group chats in terms of usage and topics. Finally, they presented a classification model based on the topic of the group, and they classify messages based on some statistics.

In a more systematically approach, ROSENFELD et al., 2016 perform a survey with a hundred of participants to anonymously collect data from WhatsApp users. To do so, they developed a software integrated Android (and consequently with WhatsApp) that enabled taking snapshots of a person's groups and messages as they appear in their phone, but that also encrypts and anonymize the data that was pulled directly from the participant's smartphone. They characterize the behavior of users inferring demographic information like age and gender, and with a dataset of 4M messages over an average period of approximately 15 months, as they do not have messages' content or recipient information, analyzed primary characteristics of the volume of exchanged messages, and other activities in chats and groups that the users are participant. They found, for example, that three thirds of user conversations are in 2 member chats (i.e., direct one-to-one messages). MONTAG et al., 2015 took a similar approach, asking 2418 users to download an app that records activities of WhatsApp and other apps to investigate general smartphone usage, and the use of WhatsApp in particular. Their results show that use of WhatsApp accounted for around 32 min per day, compared to 15 min for Facebook. Both studies outline a WhatsApp user profile, that the app is preferable among the younger population, that women spend more time on the app, and that WhatsApp acts as a key element in daily communication for most smartphone users

There are works mainly focused on acquiring data from public WhatsApp groups and analyzing their content. Specifically, GARIMELLA; TYSON, 2018 develop a different methodology that enable the large-scale collection of WhatsApp data from public groups. Their tools rely on discovering WhatsApp public groups using Google's search engine and other websites advertising public groups. Then, they join those groups and with a phone (that has root access), being members of all groups and, finally, extract WhatsApp database direct from the smartphone and decrypt it using WhatsApp private key to access all messages exchanged in those groups joined. By doing this, they managed to collect 2.5K public WhatsApp group URLs and managed to join 200 of them, mostly related to Indian topics of interest. They provided a wide analysis of the behavior of WhatsApp users in India. Their approach brought a new paradigm of exploring public groups on WhatsApp, allowing researchers to access data on a larger scale than before with other more qualitative approaches, while they become actually members of those groups studied.

Following, MACHADO et al., 2019 also use online repositories of group links for WhatsApp to find groups for political WhatsApp groups dedicated to the discussion of politics, news and current affairs with publicly available invite links during the 2018 Brazilian elections. They join 130 public WhatsApp groups and evaluate the misinformation of news sources shared within these groups. CAETANO et al., 2019 also use Google search to find invite links shared in blogs, web pages, and other online platforms to study the cascades of messages on WhatsApp. They investigate specific characteristics associated with more than 1.7M messages posted in 120 groups that discuss political subjects in Brazil and false information. Finally, BURSZTYN; BIRNBAUM, 2019 combine the searches on Google, other platforms (e.g., Facebook and Twitter), and group URLs shared within joined groups on WhatsApp, to find and join 232 partisan WhatsApp groups for both right-wingers and left-wingers in the 2018 Brazilian presidential elections. Then, they analyze the data with the goal to compare the behavior of right-wingers and left-wingers across several axes. MAROS et al., 2020 explored the public groups of WhatsApp in Brazil, but they examine the use of audio messages, a type of content that is becoming increasingly important in the platform. They extracted text from audio messages shared in WhatsApp public groups, characterizing the content properties of topics and language and the propagation dynamics in the network.

JAVED et al., 2020 also used this methodology to collected data from 227 public groups in Pakistan for investigating content related to COVID-19. They targeted the popular political parties of Pakistan and filtered messages using a set of keywords related to COVID-19 pandemic and manual labeling, comprising a set of 5K messages. They found relevant trends of pandemic in Pakistan as well misinformation present within the data.

CHANG, 2020, in his study, performed an extensive investigation on public WhatsApp groups in the Venezuelan refugee crisis. While in an initial approaching it analyses how users use WhatsApp through field interviews, inferring some demographics about use of WhatsApp among the emigrants' population, it also proceeds to use WhatsApp as data source to track migrant dynamics within the public groups and also as well to find fake news and scams circulating in the WhatsApp groups.

2.2.1 Misinformation on WhatsApp

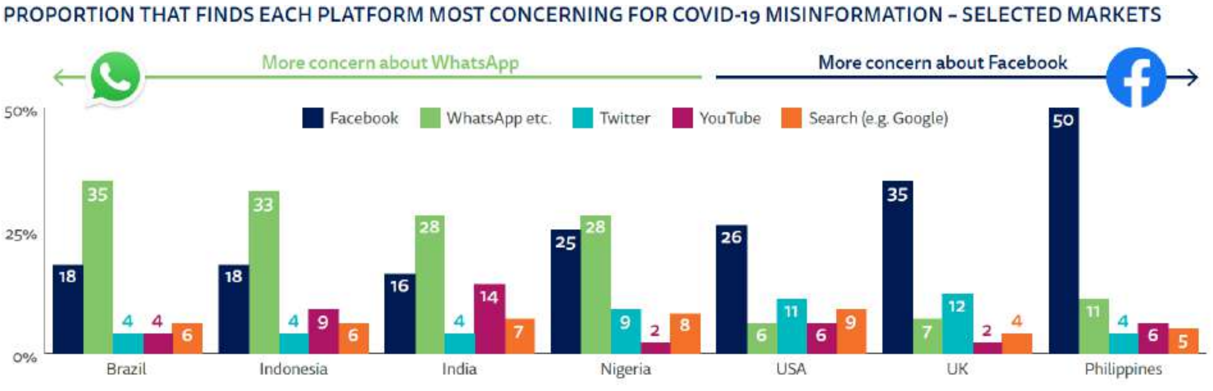
The ease of use of WhatsApp can bring many advantages for the society. LAMBTON-HOWARD et al., 2019, in collaboration with the International Federation of Red Cross, designed WhatFutures, a collaborative future forecasting engagement for global youth using WhatsApp. They present a reflection upon the design decisions of their system, and identify how decisions made around group structures of WhatsApp processes an externalization of outputs that influence engagement in the platform. In OFUSORI; KARIUKI, 2017, the authors present experiences from this electoral team in a province in South Africa that utilized WhatsApp as a secure tool for relaying information amongst themselves during the electoral period. The WhatsApp helps in the immediacy of responses of incidences related to electoral violence, and it is an effective tool for electoral monitoring and management in that situation. Moreover, MALKA; ARIEL; AVIDAR, 2015 focused on how citizens of Israel use WhatsApp during wartime, in which they noted that the platform plays a central role in the lives of its users in conflict scenarios, working as the source of news for the community and also enabling a secure channel of communication between people near the battlefield and their families.

WhatsApp, however, has gained more relevance in recent years due to the spread of misinformation in the app. Not only text messages are abused to share false news on WhatsApp chats, but also multimedia content can have misinformation, such as audios (MAROS et al., 2020) and images (GARIMELLA; ECKLES, 2020). In many places around the world, fake news, rumors and misinformation campaigns took place within the WhatsApp network becoming a problem not only in countries such as Brazil (MACHADO et al., 2019; CESARINO, 2020; RICARD; MEDEIROS, 2020), but also in India (ARUN, 2019; GARIMELLA et al., 2018; FAROOQ, 2017), in Indonesia (KWANDA; LIN, 2020), in Venezuela (CHANG, 2020), in Pakistan (JAVED et al., 2020), U.K. (VIJAYKUMAR et al., 2021), Ghana (MORENO; GARRISON; BHAT, 2017), Niegeria (CHEESEMAN et al., 2020; HITCHEN et al., 2019), Spain (ELíAS; CATALAN-MATAMOROS, 2020), just to mention some of many the of places where issues regarding misinformation on WhatsApp were reported.

In particular, WhatsApp is pointed out as a media pivot for the dissemination of certain misinformation events, such as the spreading of rumors in India that caused a series of violent lynchings (ARUN, 2019), strong indications of interference in Brazilian elections in 2018 (BURSZTYN; BIRNBAUM, 2019; RECUERO; SOARES; VINHAS, 2021; ABDIN, 2019), and an avalanche of fake news about the COVID-19 pandemic (SOARES et al., 2021a, 2021b; VIJAYKUMAR et al., 2021; RICARD; MEDEIROS, 2020). These problems brought quite attention to WhatsApp and placed it in the sight of the academic community, press (TARDAGUILA; BENEVENUTO; ORTELLADO, 2018), and even some governments that pressured Facebook to take actions against misinformation (GUPTA; TANEJA, 2018; RAJ, 2021) about the implementation of end-to-end encryption in their system (PATEL et al., 2019; MAYER, 2019). As a result, WhatsApp itself have modified some features of the system in the meantime, such as labeling popular messages as "*Forward*" and "Frequently Forwarded" for viral content[2]; lowered forwarding limits from 256 at a time, to 20 and then up to 5 contacts/groups, while viral content can be sent just to one contact at time[3]; and added a button that allows an option to search the internet for some shared content within the platform.

These questions regarding misinformation on WhatsApp is a major question in society that is also perceived by the users. Figure 2.1 shows results from the Reuters survey (NEWMAN et al., 2021) on the proportion of people that finds WhatsApp and other platforms most concerning for COVID-19 misinformation for different countries. When asked about which online social platforms they worry most about false or misleading information about COVID-19, Brazilians, Indians, and Indonesian users mostly pointed out WhatsApp as the platform where this issue is more concerning, even higher than Facebook, Twitter, and YouTube. These countries are the main audience of WhatsApp around the globe, and that had already reported problems with fake news and WhatsApp (TARDAGUILA; BENEVENUTO; ORTELLADO, 2018). But even in countries where WhatsApp is not so popular, such as the USA and UK, WhatsApp appears among the top main platforms which spreads misinformation about the pandemic. Because of that, understanding how misinformation occurred on WhatsApp has become a relevant issue that needs to be explored, however, this is a particularly hard task due to the structure of how WhatsApp chats work.

Figure 2.1: Proportion of users that find WhatsApp and other platforms most concerning for COVID-19 misinformation for different countries from Reuters.



Q_FAKE_NEWS_2021c. Thinking specifically about coronavirus (COVID-19) and its effects, which of the following, if any, are you most concerned about online? False or misleading information via ...
Base: Total sample in each market (n = 2000).

Source: Reuters Institute Digital News Report 2021 (NEWMAN et al., 2021)..

The survey conducted by BLANCO-HERRERO; AMORES; SáNCHEZ-HOLGADO,

[2] <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>
[3] <https://blog.whatsapp.com/more-changes-to-forwarding/?lang=en>

2021, to explore this aspect, analyzed misinformation from the perspective of consumers, who play a relevant role in the spread of this content. They show that users have a special high level of concern regarding fake news shared on social networks. They also observed that the perception and credibility of a piece of information consumed online is affected by age and gender of the user. Interestingly, HERRERO-DIZ; CONDE-JIMéNEZ; CóZAR, 2020 examined the motivation of young people to spread fake news on WhatsApp. They found how the young teenagers' exposure to fake news interferes with the ability to distinguish trust from information, contributing to the problem of the spread of misinformation. They are more likely to share content if it connects with their interests, regardless of its truthfulness, that trust affects the credibility of information, and that the appearance of newsworthy information ensures that, regardless of the nature of the content, this information is more likely to be shared among young people. GIMENEZ; ZIRPOLI, 2015 argues that, because of the anonymity provided by the platform and the loss of face-to-face conversation, communication occurring within WhatsApp can be abused to harass others or generate misunderstandings.

ARUN, 2019 studied the dissemination of rumors in India. Even though they say hoax and fake rumors have already spread in the country before WhatsApp, they argue that WhatsApp amplified the problem by spreading misinformation on a larger scale and much faster.

One of the contexts most exploited by scholars to study misinformation in WhatsApp is the public political debate that occurs within public groups. Following this lead, CHAGAS, 2021 aims to investigate the rhetorical elements present in memes and the misinformation that circulated in far-right political groups on WhatsApp after the Brazilian elections. They conclude that WhatsApp political groups formed a network of very high capillarity, which became one of the main forms of communication for some users. These groups, however, were notable for disseminating messages of a radical nature and strong ideology, in addition to a large volume of misinformation.

BURSZTYN; BIRNBAUM, 2019 study partisan groups on WhatsApp evaluating the use of social platforms to push political narratives during elections. Their results indicate that the electoral process can be a strong driver for the use of multimedia messages in partisan groups, especially among right-wing users on WhatsApp. Actually, several candidates in the 2018 Brazilian presidential race used mass messaging services (often financed by private companies) for their campaigns to create a massive and orchestrated dissemination of fake news on WhatsApp (RICARD; MEDEIROS, 2020). MACHADO et al., 2019 analyze a large set of WhatsApp public political groups to discuss the possible impacts of misinformation and political propaganda within WhatsApp on polarizing and impeding public debate, and fomenting acts of violence during the 2018 Brazilian elections. They conclude visuals are being heavily used within groups to distort information and manipulate users, in addition to spreading hate speech.

EVANGELISTA; BRUNO, 2019 investigate the hypotheses that the results of the Brazilian elections are related to successful campaigns built with specific communication strategies on social media platforms. They found evidence of centralized management of WhatsApp chat groups by political actors, who brought much more conservative and radicalizing elements into the political debate on the platform. More particularly, regarding the architecture and opacity of the system, they point out that the actors who produce the contents are easily made invisible. This leads to users not noticing or becoming aware they may be monitored and managed in the system. This and the impression that most WhatsApp contacts are closer to the user's personal circle produce a feeling of privacy. Those were essential factors that contributed to the success of WhatsApp for political campaigns, since it conveys an idea of proximity and safety.

CESARINO, 2020 also sought to understand the relevance of content shared on WhatsApp in the 2018 Brazilian election through a digital ethnography in large WhatsApp public groups. Her observations found a specific aesthetic of WhatsApp messages. This content of false news, conspiracy theories, offensive and slanderous material against certain people or groups, urgent and alarmist warnings, distorted or taken out of context statements spread massively across the interconnected mesh of groups on the platform. The repetition and sensationalism present within these messages is able to manipulate and mobilize users by creating an idea of threat and crisis. The misinformation found in WhatsApp's far-right political groups, through these strategies, leads the voter to a "fan" logic that blindly trusts its leader, in which everyone needs to choose a side (polarization) and if someone doesn't want their candidate to win, he is an enemy and needs to be fought.

In times of pandemic, misinformation can be especially harmful to people's health. It can be extremely abused as a political weapon, undermining the COVID-19 pandemic risk that imperils the whole population. SOARES et al., 2021b found a strong relationship between political themes and misinformation about COVID-19 on WhatsApp in Brazil. Among the topics they searched, conspiracy theories and information distortion were the more frequent strategies of misinformation surrounding the pandemic, ascribing a political bias strongly related to far-right perspectives to the health problem. Most of the topics were about how social distancing measures would hurt the economy, how some key main authorities (political or otherwise) were personally responsible for the pandemic crisis (e.g., mayors, governors, Congress, Supreme Court), how China had engineered the virus, and attacks to "leftists" and the media. RICARD; MEDEIROS, 2020 also observed that messages with misinformation about COVID-19 shared some common intentions: to minimize the severity of the disease, discredit social isolation measures intended to mitigate the course of the disease's spread and increase the distrust of public data. They concluded the messages shared on WhatsApp negatively influenced how people within the monitored groups responded to the containment measures proposed by health authorities, while, on the other hand, this misinformation intentionally influenced the public opinion, providing political gains to a specific political sphere that exploited the platform. Similar problems of health misinformation about COVID-19 on WhatsApp were also reported in Pakistan (JAVED et al., 2020), U.K. (VIJAYKUMAR et al., 2021) and Zimbabwe (BOWLES; LARREGUY; LIU, 2020).

In another direction, differently from understanding the misinformation universe on WhatsApp, there are studies that focus on measures for tackling this problem, such as the automatic detection of misinformation on WhatsApp. In work of KHURANA; KUMAR, 2018, they apply a preliminary modeling approach to estimate the spread of fake news on WhatsApp. The study proposes a framework to compare misinformation dynamics on WhatsApp based on the age group as well as topics prediction. CABRAL et al., 2021 also used of natural language processing (NLP) and machine learning techniques to develop the FakeWhastApp.BR, a system for misinformation detection in Brazilian Portuguese WhatsApp messages. Similar approaches attempt to detect automatically fake news from WhatsApp messages (GAGLANI et al., 2020; NATH; ADHI, 2019). In another approach, TARAFDAR et al., 2021, explores WhatsApp image data to implement a method for spam detection, using comparison techniques to find redundant images being shared multiple times among the messages.

KAZEMI et al., 2021, differently, give tip lines to combat misinformation on encrypted platforms in a case study of the 2019 Indian elections on WhatsApp. While MORENO; GARRISON; BHAT, 2017 made use of WhatsApp as a source of information to monitor Ghana's elections in 2016 and showed that including that data improved the effectiveness of the monitoring efforts during the analyzed period. They created a system with a dashboard containing data from various social medias, including WhatsApp. They use of this system to identify harmful content and report it to authorities to take appropriated measures.

2.3 Research Gap

Overall, previous studies are dedicated to measuring the dynamics and discourse of specific topics in WhatsApp and other messaging platforms. These previous studies show that all popular messaging platforms have been exploited for some sort of underground activities and different forms of abuse in communication systems, including misinformation dissemination, although little is done in direction to present countermeasures to fight against this problem.

Despite the undeniable importance of existing efforts, they do not attempt to provide a clear big picture understanding about the dynamics of public groups on multiple platforms and do not attempt to characterize the key differences of them. Particularly, regarding misinformation, even though more studies emerged in the last years, there is plenty of room for deeper analysis and comprehension of the ecosystem of instant messaging platforms as they became an essential part of people communication and socialization. There are still challenges to overcome in order to study this environment in large-scale and long-term, and even research methodologies are not very well-defined yet. Particularly about WhatsApp, its messages chats have been extremely abused by misinformation campaigns that take advantage of the opacity of the platform and the lack of a wide understanding of how this problem occurs within the groups. In this scenario, researchers need to continue taking steps towards providing more transparency for IMP networks and finding ways to combat it.

In this work, we fill this gap by performing, to the best of our knowledge, the largest multiplatform analysis of messaging platforms by collecting and providing an in-depth study of public groups from WhatsApp and comparing it to other messaging platforms. We present a novel and reliable methodology to gather data from public chats and explore viable approaches to tackling misinformation within those platforms, taking advantage of many findings acquired along the execution of the research and analysis of WhatsApp and other networks.

Here, it is worth to note that as a result of this thesis, some works were published over recent years (RESENDE et al., 2019b, 2019a; VIEIRA et al., 2019; MELO et al., 2019b, 2019a; REIS et al., 2020, 2020; HOSEINI et al., 2020; VASCONCELOS et al., 2020). Those works have different contributions prior to some key studies of the area referred in this chapter, and may also have influenced some of the related works presented here, given that misinformation on WhatsApp and other IMPs is a relatively new field of study and is constantly and rapidly evolving with novel approaches and results being updated constantly.

Chapter 3

The Ecosystem of Instant Messaging Platforms

Nowadays, it is hard to imagine our lives without Instant Messaging Platforms. Billions of users send numerous messages every day to contact their families, friends, business and even work colleagues. In fact, the blocking of WhatsApp by the courts in Brazil for 24 hours left a trail of losses across the country in 2016 and increased the volume of other services offered by telephone companies by up to 45% in that period, including calls and SMS messages (KAFRUNI, 2016). Along with WhatsApp, many other IMPs have appeared recently, helping users to accomplish different tasks in different contexts, since top secret communication with Signal up to gaming with Discord, and for buying and selling things with WhatsApp tools for business. These instant messaging platforms are applications in which users exchange messages in real time, being able to keep a synchronous chat.

These applications enable users to communicate effectively and efficiently with one another through Internet. They can send text messages, make voice and video calls, share media files such as images, audios and videos, documents, and other content[1]. They also brought a new feature in textual communication in the digital age: the use of stickers (LEE et al., 2016; WANG, 2016). Stickers are a specific type of picture similar to a GIF image file, usually animated, that users can send, using them for expressing emotions and humorous messages. Furthermore, most of IMPs present an end-to-end encryption to provide more safety and privacy for their users[2]. This resource consists of a communication encryption through which only the final users can read the messages.

However, this kind of online synchronous communication is not necessarily new. In 1996, the Israeli company Mirabilis launched ICQ ("I Seek You"), a text-based messenger that was the first to really reach a widespread market of online users (BONEVA et al., 2006). After that, some other popular instant messaging have appeared. American Online Corporation released AIM in 1997, which also allowed users to send messages to each other. Subsequently, Yahoo launched Yahoo! Messenger in 1998 and Microsoft released

[1] <https://www.whatsapp.com/features/>

[2] <https://faq.whatsapp.com/general/security-and-privacy/end-to-end-encryption>

MSN Messenger in 1999, attracting a whole generation of new users. Notably, there is also Skype founded in 2003 that allowed Internet users to communicate with others by video, voice, and instant messaging. In 2005, Google launched its own service, the Google Talk, which later became Hangouts.

Initially, online communications were dominated by a one-to-one paradigm, which is characterized by private and immediate exchange of messages between two equal partners of communication (CAMERON; WEBSTER, 2005; SEUFERT et al., 2016). With the emergence of OSN, some messaging systems became more popular (such as MySpace, Facebook, Twitter), adding a one-to-many layer to the interactions on Web (SEUFERT et al., 2015), in which users publish messages and broadcast them, so everyone can read, or at least a restricted list of friends or followers.

Recently, because of the democratization of access to smartphones and Internet, a new wave of IMPs, such as WhatsApp, Telegram, Viber, Line, WeChat, and Facebook Messenger, have supplemented this context with also group communication (MONTAG et al., 2015; SEUFERT et al., 2015), allowing conversations of a fixed group of users that can equally participate in the chat. Those platforms have become extremely popular in recent years. These IMPs, such as WhatsApp, are usually originally designed for mobile devices and require a smartphone, but they are also accessible from desktop computers[3].

In addition to its basic functionality for social communication, these IMPs are further served as an integrated platform that incorporates a variety of additional services (e.g., social networking, e-commerce, corporate advertising, e-payment, and social games) (OGHUMA et al., 2016). They are also being used for education purposes. CONDE et al., 2021 compare WhatsApp and Telegram as messaging tools for interaction in education, while NELLY, 2020 analyzed the use of Discord in classroom for students. More recently, even academic communities started to use IMPs as a tool in their researches or, more particularly, began to investigate the impact of their usage as methodology (OSENI; DINGLEY; HART, 2018; MAENG et al., 2016; VOIDA; NEWSTETTER; MYNATT, 2002; THIVYA.G, 2015). The instant messaging services facilitate surveys in which the contact between researcher and participant is limited either by geographic location or any other reason. It can also serve as registry of the interview. Even though, VOIDA; NEWSTETTER; MYNATT, 2002 shows there are tensions that permeate instant messaging texts and expose the collision between conventions of verbal and written communication in these research methodologies. Hence, those emergent computer-mediated communication technologies can be a very useful tool with a lot of application (MAENG et al., 2016).

Even if this kind of communication already existed before, they are getting more and more attention in the last years, with billions of users around the world. It raises

[3] <https://faq.whatsapp.com/general/download-and-installation/about-supported-operating-systems/>

questions on why it has gained such importance. CAMERON; WEBSTER, 2005 studied those emerging technologies of Instant Messaging, investigating reasons people have adopted this media in place of other traditional communication channels. They found that quick response/instant feedback and privacy are important features in these platforms that make people to abandon the SMS message old format and adopt these services. YOON; JEONG; ROLLAND, 2015 also aimed at understanding the adoption of mobile instant messaging services. Their results show that technical characteristics and ease of use play a relevant role in users' IMPs use, but they also found that social influence factors are also important in the process of decision of adoption of such technologies for users. OGHUMA et al., 2016 argued that perceived usability, perceived security, perceived service quality and confirmation determine the intention of usage of instant messaging platforms. Still exploring why people keep adopting messaging apps to communicate nowadays, SUN et al., 2017, in a different approach, explore the reasons why a user chooses to switch between different IMPs. Among the key factors that influence users' switching intentions in the context of instant messaging applications, they found that fatigue with incumbent IMP and subjective norms (perception that significant number of people are using other service) have significant positive effects on switching intention, while habit of use, affective factors and switching costs are pointed as motives to keep using the same platform. Moreover, social reasons such as to keep in touch with friends that use a specific platform is a strong consideration in both maintaining or switching messaging applications.

One of the main reasons of popularity of those communication systems in recent years, thus, is the social aspect of these services, similar to the social networks, and their key features that allow users to create groups within the platform thus enabling communication between a large number of users in real-time, a feature that does not exist in traditional SMS text messages. HOU et al., 2014, for instance, in studying the people's motives to use IMPs, stated that once users perceive to belong to a specific group, they tend to continue using the service. On the other hand, when users want to switch to a new platform, co-members in the community of the current one may keep them from leaving.

The choice of which platform to join depends on the trust of the users in the quality and security of the service. Furthermore, people want to use the platform that most of their friends already use, especially for the young population, for which IMPs increase their sense of identity and belonging to the group (BONEVA et al., 2006). The instant messages roved to be very helpful in socialization of people in contexts where they are unable to establish face-to-face communication or are confined to their homes such as the COVID-19 pandemic in which the use of WhatsApp support users to inform of announcements of the World Health Organization and regional health centers, to communicate at distance with medical professionals, avoiding frequent visits to hospitals that are high-risk areas, and also, WhatsApp can also be contacted by family members and friends during the period of quarantine and social isolation (DELAM; EIDI, 2020).

Therefore, the popularization of smartphones and wider internet access made it easier for people, especially from developing countries like Brazil, India, and Indonesia, to access the Web for the first time. However, the ease of use of those applications and their lower costs compared to SMS text message service contributed to the recent large mass adhesion of part of the population to instant messaging applications.

The increasing use of instant messaging apps and the creation of large chat groups raised some concerns about security, privacy, and harmful content that can circulate within these platforms. Nowadays, a wide range of illegal activities happens through instant messages. In this sense, THIVYA.G, 2015 argues that the currently existing Instant Messengers lack features to identify suspicious words from cyber messages and trace the suspected culprits. They proposed, then, a framework to detect suspicious messages from instant messaging systems in early stage, and that helps to identify and predict the type of cyber threat activity and to trace the offender details.

YUSOFF; DEHGHANTANHA; MAHMOD, 2017 performed a forensic investigation of instant messaging services (Telegram, OpenWapp and Line) in an analysis to extracted and track information to combat cybercrimes such as cyberstalking, cyberbullying, slander spreading and sexual harassment, which are facilitated in IMPs environment. RATHI et al., 2018 also did a forensic analysis of some encrypted instant messaging applications on Android (WeChat, Telegram, Viber and Whatsapp), exploring the implications of end-to-end encryption adopted by these platforms and some security measures of how they store data on smartphones. VALERIANI; VACCARI, 2018 explore the use of IMPs (i.e., WhatsApp, Facebook Messenger, Viber, WeChat, and Snapchat) for political conversations. They show that a large portion of instant messaging users are engaged in political discussions. They highlight the importance of these platforms, specially for citizens holding relatively extreme ideological positions, who see it as a particularly comfortable environment for political talk.

3.1 Comparing WhatsApp, Telegram and Discord

In this section, we provide the necessary background information on WhatsApp, Telegram, and Discord, and present the characteristics of these messaging platforms, highlight how they differ from one another in Table 3.1[4].

[4] This comparison was performed in 2021, reflecting the characteristics of the versions of these platforms from that date.

Table 3.1: An overview of the characteristics of the different messaging platforms, highlighting some of their differences.

Characteristic	***WhatsApp***	***Telegram***	***Discord***
Initial release date	January 2009	August 2013	May 2015
User base	2 Billion	400 Million	250 Million
Clients	Mobile, Desktop, Web	Mobile, Desktop, Web	Mobile, Desktop, Web
Registration method	Phone	Phone	Email
Options for public chats	Groups	Groups and Channels	Server
Max. #members in groups	256	200,000 for groups (unlimited for channels)	250,000 (500,000 for verified servers)
Types of content supported	Text, Sticker, Image Video, Audio, Location Document, Contact	Text, Sticker, Image Video, Audio, Location Document, Contact	Text, Sticker, Image Video, Audio, Location Document, Contact
Open source?	No	Partially	No
API for data collection?	No (only Business API)	Yes	Yes
Message forwarding?	Yes (up to 5 groups)	Yes	Only available via link and only for members
Track forwarded msgs	Label message as "Forward' "Forwarded multiple times"	Shows original source and total views for channels	Generate a link to original within same server
Encryption	End-to-end encryption for all chats	Clien-server encryption default E2EE only for "secret" chats	Not encrypted

3.1.1 WhatsApp

Launched in January 2009, WhatsApp is the largest messaging platform with over 2 billion users (WHATSAPP, 2020b) and the most used social media platform, second only to Facebook (NEWMAN et al., 2019). To use the messaging platform, users must register with their phone number. Users can also use the platform via WhatsApp's Web or desktop client, but these clients still require the user's mobile phone to be connected to the Internet. The platform supports both one-on-one chats and group chats—simultaneously with up to 257 users—through chat rooms or *groups.* Administrators of a group can add others to the group, either by directly making them members of the group or by sharing a group URL (or an invite link) with them. Members of a one-on-one chat or group can share or forward information in a range of different formats including text, image, videos, documents, contacts, locations, and stickers. In addition to chats, the platform supports audio and video calls, and all communications on WhatsApp are secured using end-to-end encryption.

3.1.2 Telegram

One of the most popular messaging apps in this scenario is Telegram. Founded in 2013, it currently has over 500 million active users and has seen tremendous growth in its user base[5] Similar to WhatsApp, it requires users to register with their phone numbers, and after registration allows them to communicate also using its Web or desktop clients. But unlike WhatsApp, users are not required to have their phone connected to the Internet while using the Web or desktop clients.

In Telegram, users can create two types of collective chat rooms: *channels* and *groups*. Channels support a few-to-many communication pattern, like a broadcast list in which the creator and the administrators of the channel can share information with the rest of the members, and do not impose a limit on the number of members per channel. Groups, in contrast to channels, facilitate a many-to-many communication pattern, where all members of the group can share information with one another, and impose a limit of 200 K members per group. Both groups and channels allow users to share and forward information in a wide range of formats. Unlike WhatsApp, not all message exchanges are end-to-end encrypted. End-to-end encryption in Telegram is only available for "secret chats", which are device-specific communication channels. Users can access the secret-chat messages only from the device on which the chat was created, and they cannot forward messages from secret chats. In addition, channels and public groups cannot be created in this 'secret' format.

Telegram has an API support for developers. With the API, users can implement bots to moderate chats and perform multiple tasks within the groups, such as give information, manage the group, or even play games. API also favors data collection, as it has access to group and some members' data.

We included Telegram in our comparison due to both its growing popularity and evidences of exploitation of the platform by bad actors, e.g., white supremacists (Anti-Defamation League, 2019) and terrorists (TAN, 2017).

3.1.3 Discord

Although it started with a focus on providing a messaging platform for the online gaming community, Discord is used nowadays by the general public for various purposes, even including education (NELLY, 2020). But their user base is still composed mostly of a younger demographic.

The platform was launched in May 2015, roughly two years after Telegram and six years after WhatsApp. It has about 250 million users. In contrast to WhatsApp and Telegram, users can register with an email only; the platform does not require users to provide a phone number. Users can create a *server* (or *guild*) and, within this server,

[5] <https://t.me/durov/147>

they can create multiple channels. Those "subchannels" are usually divided by theme or topics of interests of the server to help to organize the conversations. Those can be voice or text based. That means that multiple members can join at the same time in a voice channel to talk, while another portion of members can chat in a text channel at the same time. Administrators in Discord also have at their disposal many different tools and functionalities. Besides create channels, ban and add users, and create invite links, they can also restrict access to specific channels to some users, give them permission to chat or not, and they can exclude messages from other members. Discord's servers can have a large number of users—up to 250K by default—and some (e.g., "verified" servers of organizations, artists, or games) can host up to 500K users. Lastly, channels in Discord do not offer end-to-end encryption at all.

3.1.4 Similarities and Differences of IMPs

All three messaging platforms support public groups, and the most common way to invite other users to a public group is through a group URL (also referred to as the "invite" URL) with them. The group URLs of each IMP follow one or more distinct patterns. On WhatsApp, for instance, group URLs have the pattern "chat.whatsapp.com/`<gID>`" with `gID` representing a unique identifier of the group, which is automatically generated by the WhatsApp messenger application when the group is created. After a review of each platform's documentation and also manually examining the URLs of each platform on Web, we were able to compile a list of six patterns employed across these messaging platforms. These six patterns have the following prefixes or `host` values: `chat.whatsapp.com/`, `t.me/`, `telegram.me/`, `telegram.org/`, `discord.gg/`, and `discord.com/`. All those invite URLs are widely found on social networks such as Twitter and Facebook. This corroborates that group feature is a key feature explored by users to create public group chats to discuss and share interests in their social media. In Chapter 6 we will further investigate those groups and this public ecosystem of the instant messaging application.

An important characteristic present in the platforms is the possibility to forward content between groups and contact. This brings them closer to the social network such as Facebook and Instagram, as it allows to share content and create a network of messages spreading through their users. In Discord, it is more restricted. There and in Telegram, the forwarded message is attached with some metadata regarding the original author. In Telegram, it is possible to even see how many times it was forwarded and how many users visualized that message. On the other hand, with WhatsApp, users can send their messages up to 5 other contacts/groups, and only for highly sent messages it is limited to just 1 contact per time[6]. Furthermore, in WhatsApp, the only difference between an original message and a forwarded one is a tag "*Forward*" in the message header. There is no further information on who or when it originated. Also, none of the users know that their messages were forwarded in the systems. The forwarding option is one of the key features abused by misinformation campaigns as it has a potential scalability to the content reach thousands of users, while at the same time it provides anonymity to the author who created the content.

While all messages on WhatsApp are sent via end-to-end encryption and users can not disable even if they want it, Telegram has this option only for secret chats options, which means that channels and public groups are not protected by this feature. Discord works in a different way, they do not have encryption at all, but it is a moderate social media as the company can investigate and eventually ban accounts and whole servers that are not to agree with the terms of the Discord [7]. Thus, WhatsApp looks more secure in communication, although it is also the most closed platform in which most of the abusive content is hidden under this private structure of groups.

Probably the more notable distinction between WhatsApp and other two platforms is the group's size. While WhatsApp limit is 256 members per group, there are chat groups up to 200K members in Telegram and 500k in Discord. Moreover, in Telegram, for channels, that are broadcast groups of messages, there are no limits of members, accumulating millions of users[8]. In Discord, those groups are actually called servers, and they supply a lot of tools to the moderators of these groups to manage their chats. There, for instance, one can create subgroups of text or voice channels and give individual permissions to the members, managing to disperse the users from a unique server among its different channels. These large groups allow the information to quickly be sent to multiple users at once, which differs a lot from the direct conversations also present in those platforms. This shows a duality of the IMPs that works at the same time for private chat and mass communication.

These aspects of the platforms show why WhatsApp is a fertile and susceptible environment to spread misinformation. Users feel secure within the platform, talking to family and close friends, but simultaneously they are in contact to a public sphere of groups that rapidly share messages anonymously. For final users, it may not be a simple task differentiating both kinds of content while using the app, principally for those who are getting in contact with the internet for the first time with a smartphone.

[6] Note that it was only after 2018 that WhatsApp reduced this forward limit to 5 in Brazil. It was the second restriction, as initially users could forward up to 256, then it was reduced to 20 before it reached the currently 5 limitation.

[7] <https://www.engadget.com/discord-transparency-report-222737366.html>

[8] <https://appsgeyser.com/blog/biggest-telegram-channels/>

Chapter 4

What is the WhatsApp?

WhatsApp Messenger, or simply WhatsApp, is an Instant Messaging Platform system, cross-platform centralized messaging, and voice-over-IP (VoIP) service founded in 2009 by Brian Acton and Jan Koum, former employees of Yahoo!. As it sounds, the name WhatsApp is a pun on the phrase “What’s Up”. It allows users to send text and voice messages, make voice and video calls, and share multimedia content such as images, videos, audios, and documents. WhatsApp’s client application originally runs as an app from smartphones, but it can be also used in desktop computers, as long as the user’s mobile device remains connected to the Internet while they use the other device.

In 2013, WhatsApp had already reached approximately 200 million active users, had a team of only 50 members and was valued at about 1.5 billion dollars[1]. The client application was created by WhatsApp Inc. of Mountain View, California, which was acquired by On February 19, 2014, Facebook, Inc. announced it was acquiring WhatsApp for US$19 billion, its largest acquisition to date[2]. Already at that time, because of a great distrust of Facebook by a part of users, the acquisition caused that a considerable number of users to try and/or move to other IMPs as 8 million shows to migrate to Telegram[3] and 2 million new users were registered in Line[4]. Still, it then became the world’s most popular messaging application by 2015, and finally reach two billions users worldwide as of February 2020 (WHATSAPP, 2020b). It has become the primary means of internet communication in multiple locations, including Latin America, India, and large parts of Europe and Africa[5].

Today, WhatsApp is an instant communication application in which smartphone users can exchange messages, images, videos, audios, and make audio and video calls. Amidst a huge flow of information with more than 55 billion messages and 4.5 billion images a day around the world, WhatsApp is the second most used social platform in the

[1] <https://www.forbes.com/sites/parmyolson/2014/02/19/exclusive-inside-story-how-jan-koum-built-whatsapp-into-facebooks-new-19-billion-baby/>

[2] <https://about.fb.com/news/2014/02/facebook-to-acquire-whatsapp/>

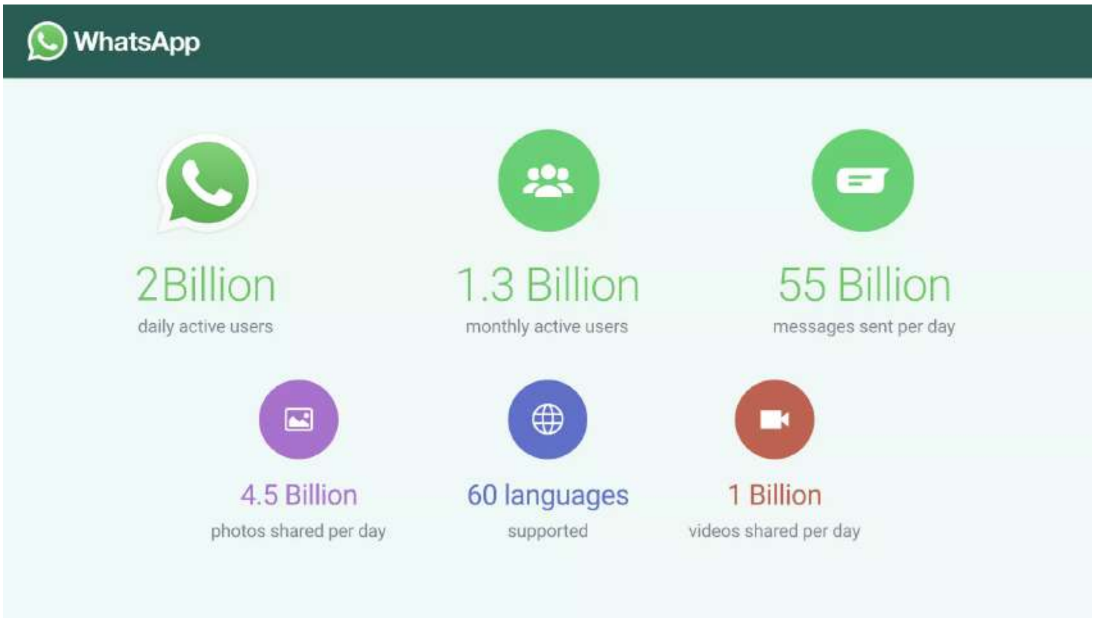
[3] <https://techcrunch.com/2014/02/24/telegram-saw-8m-downloads-after-whatsapp-got-acquired/>

[4] <https://techcrunch.com/2014/02/25/line-gets-whatsapp-outage-bump/>

[5] <https://www.wired.com/2016/04/forget-apple-vs-fbi-whatsapp-just-switched-encryption-billion-people/>

world, only behind Facebook itself.

Figure 4.1: Activity numbers from WhatsApp.



Source: <https://blog.whatsapp.com/10000631/Connectinganuser-users-all-days>.

Unlike social platforms like Twitter and Facebook, in which there is platform moderation with tools to investigate content posted in these environments, WhatsApp's end-to-end (E2EE) encrypted framework creates a very different and more complex scenario to observe. Although much of the communication in this messaging app is direct and private between two users, there are a lot of groups where the conversation involves a larger number of members. Only users involved in the conversation, however, have access to the shared content. Therefore, to monitor and collect the content that circulates on WhatsApp, it is necessary to be effectively part of interest groups.

On WhatsApp, every user account is linked to a valid phone number. Authentication with a smartphone is required during account creation. Even the web version of the application must be synchronized with the smartphone for the user to be able to use it. With a proper account, the user can modify basic information such as their name, photo, and a status message. These users can add and send messages to contacts directly from their mobile contact or, in what will be the focus of this project, chat groups.

4.1 Public Groups Chats

WhatsApp allows individuals to create and join group messages with up to 256 users total, a feature that has significantly influenced in the popularity of the app. These groups have a name, description, and a list of members, and at least one user must have the role of Administrator. A group administrator has more control and tools than other members. Among the main additional features, we can highlight the possibility of:

- change the name, image, and description of the group (or pass this permission on to other members of the group);
- add new members to the group;
- remove group members;
- set the group to broadcast mode (only administrators send messages);
- create an invitation URL for the group.

Groups with an invite URL are seen as public groups, since anyone with the URL can join and join the group. All invite link URLs from WhatsApp follow the same following pattern:

"<https://chat.whatsapp.com/<identifier>> "

WhatsApp's users utilize this link to invite others and even publicly share this URL to find more members online for their groups. This pattern can be used in searches on social networks (e.g., Twitter, Instagram, Facebook) or even on search engines (e.g., Google, Bing) to find relevant groups on a topic or topic of interest. There are even sites that specialize in aggregating WhatsApp groups so that users can find groups of interest (e.g., <https://gruposdezap.com/>)

Once inside a group, users have access to the group's membership list and the messages sent there. However, note that a user only sees messages sent after their entry. Messages prior to the date he joined the group will not be available (other applications, such as Telegram, show previous messages, but this is not possible in WhatsApp). The user can also send messages after joining the group (unless the group is put into broadcast mode by administrators.

4.2 Messages

A message within WhatsApp is usually a text message, but there are several types of message content that can be sent such as image, video, audio, and sticker. A user can also send a document, a phone contact, or even a geographical location. With the development of WhatsApp Business, users also can make transactions and buy products on WhatsApp.

In addition, in the message, you can see who sent the message and the time it was sent. In WhatsApp, there is also a “Forwarding” function, in which a user can forward a message from one conversation to another. When this happens, we see a “Forwarded” marker in the destination group, to emphasize that this message was originated from elsewhere, but it is not possible to see the original author or from which group it was forwarded. This function allows faster and more massive sharing of messages within WhatsApp and makes some contents become quite popular within the platform, but it does not allow tracking of their origins.

Messages on WhatsApp have a sequential and stacked flow, that is, new messages are displayed first (lower on the screen) while old messages lag farther and farther behind (upwards on the screen). To view an older message, you must then go back through the entire conversation (scrolling up) until you reach the desired message. Even in the automated process, this operation is necessary since the code relies on the application design to collect the data. All messages on WhatsApp are also encrypted by end-to-end security.

4.3 WhatsApp End-to-end Encryption

Security is an important topic when it comes to messaging apps. Most of them have some level of security and many, including WhatsApp, use end-to-end encryption to protect the users and their messages. Here, it will be briefly discussed some characteristics of the encryption used on WhatsApp as provided by its technical report (WhatsApp Messenger, 2021). WhatsApp defines end-to-end encryption as follows:

> *... as communications that remain encrypted from a device controlled by the sender to one controlled by the recipient, where no third parties, not even WhatsApp or our parent company Facebook, can access the content in between. A third party in this context means any organization that is not the sender or recipient user directly participating in the conversation.*

WhatsApp’s end-to-end encryption protects messages and calls made in the WhatsApp application, ensuring that the content is restricted only to the peers involved in the conversation. In this type of encryption, in contrast to client-server, no one else can read or hear them, not even WhatsApp itself. Messages and calls are protected with a unique hash, and only the recipients have access to the key needed to unlock it and read the messages. This entire process happens by default in WhatsApp, and it is not possible to activate nor deactivate these settings.

The Signal Protocol, designed by another IMP, is the basis for WhatsApp's encryption[6]. This E2EE protocol is designed to prevent third parties and WhatsApp from having access to messages or calls. Due to the generation of new cryptographic keys over time, even in a situation in which the current encryption key of a user's is compromised, they cannot be used to decrypt previously transmitted messages. This is the main reason behind the fact of a user being unable to see previous messages when they join a new group; they can access just those sent after they are already members of the group. WhatsApp also uses end-to-end encryption to encrypt the message history transferred between devices when a user registers a new device.

At the time of this research, WhatsApp uses the concept of primary and companion devices to associate a user's account. Each WhatsApp account is associated with a single primary device that is used to register a WhatsApp account with a phone number. With this device, it is possible to link additional companion devices to the account, such as WhatsAppWeb, Desktop application, and Facebook Portal. This is relevant to system security as, at the registration time, a WhatsApp client transmits its public Identity Key with its signature to be stored at the WhatsApp server for the user's identification. When linking a companion device, on the other hand, the user's primary device must first create an Account Signature by signing the new device's public Identity Key and the companion device must create a Device Signature by signing the primary's public Identity Key. Only after both signatures are produced, end-to-end encrypted sessions can be established with the companion device. This is the process to login in other devices through the scan of a QR code with WhatsApp in user smartphone.

In order for WhatsApp users to communicate with each other, the sender client establishes a pairwise encrypted session with each of the recipient's devices. Additionally, the sender client establishes a pairwise encrypted session with all other companion devices associated with the sender account. To establish this session, the initiating client requests the public Identity Key from the recipient and a single public *One-Time Pre Key* for each device of the initiating user. Then, the server returns the requested public key values. The *One-Time Pre Key* is only used once, so it is removed from server storage after being requested.

After getting the keys from the server and verifying each device identity, a long-running encryption session is built, and the initiator can start sending messages to the recipient, even if the recipient is offline. Until the recipient responds, the initiator includes the information (in the header of all messages sent) that the recipient requires to build a corresponding session. Therefore, it is possible to conclude that messages are encrypted,

[6] The Signal Protocol library used by WhatsApp is based on the Open Source library available at <http://github.com/whispersystems/libsignal-protocol-java/>

however, the message headers contain information of the sender, recipient, and other metadata involved in this process.

End-to-end encryption of messages sent to WhatsApp groups utilize the strategy to establish sessions to distribute the “Sender Key” component of the Signal Messaging Protocol. When sending a message to a group for the first time, a “Sender Key” is generated and distributed to each member device of the group. The message content is encrypted using the “Sender Key” and sent to group members.

Chapter 5

A Methodology to Collect Data from WhatsApp

WhatsApp has its own structure, very different from other social networks. Therefore, the collection architecture is very different. For instance, it is strictly necessary a smartphone with a valid account participating in the target groups as a methodology to identify and save messages for any study. Before getting into more details about implementation and development of the dataset, let's give an overview of the architecture behind the WhatsApp collector.

Apart from the WhatsApp studies that form their data corpus from interviews, surveys, and other qualitative methods, those that aim at a more systematic and automated collection of WhatsApp messages on a larger scale usually do so by focusing on data extracted from public chat groups. With the account set with the target groups of interest, the there are two different paths one can follow to actually retrieve the data from the device: one is by decrypting the WhatsApp database directly from the smartphone; the second is scraping data through the use of WhatsApp Web, that is the browser version of the application.

The first one, that uses the database from the phone, was firstly proposed by authors from GARIMELLA; TYSON, 2018. Their method involves the task of rooting Android phones (also called "jailbreaking" for iOS of iPhones), which is necessary for obtaining the encryption key from WhatsApp used to secure this message database. Although this method gives faster and wider access to all messages and some other information stored in the phone, the process of rooting a phone drastically changes from smartphone model to smartphone, which makes it very dependent on web communities specialized in such tasks. In addition, and more important, there is a discussion (BURDOVA, 2021) if rooting applies as violations regarding circumvention of technological measures[1]

, which is not allowed in the USA and also can result in a ban from the WhatsApp

[1] In 17 U.S. Code §1201 of Digital Millennium Copyright Act (DMCA) states that "*to circumvent a technological measure means to descramble a scrambled work, to decrypt an encrypted work, or otherwise to avoid, bypass, remove, deactivate, or impair a technological measure, without the authority of the copyright owner" (Legal Information Institute, 2021*)

as they detect accounts engaged in “inappropriate behavior”, which could apply for rooted systems (SIMONS, 2019).

The second approach, and the one followed by this thesis, utilizes a web scrapper with use of Selenium and WhatsApp web to navigate to each group in the WhatsApp Web interface, and then in each group extract the information of group members and messages. This approach is more complicated than simply decrypting the message database, since it relies on the rapidly changing and quite “fragile” WhatsApp Web, but is better in certain aspects (CHANG, 2020), as it does not rely on rooted phones and WhatsApp could easily change how they store and secure the message database, making the first method infeasible or substantially more difficult. On the other hand, this methodology requires that both the smartphone and the server where scripts run keep continuously connected to the Internet and changes in WhatsApp Web can also interfere in how the scrapper works. However, this method was preferable given the existence of some online libraries already implemented by users who manipulate WhatsApp Web through Selenium. In this work, we make use of WebWhatsapp-Wrapper API (HASE, 2018), which has an active community that supports the users and updates the code whenever WhatsApp Web had substantial modifications.

Next, we specify each step of this data collection, since account creation to the storing of the data. As this collection involves more processes than the code itself only, with this description, our goal is to make it clear the methodology of collection as anyone wanting to repeat the process to collect WhatsApp could understand and follow the steps to be able to perform a research using data from WhatsApp public groups.

5.1 Setup of the Account

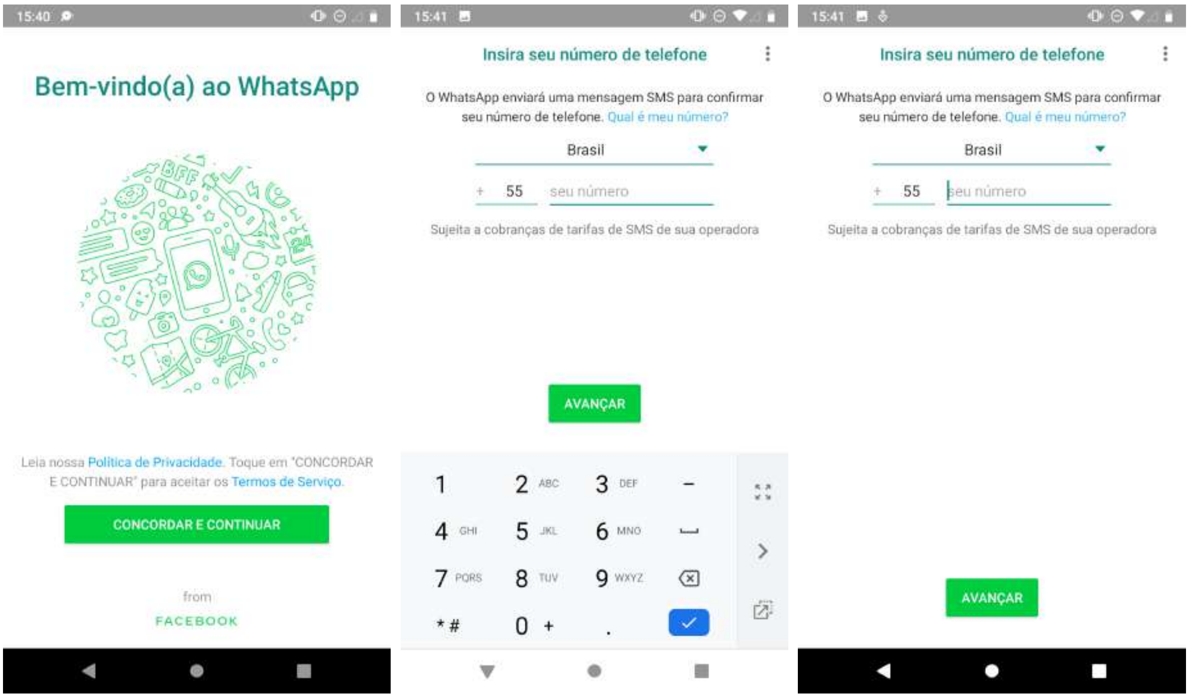
To collect WhatsApp data, a series of steps prior to the actual collection are required. These steps include, registering an account in the application, creating a profile, searching for public interest groups and logging in to the web version of WhatsApp In this section, the steps of WhatsApp settings necessary to carry out the collection are presented in detail.

5.1.1 Authenticating the App

The first step in order to collect WhatsApp is creating an account. When downloading the WhatsApp app, authentication is required to start using the app. This requires a SIM card with a valid mobile number (which is capable of receiving an SMS with the authentication code).

To register a new account, one must first download the WhatsApp app on the dedicated smartphone for collection. By opening the app, the user must agree to the app's Privacy Policy and Terms of Service. Then, it will ask to enter the mobile number to authenticate a new account. Upon entering the valid mobile number, an SMS message will be received containing the authentication code to fill in the space provided. With that, WhatsApp will be authenticated and ready to use. Figure 5.1 illustrates the steps described above in the registration process.

Figure 5.1: Step for registering a new mobile number in the WhatsApp application.



Bem-vindo(a) ao WhatsApp

Leia nossa Política de Privacidade. Toque em "CONCORDAR E CONTINUAR" para aceitar os Termos de Serviço.

Insira seu número de telefone

O WhatsApp enviará uma mensagem SMS para confirmar seu número de telefone. Qual é meu número?

Brasil

+ 55 seu número

Sujeita a cobranças de tarifas de SMS de sua operadora

Insira seu número de telefone

O WhatsApp enviará uma mensagem SMS para confirmar seu número de telefone. Qual é meu número?

Brasil

+ 55 seu número

Sujeita a cobranças de tarifas de SMS de sua operadora

(a) (b) (c)

Source: The Author.

5.1.2 Creating a Research Persona for WhatsApp

Collecting WhatsApp data requires much more interaction with other users of the platform compared to other OSNs like Instagram, YouTube, or Twitter. It is necessary to effectively participate in the groups at the same level as the other members, not only being able to observe them from a distance. Thus, it requires great attention to respect the WhatsApp Terms of Service and proceed with a more ethical observation of the project. Therefore, during the creation of the WhatsApp profile for monitoring public groups, the idea of persona used. A persona is an objective fictitious representation of a user profile, whose idealization is based on user surveys, observation of interests, desires, and needs. This is a frequently used tool for software development. This methodology was used for WhatsApp groups because it can synthesize the appearance of a conventional user of the application.

To build a persona for WhatsApp groups, you must first answer some questions related to your personality, life, and use of the app, keeping in mind the groups you want to observe. The questions used could be, for example:

- What is the person's gender?
- How old are the user?
- Is the person married or in a relationship?
- Does the person have children?
- Does the person live alone or share the house? With whom?
- What is your level of education?
- Does the person work? What's your job?
- Why does the person use WhatsApp?
- How long do they use WhatsApp?
- Does the person access WhatsApp on their cell phone, computer, or tablet?
- Which groups is she interested in?

From these questions, it is possible to imagine a fictional persona with a personality that fits into the groups and is not identified as a fake profile by other users. As the objective of this work, as it will be detailed later, is political groups on WhatsApp, the persona developed at the initial stage of this project was as follows:

A woman, approximately 40 years old. She uses WhatsApp to talk to her family, friends, and she gets information by following groups of political contents, although she does not send messages to these groups. She uses WhatsApp exclusively from her cell phone to chat with others. She shares the news and reports she finds relevant, and her group partners often include her in new groups.

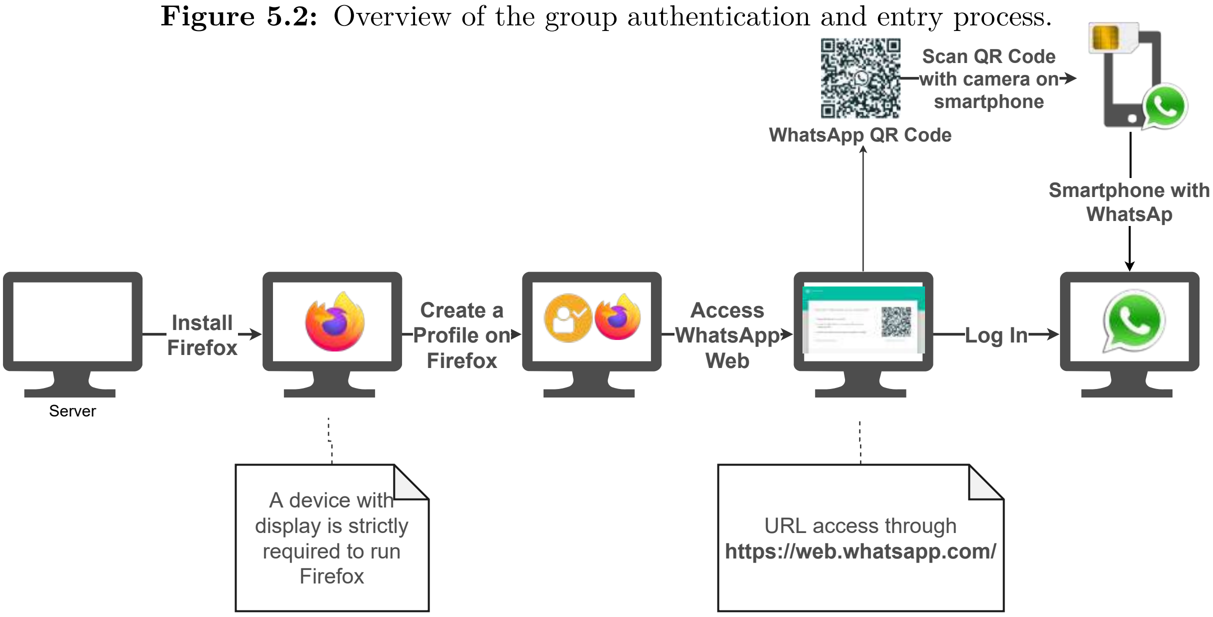
Along with the creation of the persona, it is informed, in the status of the WhatsApp profile, the purpose of that account, in which the profile represents an account for group research on WhatsApp. Following the goals of the persona developed, in addition to follow an already consolidated methodology of virtual ethnography, this process is important in maintaining selection and entry into public groups in accordance with the project's proposal and is a way to ensure that the profile respects WhatsApp Terms of Service, avoiding, among other things, being banned from the platform.

5.1.3 Using WhatsApp Web

As the data collection will be based on a web scrapper through a Selenium with a Firefox browser. Therefore, it is necessary to configure the account to log into WhatsApp Web the first time we use the tool developed in this project. Figure 5.2 shows a flowchart of how to connect in the WhatsApp Web version. With the flowchart, we can see that there is a manual step that requires the use of the smartphone with the WhatsApp account you want to collect to scan the QR Code displayed on the web page. Because of this, there is a requirement of a monitor (or any visual display) that will open the Firefox browser to perform this step. Another requisite is a camera phone to scan the QR Code. This step is required for the first time you set up the collection. Once connected, the other collection instances can already have the account properly logged.

For this phase, observe that both WhatsApp Web and the smartphone need to have constant Internet access. However, eventually, the WhatsApp account can disconnect and "logout" from the browser. This can happen due to updates from the WhatsApp itself or if the Firefox profile is deleted from memory, for example, and thus this step can be repeated whenever the account is disconnected from the browser.

Figure 5.2: Overview of the group authentication and entry process.



Source: The Author.

5.2 Joining Public Groups

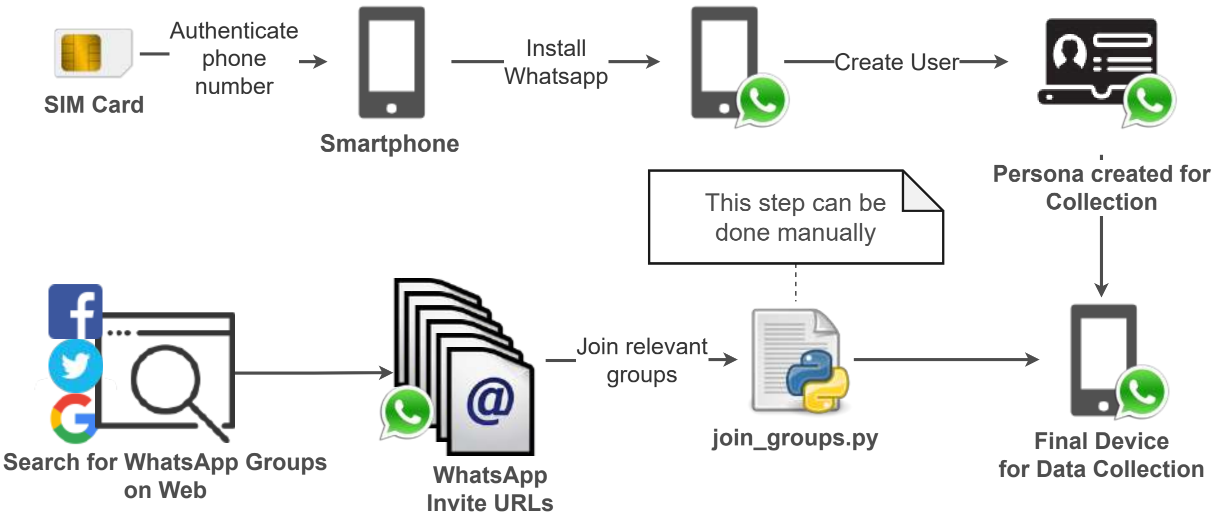
Another step prior to collection is identifying and joining the groups identified as relevant for the project. In Figure 5.3, we see a scheme of how this initial process is performed, since cell phone authentication, social media search up to the group joining.

Here, one must search for the public groups of interest that will be collected. These public groups are accessed by WhatsApp invitation URLs that can be found on other social networks, search engines or even identified by other sources. Properly joining the public groups from their invite links is rather easy, involving a few button clicks on WhatsApp, which can be automated using Selenium. However, WhatsApp platform checks for suspicious behavior among users, in this work it was more suitable to manually join each of them, since the group list was of a size large enough to be feasible to be manually joined by a real user with a smartphone and that manner it is possible to control the velocity of this process to avoid be banned. This proceeding is iterative and was repeated during all phases of the research along the 4 years of the study, finding new groups.

While the ecosystem of URLs of public groups published on the web will be meticulously discussed later, in this thesis, in Chapter 6, to a better understanding of how this universe of instant messaging platforms works, here it will be briefly described how this investigation took place in context of political groups in Brazil to find the WhatsApp public groups monitored during this research.

All invite URL of a WhatsApp public group follow the same template of "<https://chat.whatsapp.com/<identifier>>". This pattern can be used to retrieve groups on

Figure 5.3: Overview of the process of profile authentication, finding public groups and joining


SIM Card
Authenticate phone number
Smartphone
Install Whatsapp
Create User
Persona created for Collection
This step can be done manually
Search for WhatsApp Groups on Web
WhatsApp Invite URLs
Join relevant groups
join_groups.py
Final Device for Data Collection

Source: The Author.

Web. By looking for this URL in searches engines such as Google, or in social networks such as Twitter, it is possible to find a lot of groups of many different topics. Thus, in order to find relevant groups given a specific project, one can use a list of keywords to refine the results with the topic of interest. This was the approach followed in this study.

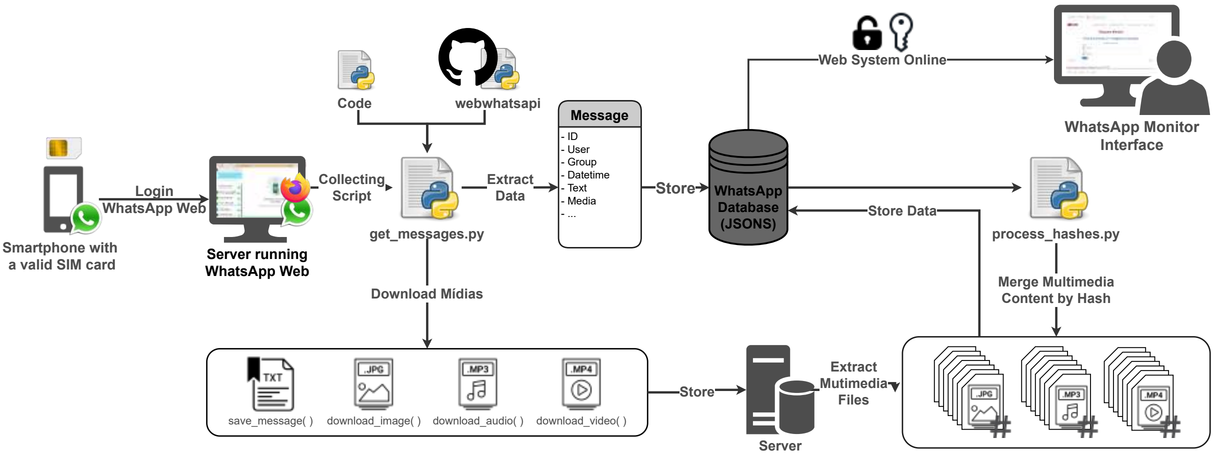
First, a list of keywords covering the Brazilian political context was developed in 2018, the year when the presidential elections took place, containing the names of the main candidates and some other relevant terms at the time. Also, we added terms of all political spectrum from both left and right wing. Then, with this list, each term was used in a search along with the WhatsApp URL pattern to retrieve the groups we joined in the next step.

Furthermore, to expand the number of groups, few other strategies were adopted: the keywords list was iteratively updated adding and removing terms as new topics emerged in political discussion in Brazil. In addition, within the groups, users usually advertise another public groups. Those new invite URLs were extracted from the political public groups themselves to increase the number of groups in the dataset.

Finally, we built a large set of URLs of WhatsApp public groups that were gradually being joined one by one through the app. During the joining process, it is important to highlight that some links result in an invalid invite in WhatsApp. This is often related to broken or mistyped URLs, but most of the errors happen because of invite links that were revoked by the owner group. There are several flaws inherent to this process, which center on the fact that if we don't collect and join links in real-time (as they are posted), links may be revoked by the time we attempt to join. Still, most of the groups remain active and able to be joined, resulting in a solid dataset of public groups of WhatsApp.

5.3 Architecture of Data Collection

Figure 5.4: WhatsApp Web connection flowchart.



Source: The Author.

Next, with all setup of WhatsApp account done and groups selected for the research, the data extraction can start collecting all data from chats in the server. The main architecture of the collection is explained through the flowchart in Figure 5.4. In this figure, we have an overview of all the steps performed by the collection, as well as the objects involved during execution: (1) The mobile phone and the Firefox browser, both connected to WhatsApp, are the source of data. (2) A script looks at the data received in WhatsApp groups and extracts the messages. (3) Messages are structured and saved in JSON. (4) Media files (image, video and audio) are downloaded and saved on the server. (5) Another script is responsible for analyzing these files and grouping messages by similarity. (6) Grouped messages are saved in JSON containing every time they were shared[2].

With this architecture, we are able to access the WhatsApp account with the Selenium based script and then navigate through the groups joined and collect the messages. Note that, in WhatsApp, when someone joins a group, this new member has access only recent messages sent after the date of joining, and it is not possible to retrieve the group message history shared before that. Given an initial date for collection, for each group, the script manages to scroll the messages back in chat until it reaches the content from the start date, then, it saves each message individually in local storage. When the message has any media attached to it, the script – using the WhatsAppWeb API library – needs to request to WhatsApp server the file of this media. This file is downloaded and decrypted

[2] All scripts used in this work for collecting WhatsApp data are accessible through the repository <https://github.com/Phlop/WhatsApp_Crawler>

(using WhatsApp hash) and also saved. At the end of this process, we can obtain some metadata for each group, as well as some information of the messages sent in the groups

5.3.1 Group Metadata

For each group, it is possible to get some metadata from it. There is available a unique group ID of that group; a title that most groups have, the description (just those whose administrator wrote one, which is not the case of many of the groups), the profile image (just a link for the actual file on server, as it was not stored), creation date, the user who created the groups and a list of all members of that group at the time of collection.

Interestingly, by looking at the structure of the data collected, it was observed that the unique identifiers of the WhatsApp groups follow the pattern `55319999XXXX-15928XXXX`. In this format, the first half is the phone number of the group creator, and the second half stands for a timestamp that represents when the group was created. Therefore, with only the group ID, one can also infer who created that group and when it was created. This unique ID is also useful to distinguish the groups, as two groups can share the same title and a single group can change the title over time.

For the other group members, we have a list of phone numbers of those users. It is interesting that all WhatsApp groups have the full list of the members of that group, this can be understood can be understood as a WhatsApp problem of security as malicious scripts can exploit this attribute to parse a huge list of groups collecting their phone numbers and build a large dataset of exposed phones. This issue was more investigated in Chapter 6.

5.3.2 Messages Data

For messages, we have the unique identifier of that message, group ID and title where the messages were sent, the user (by phone number) who sent that message, date time, text content, kind of the message (i.e., text, video, audio, image, document), and the filename of the media attached to that message.

Note that in this format, a unique piece of information that was shared in multiple groups and/or by multiple users is stored as totally distinct objects. In order to see how information flows through the network, it is necessary another step to aggregate identical messages and track this dissemination.

5.3.3 Frequency of collecting WhatsApp data

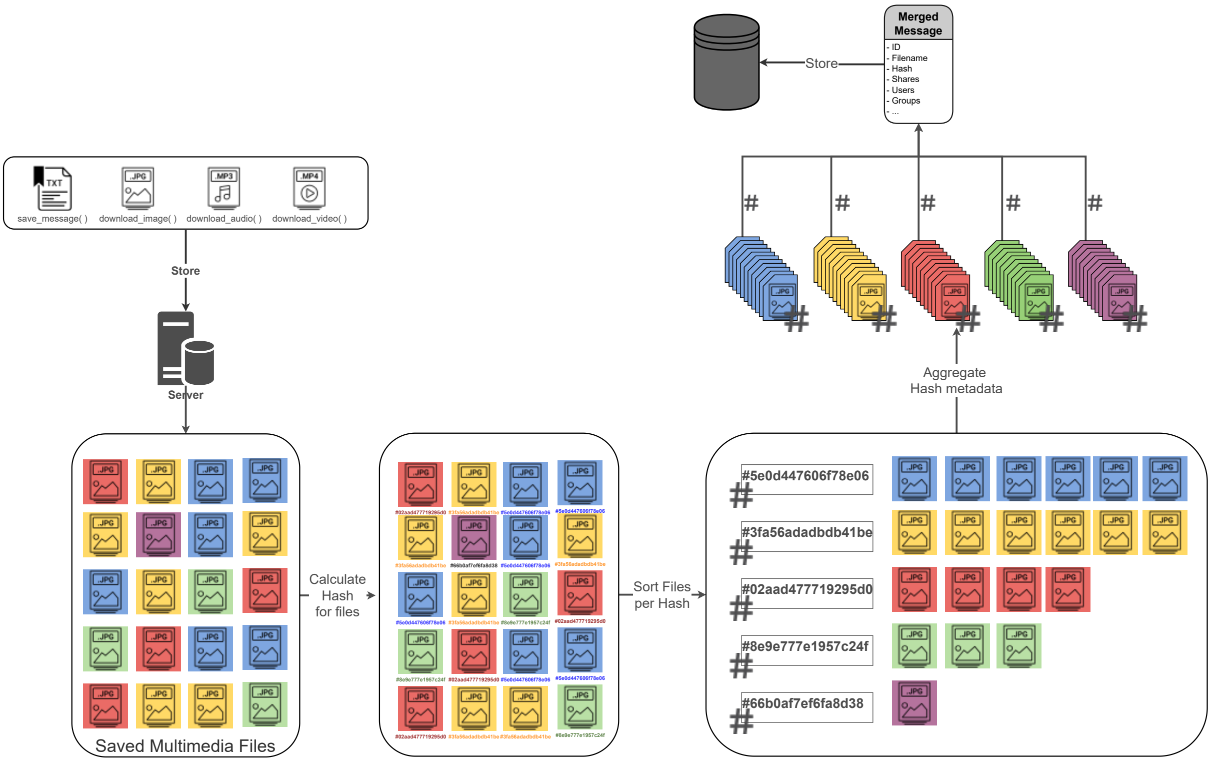
This step of going through groups and get messages can be very time-consuming, due to the design of WhatsApp Web, scrolling to an earlier message requires simultaneously loading all later messages, which imposes heavy resource (CPU/memory) usage. CHANG, 2020 performed an experiment to evaluate the CPU resources required to scroll and download messages from WhatsApp, their result shows that checking for messages every n hours has cost greater than $O(n)$ each time, since the CPU/memory are strained by having to load n hours of messages all at once, and cannot read/log messages as efficiently. That happens because, for example, in a group with 1000 messages, when one opens this group, firstly, only more recent messages are currently loaded (e.g., messages #990-#1000), if the user wants to load the message #1 of that chat, it means they need to keep scrolling all the messages from #990 to #2 before get access to #1 message. Using a sample of data, their work checked around 200 groups every three hours (for 48 hours total), and for each group-time pair, they recorded the number of new messages in that group, how long it took to read those messages, and how long it took to scroll to those messages. With this, they estimated the processing time to be around $O(n^2)$.

Another problem related to constantly collect WhatsApp is the need of constant Internet connectivity in both server and smartphone side where WhatsApp is installed. The design of WhatsApp requires that, for WhatsApp Web works, the smartphones must be connected as well. As the smartphone connection relies on wi-fi, it is highly susceptible to regular disconnections and interruptions, which directly affect the scrapper. Furthermore, WhatsApp Web frequently has changes in HTML and the structure of how it works, which also causes pauses in the collection. Finally, WhatsApp has an ephemeral design, in which the content is not stored for an unlimited time period. With "Temporary chat" functionality, all messages from a group using this are erased after a week. Moreover, for the remaining groups, WhatsApp does not store the media[3] for more than 15 days. Thus, even a regular user trying to see this content in his smartphone will not be able to get it anymore.

This shows that the problem of continuously collecting data from groups is not a trivial task. It can often lead to some data lost, specially in long terms collections.

[3] All media file is stored encrypted in WhatsApp server, when one wants to access it through the app or any other client, it makes a request on the server, then download it and decrypt in order to display the content to the user.

Figure 5.5: Flowchart with similar content grouping process.



Source: The Author.

Despite the challenges involving this active and constant collection of WhatsApp data, this research managed an extensive data collection that, so far, has collected data from 2018 to 2022 on an uninterrupted basis, building, as far as is known, the largest dataset of WhatsApp with more than four years of data from a substantial pool of public groups.

5.4 Measuring Popularity for WhatsApp Content

On Facebook or Twitter, for example, when someone sees a publication, it already contains various meta information given by the platform such as how many likes it had, how many shares, who were the users who posted that content. On WhatsApp, on the other hand, some content is also widely shared by its users, but the platform does not offer any aggregated data about this well shared content, and there is no simple way to visualize it in a single centralized piece of information. There, each message containing the same content is a unique and separated message, even when they are just forwarded by the users. Nevertheless, in this thesis, a methodology was developed capable of tracing a single post through the WhatsApp network and providing some aggregated metadata about the popular messages circulating within this platform.

To do this on WhatsApp and have the number of occurrences/shares of a specific post, therefore, it is necessary to process the entire data and count how many times each single multimedia message (i.e. text, audio, image or video) was shared, who are the users sending them, and what are the groups in which they appeared. However, comparing and merging each kind of media shared in WhatsApp or any other platform is a challenge by itself. But some strategies can be adopted in order to achieve this goal to aggregate the content shared on public groups on WhatsApp, such as using hashing algorithms to detect duplicates of the same message, which is used in this work. This process is better described in the flowchart in Figure 5.5.

With this architecture, it is possible to understand how similar but isolated messages can be grouped together. When two users share the same image, for example, they are saved separately by the collector. To track the dissemination of this single content through the WhatsApp network. Thus, this approach merges the content by similarity using a hash's system. For all media, then, the processing script (1) downloads the files related to the message during collection; (2) extracts the hash from each of these files; (3) creates a dictionary for each hash of which messages contain this same hash; (4) finally, groups all instances of similar messages, calculating metadata such as the total number of times it was sent, date, users, and the groups that shared this content.

For audio and video media types, it uses the hash of the MD5 checksum of each individual file to compare them and merge identical content. An MD5 is a 32-character hexadecimal hash number that is computed on a file in which if two files have the same MD5 checksum value, then there is a high probability that the two files are the same (RIVEST, 1992). This hashing method is used to index data in hash tables, for fingerprinting, to detect duplicate data or uniquely identify files, and for checksums of file integrity (ALLEN, 2013). Different from images, audio and video medias are much less susceptible to changes as regular users typically do not change the content or format, keeping the same file throughout all dissemination within the WhatsApp network. Because of this, the MD5 hash of the file is more "trackeable" and, then, is an effective methodology to find and merge identical content of those media types across WhatsApp messages.

While the raw data collection gives us hints in the direction of understanding what are the messages shared on WhatsApp, with those approaches of merging similar data, it is possible to get, for each message, independent of the kind of content, information of how many times it was shared, who are the users and which groups are sharing those messages. These are essential information to analyze dissemination of the content through the platform and provide valuable insights into the structure of the WhatsApp network, how the users operate on this enclosed platform.

For textual data, users can easily change the content of messages by adding or removing few characters, words, or even emojis to that message. However, we still want to track the spreading of texts messages and chains through WhatsApp in spite of those small modifications. To accomplish that, we merge similar content by computing the

Jaccard similarity (JACCARD, 1912) between pairs of messages. The Jaccard similarity between messages m_i and m_j is computed as the ratio of the number of common words in both m_i and m_j to the number of words in the union of both messages. Messages with a similarity greater than 0.75 were considered similar and were grouped together and considered as duplicates. This choice of threshold was made empirically after a manual inspection of a large sample of the messages. Moreover, given the scalability problem of comparing an exhaustive number of messages, we also establish a minimum size of text that we want to compare, so that only messages with more than 140 characters are observed to trace their spread across the network as those represent major pieces of information compared to small messages exchanged during a common conversation phrases on WhatsApp (e.g. Hello, Good Morning, bye, etc).

For messages containing a image, we use the pHash to detect copies of a same content. The same image that circulates through the groups on WhatsApp does not always come from exactly the same file. Image files suffer more compression, distortions and are more easily editable by users than other media types. Therefore, images require a more elaborate detection technique, as they are more susceptible to variations. Next, we explain the visual hashing approach used to group similar images used by this work.

5.4.1 Using Perceptual Hash to Process Image Data

For merging two pieces of information containing an image as a unique item, this approach uses the perceptual hash algorithm pHash (ZAUNER, 2010) to calculate a fingerprint for every image. Differently from cryptographic hashing algorithms (e.g., MD5, SHA), this kind of hash takes into account visual attributes of the file such as the pixels and RGB layers to generate a code that represents that image. Therefore, small divergences in the content will result in slightly differences between hashes, making it possible to compare hashes and see similar content. These hashing methods create an efficient way to store files as short digital hashes that can determine whether two files are the same or similar, even without the original image.

Using visual hashing for detecting similar images is an approach widely applied by researchers and also in industry, specially for digital forensics and cybercrime studies on Web (HAO et al., 2021). Its structure and easiness to compute allow the hashes to be used to explore the immense universe of online images and find abusive content on Web in various contexts for detection and analysis. For example, there are applications of perceptual hashes to detect malicious advertisements online based on webpage screenshots (VADREVU; PERDISCI, 2019), for phishing detection using Haar wavelet hashes (wHash) (MEDVET; KIRDA; KRUEGEL, 2008), to mark potential survey scams on Web by clustering perceptual difference hashes of their websites (KHARRAZ; ROBERTSON; KIRDA, 2018), for analyzing displayed ads to detect abusive and fraudulent services (NIKIFORAKIS et al., 2014; RAFIQUE et al., 2016) and also for building perceptual ad-blockers (TRAMèR et al., 2019). to find website vandalism and website defacement attacks (BORGOLTE; KRUEGEL; VIGNA, 2015), and combat different kinds of rogue software by grouping visually correspondent icons (NAPPA; RAFIQUE; CABALLERO, 2013) of malicious software and screenshots (DIETRICH; ROSSOW; POHLMANN, 2013) to find similar distributions campaigns of malware online through the use of average and perceptual hashes. Different perceptual hashing methods are also used to analyze the graphical user interface of apps to detect counterfeit apps that impersonate existing popular mobile apps in attempts to misguide users (RAJASEGARAN et al., 2019) and other app vulnerabilities such as authentication schemes (BIANCHI et al., 2017; SHI; WANG; LAU, 2019). The hash comparison can be used to identify fake profiles on social networks and recognizing identity impersonation in online social networks by matching duplicated profile photos (GOGA; VENKATADRI; GUMMADI, 2015).

Perceptual hash is also deployed in search engines studies such as malicious black-hat search engine optimization by comparing websites that look visually similar (GOETHEM et al., 2019), and different forms of hashing are also heavily used in real-world reverse image search (RIVAS et al., 2017; CHAMOSO et al., 2018)

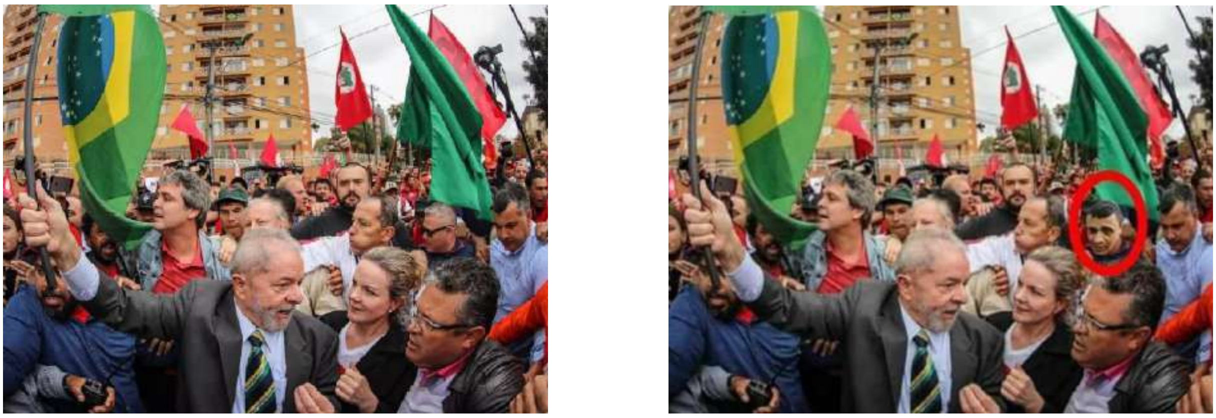
PASTRANA et al., 2019 also demonstrated an importance in usage of this kind of hashing technique to detect pornography and illicit content and, particularly, to combat eWhoring, a type of online fraud in which cybersexual encounters are simulated for financial gain; the editing of images to promote explicit contents that evade inappropriate image detectors for promoting illicit products such as sexual products or gambling websites (YUAN et al., 2019); and to flag child abuse content shared online (BURSZTEIN et al., 2019). More recently, perceptual hashing is also applied to analyze the images and memes distributed in misinformation campaigns and hate speech over the Web and social networks (AGARWAL et al., 2020; HUCKLE; WHITE, 2017; SAMANTA; JAIN, 2021; MITTOS et al., 2020; WANG et al., 2021; ZANNETTOU et al., 2018b, 2020; ABILOV et al., 2021). Therefore, this is a well established methodology to compare and detect copies of image content.

However, this technique is not without its flaws. There are some limitations regarding its application as it is possible to manipulate the image in way the hash will diverge a lot from its original source and be near to a totally different image (STRUPPEK et al., 2022) and this can be used to fool perceptual based system and even real-world search engines (HAO et al., 2021). Furthermore, in context of misinformation imagery spread online, the false information of a fabricated image can emerge exactly from just minor alterations made on the original source. Then, by using a hashing method with a great power of image generalization, we can end up detecting the original source and its edited copy as they were exactly the same content and, consequently, merging legitimate and fake content as one.

For example, Figure 5.6 shows an example of a fake image shared in 2018 Brazilian elections political context, with the original source at right and an edited false copy of it in the left. In this example, by extracting the pHash of each image, we have similar but not equal resulting hashes with a hamming distance of 2 between both images. Using average hash method, we have exactly same hashes and then 0 hamming distances between them. Using checksum, on the other hand, we have totally different hashes, as it is a crypto hashing method.

Therefore, we observe how pHash can relate both images (as they are similar), but it is still able to differentiate them. In order to avoid merging false and original images as one single piece of data as the above example, we use pHash algorithm, considering only those with exactly same hash as the same image. In this way, we can track duplicates of images keeping the peculiarities of each one

Figure 5.6: Comparison between the original source image and the edited fake version shared on WhatsApp.



(a) Original Source Image

(b) Fake Edited Image

Source: The Author.

Next, we further evaluate the impact of hashing algorithm regarding the time to data collection. Given the volume of multimedia files sent everyday on WhatsApp, another important point to choose a hash algorithm is the time required to process the actual media file attached to the message. For those experiments, a total of 905K image files, 375K video files and 80K audio files collected from WhatsApp data were processed using different hash algorithms.

In Table 5.1, 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

Table 5.1: Comparison of different visual hashing methods for processing image data.

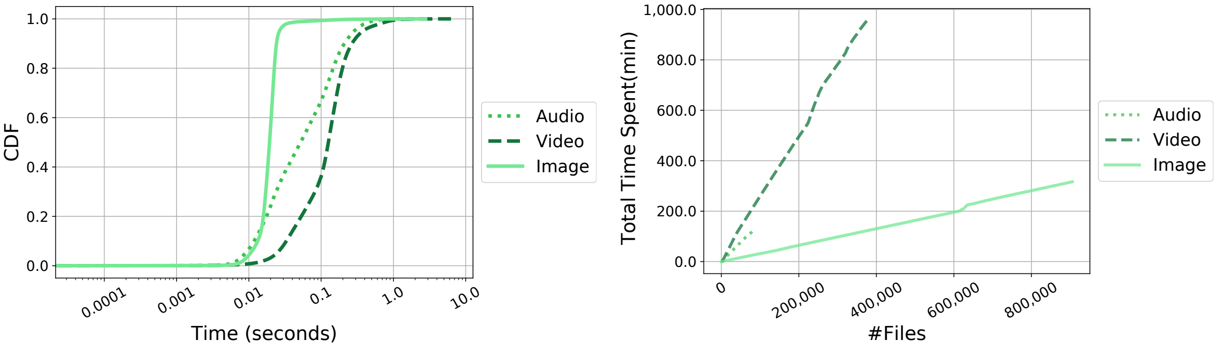
	checksum (MD5)	pHash	aHash	dHash	wHash (haar)	wHash (db4)	PDQ	Total Files
Unique Hashes	867,857	714,114	646,533	741,533	652,472	729,840	783,205	905,671
Matching Content	4%	21%	29%	18%	28%	19%	14%	
Total Time Spent (min)	316	395	359	361	1,202	1,308	14,044	

(wHash), and Facebook PDQ[4].

It is possible to note how perceptual hashes are more powerful to detect similar content compared to checksum hash. While the cryptographic hash reduce the total of unique images in 4%, perceptual hashes could match up to 29% of the files in distinct contents. Moreover, it does not take much more time to process mostly of perceptual hashes compared to the checksum method.

Figure 5.7 show the time required to process checksum hash for each different media format. Here, it is possible to highlight differences between image, audio and video files. In average, images are the fastest media to process the hash, taking mostly between 0.01 and 0.03 seconds. Video, on the other hand, are larger files, then, take more time to be processed. Figure 5.7(a) reports the total cumulative time spent to process the entire dataset of each kind of media using checksum. Even though there are more than twice as many image files as there are 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). Given the small amount of audio files compared to other two media, the time spent processing them is even smaller, taking only two hours to complete all files during the experiment.

Figure 5.7: Time needed to process the checksum hash for each multimedia type within the WhatsApp data.



(a) CDF of Average Media Time

(b) Accumulative Time Spent

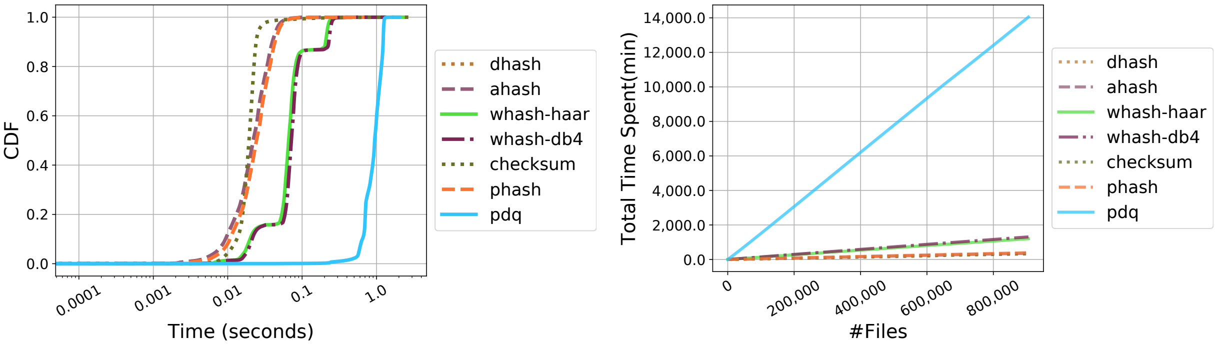
Source: The Author.

Figure 5.8 show the time required by different perceptual hash algorithms to pro-

[4] <https://github.com/facebook/ThreatExchange/blob/master/hashing/hashing.pdf>

cess the image data of WhatsApp. In Figure 5.8(a) we see the average time to process a single image file from WhatsApp data, we observe that the checksum is the fastest method to compute the hash with most of the files taking less than 0.03 seconds to be processed. However, it is not a perceptual hash. Average hash (aHash) and pHash are also fast methods, both are calculated within less of 0,1 second. The Facebook algorithm PDQ is also a very good hash method to detect images. However, the computation requires around a second. When looking at the cumulative time spent actually calculating hashes for, 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. This is not far from the checksum, which also took similar time.

Figure 5.8: Time needed by different perceptual hashes methods to process image data.



(a) Average Hash Time

(b) Total Time Spent

Source: The Author.

For those reasons, pHash was the selected method to process image data on WhatsApp, since they are a good match for detecting duplicates of the content shared on the public groups and 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 it is possible to measure distance between two pHashes, the methodology adopted by this work groups only images with exact same hash. This inflexible threshold is used do distinguish problematic images (e.g false images) that are made by making very 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.

5.5 Ethical Considerations

This work focuses only on public WhatsApp groups that discuss political topics. Only those groups available through a publicly accessible invitation URL shared on social media or other open communication channels are selected. The accounts created on WhatsApp to enter these public groups join them – manually – as normal members just like any other in that group. All profiles used here are valid WhatsApp accounts with a valid phone number (SIM chip) that legitimately participate in the groups and which even run on their own cell phones. This does not configure in any way as robot, fake, or system spoofing accounts.

Furthermore, accounts do not interact with or respond to any messages, not even in those groups that require, as a "rule", to be an active user. Also, the profile status of each account states that the account is a research persona that only observes the group. Due to these strategies, profiles can often be banned from groups, a decision that is respected during the research process. It is worth noting that the WhatsApp application itself allows any user to request and download all messages from groups he is a member of. Therefore, we are not violating any terms of WhatsApp by storing the data of groups, of which we are legitimate members. In addition, the WhatsApp company bans all accounts suspected of abusing the system, which never happened with our work, as this methodology is in accordance with its Terms of Service.

The messages gather a considerable amount of data from many users from WhatsApp groups. To ensure the privacy and anonymity of users, we do not share or disclose any Personally Identifiable Information (PII) such as cellphone numbers, username, nor even the group invite URLs used to join the groups. We use this information only to measure aggregate statistics (e.g., posts per user and number of unique users per group). Another sensitive material that can be present in our dataset are images that depict explicit violence or adult content. As it could be harmful for the users navigating in our system, we use a filter, the Yahoo Open NSFW Model[5], considering as improper all images with a score higher than 0.8, as suggested by the authors.

On *WhatsApp Monitor*, an online system that allows users to explore the entire dataset of WhatsApp messages, to avoid any misuse of even aggregated information, we limit the access of the system to a restricted number of journalists and researchers, through a login account and password. Those accounts are created only by us manually, after an analysis of the user request. Moreover, they are also informed about the data limitations and the potential bias present in the system. Since we only specifically use publicly available WhatsApp groups, our data collection does not violate WhatsApp terms

[5] <https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for>

of service.

In addition, our published studies on WhatsApp were approved by the ethics committees of MIT (MELO et al., 2019a, MELO et al., 2019b, REIS et al., 2020 and REIS et al., 2020) and the Max Planck Institute for Informatics (HOSEINI et al., 2020) in partnerships between our research group at UFMG and researchers from these universities.

5.6 Limitations

Even though adopting a solid methodology to collect data, monitoring WhatsApp for such a long period is not an easy task. Some challenges related to this process are notably relevant to mention.

WhatsApp is an ephemeral environment where new public groups often emerge as the topic of discussion changes in society. Some users may also lose interest in old groups, making them less relevant. Therefore, in order to continue collecting relevant data, it is necessary to constantly follow trends within the platform and search for new invite links and new public groups created regarding a specific context.

Furthermore, WhatsApp frequently updates its system by adding different features to chats or even how the app works, requiring us to adjust our strategy to collect data. Major changes in WhatsApp structure made scripts deprecated, requiring more changes and adaptations of the methodology. For example, enabling multiple devices connected which changed the whole infrastructure of data collection[6], the creation of temporary messages, which required greater agility in collection time[7] and new features implemented in app such as Stickers[8] and "Forwarding Many Times[9]" tag that needed to be incorporated into the collected data.

Also, such a large-scale and long term collection requires a large infrastructure; the lack of resources limits a lot the methodology such as hard drive space to store multimedia files and the need of smartphones with valid SIM cards and phone numbers to keep the WhatsApp profiles.

Besides that, we can not affirm whether our sample of groups is representative of the universe of groups on WhatsApp. Yet, WhatsApp stated that most of the conversation are private and direct between two people and only a small portion of them are from groups chats, they do not reveal any numbers regarding those groups. Therefore, our sample of

[6] <https://engineering.fb.com/2021/07/14/security/whatsapp-multi-device/>

[7] <https://blog.whatsapp.com/more-control-and-privacy-with-default-disappearing-messages-and-multiple-durations>

[8] <https://blog.whatsapp.com/introducing-stickers>

[9] <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>

groups may not reflect the characteristics of the entire set of all existing WhatsApp groups and, as we intentionally selected groups for political discussion, it probably contains a bias of activity and content from the context we study. Said that, as far we know, the data collection presented in this work is the biggest sample and comprises the longest period of WhatsApp content explored in this or any other context. Even though it could may not statistically representative, this is the closest clue we have at disposal to evaluate the WhatsApp environment, providing more transparency to such closed network.

Chapter 6

Discovering Public Chat Groups on Web

Over the past few years, online messaging platforms such as WhatsApp, Telegram, and Discord have become extremely popular (TANKOVSKA, 2021), mainly because they provide a seamless, real-time communication platform that connects billions of users from different geographies and socioeconomic statuses. These messaging platforms constitute a rich and complex ecosystem, comprising a conglomerate of various messaging platforms, each with its own unique characteristics.

In the instant messaging environment, features such as groups facilitate users to rapidly disseminate information to a large community of people. However, we lack of a general understanding of the ecosystem of public groups within the network of IMPs and the interplay between these messaging platforms and other social media such as Twitter and Facebook.

Specifically, even though we can point some technical divergence between these platforms, we do not clearly understand how they actually differ from one another in their usage or what are the characteristics and activities within the groups found in this emerging cyberspace. We also do not know if the evolution of communities formed there, whether they are ephemeral or longstanding groups, or even if there are the privacy and security issues regarding their usage. More importantly, even though groups on WhatsApp and other services proved to be very popular and used by thousands of people, there is no study in the sense of how people find those groups and how they managed to get so popular and reach so many people in such an enclosed environment. In other words, many studies focus in investigating the content within WhatsApp, but none explains how this network is built, how so many isolated users were able to constitute and join this well-connected ecosystem of messages and groups. Moreover, we do not know differences between groups from distinct instant messages platforms or how they are shared on Web nor the differences on diverse sources for finding those groups online.

As these public groups of messaging platforms are increasingly being used by non-tech savvy people or an uninformed population, the answers to these questions, including those concerning privacy, are key to limit their harm on the society.

6.1 Comparing Groups from Different Instant Messaging Platforms

For the first set of experiments, we characterize three messaging platforms' ecosystem through the lens of Twitter, a prominent social media platform. To this end, we discover public groups in these messaging platforms via Twitter and analyze their characteristics, trying to understand the pathways users follow to find and join public groups from these IMPs.

To answer the questions mentioned here, we first discover public groups in WhatsApp, Telegram, and Discord over a period of 38 days using Twitter APIs. In summary, we gather a set of 351,535 group URLs, and, for each group, collect several meta attributes (e.g., number of members in the group), once per day, to understand how the groups change over time.

We also selectively join a random sample of 616 public groups, and we gather all the messages posted in them: Overall, we collect a set of 8,255,069 messages posted by 753,329 users across the 616 groups.

Using this large corpus of data, we shed light on the discovery of public groups on Twitter and also analyze their commonalities and differences. We found the main subjects of conversation of these groups by using topic modeling, and, subsequently, we compare and contrast these topics across the discovered groups in WhatsApp, Telegram, and Discord groups. Furthermore, we conduct temporal analyses to investigate the changes in composition and activity within the discovered groups over time. Finally, we look for potential personally identifiable information (PII) exposures through these groups and discuss the privacy implications of such leaks for users.

6.1.1 Methodology and Dataset

We measure the use of different messaging platforms and identify key differences in their use. To this end, we use Twitter—a widely used social media platform—to discover groups from WhatsApp, Telegram, and Discord, and characterize the composition of and activity within these groups. Although there are some differences, the terms "groups" and "channels" are during the work interchangeably, since the distinction does not affect the analyses or findings.

Our data collection methodology consists of three steps: (1) discovering public groups from WhatsApp, Telegram, and Discord via Twitter; (2) collecting group-specific metadata; and (3) joining the discovered WhatsApp, Telegram, and Discord groups and collecting data (e.g., group metadata and messages). Below, we elaborate on each step.

6.1.1.1 Discovering WhatsApp, Telegram, and Discord groups on Twitter

All these three messaging platforms support public groups, and the most common way to invite other users to a public group is to share the group URL (also referred to as the “invite” URL) with them. The group URLs of each message platform follow one or more distinct patterns. We begin our data collection by first identifying the set of URL patterns for each messaging platform. We review each platform’s documentation and manually examine the URLs of each platform to compile a list of six patterns employed across these messaging platforms.

On WhatsApp, as discussed in Chapter 5, invitation URLs for public groups have the pattern “`chat.whatsapp.com/<gID>`” with `gID` representing a unique identifier of the group, which is automatically generated by the WhatsApp messenger application when the group is created. The remaining patterns have the following prefixes or `host` values: `t.me/`, `telegram.me/`, and `telegram.org/` for Telegram groups and channels, `discord.gg/`, and `discord.com/` are the URL formats of the invitation for Discord servers.

We search for the occurrences of the above URL patterns between April 8 and May 15, 2020 on Twitter, using two different approaches: (a) using Twitter’s Search API (TWITTER, 2020a) every hour, and (b) using Twitter’s Streaming API (TWITTER, 2020b). The former retrieves all matching tweets (i.e., tweets containing the URL patterns) that were shared during the past seven days (i.e., from the time at which the query was issued), while the latter retrieves matching tweets in real time, as they are posted on Twitter. We merge the tweets obtained via both APIs, since a preliminary investigation revealed discrepancies between the tweets retrieved using the two APIs.

Using the above approach, we discover 351,535 group URLs (belonging to the three messaging platforms) from 2,234,128 tweets posted by 806,372 Twitter users (refer the left side of Table 6.7). Per this table, we discover a larger number of group URLs from Discord (227K) than either Telegram (78K) or WhatsApp (45K). The large number of Discord and Telegram groups discovered despite these platforms being smaller (in terms of number of users) than WhatsApp, suggests that these two platforms perhaps have greater channel diversity and public accessibility compared to WhatsApp; they both also have less strict limits on group sizes compared to WhatsApp. We discover the largest number

Table 6.1: Overview of our datasets.

	Twitter			Messaging Platforms		
	#Tweets	**#Users**	**#Group URLs**	**#Joined Groups**	**#Messages**	**#Users**
WhatsApp	239,807	88,119	45,718	416	476,059	20,906
Telegram	1,224,540	398,816	78,105	100	3,148,826	688,343
Discord	779,685	340,702	227,712	100	4,630,184	52,463
Total	2,234,128	806,372	351,535	616	8,255,069	761,712

of groups of Discord, presumably owing to Discord group URLs automatically expiring after a day (SAM, 2020); users, hence, are likely sharing a large number of unique group URLs compared to the other messaging platforms.

Control dataset.

The use of Twitter as the only data source for discovering public groups of the different messaging platforms potentially introduces some bias in our sample. Where applicable, we clearly state the implications of sample bias for inferences, and also provide a control dataset to facilitate an accurate interpretation of our results. Said that, we make the best effort to mitigate potential biases that might affect our findings.

Therefore, we compare the tweets' dataset, where applicable, against a control baseline dataset. The control dataset comprises a random sample of 1% of all 1,797,914 tweets also posted between April 8 and May 15, 2020 and obtained via Twitter's 1% Streaming API to better compare it to our group collection. In this case, we use the Streaming API without limiting the results to a list of matching patterns or keywords, and obtaining a 1% random sample of all tweets.

6.1.1.2 Collecting group-specific metadata

Although we can join a messaging platform's group given its group URL, we refrain from joining hundreds of thousands of groups for three practical reasons. First, there is a limit on the number of groups a user can join, before getting banned from the messaging platform. Empirically, we find a limit for WhatsApp is between 250 and 300 groups per user and also limitations to join multiple groups simultaneously in a short interval of time, while, on Discord, it is up to 100 servers. Also, in Telegram, there is a limit of 500 groups a user can join with one account Second, the above limit translates to a need for hundreds of phones and SIM cards and accounts to join all discovered groups, limiting the scale as well as scope of the study. Third, we intend to minimize disruptions caused by joining hundreds of thousands of groups on any messaging platform. We, hence, take a more pragmatic approach to obtain metadata from each group without joining every one of them. Note that this is also different from collection methodology described in previous Chapter Below, we explain our approach.

WhatsApp. We use WhatsApp's Web client to obtain basic information about a WhatsApp group without joining it. Specifically, we automate the process of clicking on a WhatsApp group URL and opening the landing page for the group on a browser, using a Python script with Selenium library to manipulate the Internet browser. We follow the invite URL to the group until the point of refraining from the clicking the actual "Join" button on the landing page, but still scrape the resulting page to gather several details: (1) title of the group; (2) size of the group (at the time of visiting the landing page); (3) country code of the phone number of the group's creator; (4) phone number of the group's creator; and (5) the textual description of the group. We only store the hash of the phone number, although it is available to anyone with access to the group URL.

Telegram. Similar to the method for WhatsApp, we use Telegram's Web client to obtain basic information about Telegram groups without joining them. We implement a custom scraper that obtains and parses the web page for each group to gather several details: (1) title of the group; (2) size of the group and the number of members online (at the time of visiting the group's web page); (3) whether the "chat room" is a channel or a group; (4) the textual description when it is available.

Discord. For obtaining metadata about Discord groups, we use the platform's REST API (DISCORD, 2020a). For each group found from this platform, we collect the (1) title of the group, (2) number of members—both in total and online—in the group, and (3) group creator and group creation date.

We follow the aforementioned techniques to gather metadata from each group on all three messaging platforms repeatedly every day from April 8 through May 15, 2020. We commence the metadata collection for each group from the date, when we discovered it and repeat it every day unless the URL is revoked; landing pages of revoked URLs clearly indicate the revocation. By collecting data every day during this period, we were able to track the status of each group (i.e., check if the group URL is alive or revoked) and the number of members in the group over time starting from the discovery date.

6.1.1.3 Analyzing group composition and activity

For a subset of the discovered groups, we supplement the basic group metadata with details on the structure of and activity within the groups. To this end, we select a set of group URLs uniformly at random and join them using an account for each platform. Below, we also describe how we obtain data from within the groups on every messaging platform.

Telegram. Telegram, unlike WhatsApp, provides a public API for gathering data on groups (TELEGRAM, 2020). We select 100 URLs uniformly at random and join them with a new account. For each group, we collect (1) messages shared on the groups (since the group was created), (2) creation date of the group, and (3) user profiles for the members of the group. A group administrator may opt to hide the member list from the group, and we obtain, hence, the member list only in 24 groups (out of the 100) where administrators did not exercise this option.

Discord. Although Discord provides an API for developing bots to help manage groups (e.g., run commands or send automatic messages), such a bot application has limited access to the public groups. A bot is disallowed, for instance, from joining a group, albeit the group's administrator can add the bot to the group. To address the issue, we automate the process of opening the landing page of a group and joining it with a dedicated user account. We join 100 random servers (the maximum number of servers that a single user can join) and, using an application created with the user account, obtain the following data through the Discord API (DISCORD, 2020b): (1) messages on all groups on the joined servers (since the data each group was created) and (2) user profiles for the group members.

WhatsApp. As mentioned already, WhatsApp does not provide an API to join groups or retrieve messages from within a group. As a consequence, we rely on our proposed methodology in this thesis to join the groups and collect data within these groups (HASE, 2018). In total, we select and join 416 random WhatsApp public groups for this specific task. Only after joining a group it provides us with several pieces of information that are otherwise inaccessible (i.e., inaccessible without joining the groups): (1) messages shared on the groups (WhatsApp gives access to messages shared on the group, after our joining date); (2) phone numbers of the members of the group (For privacy reasons, we store only a hash of the phone numbers); and (3) creation date of the group.

6.1.2 Analyzing Public Groups Ecosystem

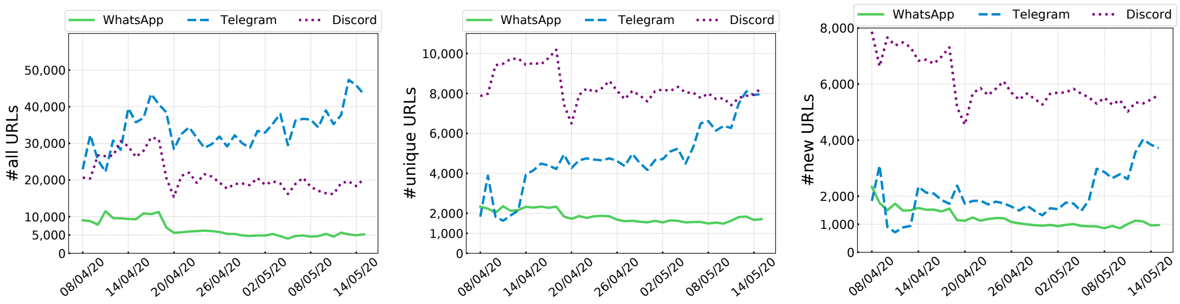
In this section, we analyze the tweets found that contain group URLs from WhatsApp, Telegram, and Discord for understanding the interplay between Twitter and these messaging platforms. The tweets also provide some context on the shared groups. We analyze how public groups are shared over time on Twitter, the prevalence in use of various Twitter features (i.e., hashtags, mentions, and retweets) when sharing groups, and main themes of these groups by performing topic modeling on the content of the tweets.

6.1.2.1 Group Sharing Dynamics

We begin our analyses with the number of group URLs discovered on Twitter for the three messaging platforms (see Figure 6.1), showing that this social network is a rich data source for discovering public groups on the different messaging platforms. We report three different metrics: (a) all group URLs discovered on Twitter; (b) the number of unique group URLs per day; and (c) the number of *new* group URLs per day (i.e., excluding group URLs already observed on previous days).

WhatsApp appears, per Figure 6.1, to be the most "private" messaging platform: We discover fewer group URLs belonging to WhatsApp than that of Telegram and Discord, despite WhatsApp being a much larger and widely used messaging platform. This observation perhaps suggests that WhatsApp users are less willing to share public group URLs on Twitter compared to Telegram and Discord users. Second, we discover the largest number of group URLs for Telegram (Figure 6.1(a)), with 33,864 group URLs, in the median, per day, followed by Discord with 19,970 URLs. In terms of unique group URLs discovered each day (Figure 6.1(b)), Discord, however, surpasses Telegram (8,090 URLs vs 4,661 URLs, in the median). These findings indicate that Telegram groups are shared more number of times than that of Discord and WhatsApp, within the same day (see Figure 6.1(a) and Figure 6.1(b)). The number of newly discovered group URLs per day (Figure 6.1(c)) indicates that Telegram group URLs are likely to also be shared across several days. Overall, we find that Twitter is a rich source for discovering public groups of messaging platforms.

Figure 6.1: Number of group URLs discovered on each day during our Twitter data collection.



(a) All

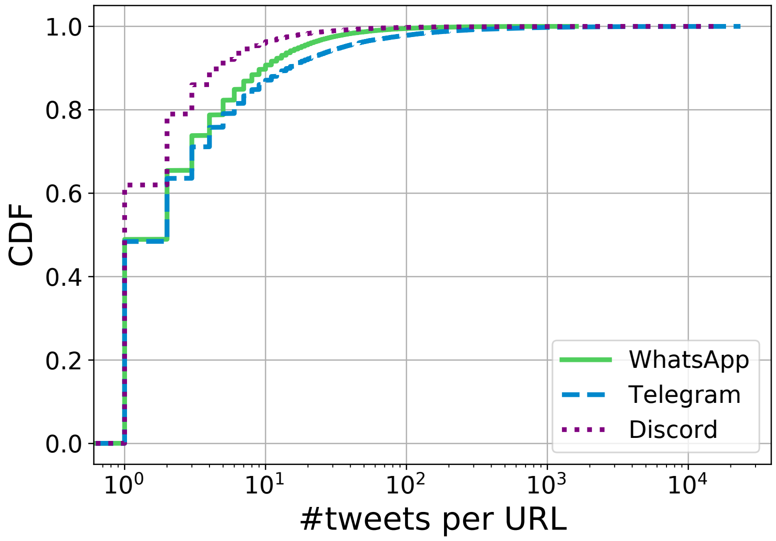
(b) Unique

(c) New (not seen in previous days)

Source: The Author.

Figure 6.2 shows the number of times that each group URL is shared on Twitter. Approximately half of the group URLs from WhatsApp and Telegram are shared only once, compared to 62% of the URLs in Discord. Overall, on average, each WhatsApp and Telegram group URL is shared in more tweets compared to Discord. We observe a few Telegram groups (14 in total) that were shared on a large number of (i.e., more than 10K) tweets. We find, via manual examination, that 11 groups focus on pornography and 2 on cryptocurrencies, and one to be a general discussion group.

Figure 6.2: CDF of number of tweets for each group URL over the entire dataset.



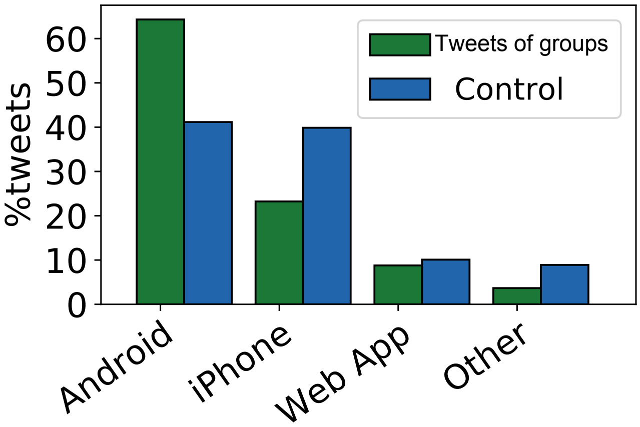
Source: The Author.

Another analyses in Figure 6.3 show from which device users post about WhatsApp groups. Majority of the tweets containing WhatsApp group URLs are sent via Twitter's mobile clients for Android (64%) and iPhone (23%). Since the IMPs are primarily mobile messaging apps, users perhaps share the group URLs via Twitter from the same device that they use the application. Users in the control dataset also favor mobile applications for Twitter; the prevalence of Android and iPhone, however, have roughly the same values—41% and 39%, respectively.

6.1.2.2 Twitter Features Analysis

For characterizing the tweets, we use three widely used Twitter mechanisms for content broadcasting and discovery: *hashtag*, *mention*, and *retweet*) as show in Figure 6.4. An *hashtag* on Twitter is a keyword associated with a tweet that conveys a topic or theme or event of interest. Users can discover tweets on a given topic by searching for a relevant hashtag, and it allows Twitter to group tweets by hashtags and broadcast

Figure 6.3: Client used by users to send their tweets.

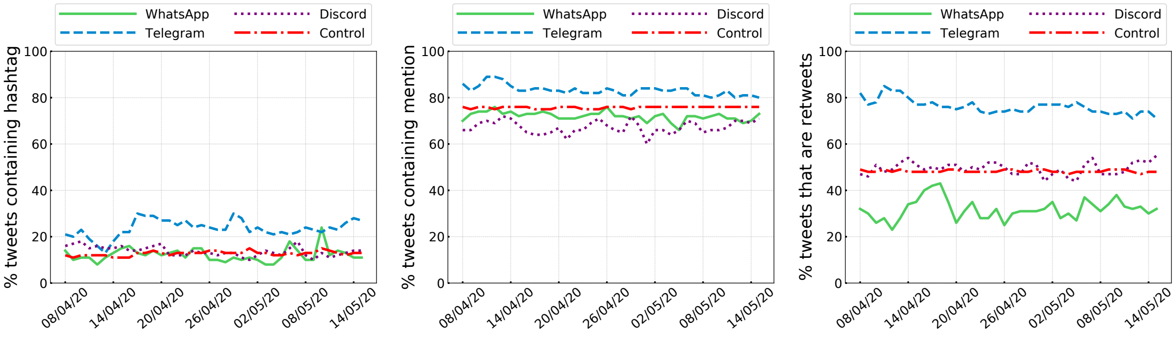


Source: The Author.

them to interested users. *mentions* support a "controlled" broadcast. A mention allows a user to refer to one or more users in the tweet who will be notified when the tweet is shared, increasing the likelihood of those users to read and also respond. In the same vein, a *retweet* is a broadcast of a specific tweet to all the followers of the "retweeting" user. Next, we analyze the prevalence of the use of these mechanisms in the tweets that include group URLs from the three messaging platforms.

Twitter users tend not to use hashtags and use mentions when sharing WhatsApp, Telegram, and Discord groups. Tweets with Telegram groups are more likely to get retweeted compared to the other platforms. Per Figure 6.4(a), only a small percentage of tweets include hashtags for all three messaging platforms. Specifically, tweets containing Telegram group URLs are more likely to include hashtags (24% of these tweets include hashtags), while for the other two messaging platforms as well as the control dataset we observe a lower percentage of tweets with hashtags (13% for WhatsApp, 14% for Discord, and 13% for control). The lack of hashtags could perhaps be due to users intentionally

Figure 6.4: Percentage of tweets that contain hashtags/mentions and percentage of tweets that are retweets.



(a) Hashtag (b) Mentions (c) Retweets

Source: The Author.

restricting the tweets' visibility to their followers. Given the relatively low limit on the size of WhatsApp groups, for instance, users might intend to share a WhatsApp group only with a few other people; tweets with WhatsApp groups, per Figure 6.4(a), contain fewer hashtags than those with Telegram and Discord groups. We also find that only a small percentage of tweets include more than one hashtag: 4% for WhatsApp, 10% for Telegram, 7% for Discord, and 5% for the control dataset.

When analyzing tweets with mentions (see Figure 6.4(b)), we observe a larger percentage of tweets with mentions compared to that in the control dataset and the other messaging platforms (73%, 84%, 68%, 76% for WhatsApp, Telegram, Discord, and control, respectively), likely because Twitter users are selective about the people they invite to their groups, despite the fact that they are sharing tweets in a public space. We also investigate the number of mentions per tweet, finding that in general only a small percentage of tweets include more than one mention; 20% for WhatsApp, 14% for Telegram, 15% for Discord, and 12% for the control dataset.

Lastly, our analysis of retweets (see Figure 6.4(c)), shows that a smaller percentage of retweets for WhatsApp (33%) than that for Telegram (76%) and Discord (50%). Twitter users are more likely to retweet posts containing group URLs from Telegram and Discord, as these platforms are probably considered more public than WhatsApp.

6.1.2.3 Topic Modeling

Next, we focus on understanding the context around the sharing of group URLs by analyzing the text of the tweets. First, we analyze the various languages that exist in our dataset. To this end, we use the language field as returned by Twitter's APIs, and observe that English is the most popular language with. Figure 6.5 shows the percentage of tweets in each language across the three messaging platforms. English is the most popular language on tweets sharing groups, with 26%, 35%, 47% for WhatsApp, Telegram, and Discord, respectively. For WhatsApp, the second and third most popular languages are Spanish (16%) and Portuguese (14%), while for Telegram it is Arabic (15%) and Turkish (8%). Interestingly, we find Discord users have a substantial number of Japanese users, as 27% of all tweets with Discord group URLs are in Japanese. These results shed light into the demographics of the users sharing the public groups and using the groups on the messaging platforms.

To better grasp the context of the shared groups, we first extract all tweets posted in English and perform topic modeling using Latent Dirichlet Allocation (LDA) (BLEI; NG; JORDAN, 2003). First, we focus on English, since it is the most popular language

Table 6.2: Topics extracted from the English tweets that include WhatsApp group URLs.

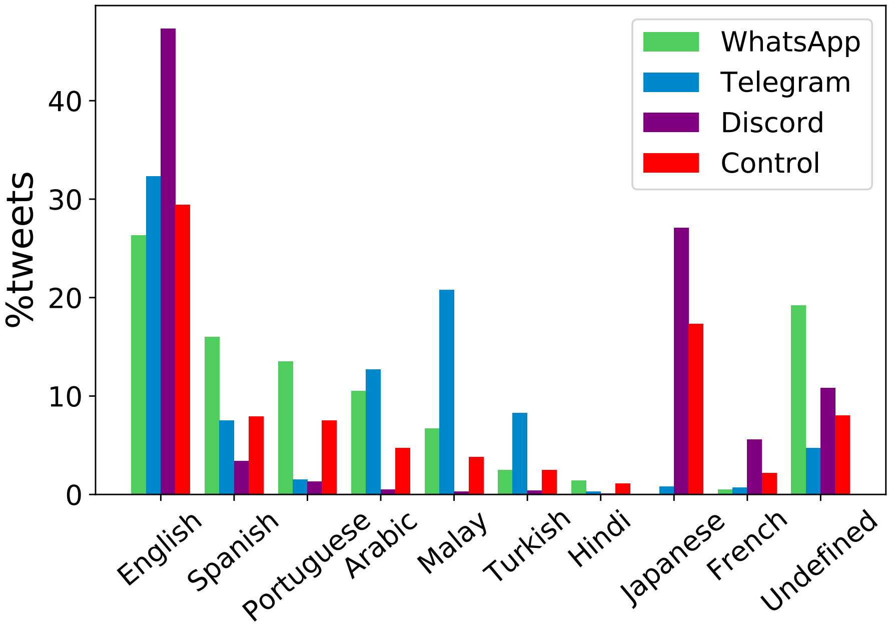
#	Label	Topic terms
1	**Forex Training**	learn, free, forex, training, join, trading, text, mini, class, animation
2	**Earn money from home**	home, earn, don, just, money, using, can, start, stay, google
3	**Instagram Followers Boosting**	join, followers, instagram, gain, want, money, online, group, learn, make
4	**Cryptocurrencies**	bitcoin, ethereum, crypto, currency, ads, year, like, line, people, new
5	**Earn money from home**	make, can, money, know, daily, home, earn, forex, cash, market
6	**Cryptocurrencies**	learn, cryptocurrency, make, join, days, period, another, want, day, accumulate
7	**WhatsApp group advertisement**	join, group, whatsapp, link, follow, click, please, chat, open, twitter
8	**Making money**	get, never, time, actually, income, chat, best, taking, account, full
9	**-**	will, new, retweet, capital, people, now, interested, writing, nigerian, online
10	**Cryptocurrency courses**	business, ethereum, free, smart, skills, eth, million, join, training, webinar

Table 6.3: Topics extracted from the English tweets that include Telegram group URLs.

#	Label	Topic terms
1	**Cryptocurrencies**	bitcoin, join, sats, get, winners, sex, hours, chat, nice, come
2	**Cryptocurrencies**	usdt, giveaways, oin, winners, ollow, enter, btc, trc, trx, hours
3	**Social Network Activity**	follow, like, retweet, giveaway, tag, join, win, twitter, friends, friend
4	**Ask Me Anything/Quiz**	ama, may, will, utc, quiz, someone, wallet, don, ust, today
5	**Advertising Telegram groups**	free, join, just, telegram, money, day, channel, don, can baby
6	**Online Sex workers**	new, worth, user, brand, xpro, performer, smartphones, girls, boobs, price
7	**Giveaways**	giving, away, will, tmn, link, honor, full, butt, video, get
8	**Sex**	fuck, want, girl, click, show, trading, pussy, powerful, can, cum
9	**Advertising Telegram groups**	telegram, join, group, channel, now, below, link, get, available, opened
10	**Referral Marketing**	airdrop, open, https, tokens, wink, referral, token, earn, new, good

for tweets including group URLs for all three messaging platforms. For each platform, we extract all the English tweets, remove stop words, and extract ten topics using the LDA method. We report the topics extracted from the tweets sharing WhatsApp (Table6.2), Telegram (Table6.3), and Discord (Table6.4) groups. For each topic, we manually assess

Figure 6.5: Percentage of tweets containing public groups of IMPs for each language.



Source: The Author.

Table 6.4: Topics extracted from the English tweets that include Discord group URLs.

#	Label	Topic terms
1	**Gaming**	patreon, free, get, today, mystery, public, gaming, gamedev, indiegames, alongside
2	**Organizing online events**	will, may, hosting, week, one, time, tonight, don, night, last
3	**Gaming**	like, oin, alpha, deal, daily, art, lots, battle, raffle, nintendo
4	**Advertising Discord groups**	discord, join, server, link, can, visit, want, just, new, hey
5	**Pokemon**	united states, venonat, bite, quick, bug, full, fortnite, pikacku, confusion
6	**Advertising Discord groups**	giveaway, follow, retweet, friends, tag, join, discord, enter, fast, winners
7	**Tournaments**	good, live, launching, now, tournament, open, next, will, free, prize
8	**Giveaways**	giving, est, away, awp, will, saturday, friday, coins, many, competition
9	**Advertising Discord groups**	discord, join, make, sure, ends, chat, token, https, music, server
10	**Hentai**	join, discord, server, come, hentai, now, new, paradise, tenshi, official

the extracted topic terms and provide a high-level label, and we also report the percentage of tweets that match each topic. The extracted topics can be categorized into three types: (1) *micro topics* that refer to topics that are specific to a single messaging platform; (2) *meso topics* that refer to topics that exist to more than a single messaging platform; and (3) *macro topics* that refer to topics that exist across all messaging platforms.

For micro topics, we observe Forex Training (6% tweets), earning money from home (21%), and Instagram followers boosting (9%) topics on WhatsApp (see topics 1, 2, and 3, respectively, sex-related topics on Telegram (23%, see topics 6 and 8), and gaming (12%) and hentai-related (Japanese anime and manga pornography, 9% of all tweets) topics on Discord (see topics 1, 3, and 10). We find several meso topics related to cryptocurrencies on both WhatsApp (18%, see topics 3, 6, 10) and Telegram groups (18%, see topics 1 and 2), but not for Discord. Finally, for macro topics, we observe that across all messaging platforms there are topics where Twitter users try to persuade people to join their groups. For instance, see topic 7 for WhatsApp topics (30%), topics 5 and 9 for Telegram (25%), and topics 4, 6, and 9 for Discord (47%).

Interestingly, during our LDA analysis in English, we do not find any politics-related topics.[1] This highlights that Twitter users are not sharing many politics-related groups from messaging platforms in English, or if they do, they do not make it clear from the tweet's accompanying text.

Finally, we repeat the same analysis for other popular languages like Spanish and Portuguese, but omit the results due to space constraints. We find some topics that do not emerge in our English analysis, mainly due to the COVID-19 pandemic (in Spanish for WhatsApp and Telegram) and politics-related groups (in Spanish for Telegram and in Portuguese for WhatsApp).

Overall, our LDA analysis allows us to obtain insights into the content of the discovered messaging platforms' groups by analyzing the text in the tweets sharing the group URLs. The extracted topics indicate that there are some similarities across the use of messaging platforms, while at the same time there are some topics where users prefer

[1] We repeated our analysis with a larger number of topics (up to 50 topics per messaging platform) and no politics-related topic emerged.

specific messaging platforms to discuss them.

With this analysis, we found that Twitter is a rich data source for discovering groups from WhatsApp, Telegram, and Discord. The results reveal that users prefer to avoid using hashtags and only mention a small number of users in their tweets when sharing content about WhatsApp, Telegram, and Discord groups. Also, by performing topic modeling in the tweets, we find differences in the groups that are shared on Twitter from WhatsApp, Telegram, and Discord. Specifically, WhatsApp and Telegram are used for cryptocurrencies discussions, Telegram for disseminating pornographic content, and Discord mainly for gaming, giveaways, tournaments, and hentai.

6.1.3 Changing and Evolution of Public Groups

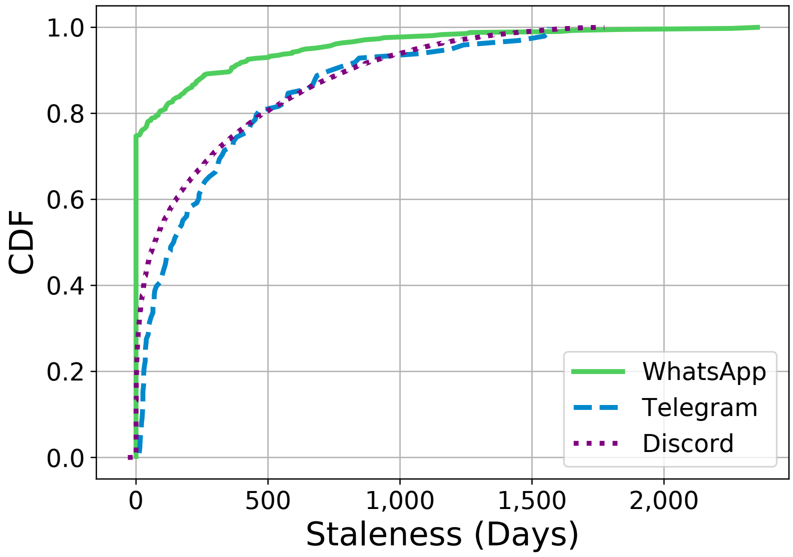
In this section, we analyze the data obtained from the WhatsApp, Telegram, and Discord groups discovered from Twitter, with a focus on understanding the characteristics of those groups, how they change over time, and the volume of information disseminated within them.

6.1.3.1 Group Creators

For all groups from WhatsApp and Discord, the information about the creator of the group is available even without joining those groups. On the other hand, for Telegram, we are only able to obtain information about the creator for the 100 groups we join. We find that 34,078 different users created groups on WhatsApp, 49,753 users created groups on Discord, and 100 users created groups on Telegram. Also, we find that most of the users create a single group (100% for Telegram, 95.9% for Discord, and 92.7% for WhatsApp), with only a small percentage of users creating 2 groups or more (5.3% for WhatsApp and 3.6% for Discord). Despite that, we find users who create a large number of groups (e.g., a single user created 61 groups on Discord and another one 28 groups on WhatsApp). The number of users creating multiple groups on WhatsApp is larger compared to the other platforms, and this is likely due to the imposed group limit (257 members). To overcome this limit, WhatsApp users are creating multiple groups with similar topics with the goal to reach a larger audience.

Group Creation Dates. Next, we analyze the creation dates for the groups. For Dis-

Figure 6.6: Staleness: time difference between the appearance of the group on Twitter and its creation date.



Source: The Author.

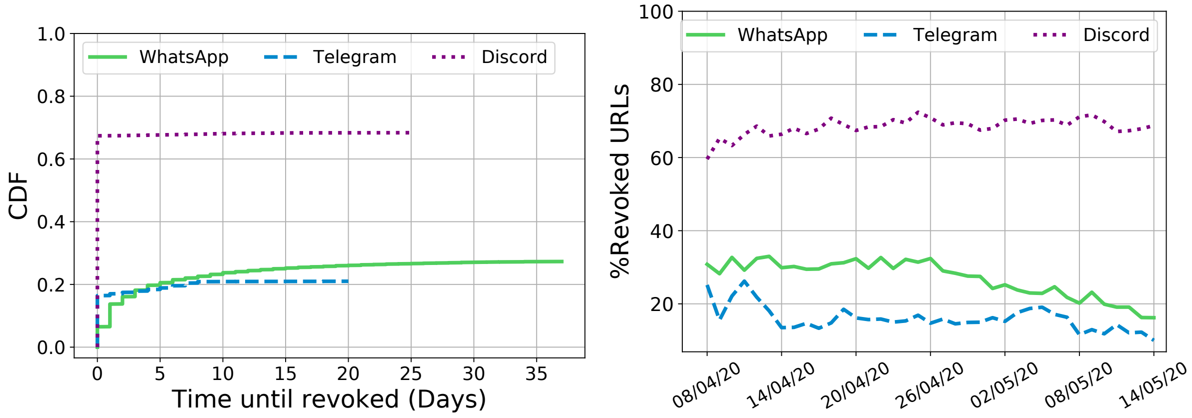
cord, the creation date is available without the need to join the groups, yet for WhatsApp and Telegram we only obtain this data after joining the groups (416 for WhatsApp and 100 for Telegram). Based on the creation date, we can calculate how old the groups are at the time they are shared on Twitter. We define *staleness* as the time interval, in terms of days, between the creation date of a group and the date at which the group is shared on Twitter. In Figure 6.6, we observe that most of the WhatsApp groups are created and shared on Twitter on the same day (76%), while for Telegram and Discord less than 30% of the groups are shared during the groups' creation day. Older Telegram and Discord groups are shared on Twitter, while shared WhatsApp groups are "fresh".

Also, only 10% of WhatsApp groups are older than one year compared to 29% and 25.6% of the groups for Telegram and Discord, respectively. The oldest group from our dataset, though, is from WhatsApp – a six-year-old group from Kuwait about the Real Madrid football team. Overall, these findings indicate that Twitter users tend to advertise older Telegram and Discord groups, compared to WhatsApp groups, and this is likely due to WhatsApp's imposed member limit (i.e., WhatsApp groups become full, hence not shared on Twitter to attract more members).

6.1.3.2 Group Countries

Since we store the country code of the creators' phone numbers for WhatsApp groups, we can investigate the group's country of origin. Note that for Discord, we do

Figure 6.7: Analysis on how long groups are accessible until they get revoked. A substantial percentage of groups are inaccessible during our first observation (especially Discord groups).



(a) Accessible period of URLs (b) Revoked URLs per day

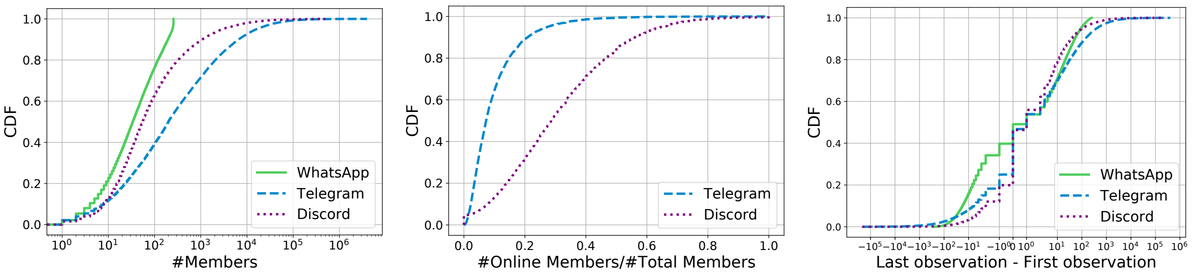
Source: The Author.

not have any information regarding phone numbers, while for Telegram we have phone numbers for only a small percentage of users (see Section 6.1.5), and hence we limit this analysis on WhatsApp. A large number of WhatsApp groups are created by users from Brazil (BR) with 7,718 groups, followed by Nigeria (4,719), Indonesia (3,430), India (2,731), Saudi Arabia (2,574), Mexico (2,081), and Argentina (1,366). Although India is the country with the largest number of WhatsApp users (340 million, followed by Brazil with 99 million (TANKOVSKA, 2021), it is only the 4th most popular country in our dataset. This is perhaps because our WhatsApp groups are only the ones shared on Twitter (Twitter has 8.15 million users in Brazil and 7.91 million in India (DIXON, 2022)).

6.1.3.3 Group Revocation

On all platforms, a group URL can be revoked either manually, by an administrator, or automatically when all members leave the group or if the group URL expires (e.g., on Discord). Once revoked, no new users can use the group URL to join the concerned group, and the landing page is devoid of any details except for the revocation notice. We monitor those URLs for their status and the number of their members, every day to analyze the behavior of the groups over time. Although we cannot precisely determine whether a revocation was manual or automatic, the lifetime of a group—defined as the time from discovery on Twitter until it is revoked—impacts our approach of characterizing groups based on the metadata from the landing page of its group URL. Figure 6.7(a) shows the accessibility time (in days) for the revoked URLs, while Figure 6.7(b) shows the percentage of revoked group URLs per day. We find that 27.3% of the URLs for WhatsApp groups, 20.4% of the Telegram group URLs, and 68.4% of the Discord group URLs are revoked at some time. This shows that Discord has much more revoked URLs, probably because, by default, group URLs auto-expire after a day, while a group URL from Telegram and WhatsApp lasts until the user manually revokes it or deletes the group. Therefore, Discord groups are less accessible through group URLs, while the URLs we find for Telegram and WhatsApp are more likely to be accessible. Looking at the lifetime, the time period a URL is accessible, we can observe that for many of the revoked URLs, the revocation is done before our first observation (6.4% of all groups for WhatsApp, 16.3% for Telegram, and 67.4% for Discord). This indicates that some groups have a very limited accessible period, indicating the ephemeral nature that messaging platforms' groups have. The ephemeral nature of messaging platforms' groups should be taken into consideration in future research focusing on collecting and analyzing datasets from messaging platforms.

Figure 6.8: Distributions of number of members per group, percentage of online members over all members, and group size change over time (between first and last observation).



(a) Total size of groups (b) Fraction of online members (c) Group size difference over time

Source: The Author.

6.1.3.4 Group Members

Since users share group URLs on Twitter to entice others to join, the size of a group over time can hint of their activity and the reasons behind its revocation. To this end, we gather the number of members in each group, for each day, that are accessible. We compare the distribution of total amount of members for each platform in Figure 6.8(a). Overall, WhatsApp has much fewer members compared to the other two, because of the group size limit of 257 members. It is also worth noting that only a small percentage of WhatsApp groups (5%) reach the limit of the size. Also, we observe that Discord has fewer members than Telegram, as around 60% of Discord groups have less than 100 members while only 40% of the Telegram have the same amount. For Telegram and Discord, we also have information about how many users within the group are actively online (provided by the platform itself via the Web client). We use this information, from our first observation, for each group to analyze the proportion of online members. Figure 6.8(b) shows that even though Telegram has more members in total, they are online in less proportion compared to Discord. We observe that around 15% of the groups on Discord have more than half of their members online, while on Telegram only a few groups have such activity. These results are likely due to the fact that Discord is a more computer/desktop-oriented platform, while Telegram is frequently used from mobile devices, hence Discord users are more likely to be online compared to Telegram users.[2].

Finally, we investigate the growth of the groups over time; Figure 6.9(a) shows the distribution of the growth of the groups, which is the difference of group sizes observed on the first and the last day (i.e., prior to revocation) of observation. We can clearly observe the impact of the limit sizes for each platform in the distribution of the growth of the groups. Discord and Telegram have groups that change in more than 100,000 members during our analysis period: e.g., a Discord group for fans of the new Nintendo game "Animal Crossing" launched in March 2020 and a Telegram channel that shares movies. We can also note that there are more groups increasing in size than decreasing (51% for WhatsApp, 53% Telegram and 54% Discord). This likely indicates that sharing the group URLs on Twitter helps the groups to aggregate more users. Still, some groups decreased in size (38% for WhatsApp, 24% Telegram and 19% Discord), perhaps an indication of a declining interest among the members of some groups over time.

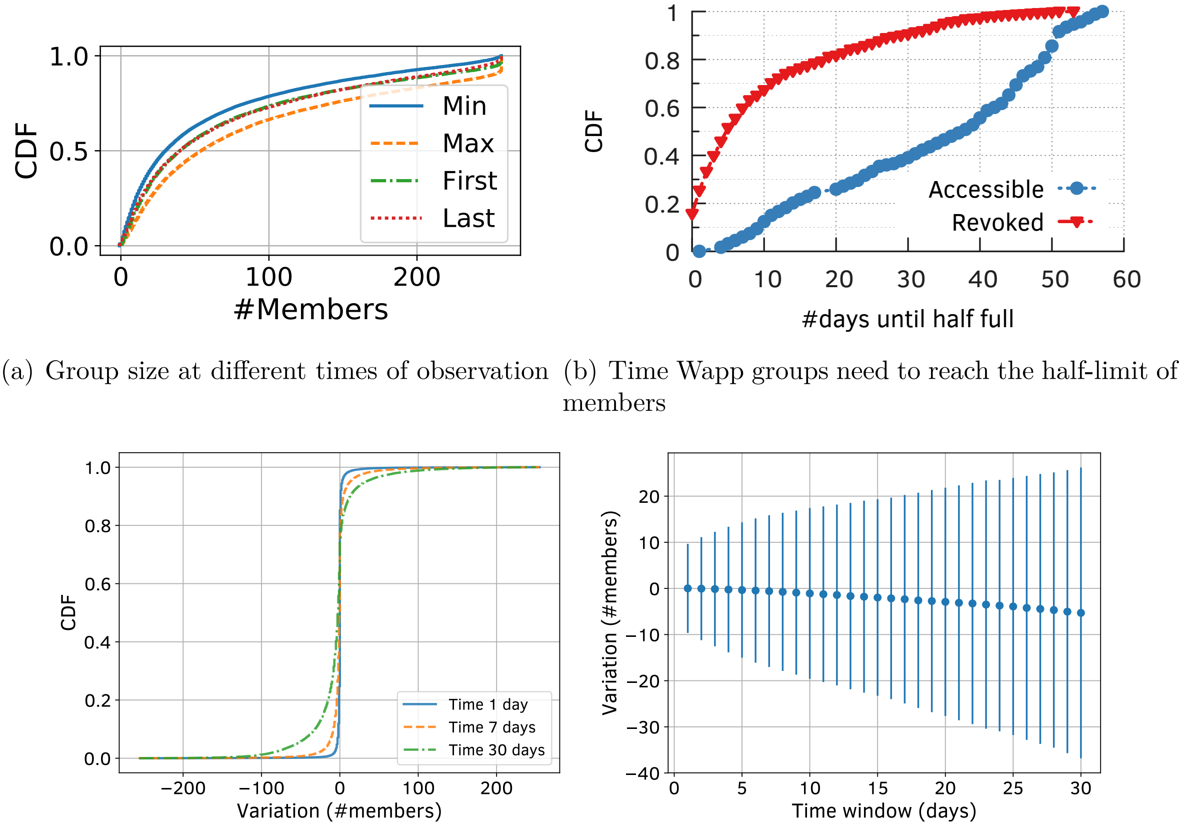
How WhatsApp groups size changes over time? Regarding the WhatsApp public groups, we explore further details of the groups posted on Twitter, by investigating how the groups changes over time.

Since users share WhatsApp group URLs on Twitter to entice others to join, the size of a group over time can hint us the evolution of this group, and for example, the reasons behind its revocation. To this end, we gather the number of members in each group, once per day from the day we discover them and until they are accessible. Figure 6.9(a) shows the distribution of each group sizes observed on the first as well as the last day (i.e., prior to revocation) of observation. The plot, in addition, shows the distribution of the maximum ("Max") and minimum ("Min") group sizes, irrespective of when a given group reached its maximum or minimum size. That first and last days of observation of a group reveals similar size suggests that groups stabilize over time. The difference between maximum and minimum sizes, however, indicate that there is substantial churn in group size, perhaps over shorter time frames. To assess the statistical significance of these distributions, we run a two-sample Kolmogorov-Smirnov test on each pair of distributions: we find that each pair exhibits statistically significant differences with $p < 0.01$.

If we vary the observation time span, from 1 day to 1 week and, finally, to 1

[2] Note that a Discord user is shown online even if the Discord Web/desktop client is running in the background.

Figure 6.9: Changes of size over time.



(a) Group size at different times of observation

(b) Time Wapp groups need to reach the half-limit of members

(c) Variation of groups in different time windows

(d) Average number of member per days of observation

Source: The Author.

month, we observe hardly any change in group size on a per-day or per-week basis, as shown in Figure 6.9(c). Group change size mostly when observations are a month apart, but even then, approximately 40% of the groups exhibit virtually no change; another 40% show a decrease in size, perhaps indicative of a declining interest among the members of any group, to remain a member, over time. Only around presented 20% of the groups discovered actually increased the size a month after the first observation. In a complementary way, the Figure 6.9(d) shows the mean and standard deviation of the variation of the groups by number of observation days. While there is an increasing deviation (i.e., groups variate size more after 30 days than in the first day), we can note a slight decrease in the mean of the sizes of groups to become even negative variation overall.

The distribution of size changes, though, do not seem to suggest that all the group URL are revoked because of all members leaving the group. We expect, if that were to be the case, the distribution of sizes observed on the last day to overlap or be much closer to the minimum size distribution.

For the increasing groups and those that reach at least in 128 members, to measure the rate at which a group grows in size, we measure the number of days (since its discovery date on Twitter) it takes for a group to become half full (i.e., have 128 members, excluding the creator), comparing accessible and revoked groups in this analysis. Per Figure 6.9(b) we observe that revoked WhatsApp groups (i.e., before they were revoked) become half full much faster than accessible groups: In the median, while accessible groups took more than a month to become half full, revoked groups only took less than a week. The plot perhaps suggests that the URLs can be revoked by administrators as the groups quickly grow in size.

6.1.3.5 Group Messages

Next, we analyze the collected messages from all the joined groups. Overall, we gather 476,059 messages from WhatsApp, 3,148,826 messages from Telegram, and 4,630,184 messages from Discord. First, we compare the types of messages in each messaging platform, as all platforms allow users to send text, images, videos, audios, stickers, and documents. Figure 6.10 reports the percentage of the messages in each type. Unsurprisingly, text is the most shared type with 78%, 85%, and 96% of all messages on WhatsApp, Telegram, and Discord, respectively. Also, it is worth noting that WhatsApp is the platform with the largest variety of multimedia with more than 20% of multimedia messages (images, videos, audios, and stickers).[3]. In particular, stickers, which are a specific format of images, represent 10% of all the collected WhatsApp messages. They are very common on WhatsApp, and there are even groups dedicated to sharing exclusively stickers between users. Note that Telegram also has a small portion of "other" types of messages, including service messages (e.g., users joining/leaving group, editing group information).

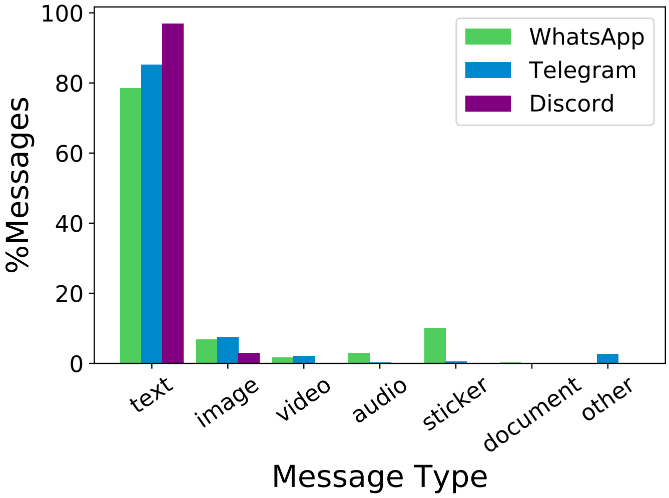
We also look into the volume of messages shared in each group and the number of messages per user. Figure 6.11(b) shows the number of messages shared per day in each group for all the messaging platforms. We report the number of messages per day, since for WhatsApp we can only obtain messages shared after we joined the group, while for Telegram and Discord, we obtain messages since the group's creation date.

The average number of messages exchanged per day across the different groups shows that the majority of the groups are relatively "quiet", i.e., members within the groups send at most 20 messages per day, specially on Telegram[4] A non-trivial percentage, however, exchange at least 100 messages per day more than 20% on WhatsApp and

[3] Note that our analysis only includes audio/video that is shared as messages (i.e., audio/video clips) and it does not consider audio/video calls within groups.

[4] If the messages were uniformly spread over the day, it would amount to less than one message per hour.

Figure 6.10: Percentage of messages in each message type. Text is the predominant type across all messaging platforms.



Source: The Author.

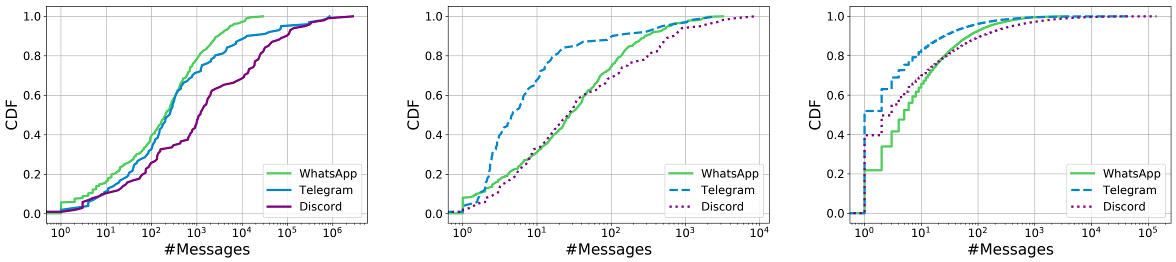
Discord and around 10% on Telegram and there are groups with impressive more than average of 1000 messages a day, which shows some of them are very active and exchange a lot of information.

We observe that Telegram groups are less active compared to WhatsApp and Discord. Specifically, approximately 60% of the groups have more than 10 messages a day, while just 25% of the Telegram groups have such activity. For all platforms, we can observe some groups with more than 2,000 messages per day.

The collected messages are shared by 12,434 distinct users on WhatsApp, 100,504 users on Telegram, and 34,543 users on Discord. This represents, respectively, 59.4%, 14.6% and 65.8% of the total number of members in the joined groups (see Table 6.7). Although we can not affirm that this represents the percentage of members sharing messages, as total size changes over time, these numbers give us a hint of the portion of active members in each platform. Discord has a higher number of active members. On the other hand, on Telegram, just a small portion of the total members share messages, probably because of channels, which allow only a small number of users to share messages (i.e., creator and few administrators).

Finally, we analyze the volume of messages shared per active member in Figure 6.11(c). We observe that most members share only a few messages, while some share a large volume of messages. In particular, 65.8% of them share up to 10 messages for WhatsApp, 70.1% for Discord and 82.9% for Telegram. When looking at the volume of the messages shared by the top 1% of the members (in terms of number of messages they shared), they are responsible for 31% of all messages collected from WhatsApp, 60% of all Telegram messages, and 63% of all Discord messages. This indicates that Telegram and Discord have a larger percentage of very active users that share a very large number of messages across groups.

Figure 6.11: Distribution of the number of messages shared in the groups of each platform and their users.



(a) Messages per group (b) Mean number of messages for each group per day (c) Messages per user

Source: The Author.

We show that the groups shared on Twitter are mostly “fresh”: they are shared on Twitter soon after they are created, yet a few groups are still being shared even though they were created more than a year ago. We discover that most of Discord group URLs expire during the first days after shared on Twitter, while WhatsApp group URLs last longer. Also, Telegram group URLs are less likely to get revoked.

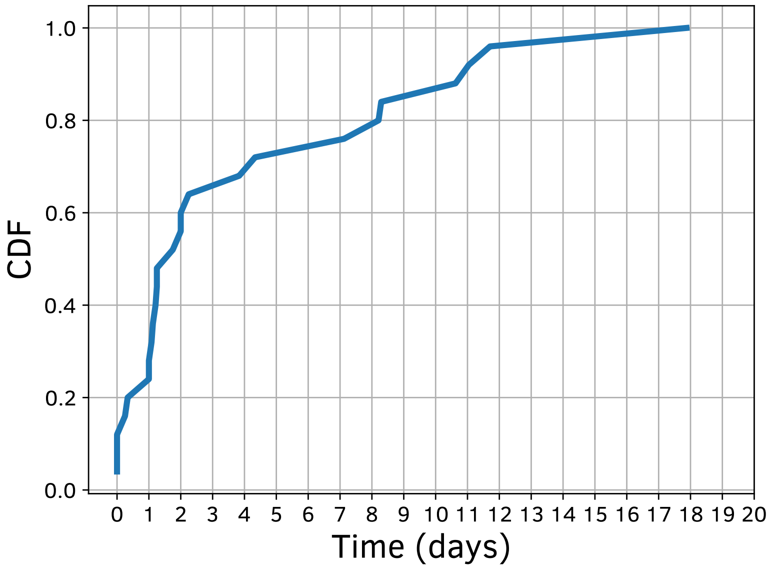
We observe that the difference of the group size limit between the three platforms indeed impacts the size of the groups, since Telegram and Discord have larger groups up to 4 orders of magnitude compared to WhatsApp. Regarding those members, we can also note that Discord members are more active than Telegram in terms of the number of online members. The selection bias and ephemeral nature of group URLs, discovered on Twitter, has implications for studies that use such URLs.

6.1.4 Composition and Activity within WhatsApp Groups

To dive in more details on public WhatsApp groups ecosystem, we shed light on the composition of (e.g., how diverse is a group, in terms of nationality of its members) and activity within (e.g., how many messages do members exchange within a group) subset of the discovered groups on WhatsApp. More specifically, we choose a subset of 400 groups selected at random from the groups discovered on Twitter. We join these groups, and we track, from that day onwards, the changes in the composition of as well as the activity within the group that are specifics characteristics for WhatsApp.

When joining the groups and gathering data, we remain a mute spectator: We never post any message to the group or interact directly with a member. Perhaps, as a consequence or for some other reasons, administrators removed us from 25 of the 416 groups we joined. Although these groups were publicly advertised on Twitter, some of them have rules or restrictions for membership—some, for instance, require members to introduce themselves upon joining or groups expect that members should be from a specific country. Group administrators may remove members who do not satisfy all the requirements.

Figure 6.12: CDF of time until our account got banned from the WhatsApp groups we joined.



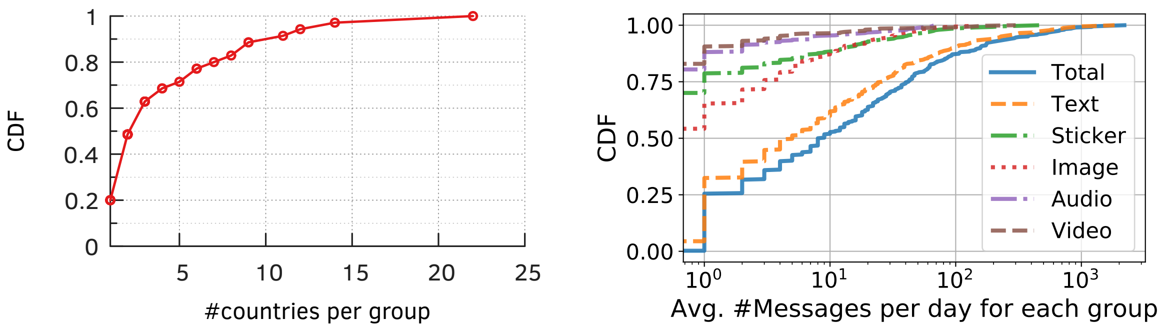
Source: The Author.

Fig. 6.12 shows the CDF of the elapsed time between when we joined the group and when we were banned. Suppose that the administrators removed us for not satisfying one or more requirements of the concerned groups. We then observe that in 20% of the instances, the administrator removed us quickly, within the first 24 hours after joining. In most cases, we were removed within two days. In a few cases, it takes more than a week for the administrators to notice and remove us; the longest time it took for us to be removed from a group to was 18 days. While the reason for quick removal could be, as we assumed earlier, that we did not satisfy some requirements, the late removals hint at other reasons: Perhaps we were removed to free up some space in the group, given the limitation on the size, for new members, who might actively engage with the group, or perhaps we were removed simply owing to our inactivity. This screening process, which might result in an account being removed from a group, has implications for large-scale collection of data on public WhatsApp groups.

Below, we analyze the data gathered from the remaining public WhatsApp groups.

Once we join a group, we can access the list of members of that group and gather the country codes of their phone numbers. The measure of the number of different country codes we observe per group can serve as a proxy for estimating the “diversity” of a group’s composition. Per Figure 6.13(a) most of the groups are more regional: Half the groups have members from no more than 2 different countries.[5] The composition of 10% of the groups is highly diverse, with members from at least 10 different countries.

Figure 6.13: Specific characteristics of WhatsApp groups composition joined in experiment.



(a) Countries per Group

(b) Type of message sent daily

Source: The Author.

Although members exchange different types of messages, text is, unsurprisingly, the most popular type of message, followed by images and stickers. Images constitute the second most commonly used format of messages, with 20% of the groups sending, on average, more than 10 images per day. Use of stickers (which are a specific format of images) are nearly as widespread as images, although stickers were only introduced recently, in 2018.[6] Stickers are more commonly used than videos and audios, and we observe some groups, in our study, sending more than 1000 stickers per day, on average. Of these groups that commonly use stickers, 27 WhatsApp groups virtually use only stickers (i.e., 90% of their messages are stickers). The format of messages has implications for how researchers can analyze the information: Stickers, for instance, are more challenging than text. We observe that a substantial fraction of messages is posted by a small number of members.

6.1.5 Privacy Implications

In this section, we analyze the users' privacy implications from using WhatsApp, Telegram, and Discord. When dealing with social media platforms, a common concern is about privacy and exposure of sensitive personally identifiable information (PII). In particular, for messaging platforms where users are engaged in direct and closed conver-

[5] We only refer to the country with which a member's phone is associated, and we cannot readily infer their nationality.

[6] <https://blog.whatsapp.com/10000653/Introducing-Stickers>

Table 6.5: Statistics on users' sensitive PII that are exposed from the three messaging platforms.

	WhatsApp	**Telegram**	**Discord**
Users observed	20,906 members 34,078 creators	74,479 members	25,701 members
Users' Phone Numbers	54,984 (100%)	509 (0.68%)	-
Users' Social Networks	-	-	7,708 (30%)

Table 6.6: Number and percentage of Discord users whose their accounts on other platforms are exposed.

Platform	**#Users (%)**
Twitch	5,256 (20.4%)
Steam	3,158 (12.2%)
Twitter	2,287 (8.9%)
Spotify	2,080 (8.0%)
YouTube	1,712 (6.6%)
Battlenet	1,338 (5.2%)
Xbox	956 (3.7%)
Reddit	785 (3.0%)
League of Legends	617 (2.4%)
Skype	169 (0.6%)
Facebook	139 (0.5%)

sations in a private and secure manner, it is important to analyze the potential PII that can be exposed by the platform.

Usually, users are joining these groups while being fully agnostic that various aspects of their private information are exposed by either the platforms' interfaces or their APIs. This raises some legitimate concerns regarding what kind of PII is exposed by each platform, and how critical and prevalent is the exposure of PII on WhatsApp, Telegram, and Discord. To this end, we collect all user-related information from each messaging platform and analyze them to understand the underlying privacy implications from the use of messaging platforms.

Each of the messaging platforms has its own peculiarities, and it requires a different approach to collect user information. On Telegram and Discord, we are able to collect user information for users that participate in groups that we also are members of. This also applies for WhatsApp, however, there is an important difference as WhatsApp exposes the phone number of group creators even before joining WhatsApp groups. To collect data related to users, for Telegram and Discord we used the available APIs to get user information for groups we joined, while for WhatsApp we scraped the information from all discovered groups.

Table 6.5 reports the number of users whose PII are exposed for each messaging platform. Looking at the total number of users which we collected data, we find 20,906 WhatsApp users within the groups we joined, and 34,078 unique users that are the creators of the rest of the groups that are accessible, totaling 54,984 users. Notably, WhatsApp not only displays the phone number of all members of the groups we joined, but also reveals the phone number of the groups' creators to non-members of the group. Those 34k phone can be accessed without the necessity of join groups. Only parsing the URLs of invite links on Twitter was enough to get the number of the creator of the groups

For Telegram, we collect information for 74,479 users, while for Discord we find 25,701 users. Note that for Telegram and Discord, the number of users is smaller than the total of users for groups we joined, representing 10.8% and 49% for Telegram and Discord, respectively. This is because on Telegram, administrators are able to restrict the access to the member list, and thus users can not see who are the members of the group. For Discord, the API blocks bots to join groups by themselves (they need to be added by an administrator) and obtain the list of members. Due to these constraints, we collect user information for users who posted at least one message within the groups we joined. WhatsApp, on the other hand, has easy direct access to all other members by the member list of the groups. As Telegram and Discord provides official APIs to manage groups and do tasks, it was possible to get information from the observed users regarding whether the user is a bot or not. We found some bots among the users (53 on Telegram and 102 on Discord), and even that it represents less than 1% of the total of users, considering that we joined exactly a hundred groups for each platform, the number of bots per group is likely to be high.

Our data collection and analysis highlights the exposure of PII information in each platform. Alarmingly, on WhatsApp we are able to obtain the phone number of *all* users that we discovered during our data collection, a total of 54,498 phone numbers. On the other hand, on Telegram we can only obtain the phone numbers of 509 users, which corresponds to 0.68% of all the Telegram users who participated in the groups we joined. The relatively low percentage is because Telegram hides the phone number of the users by default. A phone number is only shown within the platform if the user explicitly opts-in.

Finally, for Discord, since phone numbers are not required for registration, we find no evidence of phone number exposure. However, we find that Discord exposes accounts that users have on other platforms: we find 7,708 users (30%) for who we are able to obtain at least one other account that they have on other platforms, namely, Twitch, Steam, Twitter, Spotify, YouTube, Battlenet, Xbox, Reddit, League of Legends, Skype, and Facebook. Table 6.6 reports the number of users whose users' accounts are exposed for each of the other linked platforms. Since Discord is a gaming-oriented platform, some of their social networks are related to game universe, like the Twitch, the platform in which Discord users had more linked their accounts (20.45%), that is a streamer platform to gamers share their gameplays, and Steam (12.29%), that is a widely known video game digital distribution service. It is important to highlight also the high presence of accounts of Twitter(8.9%), YouTube (6.66%), and Facebook(0.54%), that can be considered very high sensitive PII information. Our analysis shows that Discord exposes at least one social media account for 30% of the Discord users monitored.

Overall, these findings have important implications to user's privacy. The exposure of PII from all these messaging platforms can be potentially exploited by malevolent actors that aim to target users. For instance, state-sponsored actors (ZANNETTOU et al., 2019a, 2019b) that have considerable resources and can perform a much larger data collection than our study, can create profiles for all those users and target them on the same or on other social media platforms. A potential attack vector is the creation of user profiles based on the topics of the groups they participate and then their targeting on other social media platforms via posts or advertisements with the goal to manipulate them or change their ideology. Also, our results highlight the need to raise user awareness about the privacy implications from the use of messaging platforms like WhatsApp, Telegram, and Discord.

6.1.6 Findings

In this experiments, we performed a large-scale characterization of public groups from WhatsApp, Telegram, and Discord shared on Twitter, a popular microblogging platform. In contrast to other works on IMPs, here, we first focus on the discovery and evolution of the ecosystem of public groups of the messaging services, that most prior work rely on.

We searched for group URLs (or invite links) on Twitter for all three platforms for a period of over a month and obtained a set of approximately 351K URLs to groups. By performing topic modeling on the tweets including group URLs, we were able to understand the content of these groups and the differences that exist between these messaging platforms.

Although these platforms are also designed for private conversations, we find that Twitter is a rich data source for discovering public chat groups. This is a strong evidence how communities are organizing on web through public groups within messaging platforms looking for a closer and safer environment to talk and share interests. Our findings highlight several points that need to be considered by the research community focusing on similar platforms. First, we show that by taking a multi-platform view of the Web ecosystem, we can extract meaningful insights that otherwise will be hard to deduce if we were studying, for instance, WhatsApp in isolation.

We meticulously monitored discovered groups from three platforms, gathering measurements once per day, and used these coarse-grained measurements for investigating the changes in the characteristics of the groups (e.g., number of members) over time. Our analysis highlights the ephemeral nature of groups, as during the course of this study 27% of the groups become inaccessible for WhatsApp, 20% for Telegram, and 68% for Discord. This phenomenon prompts the need to design and develop robust, scalable, and real-time data collection solutions that will enable the research community to obtain a more holistic and complete overview of the messaging platforms' ecosystem. We also joined a sample consisting of hundreds of those groups for all three platforms, and provided a characterization of the activity within the groups.

Finally, we analyzed the exposure of sensitive PII on all three platforms, particularly phone numbers for WhatsApp and Telegram, and linked social media accounts for Discord users. For WhatsApp, even without an account, we could collect an impressive number of over 34K phone numbers. Moreover, after joining groups, we obtain another 20K phone numbers. We also found exposed phone numbers for a small portion of users on Telegram (less than 1%). Finally, on Discord, we were unable to find phone numbers, however we collected at least one linked social network account for 30% of the users analyzed. These privacy implications are alarming, since messaging platforms are often used because of their perceived security in communication and privacy.

Our results highlight the need to raise awareness of the public related to these privacy implications and design guidelines on how messaging platforms can adjust to better safeguard users' privacy. IMPs are often used for private and personal conversations. Telegram and WhatsApp even have end-to-end encryption implemented to protect data for their users. Therefore, many of their users may not realize what information they have exposed by participating in those public groups.

Naturally, our work has some limitations. First, we rely solely on Twitter to discover groups from WhatsApp, Telegram, and Discord, and hence we are unaware of numerous publicly available groups. Despite this fact, Twitter is a very large and mainstream social network that we use to make the best effort to attempt to discover many groups from WhatsApp, Telegram, and Discord, and mitigate potential biases. Second, we join and collect data from only a limited number of groups from WhatsApp, Telegram, and Discord, mainly because these messaging platforms have specific constraints that prevent us from scaling up our data collection. Namely, WhatsApp requires a large number of mobile phones and SIM cards, Discord requires the creation of multiple user accounts, while Telegram's API is rate-limited. Overall, this is a limitation that exists in every study that collects data from messaging platforms.

To the best of our knowledge, by collecting and analyzing the dataset of over 300K public groups from messaging services, including an analysis of their evolution over time, and comment on the implications of our observations for a rich body of work on social media, we perform the largest characterization of this environment and the interplay with other platforms. We specifically focus on the role of Twitter in discovery, since it facilitates easy discovery of a large number of groups to both the users who intend to join these groups and the researchers who attempt to study, via the groups, the dynamics of information propagation in WhatsApp, the underlying private social messaging platform.

After all this investigation over the public aspect of IMPs and the ecosystem of public groups on WhatsApp and other platforms, we can better understand how this platform is used for mass communication and can dive in the content shared within these messages to explore how misinformation propagates in this network.

6.2 Comparing Different Sources for WhatsApp Public Groups on Web

This investigation gives us a big picture of the behavior of public groups, channels and servers created by IMPs on the Web. This give us the first stepping in understanding how this ecosystem is formed and highlight the importance of promoting invite links in social networks to expand the interconnection of users within these messaging platforms. Even though these results show us general characteristics of these networks through the lens of Twitter, we still have other different sources for advertising groups online, specially in case of WhatsApp, which the groups are recognized as having an essential role in recent events taking place in context of Brazil.

Given that there is no search tool in WhatsApp itself to look for groups within the platform, people rely basically on Web and social networks to find public groups of topics they have interests to join and participate. On the other hand, we know there is a range of themes and discussions very popular on WhatsApp in Brazil.

Despite WhatsApp being one of the most used social media in Brazil and also in the world. However, we lack of efforts that aim to investigate this WhatsApp ecosystem due to the challenges of its structure and there are even fewer studies towards an analysis in large scale of this network.

Therefore, exploring how this environment is constituted is a key element to further analyze this platform before diving in the content shared within WhatsApp, and also very important to comprehend how it has been abused by misinformation campaigns to spread harmful content through this network. More specifically, we need a more panoramic view of WhatsApp to better understand how these public groups on WhatsApp are organized, what the interplay between this network and remaining universe online and also which topics attracted more users to creating and share their groups.

In this sense, this section expands the findings from previous experiments on Twitter and performs a first large-scale analysis of the ecosystem of Brazilian public groups on WhatsApp, seeking to understand and quantify the main subjects discussed by Brazilians through the application and also how these public groups are disseminated on the Web. Our objective is to understand how users arrive at public groups, differentiating the main sources of public groups and the different subjects that exist on this platform, thus pointing out how the public uses WhatsApp and which subjects attract the most attention.

For this, an extensive collection of public WhatsApp groups released from different sources on the Web was carried out. Shared groups on social networks such as Twitter and Facebook were collected, in addition to group repositories, which allow users to disclose or search for public groups. In all, we analyzed a set of more than 270,000 public WhatsApp groups.

In this work, we propose a methodology for large-scale categorization of WhatsApp groups. Our analyses show a great diversity in the universe of public WhatsApp groups, with popular subjects that go far beyond the political and health contexts already explored, with hundreds of groups dedicated to topics in pop culture, entertainment, games, sports, television, music, social networks and memes. We also noticed that WhatsApp is widely used as a space to meet new people, especially during the COVID-19 pandemic period, when remote and digital relationships gained greater prominence in our daily lives. Not only that, we see that groups about education and work are also popular, showing that the app is not only used for chatting but also as a working and learning tool for users.

Furthermore, we found occasional differences between the different sources where groups are posted on the Web. On Twitter, for example, we noticed a large presence of groups related to pornography, while on Facebook we found more groups on religious topics. Going deeper, we discovered how the dissemination of groups on the Web serves to attract users from other platforms to a less moderated environment, in order to discuss topics that are not accepted in other networks.

A significant part of the public groups identified have topics related to fraud and embezzlement, such as the sale of counterfeit bills, dubious promises to earn money, in addition to trading and selling influence of *likes* and followers on social networks. On Twitter, for example, there is a significant and automated dissemination of cloned card sales groups, with more than 7.5% of messages shared by groups on the subject. In practice, the public space is used to persuade users of other platforms to join the WhatsApp group ecosystem, a more restricted space with little moderation, facilitating the practice of illegal activities.

6.2.1 Sources of Public Groups on WhatsApp

Generally speaking, the goal of a public group creator is to reach as many members as possible. For this, a group invitation link is generated and shared by the administrator, thus making the group publicly accessible to other users. Although WhatsApp allows the creation of public groups, it does not provide any resources to be found by other users, causing administrators to seek other means of dissemination. A common alternative is the dissemination of invitation links on social networks, as they allow you to reach a large number of people in a very practical way. Another widely used possibility is to add the group to online repositories aimed at disseminating WhatsApp groups. These repositories work like an online catalog, where the administrators themselves inform the category and description of the group, making the group accessible to anyone who wants to join and other users search for those groups they interest most.

In order to better investigate the interplay between those sources and WhatsApp, for data collection, we follow the same path taken by a user when searching for WhatsApp groups on the Web, listing the available group invitation links found from different sources of groups on the Web. Particularly, four sources were selected for the collection of Brazilian public groups: (1) Facebook; (2) Twitter, which are two traditional social networks widely used for publicizing public groups; (3)"Zap Groups"[7] and (4)"Whats Groups"[8], two of the largest online repositories dedicated exclusively to index public WhatsApp groups. Below, we describe in more detail the process of collecting data from each of the selected public group sources.

During the period of one month in December 2021, we collected the data of all posts on Twitter and Facebook that contained at least one WhatsApp group invitation link and used the Portuguese language similarly to previous experiments from this chapter, but more exclusively for WhatsApp and Brazilian groups. We, hence searched for posts that had at least one invite link to a group during period. It is worth noting that language/nationality information is missing from Facebook posts and for part of Twitter. Therefore, in order to filter only Brazilian groups, in addition to those publications that explicitly mentioned the language, we also selected posts that contained terms in Portuguese. In total, 20,000 publications were selected for Twitter and 118,000 for Facebook.

Different from previous experiments, here we expanded the collection with other sources of public groups. The websites "Grupos de Zap" and "Grupos de Whats" have a very similar structure to each other. In both, users register a public group informing its title, description, and category. Within a predefined set, the creator selects the category that best represents his group, allowing the search for groups of interest. These categories

[7] Zap Groups: <https://gruposdezap.com/>

[8] Whats Groups: <https://gruposwhats.app/>

Table 6.7: Overview of the dataset of discovered Brazilian public groups on WhatsApp from different sources on Web.

Corpus	Total	#Unique Groups	Active Links(%)	Users
Twitter	20.027	3.360	49,7%	8.174
Facebook	118.025	10.441	69,4%	9.303
Grupos de Zap	31.371	29.203	24,6%	21.669
Grupos de Whats	230.308	230.218	32,9%	–

are even very similar between the two sites, even including categories with the same name. In order to extract the groups from these repositories, a web collector was built that runs through the site, raising all the reference URLs to the registered groups and, then, for each group found, the information available on the site is collected, including title, description, category, date of inclusion and, for "Zap Groups", the creator and number of views for each group. This collection was carried out in January 2022 for all groups registered on the website from 2018 until that moment.

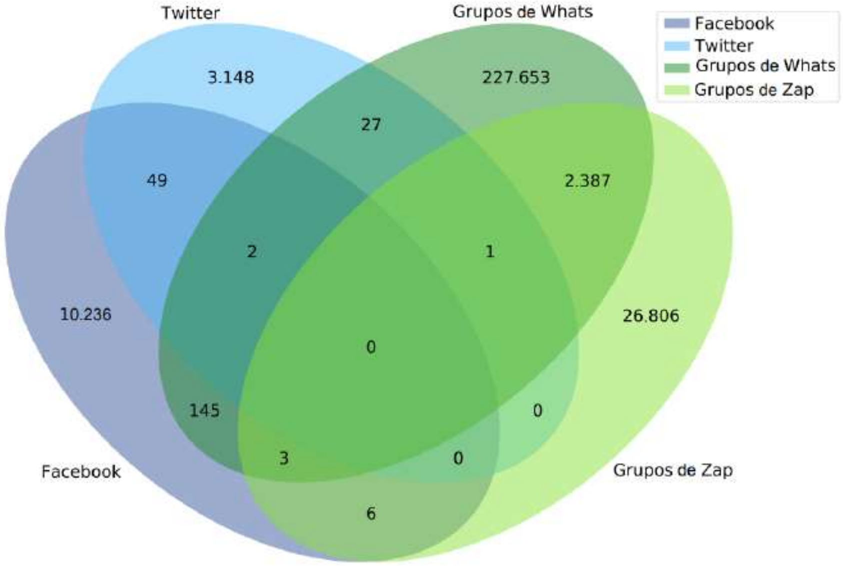
6.2.2 Dataset of Discovered Groups

With that process, a dataset [9] with more than 270,000 distinct public groups of WhatsApp from four disclosure sources, as described in Table 6.7. We observed that online repositories have a relatively larger number of groups than social networks, although we need to take in account that these groups represents a short period for social networks compared to the whole database from online repositories since 2018. In concordance to other previous results made on Twitter exclusively, we can observe there is a significant number of WhatsApp public groups in Brazil, with thousands of groups available online in distinct sources. In addition, the Table 6.7 provides the number of unique users who shared groups and the total number of groups still active, as verified in March 2022.

When we observe the intersection of groups between sets with the diagram in Figure 6.14, we notice the a small intersection between them. This indicates a considerable difference for each group source and a kind of independence between them, in which groups found on a specific social network are unlikely to appear elsewhere. This observation is relevant considering that works that look for public groups must seek for different sources of data in order not to ignore a significant portion of the data due to the bias of origin of the groups.

[9] The public groups found in the repositories are available at: <https://doi.org/10.5281/zenodo.7017909>

Figure 6.14: Venn Diagram of intersection of groups discovered in each source.



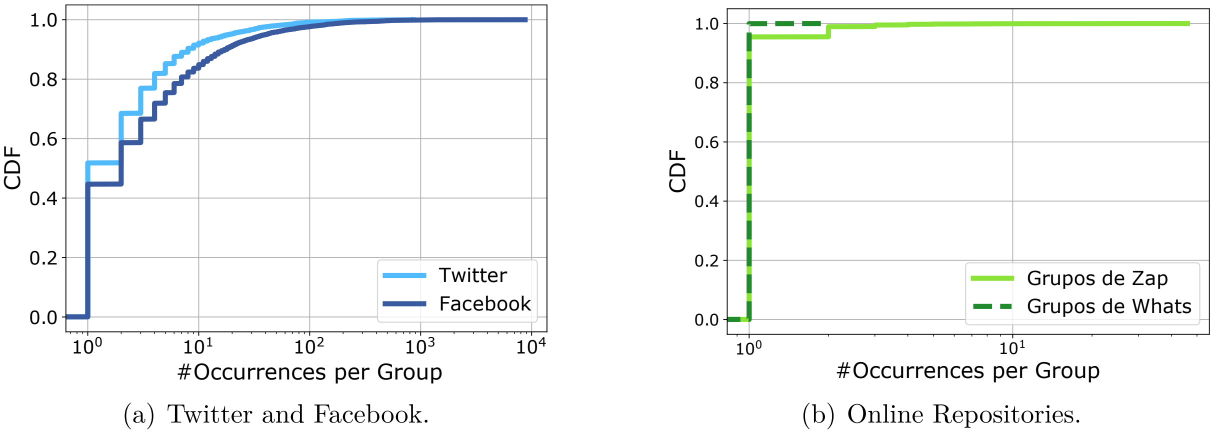
Source: The Author.

With this data, initially, we analyzed the volume of groups sharing, as well as the number of groups still active. Then, we carried out an extensive investigation of the topics discussed in the groups shared in the repositories and in social networks. To explore the categories of groups shared on social networks, we used a statistical model with latent variables to detail the topics in the groups.

There is a difference between the total content collected and the number of unique groups found in each source. This is because the invite link to the group can be shared more than once in each dataset. In Figure 6.15, we observe the cumulative distribution function (CDF) with the amount of occurrences for each invite URLs found. We've noticed that on social media, it's common for different publications to post the same group invite link, whether it's a single user spamming his own group or even multiple users sharing a single popular group through the platform. About half of the groups are advertised multiple times on social networks, while in repositories, more than 95% of groups were registered only once. This suggests that, unlike group repositories where there is a more horizontal and equal position between the groups, social networks have a wide variation in the visibility that each group receives on the platform, which means that exists a hierarchy structure that some groups may become more popular simply because they are more shared.

Furthermore, in online repositories, we have the date of inclusion of the group. With that, we can see in Figure 6.16, the number of groups created over time. The website "Grupos de Zap" has had groups since 2017, while "Grupos de Whats" is more recent, with the first groups dating from 2019. Although the creation of groups on "Grupos de Zap" remains stable, after 2020, in addition to the "Whats Groups" groups, we have a

Figure 6.15: Volume of occurrences for each WhatsApp group invite link in each dataset.



(a) Twitter and Facebook.

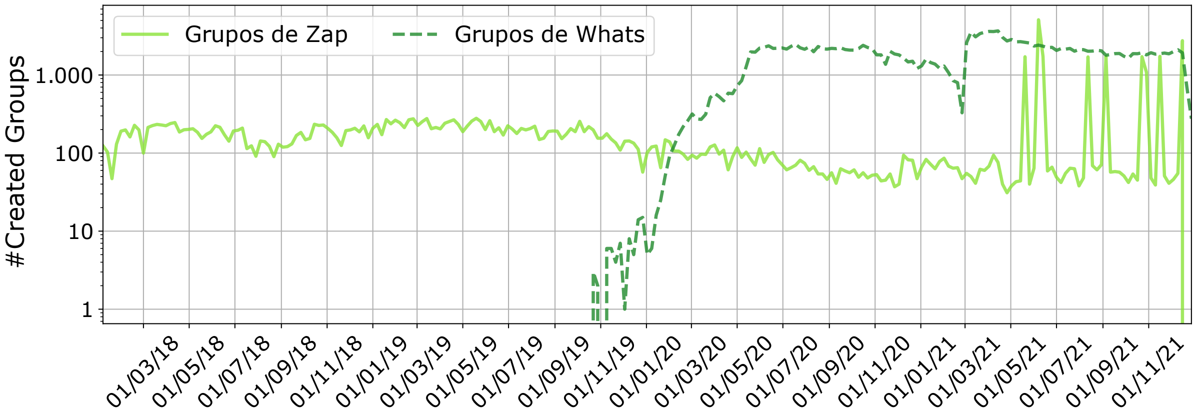
(b) Online Repositories.

Source: The Author.

growing interest in public WhatsApp groups, with weeks in which more than a thousand new groups were created on both platforms, further highlighting the importance that this ecosystem has attracted for the Brazilians. The sudden increase in the creation of groups on these sites also coincides with the emergence of the COVID-19 pandemic, a period when long-distance relationships became more common, and remote means of communication, such as WhatsApp, passed to represent an important part of our daily lives. to-day.

In addition to creating and publishing the invitation link to your group, admins can also revoke the invitation, making the group inaccessible. Therefore, not all links remain active after a certain time. In the "Active Links" column of the Table 6.7, we observe the percentage of invitations still valid, verified in March 2022. In practice, it is noticed that most of the public WhatsApp groups are available for a limited period. On Twitter, just over half of the invitations were revoked after 3 months of collection, while on Facebook 30% of the groups are already inaccessible. In the repositories, an even smaller percentage is found, given that they have registered older groups. These results

Figure 6.16: Number of groups registered weekly in the online repositories of public WhatsApp groups.



Source: The Author.

show a dynamic and ephemeral nature of public groups, in which new groups are created and others become inaccessible in a short period of time.

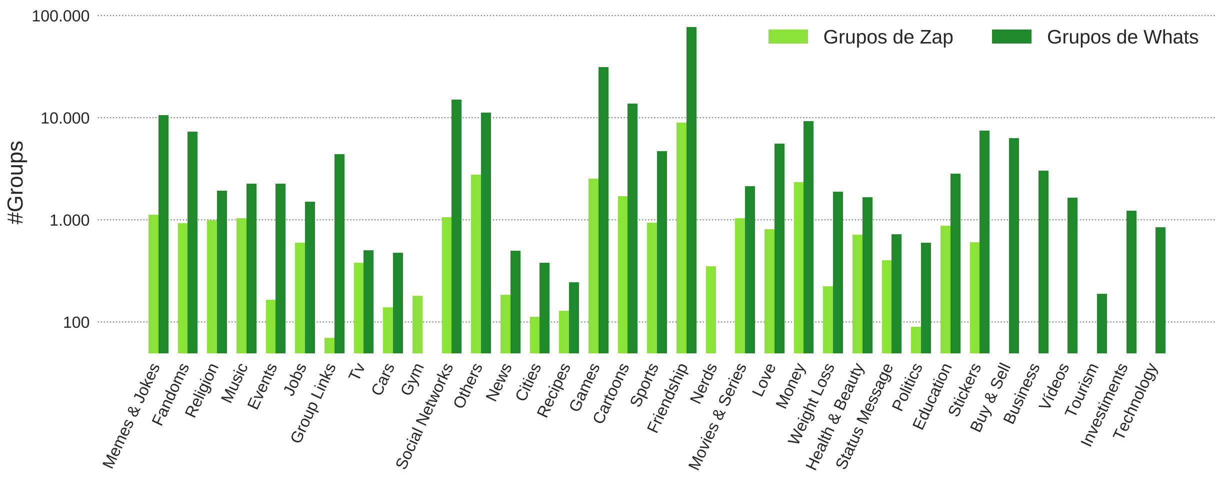
6.2.3 Main Topics for Brazilian WhatsApp groups

The online repositories carry the real category label of each registered group, provided by its own author during creation on the website, displayed by Figure 6.17. It is noteworthy that the two analyzed repositories have a large intersection between the existing categories names, requiring only a few adaptations in order to compare them (e.g., "Romance" and "Love" were concatenated; "Catholic" and ' "Spiritism" were translated to "Religion"; "Football" was added to Sports; and less expressive categories added to "Others"). Saying that, we separated the groups into 35 final topic categories. In both repositories, it is interesting to note that the Friendship category is the most popular. This category encompasses groups about relationships, meeting people, dating groups, flirting, and general conversation. As groups with adult content are not allowed on these websites, we observe that users creating dating groups fitted their groups to this category, although some of them are also intended for adult audience.

Some categories also are very representative in terms of the number of groups created, such as the Games category (33,954); Cartoon category (15,603), considering cartoons and anime; Memes and Jokes (11,846); Sports (5,688), composed mostly of football groups. In general, we see that topics related to pop culture are very present, revealing the main themes of public discussion from WhatsApp.

We also have many groups on Social Networks category (16,192), with several

Figure 6.17: Number of groups created by category in online repositories.



Source: The Author.

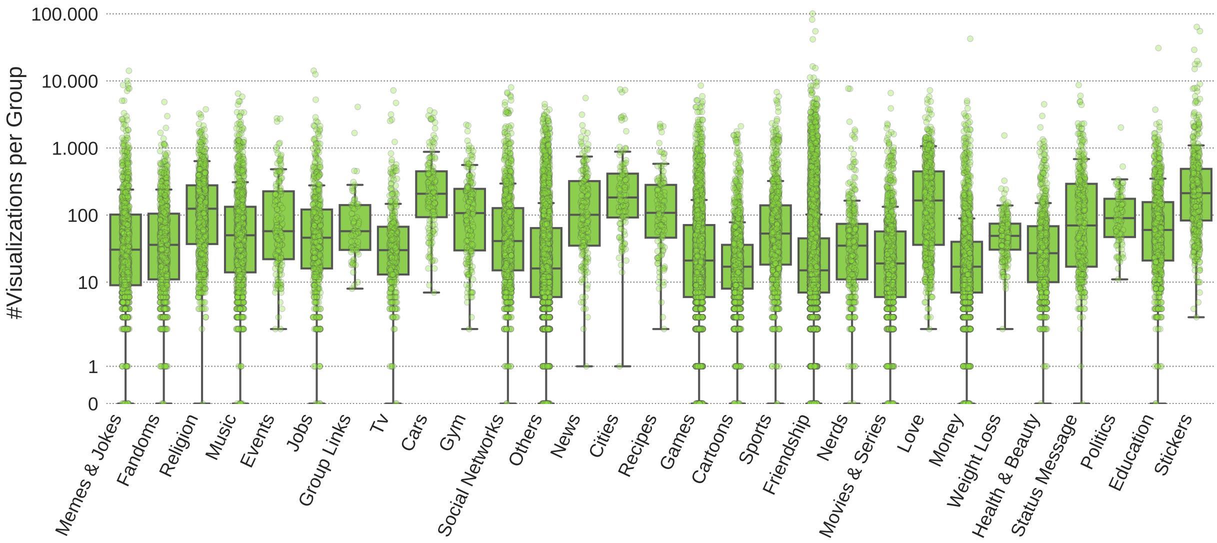
WhatsApp groups dedicated to other platforms such as YouTube, Facebook, TikTok, Twitter and Instagram. A large part of them are about how to gain followers and likes on these networks. Although it is a practice that does not comply with the rules of the platforms (CASSITA, 2019), it is common for users to join together in the search for followers and likes on other networks. We even noticed that some of these groups offer paid and automated services for selling likes and followers.

Other information available is the number of views of each group in the repository "Grupos de Zap". Despite the limit of 256 members on WhatsApp, a group can be viewed numerous times on the site, which gives us a sense of how attractive a category is to users. In Figure 6.18, we have a *boxplot* with the distribution of views by category. It is noted, a large number of views in all categories, which suggests that the repositories not only store, but that, in fact, attract a lot of users who search for those groups. These results show that groups are objects of interest for many users, and that these repositories are fundamental tools for them to be found within the universe of public groups existing in this ecosystem.

It is important to note that the categories with the most views (i.e., considering the median of the distribution) do not necessarily correspond to those with the highest number of groups. While Friendship, the category with the most groups, has an average of around 50 views, others have an average of over 200 views, such as Romance, Stickers, Cities and Cars. Hence, we can see some differences regarding what kind of content attracts more attention from users wanting to join the groups and what authors want to create more groups.

Moreover, Education, Stickers, Money, and Employment categories have groups with particularly high number of views, some individual groups with more than 10 thou-

Figure 6.18: Distribution of number of visualization by each group topic.



Source: The Author.

sand views within the repository. The Jobs category refers to groups offering and seeking work opportunities and also selling products. The Education section offers courses and online classes. The large number of views of these two categories reveals an important use of WhatsApp, which goes beyond entertainment with a simple chat and discussion space. Users also utilize the public groups for working and learning purposes, making them also a tool of social services and real-world implications. Even though we don't have as many groups, these categories are highly viewed, which shows the great interest of users in these topics on WhatsApp.

Sticker groups on WhatsApp are also extremely common. Since users can create their own stickers on chats, this space is often used to share creations of popular characters or memes on the internet. Even though there are not many groups, the Stickers' category is one of the most popular among the groups, showing a very specific behavior based on a feature of WhatsApp itself.

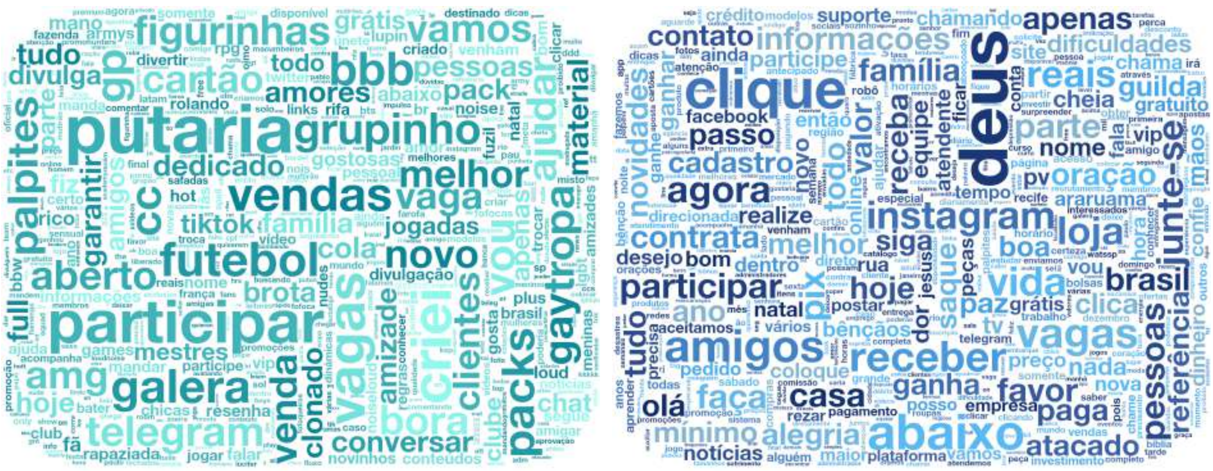
Furthermore, we look at the Politics and Health categories within the online repositories, as these topics are the focus of several studies on public WhatsApp groups. In health, despite there are more than a thousand groups, it has a low average number of views. Political groups represent a small portion when compared to the total universe of groups in the database. However, there are still a fair amount of groups dedicated to the subject, with about 700 groups, in addition to a significant number of views (i.e., median close to 100). Given the prominence that political messages have in society, we see that these topics manage to generate a large volume of content on WhatsApp, even with a small number of groups, deserving special attention.

After that, we take our attention to the groups found on the social networks. When sharing a group on social media, the user usually also provides some context information that helps describe what the group is about. With that in mind, even though not having the actual label for those groups, these messages help us to have a more general view of the content of the groups posted on Twitter and Facebook. Here, we repeat some analysis regarding groups discovered for this specific dataset of Brazilian public groups, focusing in differentiating each source of groups. First, we observed the cloud of terms from messages containing WhatsApp groups (Figure 6.19). With the cloud of terms, it is already possible to identify some relevant differences between the platforms: while Twitter highlights some terms with sexual connotations, on Facebook it is possible to see more religious words.

On Twitter, pornography-related terms are much more frequent, with terms like (e.g., “whore”, “porn”, “xvideos”), while on Facebook (Fig. 6.19(b)), this type of content is not allowed by the Terms of Service. This not only shows that WhatsApp has a large amount of adult content groups (which is hidden through the other data sources), it also points to Twitter as a relevant source of these types of groups.

It is noted that many messages analyzed use a metalinguistic language about the promotion of the groups by inviting people to join WhatsApp groups with words such as

Figure 6.19: Word cloud of messages containing WhatsApp groups.



(a) Twitter

(b) Facebook

Source: The Author.

"click", "participate" and "access", showing that these messages, in fact, aim to invite users to be members of their groups. In both figures, it is also possible to identify mentions of other social networks, such as Telegram, TikTok, Instagram. This demonstrates a strong relationship between the numerous popular social networks and WhatsApp groups, suggesting that users do not only advertise their WhatsApp groups, but probably also several profiles from other networks, with a large interplay between them.

We see the word "stickers" highlighted on Twitter, a category of groups that was also one of the most popular found in the analyses of online repositories, showing, again, that groups about stickers on WhatsApp are also very popular on social networks.

Finally, several terms were also found suggesting the existence of groups that promote and sell content and/or services (e.g., "sales", "pix", "vacancies", "pack"). With this, we see that WhatsApp is a business tool for many users. In addition to a conversation application, it can be used to buy, sell and offer various products and services. In view of the analyzed period, the presence of this category is quite interesting, since the COVID-19 pandemic led many companies and service providers to look for ways to undertake in the digital world, making WhatsApp become a popular channel of communication. However, among these groups, there are many promoting schemes and frauds persuading users to join suggesting a worrying existence of spam, scams, phishing and other types of harmful and dangerous content within WhatsApp network, hidden from public audience and without any moderation.

6.2.4 Findings

We have analyzed, in this part of the thesis, a large volume of instant messaging groups shared Web. This is an important result to elucidate better how the WhatsApp network is constituted and the interplay between this platform and other traditional social networks Web. We support a large scale characterization of Brazilian ecosystem of public groups on WhatsApp and enable further steps in direction to dive in this cyberspace and find means to explain its impact in the society.

We find that a large number of the groups are linked to general relationship topics, such as friendship and dating. Note that WhatsApp provides to its users a favorable environment for socializing, especially during the pandemic, where long-distance relationships have become more present in people's daily lives. Groups for discussion of pop culture in general, such as games, sports, music and movies, proved to be very popular, as well as those with themes of adult content, religious themes and trading cards. Also, it was observed that groups about job opportunities, education and, above all, purchases/sales, are very frequent, showing that WhatsApp is not only an informal space for conversations, but also a commercial and work tool for many users. Furthermore, exchange stickers, a peculiarity of WhatsApp, has also become a common topic across platforms. Finally, there are groups for scams, schemes and frauds that persuade users to join promoting harmful practices within the platform.

We found that existing public groups vary depending on the source where it is posted (e.g., pornography is only allowed on Twitter). The analyses demonstrate that WhatsApp has a very diverse public group ecosystem that also expands to other Web instances. In practice, it is possible to observe that the public space of platforms, such as Twitter and Facebook, is used to attract users to enter into a more restricted and less moderated environment, such as WhatsApp. This facilitates the existence of abusive practices that would be difficult to sustain in other networks.

Moving in this direction, moderating the process of creating groups on WhatsApp becomes a relevant issue, along with the process of identifying the existing topics on the platform. With this work, we intend to expand the study on WhatsApp, providing better context about public groups and tools to help the search and categorization of themes in these communities. For next chapters, in addition to expanding the built databases, we are going to apply the knowledge obtained to better understand the structure of groups existing on WhatsApp and explore those that abuse of the platform to spread misinformation campaigns from the perspective of inside the network, deep investigating the actual content shared within those groups and the connections created through them.

Chapter 7

The Information Pathways through WhatsApp

In this chapter, we will investigate how the functionalities presents in WhatsApp and its network structure impact in the dissemination of information within this social media. After discovering the universe of public groups being shared on the Web and the extensive use of WhatsApp to propagate political messages and news in public groups, one may wonder how it can be possible within an app that is mainly used for private communication. In fact, in this chapter, we will observe that a piece of information in WhatsApp can navigate through a network and reach numerous users, and can become a popular trend between users. We will understand better how some features implemented by WhatsApp and how the format of groups can assist a message to get viral in such closed environment and how it differs from other platforms that use group-based communication.

Nowadays, many people prefer to use these messaging apps to directly speak with each other. This popularity, which is particularly seen in countries like Brazil and India, is partially explained by some distinctive traits of such kind of system. First, IMPs are cheaper tools than old Short Message Service (SMS) messages of mobile phones, which people were used to. Second, they offer multiple features that were absent in SMS, such as various forms of communication, including texting, voice and video calling, chat groups, and the possibility to send and forward messages to multiple users. On top of that, all messages are encrypted and hence anonymous, which provides more security to the communication. All those factors influenced in the shift of instant communication to this new IMP cyberspace.

However, the characteristics of these platforms are still not fully understood and their consequences are still unknown by their users and even researchers. One controversy raised by introduction of IMPs is about anonymity and virality (PATEL et al., 2019; MAYER, 2019), in which many have concerns that harmful content of hate speech or fake news may spread within those platforms without any moderation, unable to find the authors nor even possible stop the sharing of this content.

Actually, for Brazilians who lived through the presidential elections of 2018 and 2022, who had a smartphone and participated in WhatsApp groups, it was possible to observe the capacity for viralization and potential capillarity that content from those platform reach in society. Indians also suffered with real life mob lynchings caused by misinformation campaigns and rumors spread through WhatsApp. Family groups, work groups, friends, political groups and many other scenarios on WhatsApp have been flooded by the spread of misinformation.

WhatsApp itself realized this position of spreading of misinformation and rushed to self-regulate to mitigate this viral problem: after having increased the size of groups from 100 to up to 256 participants in 2016, WhatsApp decided to limit simultaneous message forwarding to five recipients at a time in 2019[1] and, subsequently, identify and further limit the sharing of widely shared content, to just a single user after the COVID-19 pandemic in 2020[2]. However, the company seems to have turned in the opposite direction with the creation of communities and increasing groups max size[3], facilitating even more the wide spreading of unmoderated content. On Telegram, messages can only be forwarded to one chat at a time, but the size of groups is unlimited, which also helps to magnify the reach of messages.

The forwarding button combined with the large public groups are important tools in information propagation within IMPs. Using this forward method, a user can select a message from a conversation and promptly spread it to a large number of contacts or groups at once, while the existence of groups with hundreds of members, or even millions in Telegram, is a quick path that facilitates the mass communication.

Although we already know that messages in such platforms spread rapidly and broadly, we do not know if (and how) their dissemination can be controlled by the features of information diffusion provided by those systems. We lack of information of how exactly this spreading happens within this network, how a single message exchanged in a chat is amplified and circulate throughout WhatsApp.

In order to evaluate this massive viralization on WhatsApp, we perform a seep investigation about how the features of public groups and forwarding messages impact in the flow of content through this distinctive network topology, highlighting the influence of them in making a content getting viral. Furthermore, we analyzed data from more than 5,000 public WhatsApp groups and involved over 360K users from our collection methodology in Brazil (Chapter 5), as well data from India, and Indonesia shared in collaboration from the authors of GARIMELLA; TYSON, 2018; GARIMELLA et al., 2018.

We used this data to reconstruct the network of public groups on WhatsApp, in which users are connected when they have groups in common. We analyzed some network characteristics for each set of data, and evaluated the propagation of information

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

[2] <https://about.fb.com/news/2020/04/whatsapp-message-forward-limit/>

[3] <https://blog.whatsapp.com/reactions-2gb-file-sharing-512-groups>

through these networks using an epidemic model. Moreover, we expand this analysis by evaluating how information propagates in general Instant Messaging Platforms by exploring also the Telegram network. Besides a deeper analysis of the WhatsApp group network, we also compare it to the network composed by groups on Telegram. We also perform several simulations to assess how their sharing and topological features affect information dissemination.

More specifically, we study the anatomy of this emerging way of communication by analyzing their groups networks. Our goal is to answer the question of how the tools of the system contribute to the virality of (mis)information and whether limitations are capable of preventing the spread of content. In short, our goal is to understand the effects of the group-based network topology and the sharing features (forward and broadcast) of IMPs in the diffusion and spreading of (mis)information within their networks. More specifically, we aim to answer the following questions:

- What is the impact of public groups on the propagation of information in different Instant Messaging Platforms' network structures?
- Is it possible to control the spread of false content with the forward limitations imposed by the application?
- What is the importance of the features and topological structure of the network to the dissemination of information on these platforms?

To address these questions, we use the Susceptible-Exposed-Infected (SEI) epidemiological model (LI; ZHEN, 2005) to the problem of estimating the virality of malicious messages in IMPs groups. In this model, ***Susceptible*** (S) is the initial condition in which the user did not have any contact with the false content; ***Exposed*** (E) are the people who received the false news through any of the groups they participate; ***Infected*** (I) is the final stage where a user who was exposed to the content shares this message in the network.

Since the platforms have different size of groups, restrictions on how many people a message can be forwarded, and have different network structures, we have included those characteristics in our proposed simulation model to better investigate the process of information diffusion. Thus, we adapted some parameters of the original Susceptible-Infected-Recovery (SIR) model to simulate the functioning of those IMPs networks using different configurations of forward and virality limits. In this work, he have added a new class of vector in the simulations: the "*exposed*". They represent a prior stage of contamination in which represents those who only received a malicious messages in a group and, therefore, is exposed to that content. They only become a proper "*infected*" users when they get in contact and re-share that content for other groups (which may mean that they believed in that content). To simplify the model, we firstly remove the "*recovery*" as we do not have evidences of user "recovering" from sharing the content (particularly for misinformation), however, we also performed experiments exploring this possibility at the end of the work.

The rest of this chapter is organized as follows. Before start the analysis of the data, Section 7.1 provides a discussion of the main characteristics of WhatsApp and Telegram platforms that contribute to information spreading on WhatsApp. In Section 7.3, there is the description of the dataset. In Section 7.4, we analyze the network structure of public groups on WhatsApp and Telegram. The experiments are in Section 7.5, where we simulate the spread of misinformation within those networks under different circumstances via a Susceptible-Exposed-Infected (SEI) epidemiological model. Further evaluation of the data and more results for time spreading collected for three countries (India, Indonesia, and Brazil) on WhatsApp is presented in Section 7.5.5. Finally, Section 7.6 we present the final remarks of this piece of work.

7.1 Virality Features of Instant Messaging Platforms

Before we start to analyze the data collected and its implications, we must understand the instant messaging platforms' (IMPs) operations and the differences among them and to other kinds of networks that contribute to its become a viral tool to disseminate content.

At the beginning of the Internet era, instant messengers such as ICQ or MSN Messenger were more focused on one-to-one communication, in which only two users exchanged messages directly to each other. After the turn of the millennium, a new trend in online communication emerged together with popular platforms such as Facebook and Twitter: the one-to-many model, in which users can broadcast information to users who follow their feed. Current IMPs models fit in a different category of group-based communication (SEUFERT et al., 2016), in which multiple users in these platforms join together in groups where all of them are able to send and receive information from each other. These groups compose communities in a form of many-to-many communication, but in a restricted environment compared to a full broadcast for the Web. With the widespread use of smartphones and easy access to the internet, mobile apps for IMPs such as WhatsApp and Telegram became increasingly popular, as they provide services that are cheaper and richer than SMS. More important to the context of this work, they are not limited to one-to-one or one-to-many conversations, as most of them provide group conversation features.

Comprehending this evolution of communication paradigms and some of the features developed by these platforms is essential to the understanding of how services such as WhatsApp and Telegram turned into tools to disseminate fake news and became protagonists in misinformation campaigns around the globe. While one-to-one communication is more private, individual, personal, encrypted, one-to-many communication tends to be public, in mass, and viral. The group-based communication, on the other hand, incorporates the characteristics of both models. Messages on these IMPs can involve only two individuals, but, at the same time, can become viral and reach a massive number of users through their network.

These applications have key features that contribute to a message to disseminate in the network and then become viral. The groups' creation allows hundreds of users to talk simultaneously, and each one of them has access to forwarding tools to quickly share the content to other users. In the forwarding process, the encrypted nature of these apps makes it hard to track the original source or author of a message. The users access the information that has been forwarded, but they may not distinguish if this received content is originated directly from one of their known contacts or if it comes from a more extensive path created by a totally unknown creator. Then, being a message received from a friend or a complete stranger, from the perspective of the final users, all messages are equally, and always received from a direct contact within a chat that this final user participates, even though if it has been already shared by multiple users in the network. In addition, it may have no clue of how many times it had been shared before reaching the user's phone.

Misinformation campaigns take advantage of this space that blends public and private information to disseminate false content misleading the users, where the real authors are still anonymous, hidden by the closed nature of the network. Then a question that remains with this aspect of IMP is that should an encrypted message be able to go viral?

Next, we discuss more specifically some of the features that contribute to this scenario, comparing the differences of how they are used by Telegram and WhatsApp.

7.1.1 Group Creation

The group creation feature is one of the most important aspects of IMPs. The groups are usually created around specific topics, from more personal such as family and work, to more general such as sports and politics. While creating a group, the user can add friends and other users to it. Although users can manually add their friends and contacts to groups, both Telegram and WhatsApp allow the creation of invite links for their groups. Once a person clicks the invite link, they can freely join the group. This is an effective manner to advertise groups and reach more users, as the group creator does not need to add each contact one by one. More important to our context, the invitation URL can be easily shared on other social networks, helping to connect unknown users who share similar interests on Web.

Some of these groups may become very popular and attract a lot of users, but the number of members is limited to a constant that is defined by the IMP. In WhatsApp, groups cannot have more than 256 users, i.e., after reaching this maximum, no one can join it nor be added. Telegram, on the other hand, has a much larger limit and groups can have up to 200,000 members. Moreover, Telegram has a different form of group, which is called *channel*, a type of broadcast group where only the admin can send messages (one-to-many communication) and can have an unlimited number of members. The limited size of WhatsApp gives groups a more personal nature, even though they can also be composed by many unknown users. On the other hand, Telegram groups create large communities that can quickly spread messages for thousands of people in the network.

7.1.2 Forwarding Messages

Both Telegram and WhatsApp also have the functionality to forward a received message to other users or groups. This tool helps a user redirect a message received from one contact to a third different user. This is especially common to multimedia messages of image, audio, and video, and it is one of the most responsible tools for the propagation of misinformation through the networks of these platforms. With one click, users can share a message with multiple other users, allowing that message to become viral faster, principally when it reaches multiple groups, each one with hundreds of users.

In WhatsApp, the messages can be shared to up to 5 other users/groups and only 1 user for viral contents[4], also the users can create a broadcast list to send a message up to 256 other users/groups. When someone redirects a message using this tool, the message is flagged as “Forwarded”, but users may not know who is the original author while the authors do not know that their messages were forwarded. This allows the rapid spreading of messages in the network and covers the source, creating a fertile environment for misinformation. In Telegram, the messages also can be forwarded to other users, but

[4] The forward limit in WhatsApp was initially 20 but was reduced after reports of the use of WhatsApp to spread rumors and fake news in India and Brazil.

just one at the time. However, it specifies the source name when the message is forwarded. This means that the person to whom the message was forwarded will know the creator of that message, which provides more transparency to the platform. Later in this chapter, we will study this attribute in more depth.

7.1.3 Encrypted Content

Another feature that it is needed to take into account in the process of dissemination of information on the Instant messaging platforms is the encryption. WhatsApp has encrypted messages as default (and there is no way to disable it). Even though Telegram has the “secret chats” that are encrypted one-to-one conversations, the default chat has no encryption and group chat does not support it at all. While encrypted messages give more privacy to users, it also hinders the track of dissemination of the content within the platform. This feature can be exploited by malicious users to become anonymous within the platform. The privacy and anonymity of encryption allow an unmoderated and free environment for occurrences of many cybercrimes and harmful content, such as child sexual exploitation, terrorism, extortion, hate speech and also misinformation which is already been increasingly reported by Facebook within its platform (MAYER, 2019).

This is not an easy problem to solve. While there are organizations and governments requesting that companies do not proceed with plans to implement end-to-end encryption across its messaging services without including system backdoors with means for lawful access to the chats content (PATEL et al., 2019), it would also impact in safety and privacy for all users. Others argue that there are forms of content moderation may be compatible with end-to-end encrypted messaging, without compromising important security principles or undermining policy values (MAYER, 2019).

Furthermore, it is important to discuss the encryption as it implies more privacy and security to the users, but it does not necessarily always mean more private. As observed with other features, Telegram and WhatsApp do not only work as applications for personal conversation, they also work as platforms of mass communication, with content that can go viral and become public, yet it is done in an encrypted way. The encryption in IMPs, therefore, provides security of data, but does not prevent the spread of the information on the network. While it is a good way to block outside access to user data, it can give a wrong feeling of privacy on users, and makes it impossible to track creators of misinformation circulating in the network.

7.1.4 The Public and Private Ambiguity

In social media, where much of the information does not sustain the users' attention, the quality of the content is not a sufficient condition for a message to become viral (WENG et al., 2012). Even high-quality has little advantage over low-quality information in this environment, (QIU et al., 2017) since information overload contributes to a degradation of the user's discriminative power of quality. Also, the popularity of a piece of content may be distorted by some other previous related information and social influence such as example number of views, downloads, or even the arrangement/ranking that items are shown (SALGANIK; DODDS; WATTS, 2006). On IMPs, it is even harder to characterize the shared information. Information, misinformation, propaganda, private messages are all disputing the same space on the user's screen, and they are promptly replaced by a new piece of information due to the linear and continuous format of chats.

By analyzing the platforms and their virality features, we realize that there is an ambiguous utilization of these platforms, which allows the encrypted content to propagate on the network without anyone taking responsibility or credit for them. OGHUMA et al., 2016 argued that perceived usability, perceived security, perceived service quality and confirmation determine the intention of usage of instant messaging platforms. Groups that seem to belong to the users' private lives, like school and family groups, are then connected to an extensive external network. For that reason, it is difficult for users to identify and distinguish the origin of some messages. However, because these messages arrive from personal contacts whom they trust (e.g friends, family, work colleagues), they are more likely to be perceived as reliable messages even when they are not. In other words, IMPs allow messages from unknown sources to be vastly disseminated by known and trusted contacts.

Part of this problem emerges because these messaging services, in practice, operate at the same time as a personal communication tool and as media companies. As a personal communication tool, they provide one-to-one chats and ensure to the user anonymity and security by encrypting transmitted data.On the other hand, as a media platform, they work similarly an online social network as they are able to transmit and disseminate information to the public and enabling virality functions (NAPOLI; CAPLAN, 2017). In terms of content curation and dissemination, then, IMPs also share some similarities with traditional media companies. However, unlike traditional media companies, these features allow any content creator to broadcast anonymous and potentially malicious messages to thousands of people without any ethical or legal regulation of their content.

Some studies point this ambiguous use of WhatsApp as a personal private tool for closed conversation and at the same time as a viral mass communication with public chats. For example, MALKA; ARIEL; AVIDAR, 2015 studied how Israeli citizens use *WhatsApp* during wartime, in which they observed that the platform plays a central role in the lives of its users in conflict scenarios, functioning as a multifunctional channel of communication to get the news from the conflict as well interpersonal chats to be in contact with those in battlefield. Similar duality between mass communication and private chat is also being highlighted on Telegram during the Ukraine war. Both Russian and Ukrainian made use of the app to get updates and news about the war, and the Ukrainian president Zelenskyy itself broadcast messages from its channel on Telegram (ALLYN, 2022) while also raising concerns about privacy within the platform.

Since WhatsApp claims to keep the app more personal and private (WHATSAPP, 2020a), the flow of messages can easily get viral within its wide and well-connected network, which may confuse the user about the intention behind that message and favor the viral spreading of misinformation on this platform.

All of those issues, summed to the fact that many users are getting the first access to the Internet through the first smartphone (NEWMAN et al., 2021). Those people do not know how to navigate the rest of Web beside the IMP they used to communicate. They are not aware of the vastness that permeates the user's network and how far a message can travel before reaching them. In this scenario, malicious users can abuse of the platform to instigate users to share their content and believe in false news by thinking that all they are receiving are part of a small and secure community of closed friends and family that they are members. This intimacy sense of the platform, combined to viral features it provides, turn it in a dangerous place where misinformation campaigns become very effective and hard to combat.

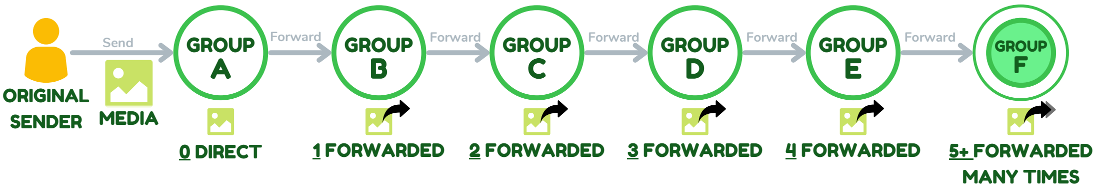
7.2 How Forwarding Chains Work on WhatsApp

When someone sends a message in a chat on WhatsApp, other users can select this message and easily share it by using the Forward button on WhatsApp. The forward feature allows users to share messages from an individual or group chat with another individual or group chat. Forwarded messages are then indicated with a "*Forwarded*" label and a symbol, which alerts users that the message originally came from someone else.

There are, however, some limits and exceptions in the Forwarding process. Users can forward a message with up to five chats at one time. If a message has already been forwarded, you can forward it to up to five chats, including a maximum of one group chat[5]. WhatsApp also implemented the concept of messages that have been "*Forwarded Many*

[5] <https://faq.whatsapp.com/193766651706684/>

Figure 7.1: A forwarding chain spreading messages until it reaches the label "Forwarded Many Times".



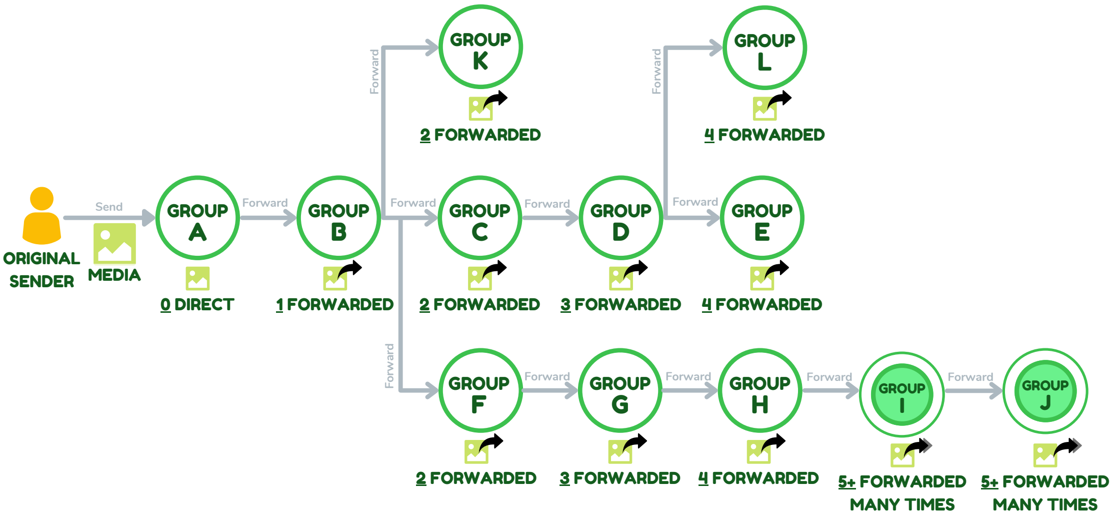
Source: The Author.

Times". These messages are labeled on WhatsApp to indicate they did not originate from a close contact, more specifically, when a message is forwarded through a chain of five or more chats, meaning it's at least five forwards away from its original sender as represented in Figure 7.1. When a message is forwarded many times, it can only be forwarded to one chat at a time (WHATSAPP, 2020a).

Therefore, for a message to become a "viral" message by WhatsApp, it needs to travel a path of at least five hops from the author to reach the final user. Saying that, however, there are many examples of forwarding forks and interruptions that can make this count harder than it looks. Next, we will show some examples of this occurring during sharing media, making this a non-linear path.

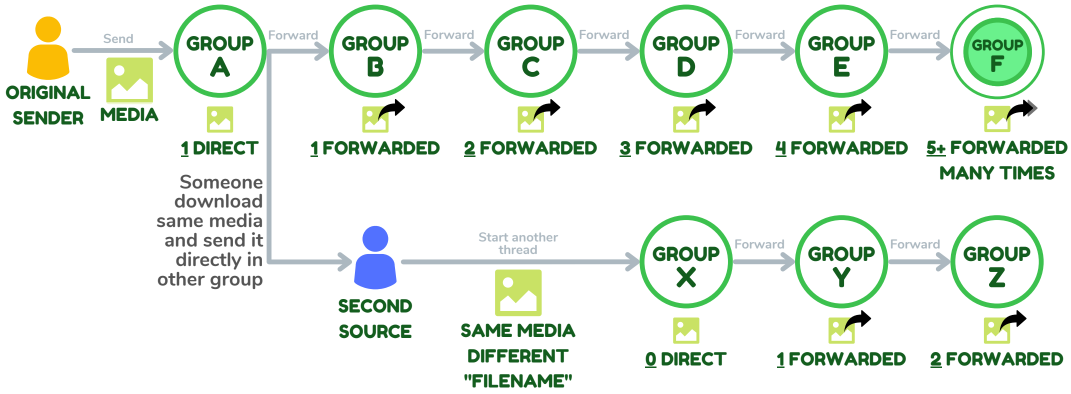
For example, since users can select multiple chats to forward a message (up to five), each "path" created by this fork receive its own counter, meaning that forwards made in parallel paths will not be taken into account for the "Forwarded Many Times" labeling, as shown in Figure 7.2.

Figure 7.2: Flowchart with a scenario of a forked thread of a spreading message chain on WhatsApp.



Source: The Author.

Figure 7.3: Flowchart with a scenario of a spreading message with a second user breaking the chain and starting a new thread with same media file.



Source: The Author.

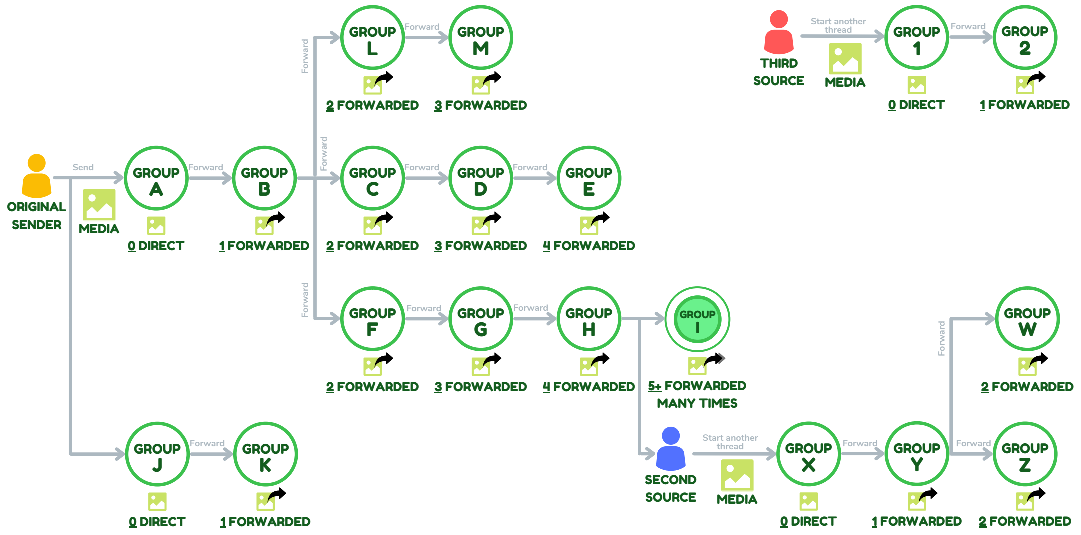
Another issue is that the counter was created only for continuous forwarded chains. Therefore, if a user in this process downloads a message content and resends it directly to someone else (instead of just use the forward button), it will reset the counter, interrupting the viral tracking, as shown in Figure 7.3. Since downloading a multimedia message and sending it directly from your phone's gallery or even copy-pasting a message is very easy for most of the users, it is very simple to get around the restrictions and fool the viral labeling of WhatsApp. Moreover, given that WhatsApp messages are E2E encrypted, the new message sent will be a completely different one, and WhatsApp will not even notice that both are the same.

Then, the network of a spreading message may build a complex flow, such as the example in Figure 7.4, in which even though there are 19 shares of the same piece of media, only a single one would be flagged as "Forwarded Many Times" in this scenario.

WhatsApp has implemented these symbols and tags to help users to distinguish whether a message is more personal or not, in special, the "Forwarded Many Times" is used to label a message as something even more disseminated through the platform. Because of this, in the next sections, we will refer to messages marked "Forwarded Many Times" as *viral messages* for more simplicity.

However, even though the system logs behind WhatsApp forwarding and its limits are very straightforward, they may present many instances that need a better understanding during the actual use of the app, and some of them can be occult for the users; especially regarding the feature of "Forwarded Many Times" flag which can be misleading of what that means.

Figure 7.4: Flowchart of a larger forwarding chain on WhatsApp with just one message flagged as “Forwarded Many Times”.



Source: The Author.

7.3 Extending the Data Collection

As discussed before, we do not have easy access to the communication network of IMPs, even though there are billions of users exchanging messages daily through these platforms. Since we found a large number of public groups shared on the Web using an invitation URL, the most effective method to collect data is by joining these groups and monitoring their users and shared messages. GARIMELLA; TYSON, 2018 have also collected large-scale data on WhatsApp for public groups from India and Indonesia. By applying our methodology of merging similar content in both our Brazilian data and their substantial dataset for India and Indonesia, we can build a network of users and groups and track the information flow for all three distinct countries. This allows us to evaluate the dissemination of information in three similar contexts of elections, but independent events, that all occurred on WhatsApp as the media of spreading.

In this step of the thesis, we follow a similar methodology to collect data from another big IMP – Telegram. We also focused on groups dedicated to political discussions, which played major roles in misinformation dissemination in Brazil. For Telegram, there is an official Application Programming Interface (API)[6] that provides full support to interact with the system. The API allows users to build their own customized Telegram clients, develop bots to connect and moderate chats, and also collect messages, users and some other metadata of the chats – groups and channels – of which they are members. First, to find groups of interest, we searched the Web for invite URLs that match a list of keywords related to elections in Brazil. Since Telegram is less popular than WhatsApp,

[6] <https://core.telegram.org/>

Table 7.1: Overview of the instant message platforms datasets.

Platform	Country	#Users	#Groups	Total Images	Period ∼ 2,5 months
WhatsApp	Brazil	17,465	414	416k	2018/08/15 - 2018/11/01
	India	362,739	5,839	810k	2019/03/15 - 2019/06/01
	Indonesia	8,388	217	21k	2019/03/15 - 2019/06/01
Telegram	Brazil	62,836	49	17k	2018/08/15 - 2018/11/01

we discovered approximately a hulinks, links from which 50 were selected following our relevance criteria. Then, we used the Telegram API to join and collect all messages shared by those groups. Although Telegram gives users access to messages that were posted before their joining dates, we restrict the data collection to the same period as our WhatsApp data. Thus, we collected all messages that were shared in these groups during the second semester of 2018. For each message, we have: (i) the user ID, (ii) the name of the group in which the message was posted, (iii) the timestamp, (iv) the text of the message, and (v) the media type of the message (e.g. text, image, audio and video).

For all set of data collected, we analyzed the messages around the election day, considering approximately three months of data for each country. We kept the same time span for the four datasets to facilitate the comparison among them. The dataset overview and the total number of distinct images collected during this period are described in Table 8.2. As expected, Brazil and India have a much larger volume of data (258k unique images from Brazil and 509k in India) shared on WhatsApp compared to Indonesia (16k unique images), as they have many more groups and users registered in our data collection system.

Our methodology collects a large dataset from public groups, but we are aware that, at least for WhatsApp, most of the conversations occur in private channels. A key limitation of this is that our results refer only to the content that circulates on the public layer of these platforms. Nevertheless, there is evidence suggesting that public groups make up the key backbone of the misinformation campaigns on WhatsApp (PONNIAH, 2019).

Furthermore, this project brings a considerable amount of data that can help to elucidate how WhatsApp and Telegram are being abused for mass communication (MAGENTA; SOUZA, 2018) and the amplification backbone composed by public groups that distribute messages in bulk for thousands of users. At least, our results provide a ‘lower bound’ on the ability of messages to spread on WhatsApp and Telegram, since the network we consider here is a subset of the entire network. Thus, the issue may be much bigger if considering the entire network

7.4 Network Topology

In this section, we dive in the network structure of public groups from instant messaging platforms by reconstructing the graph of users and groups. Also, we analyse the main characteristics of these networks, comparing it to other synthetic and real social networks. Moreover, we discuss some of the features of each platform and their peculiarities that make it easier for some content to go viral.

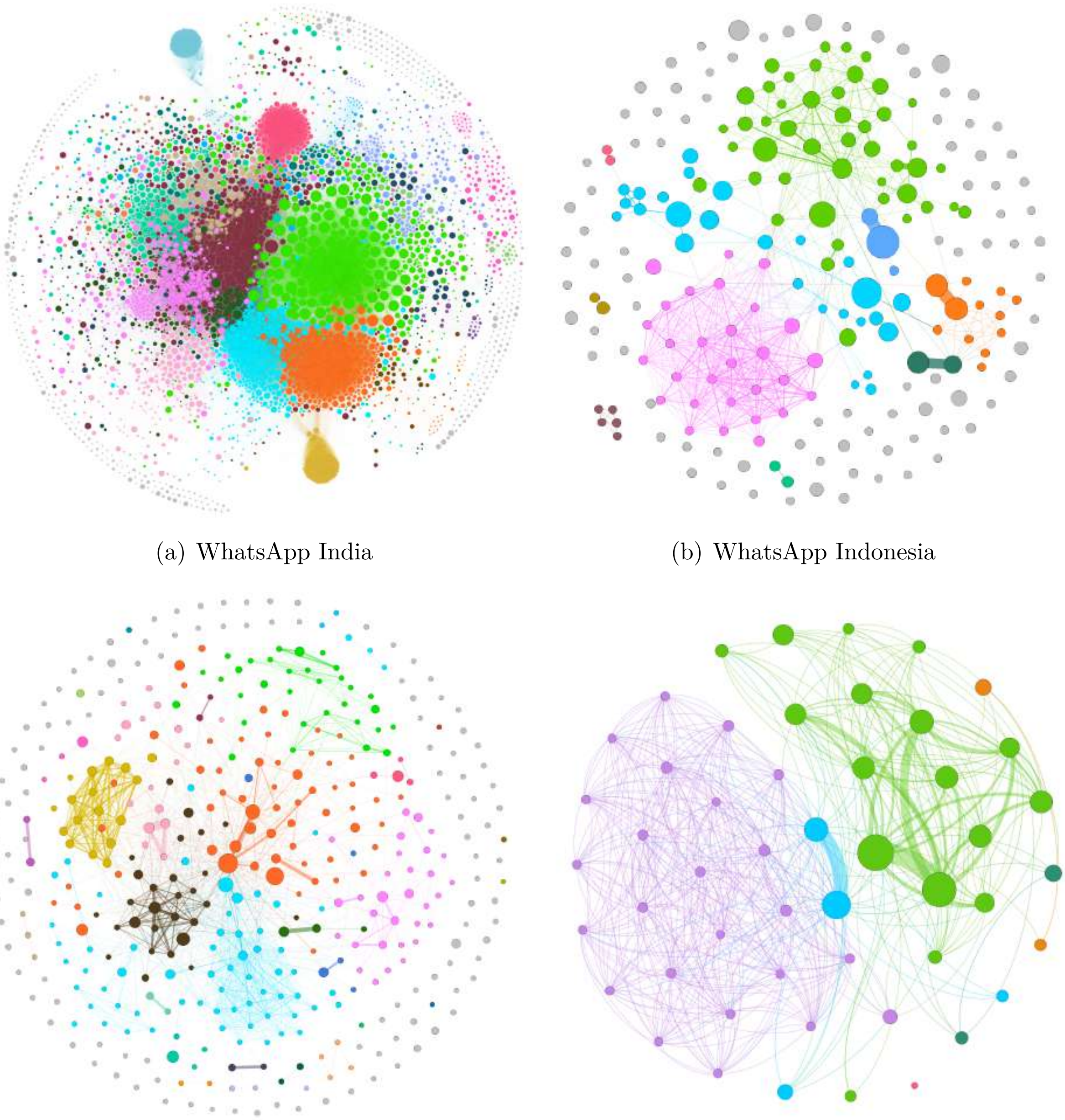
The IMPs have a very peculiar network topology, which is composed of traditional social links and also very large cliques – i.e., public group chats – where every user is connected to each other and, therefore, exposed to every message sent to that group. The size of groups in Telegram is also unlimited and one message sent to only one group can reach an arbitrarily large number of users. Thus, because groups connect all their members simultaneously, they can work as a mass media communication platform.

To create a network from WhatsApp and Telegram groups, we connect two groups if they share a common user. In Figure 7.5 we show these networks for both platforms. The size of the node is proportional to the number of members in that group. We colored nodes according to its community in that graph, following the modularity algorithm (BLONDEL et al., 2008). By observing the graph of groups, we can note already some similarities with other online social networks. Note that there is an evident largest connected component in all graphs and some group clusters. Also note that some groups position themselves as bridges and hubs, connecting different communities in the network structure.

These networks suggest that we should not think of IMPs such as Telegram and WhatsApp as simple apps that send messages. We need to consider those platforms as well-connected networks, and each of these groups represents hundreds or thousands of users that may have many more connections than we can see with this data. The possibility to create public groups allows multiple and socially distant users to connect to each other across the network, forming a complex social structure able to flow high volumes of information by sharing messages in large chats. Although these group networks formed from WhatsApp and Telegram data resemble many other social networks, little is known about the differences in information dissemination.

From another point of view, we also modeled the relationship between users by building a user network, in which each node is a user and an edge is added between two nodes if the corresponding users have shared image content in at least one group in common. Node size represents the number of groups in which the user shared images. The user networks are naturally larger and harder to visualize. For illustration purposes, Figure 7.6 shows a subgraph of the network built from the data from WhatsApp Brazilian groups with 5,700 nodes/users. The network structure of the groups is evidenced by the clusters formed. We note a large number of users blending together and connecting to

Figure 7.5: Public groups network for WhatsApp and Telegram. Each node is a group and edges represent members in common.



(a) WhatsApp India

(b) WhatsApp Indonesia

(c) WhatsApp Brazil

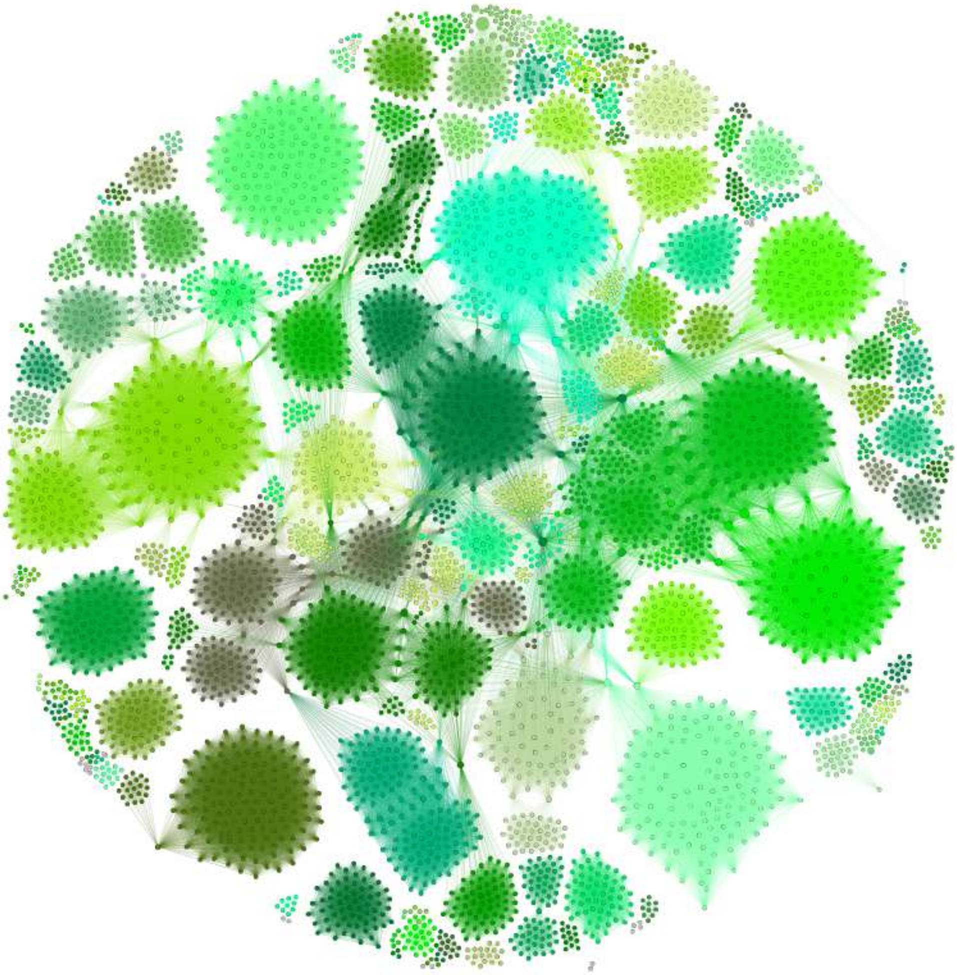
(d) Telegram Brazil

Source: The Author.

each other inside those groups. Most users indeed form a single cluster, connecting mostly to other members of the same community. On the other hand, there are also a few users who serve as bridges between two or more groups linked by multiple users at the same time. Furthermore, a few users work as big central hubs, connecting multiple groups simultaneously. Lastly, some groups have a lot of users in common, causing these groups to be strongly interconnected, making it even difficult to distinguish them.

The composition of users and groups is one of the first analysis in the direction to understand the topology of those networks. In Figure 7.7 and 7.8, we show the distribution

Figure 7.6: User network in which nodes are users and edges connect users with group in common for Brazilian WhatsApp groups.

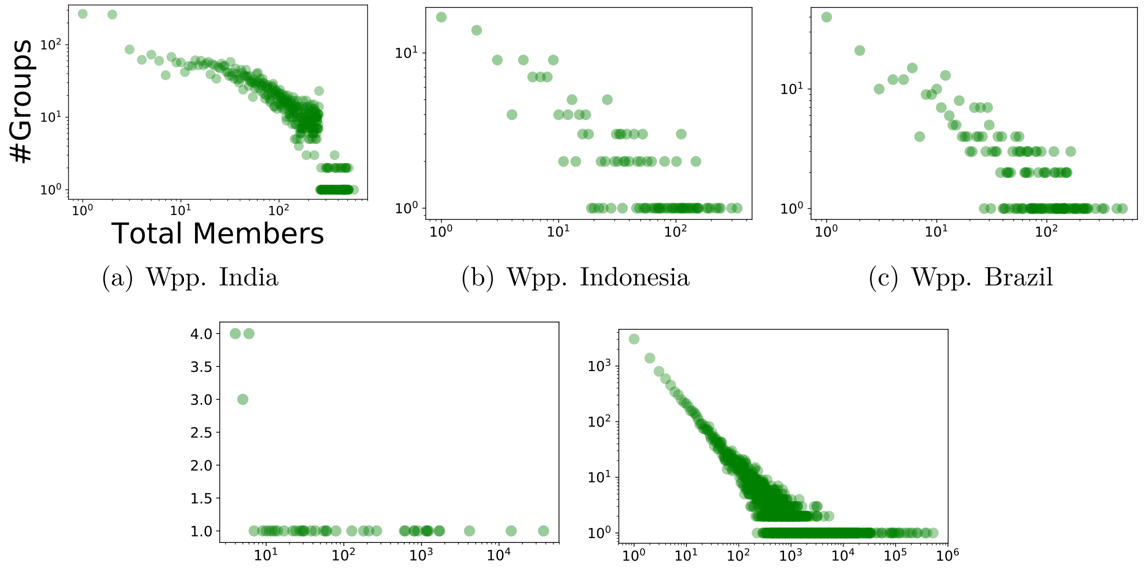


Source: The Author.

of groups per user and users per group, respectively. In order to compare the peculiarities of WhatsApp and Telegram with other popular platforms, we compare these networks with the Reddit network that was made available by TAN; LEE, 2015, which modeled the subreddits as groups and users as members. Note that even though Reddit has the same group characteristic, there are other different features of IMPs that lead to very different network structures.

Considering the WhatsApp curves, we see that they are similar to a well-behaved power law curve, which naturally yields a larger variance. Note, in Figure 7.8(a), that in India we have users who participated in more than 300 groups. Figure 7.7(d) shows that the number of users in a single group on Telegram can vary between 1 to almost 100,000 users. Even with this small dataset regarding public groups on Telegram in Brazil, we notice that these popular groups can be seen as hubs in the network, and they have a huge impact on the spread of contents through the network. For Telegram, even in a smaller data collection, we can see some larger groups, although none of them reach the limit of 200,000 on the platform. By comparing this data with the one collected from WhatsApp

Figure 7.7: Distributions of the number of members per group on WhatsApp, Telegram, and Reddit.



(a) Wpp. India (b) Wpp. Indonesia (c) Wpp. Brazil

(d) Telegram Brazil (e) Reddit

Source: The Author.

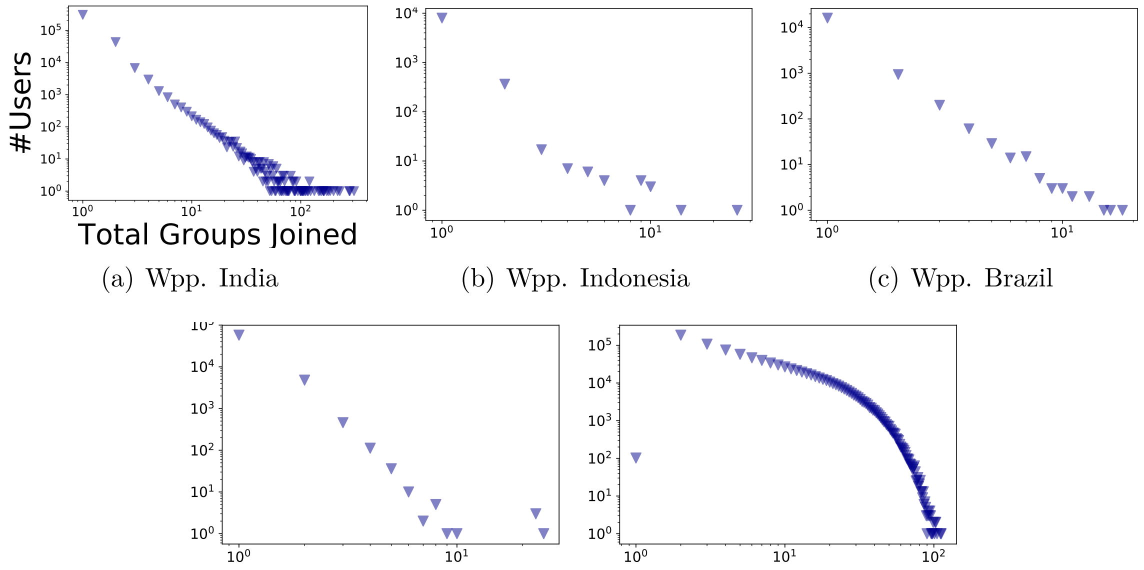
Brazil, which represents the same scenario (Brazilian elections of 2018), we see that the maximum of 256 members in groups is a determining element in the network, effectively capable of limiting the group size. In other words, although Telegram is less popular than WhatsApp, Telegram has groups with more users, what indicates that if the WhatsApp limit was bigger we would also have bigger groups. This is also true for India (Figure 7.7(a)), where there are over 300k users and more than 5k groups[7].

On the other hand, in Reddit, where there is no limit, it is possible to see that the group size can be as large as 10^5 members, as shown in Figure 7.7(e), which creates big hubs of users. As both platforms have no limit on the number of groups users can join, we expect to see no differences in the total number of groups that users participate. However, in Figure 7.8(e) it is possible to notice that in Reddit, the distribution has an exponential decay, with a limit on $\approx$ 100 groups.

For the number of groups per user, none of the platforms have known limits for how many groups a single user can join. Despite the different limit of members in Telegram and WhatsApp groups, we can see that the distribution of groups per user in our data is similar. The different sources of data from Telegram, WhatsApp and Telegram presented a similar distribution of groups per user as shown in Figure 7.8, with many users joining a single group of just a few groups while a minor number of users joining a large number of groups simultaneously. The network of Indian WhatsApp (Figure 7.7(a)) has the users with more groups, including users who are members of up to 300 groups, this is a

[7] In our data, some groups have more than 256 members, because our data is a temporal snapshot and members can leave and join groups during this time.

Figure 7.8: Distributions of the total groups joined per user on WhatsApp, Telegram, and Reddit.



(a) Wpp. India (b) Wpp. Indonesia (c) Wpp. Brazil

(d) Telegram Brazil (e) Reddit

Source: The Author.

larger maximum than Reddit (Figure 7.8(e)), where users do not surpass the maximum of joining one hundred groups. This pattern, which is present for all networks, reveals that, independently of the network topology or group limits, users do handle a large number of groups at the same time. All platforms have just a few users who participate in more than 100 groups. This suggests that users have a limit of information they can follow, so they usually prefer to join only a few groups, even though those platforms do not impose a limit to the number of groups they can join.

Next, we compare the characteristics of the WhatsApp and the Telegram group networks (Figure 7.5) with other social networks: (i) synthetic networks generated by well known social network models, including the Barabasi-Albert (BA) scale free model, the Erdős-Rényi model, the small world model (WATTS; STROGATZ, 1998) and the Forest Fire network model (LESKOVEC; KLEINBERG; FALOUTSOS, 2005), for which we used the same number of nodes in the Indian dataset in order to create a comparable network; (ii) the network of subreddits from Reddit (TAN; LEE, 2015) and (iii) the Flickr network (MCAULEY; LESKOVEC, 2012), which, differently from the WhatsApp, Telegram or Reddit group networks, represents the network of images shared by users on the platform. The results are shown in Table 7.2. First, observe that WhatsApp shares common characteristics with other real-world social networks: high clustering coefficient, giant largest connected component, and small average path length, which are all typical properties of social networks. On the other hand, the Telegram network has less than 50 nodes, but it has a high density and also shares some characteristics with other social networks, such as high clustering coefficient and giant largest connected component. Finally,

Table 7.2: Network metrics for WhatsApp compared to other networks.

Network	#Nodes	#Edges	MD	CC	D	APL	Density	LCC	PC
W. India	5,839	407,081	139.44	0.59	11	3.17	0.0239	92.6	0.295
W. Indonesia	217	699	6.44	0.38	9	3.09	0.0298	55.3	0.290
W. Brazil	414	14,00	6.76	0.32	8	3.19	0.0164	65.2	0.346
Telegram	49	464	18.94	0.88	3	1.72	0.3946	98	0.014
BA Scale free	5,839	792,300	271.38	0.10	3	1.95	0.0465	100	0.008
Erdős-Rényi	5,839	1,534,952	525.76	0.09	2	1.91	0.0901	100	-0.001
Smallworld	5,839	604,250	206.97	0.34	3	1.98	0.0355	100	0.007
FireForest	5,839	12,930	4.43	0.42	17	5.25	0.0008	100	-0.066
Reddit	15,122	4,520,054	597.81	0.82	6	2.03	0.0395	99.8	-0.045
Flickr	105,938	2,316,948	43.74	0.09	9	4.80	0.0004	99.8	0.247

(i) **MD** - Mean Degree
(ii) **CC** - Clustering Coefficient
(iii) **D** - Diameter
(iv) **APL** - Average Path Length
(v) **LCC** - Largest Connected Component (%)
(vi) **PC** - Pearson Coefficient

all WhatsApp networks show a high assortativity coefficient, which indicates that nodes tend to be connected with other nodes with similar degree values. In epidemic analyses, it can help us to understand the spreading of infection across the network, as a misinformation campaign targeting high degree groups is likely to spread to other high-degree nodes.

7.5 Information Propagation on WhatsApp

We use the epidemiological model of Susceptible-Exposed-Infected (SEI) (LI; ZHEN, 2005) to estimate the virality of malicious messages in IMPs by assuming misinformation as an infection that spreads to users through the group network. In our scenario, the nodes are members of various groups and the infected nodes can spread the infection to a entire group at once, exposing all their participants. In this model, *Susceptible* (S) is the initial condition in which the user did not have any contact with the infection; *Exposed* (E) are those who received the misinformation through any of the groups they participate but did not share it; *Infected* (I) is the final stage in which a user who was exposed to the content shares this message in the network. This model has two basic parameters: *virality* (α) and *exposition* (β). Since WhatsApp and Telegram have clear restrictions on how many people a message can be forwarded to, we also implemented a third parameter *forward limit* (φ) to test the sharing restrictions that could be implemented by the IMPs.

The **virality** (α) of malicious content controls the rate of infected users. This indicates the probability of an infected user to share the content with neighbors. We consider that users are infected when they forward or broadcast this content, as it indicates a degree of belief in the shared message. The **exposition** parameter (β) refers to the rate at which exposed users become infected. It represents the probability of an exposed user to transform into an infected one. Lastly, the **forward limit** (φ) indicates the maximum number of groups an infected node can spread the infection to. The model can be summarized by the following equations:

$$\begin{aligned}\tfrac{dS}{dt} &= -\beta IS\\ \tfrac{dE}{dt} &= \beta IS - \alpha E\\ \tfrac{dI}{dt} &= \alpha E\end{aligned}$$

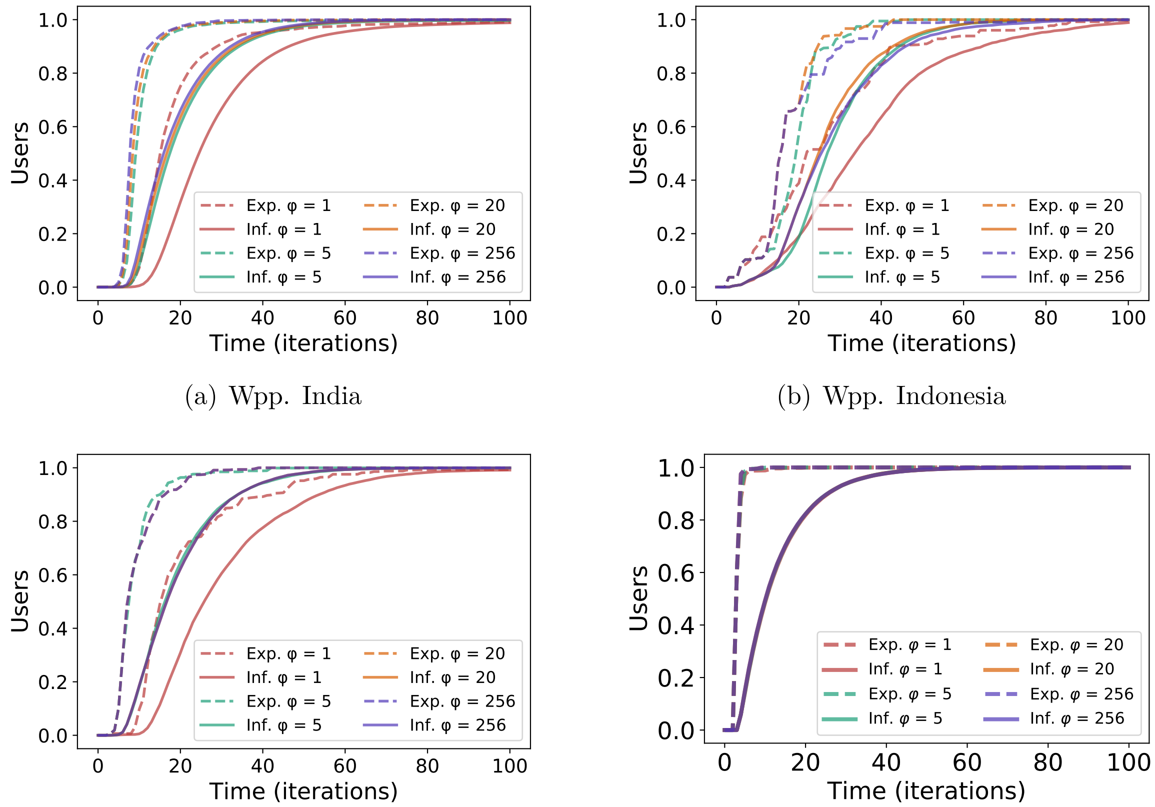
Note that $\frac{dS}{dt}$, $\frac{dE}{dt}$, and $\frac{dI}{dt}$ represent the rate of susceptible, exposed and infected users at each interaction, respectively. While $\frac{dS}{dt}$ corresponds to a decay rate, since the number of susceptible users decrease with time, $\frac{dE}{dt}$ and $\frac{dI}{dt}$ represent the increase in the number of exposed and infected users. Eventually, the number of users infected will reach 100% of the users in the network. The simulation starts by randomly selecting one user to be the first infected node that will start spreading the information. For each user exposed, they have a probability given by α to share the malicious message. When these infected nodes decide to forward, the parameter φ will limit the maximum number of groups they will send the content to. Once the information is forwarded, each user in the groups that received the message is exposed. Then, each exposed user has also a probability of β to become an infected node and share the content. This process is repeated until all users are infected.

7.5.1 Experimental Results

We perform several experiments using our SEI model to compare the information dissemination in different scenarios. Since it would not be possible to reach isolated nodes using the whole structure, only the largest connected component was considered.

Figure 7.9 shows the fraction of users infected over time for all countries when the forward limit (φ) is varied, i.e., to analyze how the restrictions implemented by IMPs can impact the spread of information. We set the forwarding limit to 5 groups (the real scenario of WhatsApp), 20 groups (the previous limit), 256 groups (the limit for broadcasting), and to 1, which is the Telegram limit and a more restricted limit compared to WhatsApp. First, note that the rate of users exposed in the network grows very fast,

Figure 7.9: SEI model varying the forward limit (φ). $\alpha = \beta = 0.1$.



(a) Wpp. India

(b) Wpp. Indonesia

(c) Wpp. Brazil

(d) Telegram Brazil

Source: The Author.

regardless of the forwarding limits. A maximum of 60 iterations is enough to infect the entire network. Moreover, limitations on forwarding slightly diminish the velocity of spreading, but does not stop it completely, especially for exposed users. We also simulate those scenarios using the Telegram network. The results do not show significant changes between the different forward limits as users are quickly exposed and infected. For all scenarios, the infection exposes all users after a few iterations of the SEI model. This occurs due to the large size of groups, as a single "share" in those groups is enough to expose thousands of users.

7.5.2 Varying Virality

We also evaluate the time needed for (mis)information with different viralities to infect all users. Figure 7.10 shows the timelimits,d to infect 100% of the users by varying α from 0.001 up to 1.0, with different forwarding limits. Observe that in situations of mass dissemination (high α), it is difficult to stop the infection because of the strong connections between groups. However, note that the limits in forwarding and broadcasting help to slow down the propagation, mainly in larger networks (e.g. India). In short, **limits on forwarding and broadcasting can reduce the speed of the dissemination by one order of magnitude, even considering all values of virality** (α).

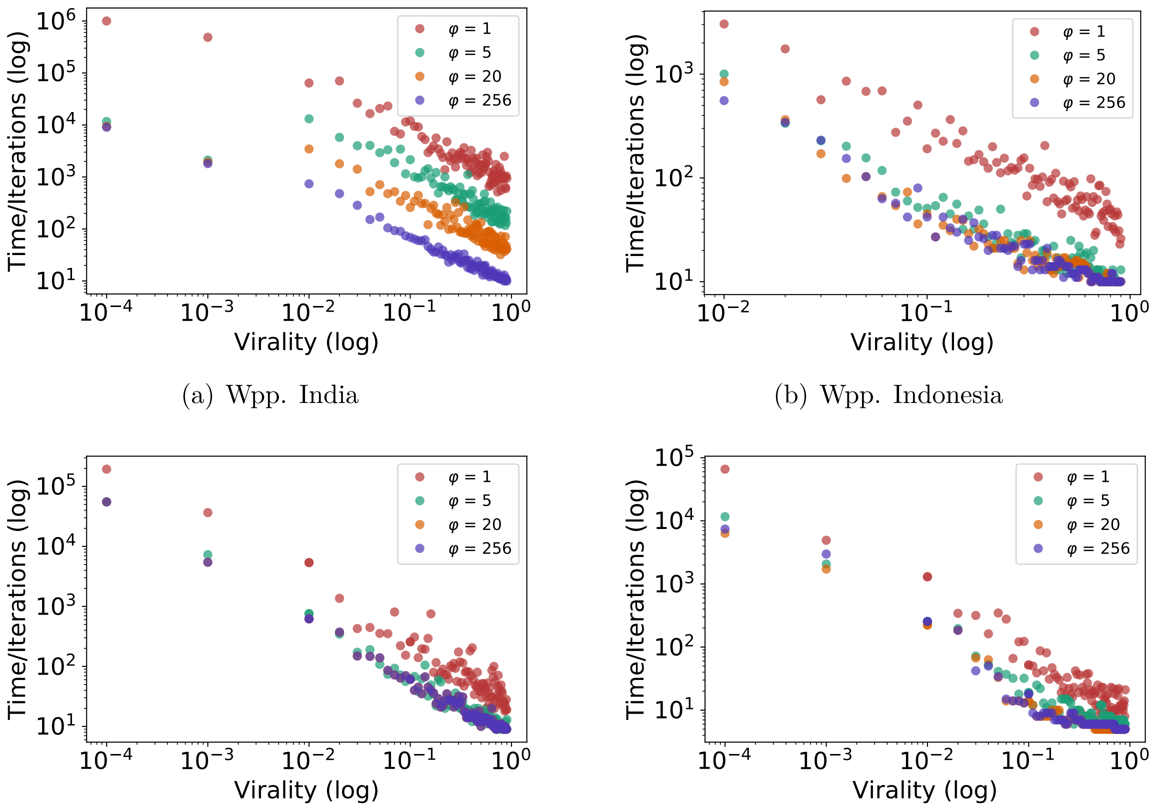
Also, observe that the differences between the forward limits of 256, 20 and 5 are more evident in the Indian WhatsApp network. The other remaining networks show just small differences on the time required to infect all users. In those networks, as there are almost no users in our data who are members of more than 20 groups, the forward limit does not have a big impact on the speed of the infection. Moreover, just a few users are part of more than 5 groups, which explains the similar results for when the limit is set to 20 and 5. In this case, just a few users are affected by those restrictions. However, when we look at the results when the forward limit is 1, there is a substantial difference in the spreading time for all scenarios as there are many users that participate in two or more groups. These results suggest that the forward limits have a similar impact independently of the virality of the infection. Experiments with both low α and high α slow down the velocity of dissemination in India, but have minor impacts on the other networks.

Furthermore, those simulations suggest that in dense networks, in which many users are part of multiple groups, the forward limit has a considerable effect on the speed of information dissemination. On the other hand, in sparse networks, where just a few users participate in a large number of groups, the forward limit has a negligible effect on the time required to infect all users. Additionally, this shows that limits on forwarding drastically affects those who join multiple groups, which represents just a small portion of all users. The majority of members just participate in a single community and feel little or no impact by this limitation.

7.5.3 Virality Decay

As users may lose interest in some topics through time, it is natural for the virality factor to decay. So, we expect a time limit for the content to spread (i.e., the content circulates until it loses attention and stagnates). We simulate this decay by making our virality parameter α to decrease over time. To do that, we add a time period called "lifetime", which denotes the maximum duration of an infection in the simulation before it is entirely extinguished. A short lifetime represents a large decay of virality, meaning that it quickly goes to zero and stagnates, while a bigger lifetime represents a slow decay in the virality rate, which makes the virus to live longer and facilitates its spread through

Figure 7.10: Time to infect all users in the network on simulations of SEI model by varying the virality (α) from 0.001 up to 1.0.



(a) Wpp. India

(b) Wpp. Indonesia

(c) Wpp. Brazil

(d) Telegram Brazil

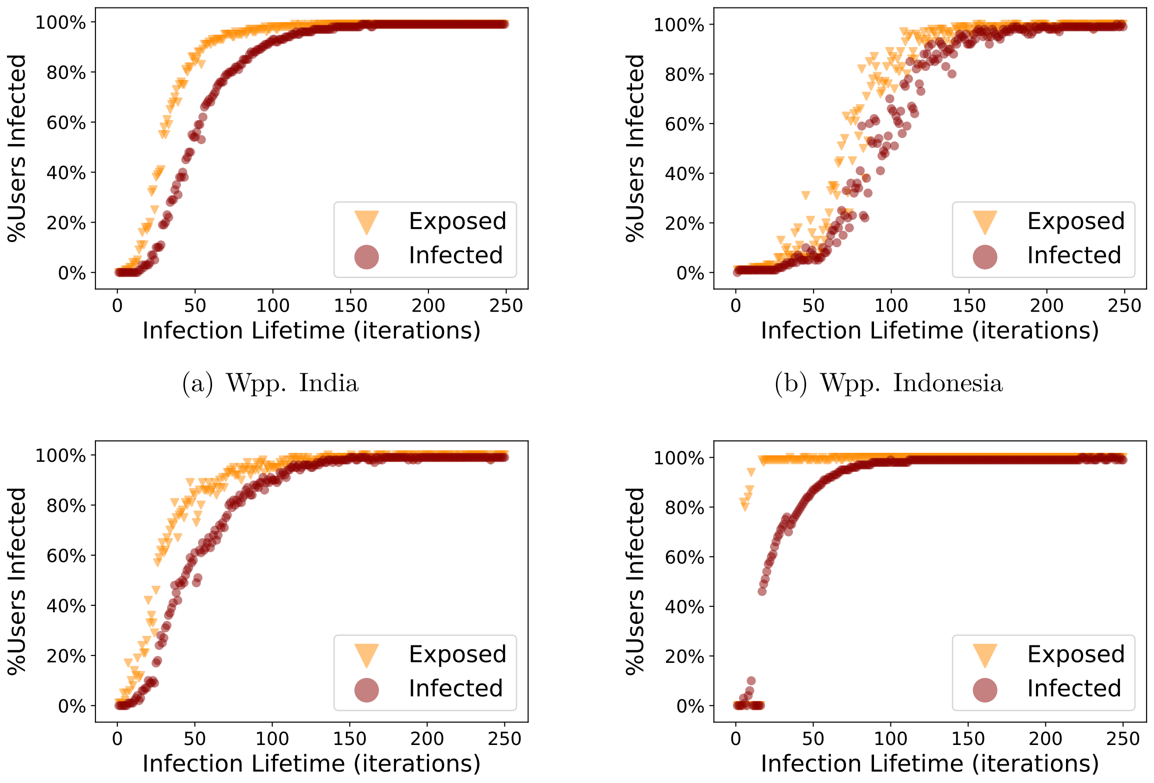
Source: The Author.

the network.

Figure 7.11 shows the percentage of users infected as we increase the lifetime of the infection. Each data point in the plot indicates a simulation where we fixed the values α, β and increased the lifetime that an infection could last. We observe that, for WhatsApp, an infectious content that lasts 100 iterations or more is powerful enough to expose more than half of the population for all networks analysed. Furthermore, when this content persisted in the network for at least 150 iterations, it infected almost 100% of the users. For Telegram, even shorter lifetimes are enough to expose a great part of the network. Since there is a larger group size in this platform, when the infection reached those groups, it exposed the majority of the users on the network at once.

Note that there is a window of possibility to identify infectious misinformation already spreading (say, around 50 iterations), where a large enough sample of the users were exposed to the content but were not infected. If malicious content is detected during this period of time, it can be used to nullify its virality (e.g. disabling forwarding on that piece of content), thus preventing further contagion.

Figure 7.11: Users infected by time in simulations of the SEI model using max lifetime for infections. $(\alpha) = 0.1$. Forward limit $(\varphi) = 5$.



(a) Wpp. India

(b) Wpp. Indonesia

(c) Wpp. Brazil

(d) Telegram Brazil

Source: The Author.

7.5.4 Adding Fact-checking as Recover

The Susceptible-Infected-Recovered (SIR) epidemic model, or SEIR if we consider the stage of Exposed, utilizes Susceptible, Infectious, and Recovered curves to measure the spreading of infections. In the models presented up to this point, we eliminated the recovery from the simulation to assess the reach of information dissemination when it is not stopped. Although it could not be clear what recovery represents in the misinformation scenario, there is a wide known means to combat fake news on the Web: fact-checking. The fact-checking agencies operate to uncover the facts within false stories and elucidate the truth behind them. Some works argue that the use of fact-checking helps in the combat of misinformation propagation. REIS et al., 2020 compared the sharing of misinformation within WhatsApp groups to fact-checking for these false stories and observed that more than 40% of the misinformation shared on groups on WhatsApp were already checked as false by some fact-checking specialized agency. This suggests that one can use the checked content to counter the spreading of misinformation.

In one more experiment, we explored the use of fact-checking as a means to add a recovery curve to our model. The fact-checking works as a "vaccine" that can grant immunity to the users in contact with it, after that users are resistant to the infection and can not be infected again. Therefore, in our model, we consider for "immune" users that they cannot be "exposed" to the infection either. In practice, we create a recovery rate called r that represents the probability of the user to recover from the infection. Then, during each iteration, we test all the users against this rate r to verify if it will be allocated to the Recovered set. It is known that misinformation spreads faster than other kinds of content (VOSOUGHI; ROY; ARAL, 2018). Then, in this experiment, we tested three different recovery rates to measure a propagation. The first, we run the simulation with recovery rate equals to the virality ($r = \alpha = 0.1$) and two with recovery smaller than virality ($r < \alpha$), testing ten times smaller ($r = 0.01$) and a hundred times smaller ($r = 0.001$).

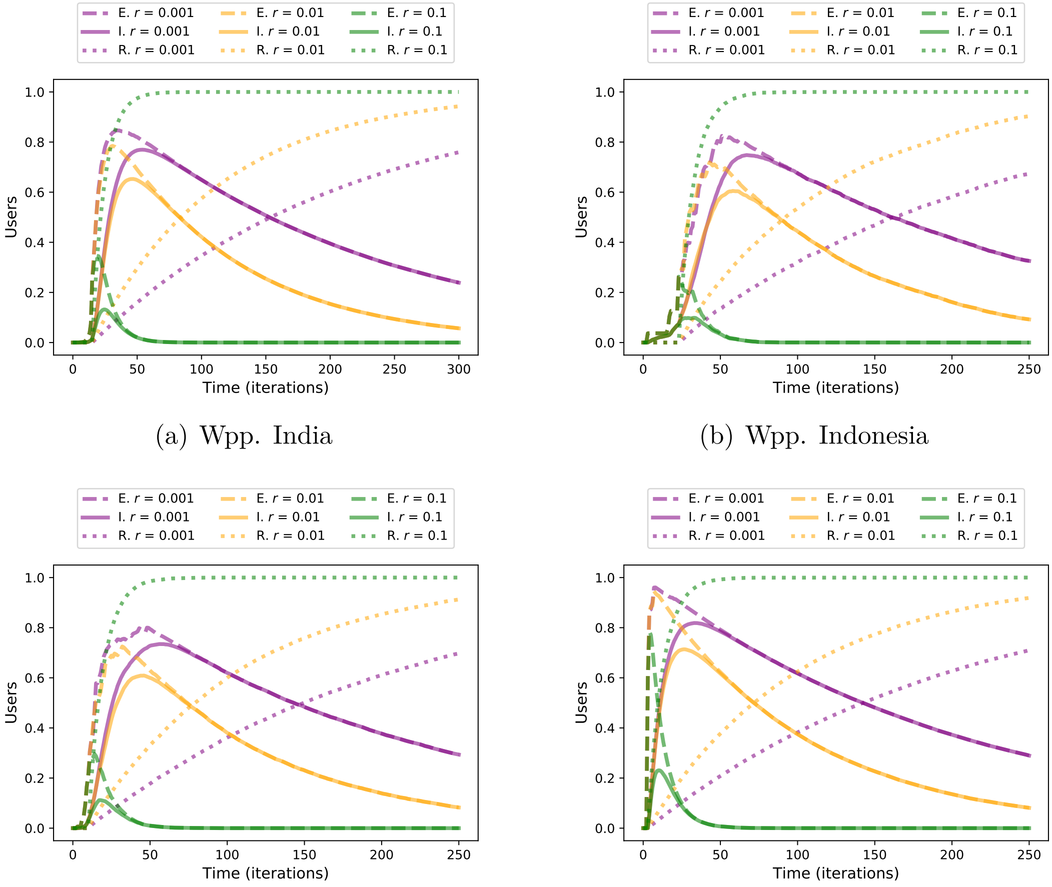
In Figure 7.12, we see results of applying the recovery rate in the SEIR model. It is possible to note similar curves for all IMP in which the Recovered curve grows while it interrupts the spread of the infection. The major difference occurs at the beginning of infection, in which we observe that Telegram quickly exposes more users before the Recovered starts to grow. Even for recovery rate probability a hundred times smaller than the actual virality α, the addition of this element to counter the infection is compelling to prevent the infection from reaching all 100% users. When recovery rate is set to the same probability of the virality rate of infection, it is able to quickly hold the propagation by drastically reducing the number of infected users. Those results illustrate the importance of combating misinformation in its source of dissemination, it suggests that the sharing of contents of fact-checking or any similar countermeasures should be encouraged as it seems to be very effective when presenting high probability (similar to the virality of the malicious content).

7.5.5 Real Time Dynamics on WhatsApp Network

Finally, we look at how actual content disseminates through WhatsApp network and apply these real time dynamics in our infection model to estimate the time of dissemination of misinformation in IMPs. By observing all occurrences of a single piece of information, it is possible to analyze some dissemination characteristics of this kind of content (e.g., we can point where it originated, how long it takes to be reshared, how long it lasts on the network and how many users this content reached).

Multimedia messages usually are shared unaltered across the network and, then, they are easier to track than text messages. Thus, we choose to select the images posted on

Figure 7.12: Users infected by time in simulations of the SEIR model using recovery rate (r) to add immunity to users. (α) = 0.1. Forward limit (φ) = 5.



(a) Wpp. India

(b) Wpp. Indonesia

(c) Wpp. Brazil

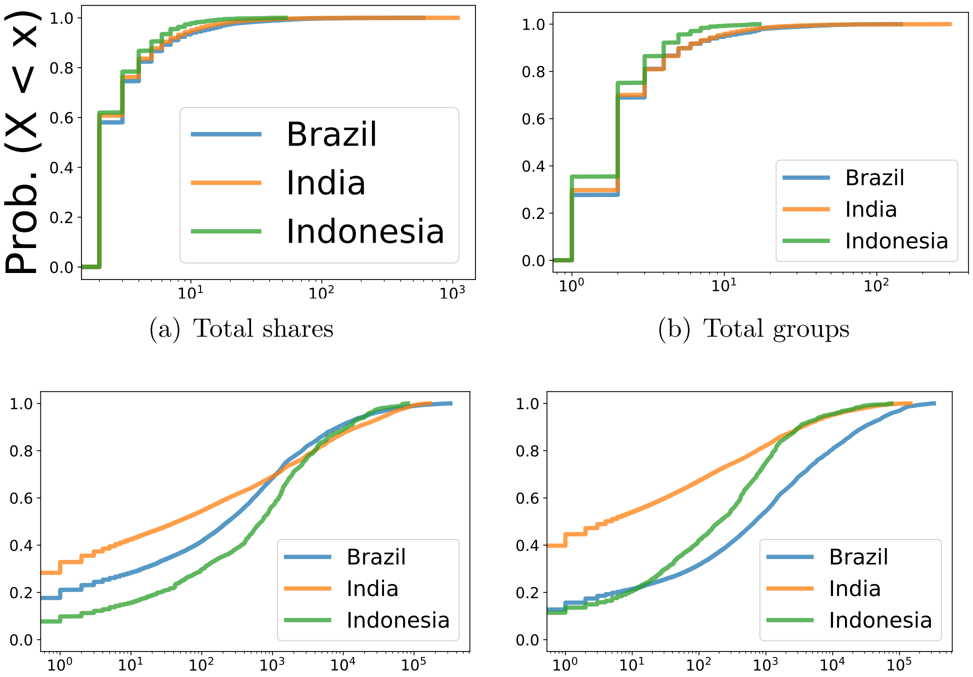
(d) Telegram Brazil

Source: The Author.

WhatsApp to analyze real time characteristics about the propagation of content across the network. Since we group together the sets of similar images and all their sharing information using perceptual hashes, we are able to determine time intervals regarding spreading of those messages. For this analysis, we count image popularity for these datasets used during experiments and simulations with SEI model and calculate its spreading across the network. In total, for WhatsApp datasets from Brazil, India and Indonesia, more than 1 million of images were identified, from which 784k were unique image objects. To evaluate spreading metrics regarding time and coverage, we just consider the images that were shared at least twice, since we cannot see the effect of spreading of images only posted a single time. This set consists of 2,384 images in Indonesia, 103,031 images in Brazil, and 44,731 images for India, which represents approximately 20% of all images.

With this data, we calculate the total number of shares of each image and how many groups they have appeared. Figures 7.13(a) and 7.13(b) show the Cumulative Distribution Function (CDF) of the total number of shares and the number of distinct groups each image appeared in. It is possible to note that there are some very popular images broadly

Figure 7.13: CDF of sharing coverage and time dynamics metrics of images shared at least twice on WhatsApp.



(a) Total shares (b) Total groups

(c) Lifetime (d) Inter-event

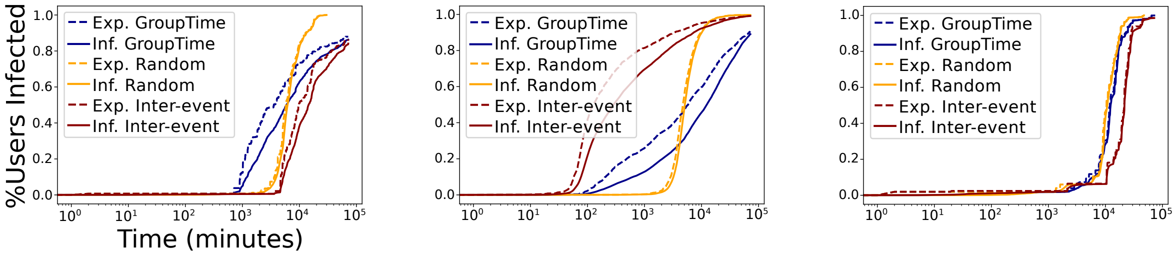
Source: The Author.

shared more than 500 times in Brazil and a thousand times in India, moreover, they reached more than 100 groups in both countries. Even though a large part of images were shared just a few times, the more popular ones demonstrate that WhatsApp can be used not only for particular conversations but also as a mass communication media with a potential virality of its content.

We also analyze their “lifetimes” in Figure 7.13(c). The lifetime is given by the time difference in terms of minutes between the last and first occurrence of a single image in our dataset. While most of the images (80%) last no more than 2 days, there are images for both Brazil and in India that continue to be reshared even after 2 months (10^5 minutes) of their first appearance. Further analysis, in Figure 7.13(d) shows the distribution of the “inter-event time” between posts of the same image. This represents how many minutes it takes for the image to be re-posted after an appearance in the network. We observe that the inter-event time of images in India is much faster than in Brazil and Indonesia, i.e., more than 50% of posts are done in intervals of 10 minutes or less, while just 20% of shares were done in this same time interval in Brazil and Indonesia. A manual investigation for reasons behind the short period of time between posts suggested that in the data from India, there is more automated, spam-like behavior compared to those in Brazil and Indonesia.

In conclusion, these results suggest that WhatsApp is a very dynamic network

Figure 7.14: Real Time SEI model using "incubation time" before spread infection and each iteration equals 1 minute (log). $(\alpha) = (\beta) = 0.1$. Forward limit $(\varphi) = 5$.



(a) Wpp. Brazil (b) Wpp. India (c) Wpp. Indonesia

Source: The Author.

and most of its image content is ephemeral, i.e., the images usually appear and vanish quickly. The linear structure of chats makes it difficult for an old content to be revisited, yet there are some that linger on the network longer, continuing to spread over weeks or even months.

Finally, we used these dynamics of propagation calculated from our data of WhatsApp to realize one more simulation to measure viralization velocity in terms of real time. In the SEI model, the spread of information was measured in terms of the *number of iterations*. In this experiment, we adapt the SEI model to use these time dynamics and to present a measure of the spread in terms of minutes. For this, we add an "*incubation time*" based on the time real data takes to spread over the network. In this version of the model, each iteration represents 1 minute, but when an infected node intends to spread, it has to wait a specific amount of time before doing it. This time is sampled from a distribution of "waiting times", which can be:
(i) **Random**: a uniform distribution with domain between 1 and 1440 minutes (1 day); (ii) **Inter-event Time**: the empirical distribution of inter-event times computed in Figure 7.13(d); (iii) **Group Time**: this strategy is based on the following idea – it usually takes longer for a message to reach 100 groups than to reach 2 groups. To implement this, in this strategy, we make the incubation time in the initial steps smaller than in the subsequent steps, also based on the time information from dataset gathered from WhatsApp.

During the simulation, we track the number of groups the infection has already spread and, for each step, we have a different time distribution according to how long it took for the actual images in WhatsApp to reach those number of groups in our data.

Figure 7.14 shows experiments considering the three strategies to compute the time to spread. In India, where we have the bursty inter-event times, we see that with the *inter-event time* strategy 60% of users are exposed to the content in the first 200 minutes of infection. In Brazil, *group time* is faster than *inter-event time* and infected around half of users in the first 2 days (3000 minutes). Finally, in Indonesia all three strategies have very similar behavior, taking over 2 weeks to infect more than 80% of the users. Nevertheless, a content is still viral when all three strategies are considered, i.e., a misinformation can spread in most of the network before one month of infection.

7.6 Findings on Dissemination on WhatsApp

With the infodemic of misinformation that the world has faced, it is imperative to find countermeasures to combat misinformation viralization. The enclosed nature of instant message platforms such WhatsApp and Telegram and the ease of transferring multimedia and sharing information to large-scale groups makes IMPs an extremely hard environment for the deployment of countermeasures to combat misinformation. They are open to an ambiguous use of information, allowing at the same time the viral spread of a content through encrypted personal chats. Together, those two features can be widely abused by misinformation campaigns. This duality between the public and private use of the platforms (WhatsApp and Telegram) confuses the users. They can at the same time create a secure and encrypted chat with a friend and share a message that already went viral by a great part of the users. With our results, we help to remedy the problem of the lack of knowledge about the social network structure necessary to determine whether an information could be viral.

Both WhatsApp and Telegram have features that help a viral sharing of content for multiple users within the platform. While WhatsApp allows the sharing by forwarding messages up to 5 members and does not show the creator of messages, Telegram is limited to one forward per time and shows the source of the message. On the other hand, the groups on WhatsApp are limited to 256, while on Telegram, public channels can reach up to 200,000 participants (or even unlimited for broadcast channels). Furthermore, both have encrypted features that provide more security to user data, but also cover the real identity of authors of messages. Our analysis tries to understand how some of these features help in the virality of the content. While the creation of groups allows hundreds of users to connect to each other, they also can be easily shared on other social networks by an invite link. This establishes a fertile scenario for the emergence of a well-connected network, with a singular topology of groups that facilitate the spreading of information. The forwarding functionality provided by the app plays an important role in the dissemination. It creates a powerful shortcut that helps a message to quickly be broadcast to multiple users at the same time, while encryption hides real authors and makes it hard to track the path a message takes on the network before it is received by the final user. Thus, one cannot know if a piece of information has come directly from a short path or if it has come a long way through this well-connected network.

By collecting data from real scenarios for both platforms, we evaluated the limits of the propagation of information on them. We specifically evaluate the potential that the forwarding functionality provides in each of these apps. Our results show that content can spread quite fast through public groups in both these platforms, potentially reaching private groups and individual users later. Our empirical observations about the network of WhatsApp and Telegram public groups in three different countries provide a means of inferring the information velocity of spreading in real-world scenarios. Using a SEI model, we investigate a set of questions about the limits of virality imposed by IMP's features regarding information propagation. While the limit on the number of users per group can prohibit the creation of giant communities to spread information through the network, this limit, however, is not able to prevent a content to reach a large portion of the entire platform. More important, our analysis shows that low limits imposed on message forwarding and broadcasting indeed offer a delay in the message propagation of up to two orders of magnitude in comparison with high limits, although this measure alone is not sufficient to block entirely the infection. Also, while the forward limits has a minor impact for the majority of members (as they join just a single group), it is an effective way to slow down the dissemination for users that are members of multiple groups at once.

Additionally, by tracking the dissemination of images on WhatsApp, we verified that most of the images (80%) last no more than 2 days in WhatsApp. With the results of velocity of spread of our SEI model, in India, even content with this short lifetime can be already enough to infect half of users in public groups. Although there are still 20% of messages with a time span sufficient to be viral in the three countries using any of our strategies to estimate time of infection.

Note, however, that depending on the virality of the content, those limits are not effective in preventing a message to reach the entire network quickly. Misinformation campaigns headed by professional teams with an interest in affecting a political scenario might attempt to create very alarming fake content that has a high potential to get viral on IMPs. In another experiment, we applied a recovery curve to the model to simulate the effect of fact-checking in the diffusion. Even for a recovery rate with low probability, the addition of a possible scenario of recovery prevents the dissemination to reach the whole network. This illustrates the importance of the task of fact-checking and other countermeasures to block the spreading of misinformation. Thus, with better comprehension of how the system functionalities impact on dissemination of information in such platforms it is possible to propose counter-measurement that could be implemented, preventing coordinated campaigns to flood the system with misinformation. Platforms could seek means to aid the public, such as providing transparency about the shared content and its author or stimulating the users' autonomy on seeking more reliable information outside the network.

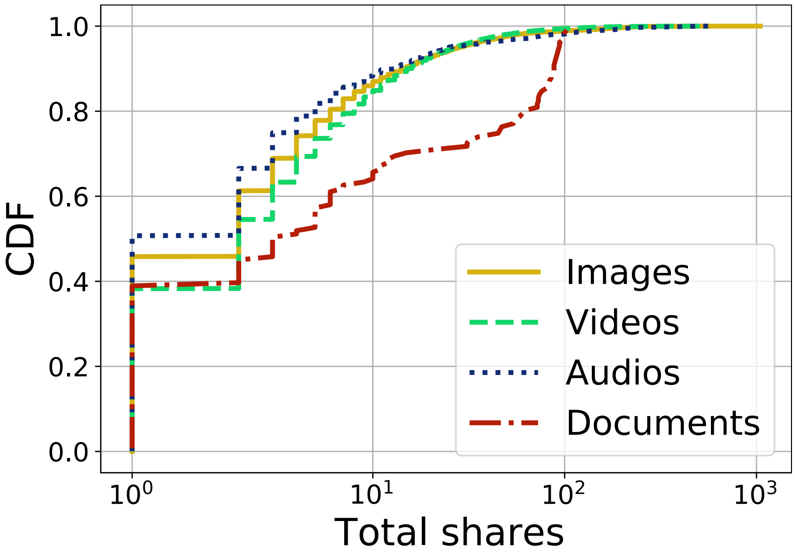
7.7 Analyzing Forwarded Messages on WhatsApp

As we observed, the forwarding tool of WhatsApp play an important role in the scenario of dissemination of (mis)information in this ecosystem. Many users, in fact, forward helpful information, as well as funny stuff, and content they find meaningful, however, often these messages also carry with them unwanted content of misinformation, conspiracy theories, or even hate speech. This is potentially problematic within WhatsApp network since users may want their messages to go viral because that, the very act of forwarding a viral message, may be interpreted as an implicit endorsement of the content (HARVEY; STEWART; EWING, 2011) and the credibility of the message is, then, enhanced by a close contact (BAKARE; ABDURRAHAMAN; OWUSU, 2022).

In this section, we explore our data collection in terms of the forwarding messages sent on WhatsApp. By investigating how users make usage of forward tool of WhatsApp and keeping track of sharing chains of messages on WhatsApp by our duplicates detection approach, we can understand the impact of this kind of service provided by the application and, moreover, we can be able to analyze how the measures taken by WhatsApp of tagging and limiting forwarding messages affect this kind of content that circulates within its network.

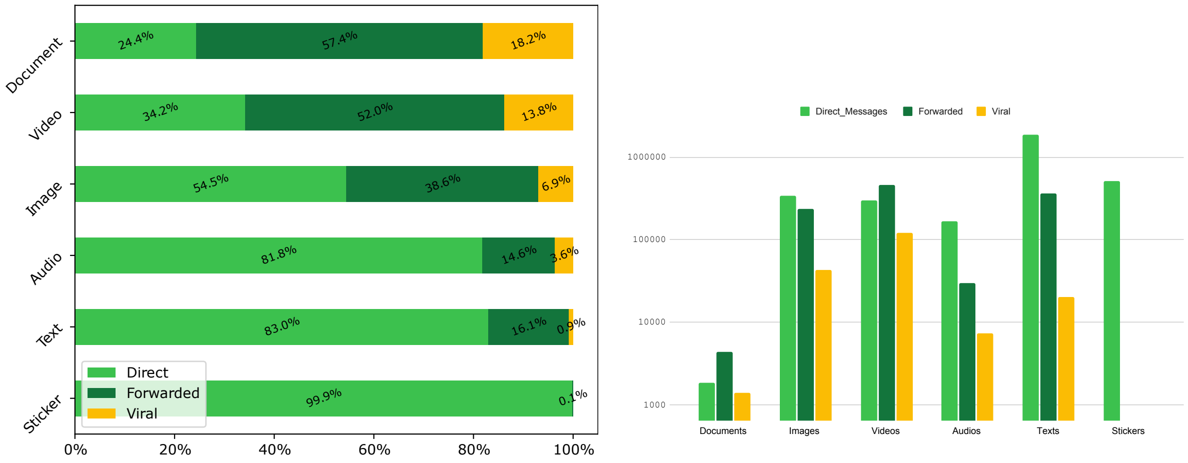
As we merged together messages with same content, we first analyze the volume of shares of each distinct content (Figure 7.15). For all kinds of multimedia formats analyzed, more than half were shared more than once in our WhatsApp dataset. This evidences that messages are often disseminated within the platform and there is a need to investigate then sharing patterns of how this content is spread. Although documents

Figure 7.15: Distribution of total number of shares on our dataset per media type.



Source: The Author.

Figure 7.16: Percentage and number of forwarded and viral media message sent per type.



(a) Percentage (b) Absolute (log)

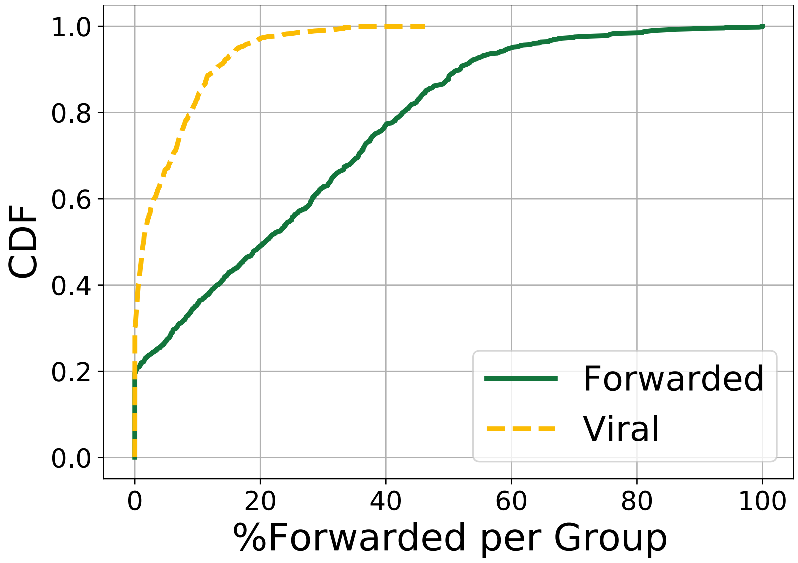
Source: The Author.

were generally shared more than other media, it was the type that had the fewest total shares (about 100 shares) and least common format in our dataset. We have cases in our dataset of images shared more than a thousand times within the groups monitored. Then, we observe that multimedia messages are not only very frequent in our data as also they are recurring on WhatsApp, which means that a large portion of the content in WhatsApp is not unique but a copy of another one already existing.

Next, we analyze messages tagged as forwarded by WhatsApp. In Figure 7.16, we calculate, for each type of media, the portion of messages that were sent directly by other users, the portion that received just the flag "*Forwarded*" and those viral messages that were labeled as "*Forwarded Many Times*". While text messages have less than 25% of the messages forwarded and stickers almost none are forwarded messages, the remaining media messages have a large number of content being forwarded, with 75% of all documents received are forwarded, 65% of videos and almost half of the images. This also shows that a considerable amount of content is forwarded from another source, and they are not original messages on those chat groups. Furthermore, when looking at the "Forwarded Many Times" flagged messages by WhatsApp itself, we see that a large portion of these media also receive this viral status, suggesting that many messages traveled a long path through the WhatsApp network before reaching the end user, distancing themselves from the original authors who created them.

We further investigate the importance of forwarding on group-based communication. In Figure 7.17 we show the distribution of the portion of messages that are labeled as forwarded and viral within each group. About 10% of the groups have more than half of all of their messages being forwarded from somewhere else. We observe that sharing messages is an essential component of WhatsApp communication. A large portion of the content of these monitored groups was not generated in the same place they were

Figure 7.17: CDF of the portion of forwarded content in each group.



Source: The Author.

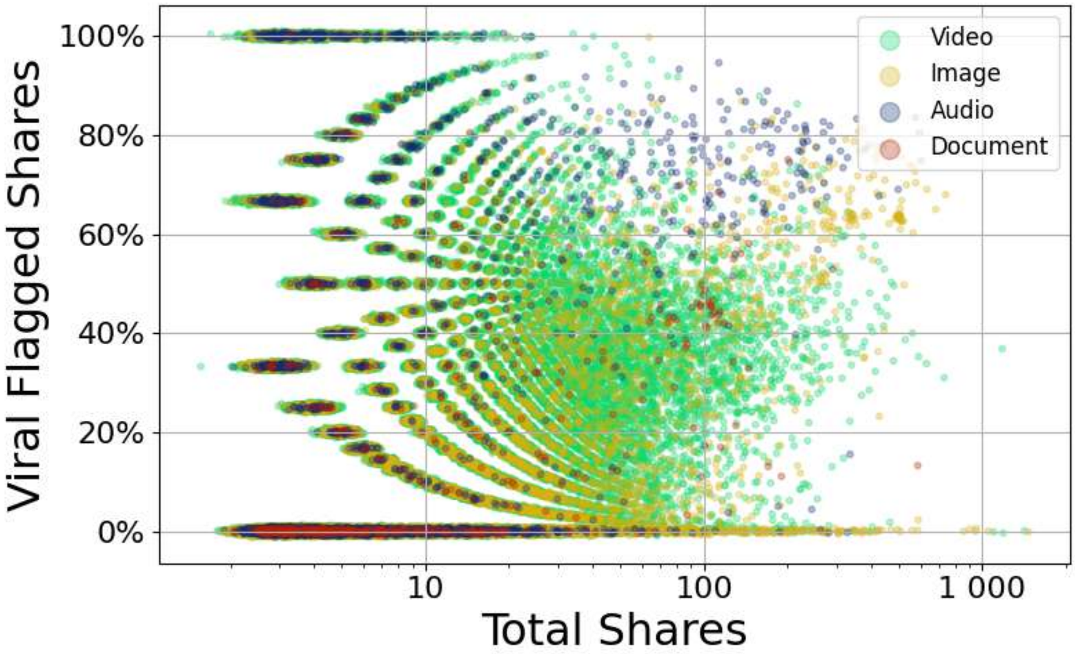
posted, which shows that users frequently employ the forwarding functionality to share messages between chats, enabling the virality in this ecosystem. There are even some groups where most of the content (more than 80%) was not actually created within that group, suggesting that communication in these types of group is based only on sharing things they see elsewhere with the other members, working as a repository for dumping external messages.

These results of highly shared and broadly forwarded messages on WhatsApp reinforce the public and massive nature of this platform can assume, opposing of the exclusive and private idea that instant messaging services usually present. There is a considerable volume of repeated content being sent numerous times and circulating widely through WhatsApp reaching beyond individual groups boundaries and spreading to hundreds of other users across the network.

Next, we will analyze the ways in which users break the forwarding chain within WhatsApp and the impact of this on the detection of information spreading counted by the application.

An important question regarding messages tagged as forwarded or viral and the number of shares, is that not all occurrences of media receive the label “Forwarding Many Times”. in order to analyze WhatsApp’s ability to flag the spreading of a set of instances of a given media, we evaluate the portion of each media that received a viral tag by WhatsApp given the total of shares of that message within our data, as shown in Figure 7.18. Here, we have the relationship between the percentage of content labeled as viral and the total of shares for all multimedia messages collected. We see that many shared multimedia messages from our dataset, in fact, received a tag of Forwarded Many Times among their occurrences in our data, regardless of their number of shares, but this

Figure 7.18: Portion of viral flags and total number of shares per media.



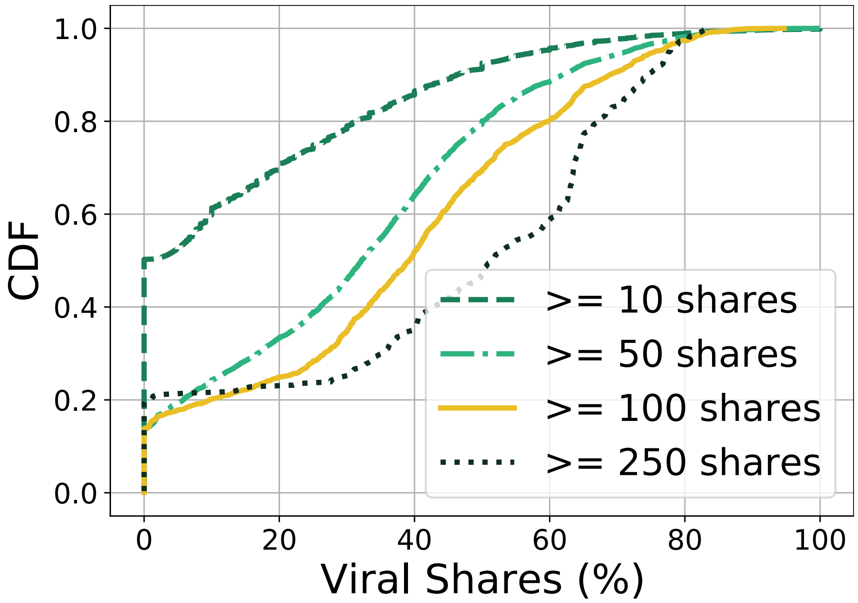
Source: The Author.

labeling hardly reached all instances of that message. We can observe in this scatter plot that after some point (media with more than 60 shares), there are no more points with 100% of the data flagged as viral. On the other hand, we also have fewer points with less than 50% of instances flagged by WhatsApp.

To better visualize WhatsApp's ability to flag viral data, we also observed the distribution of the portion of viral data considering the total number of shares of those messages in Figure 7.19. It is possible to note that the more shares the content has, the higher is the percentage of viral instances labeled by WhatsApp. Considering all media with more than 10 shares about half of them do not have any of their occurrences labeled as viral, on the other hand, for the set of media shared more than 250 times is the one with the largest part of their occurrences having the viral tag. Although we have some widely spread media through WhatsApp groups with many shares, however, not all of those shares received the label "Forwarded Many Times". For all media with more than a hundred shares in our dataset, none of them has 100% of their occurrences flagged. Actually, considering these widely shared media (i.e. those shared more than a hundred times), for 14% of them, none of their occurrences (0%) were labeled as "Forwarded Many Times", and only 31% of them have more than half of their shares properly flagged as viral, which represents that in most of the messages, users cannot distinguish whether the content came from more personal contact or it is a viral message on WhatsApp as the presence tag would suggest.

Furthermore, when looking at the distribution of this percentage of viral flagged content for each type of media separately (considering only the set of media with a hundred

Figure 7.19: CDF of total of viral shares ("Forwarded Many Times") per media for different number of shares



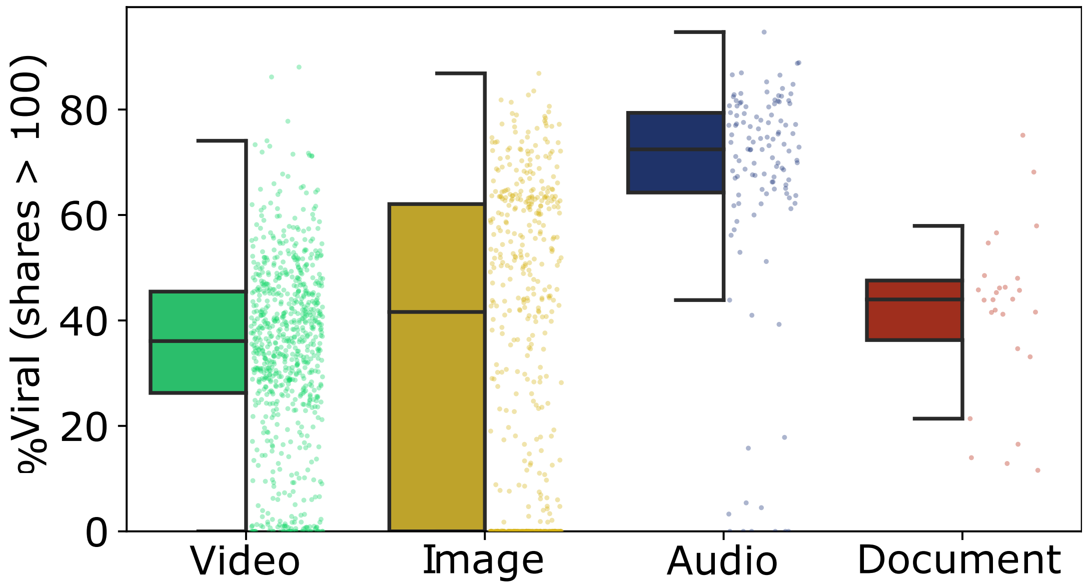
Source: The Author.

or more shares) in the boxplot of Figure 7.20, we observe some interesting differences. Most audio data have more than 50% of their occurrences tagged as viral, while for video and images, there are much fewer instances tagged as "Forwarded Many Times". In particular, for images, we have a considerable volume of media that none of the shares were labeled. This suggests that images are harder to track the viral dissemination while it is easier to keep the forwarding chain for other media formats, specially audios, within WhatsApp.

These experiments help us to note that even though WhatsApp is able to detect that those multimedia are getting a viral status at some time, a large part of their actual shares are unnoticed by WhatsApp and lack of proper labeling. Thus, they circulate throughout the platform indistinctly from any other "normal" message.

Besides that, there is another challenge in tracking one single piece of information circulating on WhatsApp. Since each image is isolated and the content is encrypted within the system architecture, a same message might be shared from different sources. Therefore, WhatsApp will not notice those are the same. This could create two different chains of clones for the same content. When someone forwards a media message on WhatsApp, the app can identify that this is the same message by keeping some metadata of the message, including the *media_url* which represents where the encrypted file is really stored on WhatsApp server. This *media_url* is what allow WhatsApp to track the chain of dissemination of a single piece of information and count the forward steps of that message. However, as explained in Section 7.2, users can simply download the images, audios, videos and documents received on WhatsApp in their gallery or even get that exact media from somewhere else (e.g. from the Web or other application) and sent it

Figure 7.20: Boxplot of the portion of viral shares (“Forwarded Many Times”) per media type for those shared more than a hundred times.



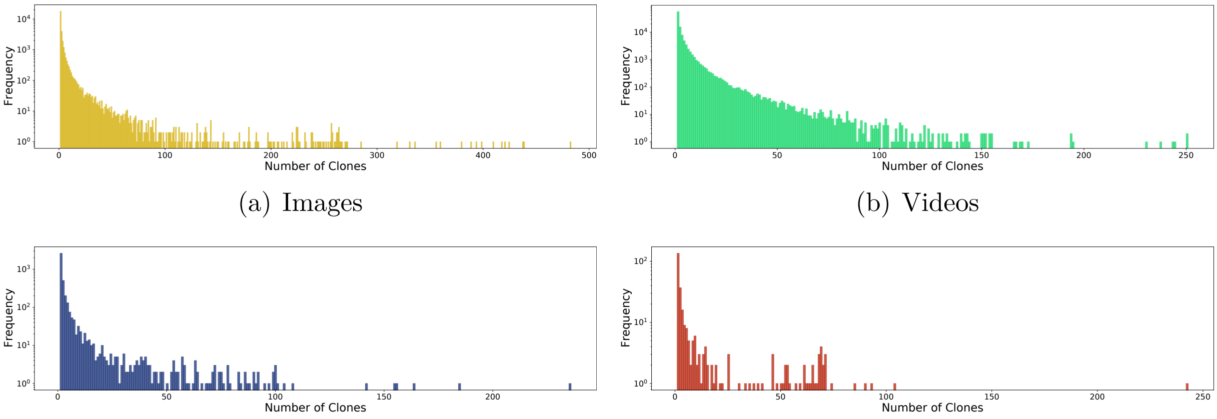
Source: The Author.

directly to other users instead of using Forwarding functionality of WhatsApp. For those cases, in which the same content is sent by a different source other than forwarding, a new *media_url* is generated, creating a totally separated chain with a clone of that content. In this scenario, there is no way to WhatsApp see that both messages are the same due to the encryption architecture changing all metadata of the message and also creating another media storage. Nonetheless, as we observe the data from user side in this work, we are able to merge these same multimedia messages in our dataset using hash similarity. Therefore, we are able to track multiple shares of the same content even when they are created by different users and sources, which we call clones. Clones are exact identical content that were sent in the chat groups, but WhatsApp is not able to relate them as the same (i.e. they have different sources and, thus, distinct *media_url*).

The creation of those clones, difficult WhatsApp to track the viral messages and label them as “Forwarded Many Times”. In Figure 7.21 we analyze how many clones exist for each distinct viral medium. Note that clones are different from number of shares as a clone is unrelated from other appearances of that media and each clone may have they own shares. We can observe that yet many media present only one single object on our data, the high presence of clones for viral content indicates the existence of multiple different sources for that media. This suggests that users often reproduce viral data directly instead of just using Forward mechanics of WhatsApp app.

Since WhatsApp restricts the number of forwards a user can make according to the size of the forwarding chain a message, by copying it and directly sending the content as they were an author can be a tactic used by users to circumvent the limits imposed by

Figure 7.21: Number of clones, duplicated media within messages on WhatsApp per type.



(a) Images

(b) Videos

(c) Audios

(d) Documents

Source: The Author.

WhatsApp and share their viral content.

It suggests that WhatsApp's actions to avoid content going viral within its network cannot prevent users from widely sharing multimedia content. The ease with which a user can bypass the forwarding counting system by creating a large number of clones of the same piece of media can help us to understand how this platform has been abused, supporting harmful content such as misinformation, hate speech, or conspiracy theories to spread virally without any warning to WhatsApp users.

Finally, we investigate the potential of our merging methodology used to group duplicate messages in order to detect viral content in our WhatsApp dataset. For that, we firstly take in account each unique multimedia message that received at least one "Forwarded Many Times" tag among all of their occurrences, which represent 136,924 distinct media. Using the aggregation approach, we calculate that those media were shared a total of 1,037,113 times within our data. Next, for those messages, we count that 612,855 (59%) did not receive a viral tag by WhatsApp, but based on our analysis, they could have been flagged as such, since they actually represent a copy of another message that was tagged as viral by WhatsApp.

7.7.1 Summary

When a messaging service provides forwarding tools to users, it enables the transformation of the platform into a network for mass media communication, as it grants to the messages the ability of being shared between peers and reaching a thousand of people very quickly. Due to that, however, on WhatsApp, users may have their messages and content become viral without their consent or without even notice that it happens. At the same time, users can also receive in their account messages that they have no idea where they come from, whether it is just something a close contact or friend created or a viral content that was disseminated through the network.

To tackle this issue, WhatsApp implemented labels in its messages, alerting users when they are “Forwarded” and “Forwarded Many Times”. Those tags seek to help users identify that a message is not originally sent by the contact that sent it to them or even suggesting that it may be created far away from that chat, going viral through the app. Nevertheless, this label may not be enough to track the spreading of a viral message and, in cases when it is not properly attached to received messages, this can be misleading for users that expect that widely spread messages will be tagged as such.

In this study, we analyze a set of 10 million messages exchanged in a context of public groups of political discussion in Brazil during five months of presidential elections in 2022 and also investigate how WhatsApp users make usage of forwarding tools to propagate their content within the platform. Using this dataset, we explore the impact of forwarded messages in WhatsApp environment, showing the large volume of multimedia message that circulate within WhatsApp network due to this functionality. Furthermore, we find indications that WhatsApp counter method fails to identify the dissemination of viral pieces of media, as we find numerous examples of occurrences of a single multimedia content without the viral label and that there are multiple clones of messages that WhatsApp is totally unable to track due to its encryption architecture. As a consequence of that, there are a considerable number of messages that freely circulate throughout this network, totally hidden from WhatsApp tools for detecting and moderating this viral content.

As a result, we show that the issue of combating misinformation and other forms of abusive content within the WhatsApp ecosystem is not an easy task, as even WhatsApp tools developed to prevent them can be frustrated by simple actions of users. If someone breaks the forwarding chain and resets it by sending a message directly himself, WhatsApp loses the counting and stops to label that as a “Forwarded Many Times”. The architecture of the system allows users to quickly share a message and reach hundreds of other users with a click of a button, but is often unable to track its origins or prevent that harmful content goes viral.

By shedding light on how messages disseminate on WhatsApp, we show some of the challenges in counter content viralization within instant messaging platforms. However, by using perceptual hashes to detect copies of messages being shared through public groups, we show that there are means to properly find the spreading of forwarded chains between messaging chats and, consequently, it is possible to moderate the content in this closed environment, finding and countering that potential harmful messages before they spread further among the WhatsApp ecosystem. With our merging approach, we discover that

69% of viral multimedia messages circulating within the monitored public groups lack of label, but could be flagged as such to prevent even more dissemination.

This is an important discussion about WhatsApp and also other instant messaging platforms, as we want to keep them personal and private to grant security to their users to communicate and even inform themselves in a healthy environment while, on the other hand, we see they are being abused by misinformation campaigns around the world and needs to be further addressed to prevent malicious actors from taking advantage of their issues to spread harmful content to thousands of users who use them daily.

Chapter 8

Combating Misinformation on WhatsApp

So far, at this point of the study, we have investigated WhatsApp social media in various aspects. We identified how public groups are shared on the Web, how WhatsApp is extensively used to consume news and spread political content, and the dangerous potential for viralization of information within this platform, especially at a time after the COVID-19 pandemic spreading along with a digital "infodemc" with a huge amount of fake news being spread by this messaging app. We already know that WhatsApp is an essential tool to the dissemination of misinformation in the world, and that it can directly impact in the society. We also know that the specific design of WhatsApp (and other IMPs) hampers the combat of misinformation on the platform, as it is mainly private and encrypted.

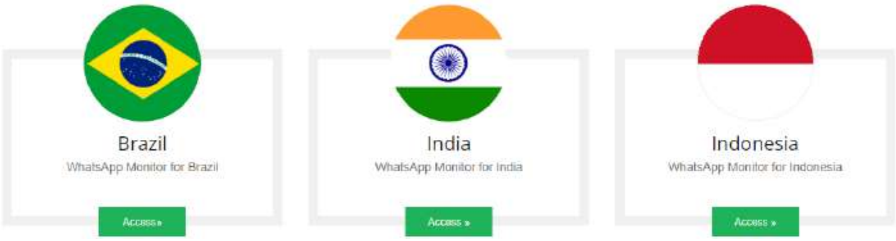
After point the problems and challenges about WhatsApp and how this platform is abused to disseminate harmful content, next we also propose some measures that can be followed in order to effectively combat the misinformation within this social media. In this chapter, we will discuss actions that are being taken to prevent the spread of misinformation on WhatsApp by this research, as well other approaches that could be adopted to diminish the advance of fake news in WhatsApp.

8.1 *WhatsApp Monitor*

As elaborated throughout this thesis, one of the key challenges when working with instant messaging platforms is that it is hard for researchers to have access and investigate the actual content people share through WhatsApp at scale. Although we observed there is a well connected network with viral messages exchanged between multiple groups and users, the data are isolated within different group chats and we lack of an informational apparatus to aggregate and provide us a big picture of this environment.

We addressed this issue as our first approach in tackling misinformation on What-

Figure 8.1: Land page of *WhatsApp Monitor* with the countries implemented.


Brazil
WhatsApp Monitor for Brazil
Access »
India
WhatsApp Monitor for India
Access »
Indonesia
WhatsApp Monitor for Indonesia
Access »

Source: The Author.

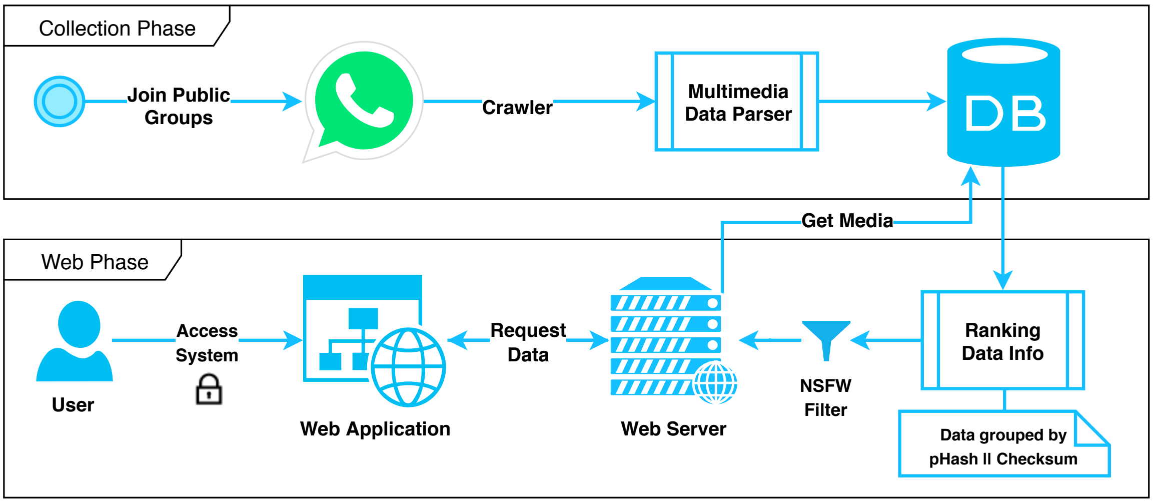
sApp. We propose here the *WhatsApp Monitor* (<http://www.whatsapp-monitor.dcc.ufmg.br/>), a web-based system that helps researchers and journalists to explore the nature of content shared on WhatsApp public groups from three different contexts: Brazil, India, and Indonesia (Figure 8.1).

Our tool contain various categories of multimedia messages such as image, video, audio and text that have been posted on a set of more than a thousand WhatsApp political public groups and displays the most shared content per day in an online interface. The system was already employed to monitor content during the Brazilian general elections in 2018 and 2022, it was also used in context of 2019 elections in India and Indonesia, and for exploring news and other (mis)information about the COVID-19 pandemic, specially between 2020 and 2021. In Brazil, the system is kept up to date since 2018 and has captured key events on the platform, such as the truckers' strike in 2018, the presidential elections of both 2018 and 2022, and the COVID-19 pandemic, in which our tool was an important source of fact-checking efforts on WhatsApp. Currently, this system is one of the major efforts to estimate the spread of disinformation and assist fact-checking efforts within WhatsApp scenario.

The *WhatsApp Monitor* is responsible for grouping, collecting, processing, ranking, and displaying all content from WhatsApp and condenses it in an online interface, accessible on an Internet browser through login and password, in a way that the end user can navigate through the collected data and do their own analysis with the help of the system. The purpose of *WhatsApp Monitor* is to inform and anticipate communicators about the type of information shared on public groups. The goal is to provide, for a restrict set of researches, journalists or fact-checking agencies, access to this system to perform fact-checking on the information shared on those groups, revealing highly popular content spreading on WhatsApp, but which are hidden from the experts' lens due to the enclosed structure of the platform.

Next, we will go into the details of the functioning behind the system, then we'll cover its user interface. Afterwards, we will discuss the impacts of using the tool created and an overview of the collected data through the years of study. Finally, we will make our final remarks about the work.

Figure 8.2: The system framework behind the structure of *WhatsApp Monitor*.



Source: The Author.

8.1.1 System Architecture

Since all collections' methodology was already explained in earlier chapter of this work, we will focus here more in another points related to the developing of the system of *WhatsApp Monitor*.

The web system uses data collected from selected public WhatsApp groups that discuss political topics. These public groups are operated either by individuals affiliated with political parties, local community leaders, or common users with an interest in this topic and can be freely accessed by anyone with an invite link: a WhatsApp URL in the pattern `chat.whatsapp.com/<groupID>` shared on social media for anyone wishing to join the group.

The first step of the work is, therefore, to select the relevant groups that will be monitored by the system. While an extensive list of groups and data from India and Indonesia were acquired in partnership with MIT researchers that also study WhatsApp groups (GARIMELLA; TYSON, 2018), for Brazil, the process to find, join, and collect data from the groups is the same as fully described along the methodology of Chapter 5. As a result, we set up a system that monitors 1101 selected WhatsApp political public groups in Brazil[1].

Then, with all the configurations prepared and after properly joining those groups, we started extracting data, collecting and processing all chat messages on the server. The main architecture of the collection is briefly explained through the flowchart in Figure 8.2. We daily deploy our collection methodology of downloading data shared within each group and persisting them in a database ranked by the total number of shares for each item.

[1] Accordingly the status of the system at October 2022.

In summary, the messages are extracted and saved in a structured way, while we also download their respective files for the multimedia messages (image, video and audio). The content is processed and then merged based on their similarity. After that, it is stored in the database of *WhatsApp Monitor* every day. The online system, then, accesses this database and displays it in the web interface according to the user's navigation and requests.

To identify duplicated images and count their popularity, we use the hashing algorithms to calculate a fingerprint for every image, and also find duplicates of files, as better described in Chapter 5. This allows us to group a set of the same content and process aggregated metada about the messages shared on WhatsApp. We also analyze similarity for identical URLs and use the Jaccard index to compare text messages.

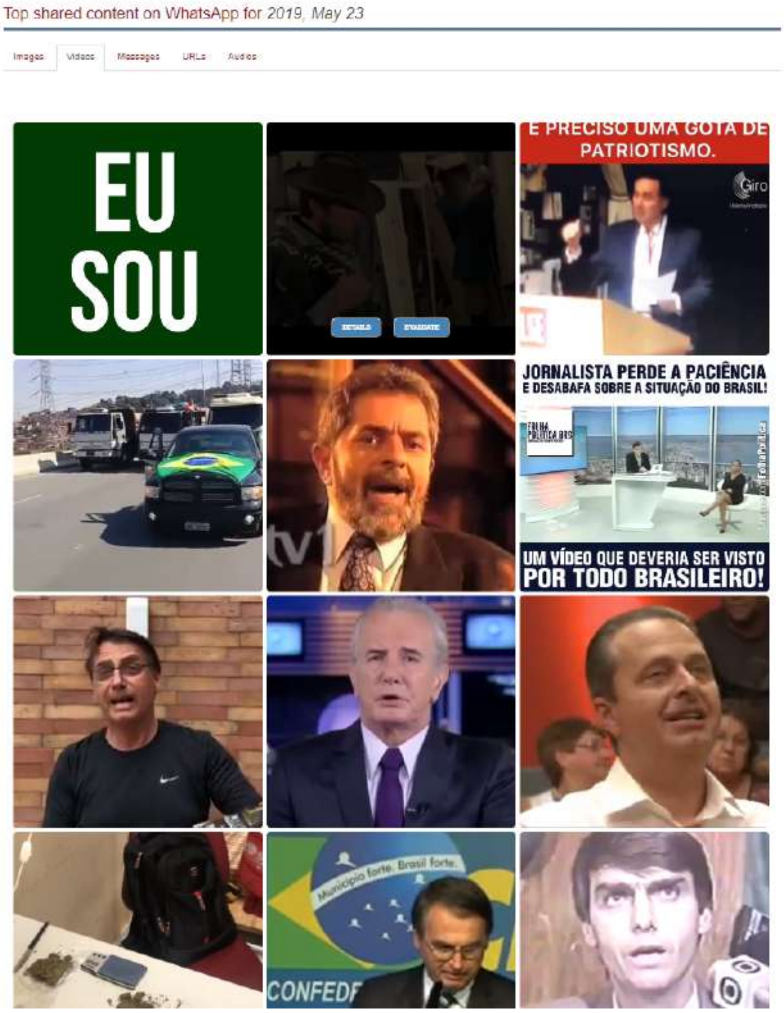
To insert the content to the *WhatsApp Monitor* database, all this processing methodology are repeated every day, that is, for each day all messages from that date is processed, and we extract the additional attributes of how many times a single piece of information was posted on WhatsApp, how many different users posted it and how many groups it appeared. For each multimedia message format, this set of messages is finally ranked by the number of shares, so users can browse the most popular content shared every day using the interface of the WhatsApp Monitor.

All of this data is stored in a system specific relational database manipulated through SQL queries. The web application is hosted on a server at the university, which uses this database and displays all the content in an online interface that users can access wherever they are via a Web browser with an Internet connection. In this system, the user can navigate through the days, weeks, or months and see the whole content posted on the WhatsApp groups we monitor during that period of time.

The *WhatsApp Monitor* gathers a considerable amount of data from many WhatsApp groups. To ensure the privacy of users, we do not share or disclose any Personally Identifiable Information (PII) such as phone numbers. Another sensitive material that can be present in our dataset are images that depict explicit violence or adult content. As it could be harmful for the users navigating in our system, we use a filter, the Yahoo Open NSFW Model (MAHADEOKAR; PESAVENTO, 2016), considering as improper all images with a score higher than 0.8, as suggested by the authors. Another security measure we take to avoid any misuse of even the aggregated information within our system is limiting the access of WhatsApp Monitor to a restricted number of users (i.e. journalists and researchers) with a login account and password. Those accounts are created manually exclusively by us after an analysis of the user request. Moreover, when granting access to the system, they are also informed about the data limitations and the potential bias present in the system. Still about further privacy discussions regarding our application, our data collection does not violate WhatsApp terms of service, since we solely use publicly available WhatsApp groups, joining them with authentic profiles using authentic phone number and specifying in our account that it is actually a research profile.

8.1.2 User Web Interface

Figure 8.3: A screenshot of main interface of the WhatsApp Monitor web system.



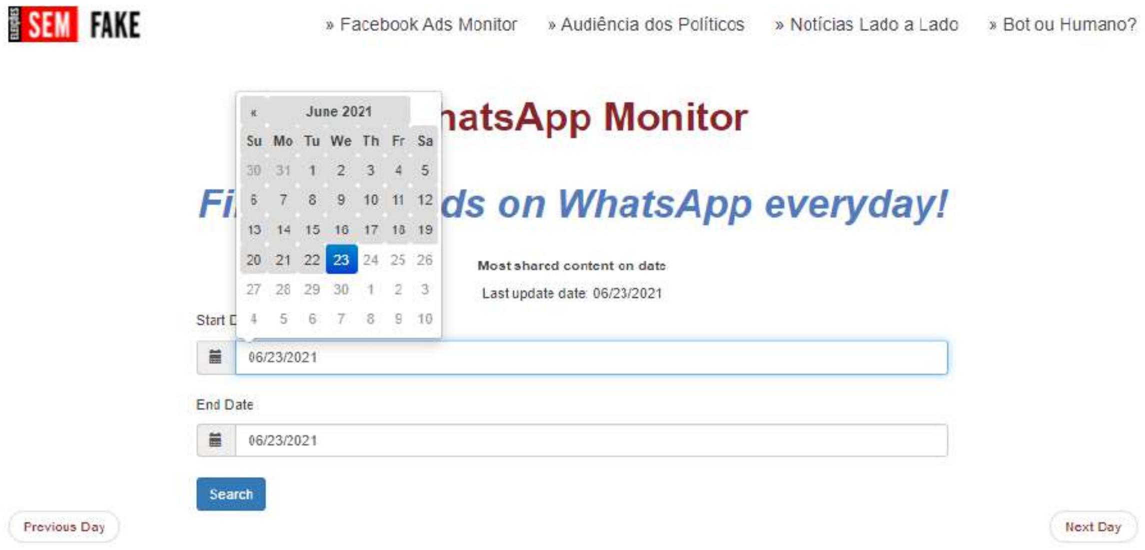
Top shared content on WhatsApp for *2019, May 23*

Images
Videos
Messages
URLs
Audios
EU SOU
DETALHES
EVALUATE
É PRECISO UMA GOTA DE PATRIOTISMO.
Giro
JORNALISTA PERDE A PACIÊNCIA
E DESABAFA SOBRE A SITUAÇÃO DO BRASIL!
UM VÍDEO QUE DEVERIA SER VISTO
POR TODO BRASILEIRO!
tv1
Município forte. Brasil forte.
CONFEDE

Source: The Author.

We provide an online system in which users can monitor daily trends shared in WhatsApp public groups for a particular context (countries such as Brazil, Indonesia, and India, or domains such as politics and news). Our system displays information about the text media (only those with more than 140 characters), audio, images, and video of groups related to political and news topics monitored daily. We then ranked each media type among the most popular ones. Figure 8.3 shows a screenshot of how the content is displayed in the *WhatsApp Monitor* interface as the user logs into the system. It shows, in order, the contents with the most shares divided by category of media. It is worth

Figure 8.4: *WhatsApp Monitor* dashboard where users can navigate between dates or chose a period to explore the content from monitored public groups.



Most shared content on date

Last update date: 06/23/2021

Start D 06/23/2021

End Date 06/23/2021

Source: The Author.

mentioning that the user can also choose between Portuguese or English as the interface language in the upper-left corner of the screen.

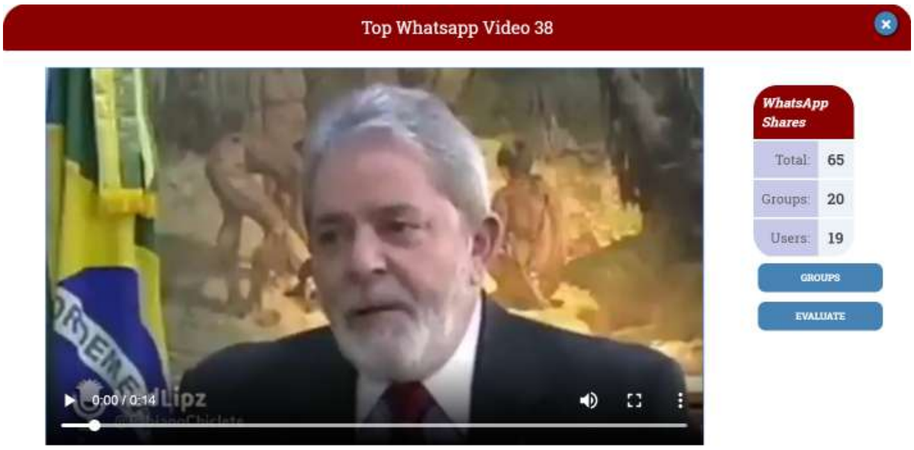
Our system replicates three distinct instances: a Brazilian, an Indian, and an Indonesian version[2]. Once an instance is chosen and the user is logged in, they are taken to a dashboard where they can navigate between dates and observe the most shared content, as shown in Figure 8.4. The system also allows users to select the period they want, comprising, for example, a day, a week or an entire month of data. After choosing a start and end date for the search, the system retrieves from the database and reports the most popular content for the entire selected time window. This allows journalists and researchers to investigate a specific period or even events that last more than a day, combining thousands of messages in a summarized and ranked interface, in which some publication and content patterns may emerge that, without the system, it would be difficult to notice.

Furthermore, in *WhatsApp Monitor*, five types of content were identified and persisted: images, videos, audio messages, external links, and text messages (only those with more than 140 characters). Our system daily displays the content divided by media type and shows them ranked by number of shares. This allows journalists to get an idea on a daily basis about critical content shared on WhatsApp that might be worth being fact-checked. Figure 8.5 depicts a screenshot of an example of media exhibited in *WhatsApp Monitor* after user select it.

To give more details about each content shared on WhatsApp, by clicking on an object in the dashboard, we make available the information compiled from the total shares

[2] The Indian and Indonesian version only have static data from 2019 during the general elections, while Brazil continue to be updated daily.

Figure 8.5: Once a user click on a specific content, *WhatsApp Monitor* displays a popup with more information of the media selected.



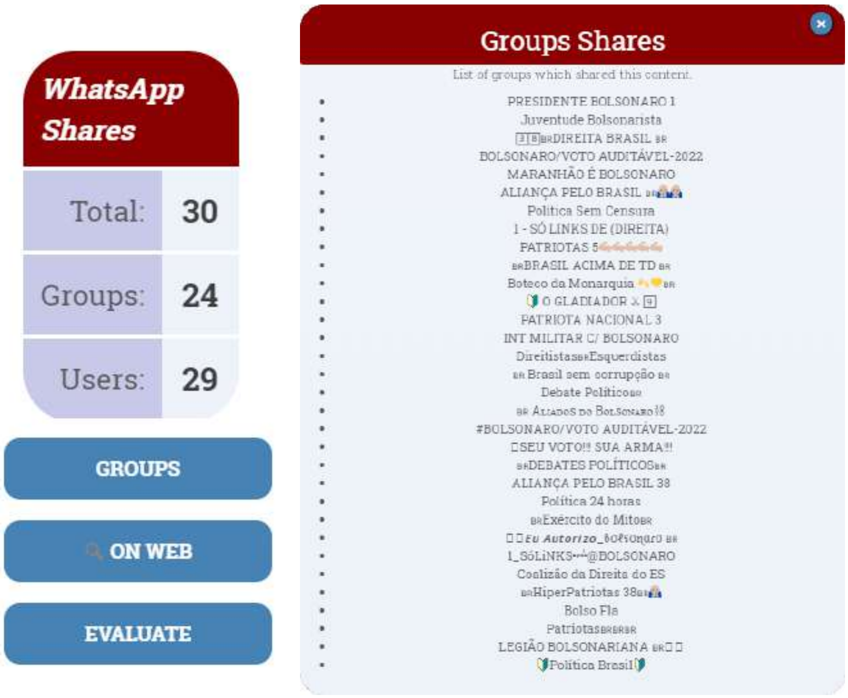
Source: The Author.

among the monitored groups for each selected content. A user has at their disposal the total number of shares, how many groups that content appeared, and how many unique users posted about that content. With this, it is possible to distinguish, for example, between a spam that is widely shared but posted to a few users, and a viral piece of news that the total of shared is distributed among several different groups and users. In addition, by clicking on "GROUPS", we display the names of the groups the media has appeared in, to help people to identify some context about the content (for example, a posted content is more left-aligned or right-aligned). Figure 8.6 shows the web elements that provide such information to the users.

Moreover, we provide a tool to enable fact-checkers to verify information quickly for each media content since system conceiving in 2018. In *WhatsApp Monitor*, by clicking on an image, one has a simple way to double-check these messages by searching "On Web" button that uses Google's reverse image search to track external sources where that exactly the same image was shared. Interestingly, the WhatsApp application itself implemented a very similar functionality, called "Search in the web", later in August 2020, in which users can search the web for viral content[3]. According to WhatsApp, when you tap a magnifying glass button in chat (only enabled for messages that have been marked "Forwarded many times"), it searches the content on Google to help people find news or results other sources of information about the content searched. This feature works by allowing users to upload the message via their browser without WhatsApp ever seeing the message itself. Accordingly WhatsApp, the *Search the web* functionality on WhatsApp was available in Brazil, Italy, Ireland, Mexico, Spain, UK, and US for those on the latest versions of WhatsApp.

[3] <https://blog.whatsapp.com/search-the-web/?lang=en>

Figure 8.6: Details from medias on *WhatsApp Monitor*.



WhatsApp Shares

Total:	30
Groups:	24
Users:	29

Groups Shares

List of groups which shared this content.

- PRESIDENTE BOLSONARO 1
- Juventude Bolsonarista
- 38BRDIREITA BRASIL BR
- BOLSONARO/VOTO AUDITÁVEL-2022
- MARANHÃO É BOLSONARO
- ALIANÇA PELO BRASIL BR
- Politica Sem Censura
- 1 - SÓ LINKS DE (DIREITA)
- PATRIOTAS 5
- BRBRASIL ACIMA DE TD BR
- Boteco da Monarquia BR
- O GLADIADOR ⚔ 9
- PATRIOTA NACIONAL 3
- INT MILITAR C/ BOLSONARO
- DireitistasBREsquerdistas
- BR Brasil sem corrupção BR
- Debate PolíticoBR
- BR Aliados do Bolsonaro 18
- #BOLSONARO/VOTO AUDITÁVEL-2022
- SEU VOTO!! SUA ARMA!!
- BRDEBATES POLÍTICOSBR
- ALIANÇA PELO BRASIL 38
- Política 24 horas
- BRExército do MitoBR
- Eu Autorizo_bolsonaro BR
- 1_SóLiNKS@BOLSONARO
- Coalizão da Direita do ES
- BRHiperPatriotas 38BR
- Bolso Fla
- PatriotasBRBRBR
- LEGIÃO BOLSONARIANA BR
- Política Brasil

(a) Shared Details. (b) Groups Details.

Source: The Author.

8.1.3 How this System Has Been Used

For the *WhatsApp Monitor* Brazil, the database continues to receive daily updates, being available to its users since 2018, with a range of almost 4 years of uninterrupted content that would otherwise be impossible to retrieve, as much of the content does not exist anymore on WhatsApp servers and probably not even on the phones of the members of monitored groups. As far as we know, this system is the biggest record of the contents that circulated on WhatsApp in that (or any other) period of the political context. By making it available in a free, organized and systematic way for other people to access our data, we believe we are helping in the general understanding of what happened within this closed social platform and evaluating its impact on society.

8.1.3.1 Social Impacts

From the Brazilian elections of 2018 to the next presidential elections in October 2022, we provided access to the system to more than 300 users, including journalists, researchers and *fact-checking* agencies that explicitly mentioned our system as a data source during their checks. Additionally, dozens of news made reference to our system or used our data during the Brazilian elections and during the COVID19 pandemic to better understand the discussions taking place within WhatsApp. More specifically, news articles from BBC, The Guardian, El País, The Intercept, O Globo, Estadão, Folha, Uol, among several others that use the system to investigate WhatsApp and produce the report. A detailed list of news articles covering our research is provided in Appendix B

The *WhatsApp Monitor* was able to group similar content from a large volume of data and sort it daily into use by many journalists and agencies, facilitating the process of fact-checking. Notably, our system is referenced as a partner of Comprova[4], a collaborative journalistic project of *First Draft* with a focus on verifying stories published on social media and WhatsApp during the Brazilian presidential elections and by the fact-checking agency Lupa[5] by journalists from the Piauí Folha group.

We really hope that our system can be useful and help fact-checking agencies to combat misinformation and positively contributes to more and real improvements to our society.

8.1.3.2 Scientific Impacts

In addition to the impacts and collaborations mentioned above, several studies, carried out outside our research group at UFMG, also used the *WhatsApp Monitor* as a methodology and data source to advance research in the areas of WhatsApp and Fake News. The system helps these researchers to develop their work, providing more transparency and ease to navigate past periods of WhatsApp (SOARES et al., 2021b; SILVA, 2021; SOARES et al., 2021a; GOMES; NAKAGAWA; CARDOSO, 2020; RECUERO; SOARES; VINHAS, 2021).

MüZELL, 2020, in his master's thesis, he drew on system data during the 2018 elections, identifying strategies and patterns about how the political campaign took place on WhatsApp and how they impacted in the election. With the help of *WhatsApp Monitor*, he was also able to point out various popular misinformation circulating at that time. ALMEIDA et al., 2019, to find out what contents were aired in the 2018 elections, also used the *WhatsApp Monitor* to explore characteristics of the circulation of this platform, considering the closed space for exchanging information. OLIVEIRA; CASALECCHI; BACHINI, 2020 also investigates the *WhatsApp Monitor* data from the elections to the

[4] <https://projetocomprova.com.br/partner/monitor-de-whatsapp-ufmg/>

[5] <https://piaui.folha.uol.com.br/lupa/tag/ufmg/>

COVID-19 pandemic as the platform makes it difficult for the actors involved in the exchange of messages to be aware of the information and subjects that circulates there and for these to be questioned, thus causing a perverse effect of skewed and little-diverse circulation of information on WhatsApp public groups. SOARES et al., 2021a, with the help of our tool, also studied the misinformation about COVID-19 on WhatsApp, observing how the pandemic is framed as a political debate. In another direction, the study by TOMáS; TOMáS; ANDREATTA, 2020 investigated the fake news against public universities in Brazil, using our system to analyze a series of images that went viral on WhatsApp that attack public universities as a place of depravity and waste of resources.

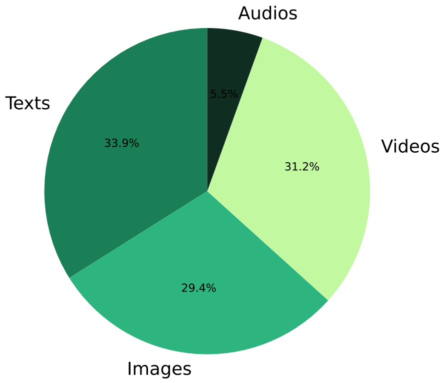
All these studies relies on data collected and exhibited by *WhatsApp Monitor* to investigating misinformation on WhatsApp environment and fully explored the potential of our tool to perform their researches. These results found that the system is able not only to help journalists, but can also be useful to researchers as they can concentrate their time analysing the data instead to worry in gathering or extracting data from WhatsApp.

8.1.4 Overview of WhatsApp Monitor Data

Type	Total	Unique
Large Text	4,970,345	1,604,185
Image	4,286,029	1,892,573
Video	4,583,715	1,694,633
Audio	808,972	518,160
Total	**14,649,061**	**5,709,551**

Table 8.1: Overview of collected messages by type along four years of WhatsApp data.

Figure 8.7: Composition of WhatsApp dataset by type.



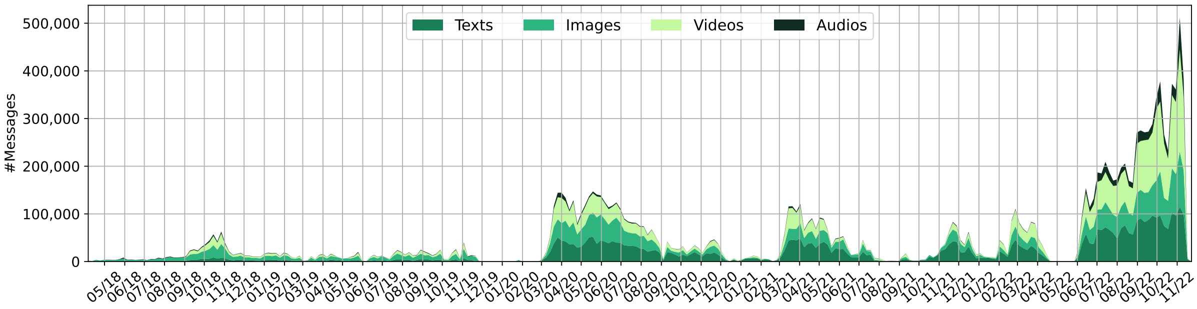
Source: The Author.

Considering the messages from the database of the *WhatsApp Monitor*, which has been online for a long period of time collecting data from public groups, there are about 15M messages collected. Since the objective of the system is to help researches and journalists to understand popular trends on WhatsApp, we filter in this set only multimedia messages and texts with 140 or more characters that represent larger text documents, sharing chains, and piece of news that circulate on WhatsApp. Small chat messages are not considered in the *WhatsApp Monitor* as they usually lack of context to provide any meaningful information alone. We can observe the large relevance of multimedia messages, specially images and videos, in the context of WhatsApp, being responsible for about a third each of the content displayed in the system.

It is interesting to note that there is a volume of recurring (repeated) messages almost three times as high as those that are unique. This suggests that messages on WhatsApp, thus, are very often shared to more users and only a third of the content is "original" while the other two thirds of the remaining messages are just duplicates or re-shares. That also highlight the importance of the step in the methodology of grouping the content based on similarity. This process can track the messages and find their duplicates, counting its dissemination on WhatsApp network, which represents, as the results shows, a large portion of the content.

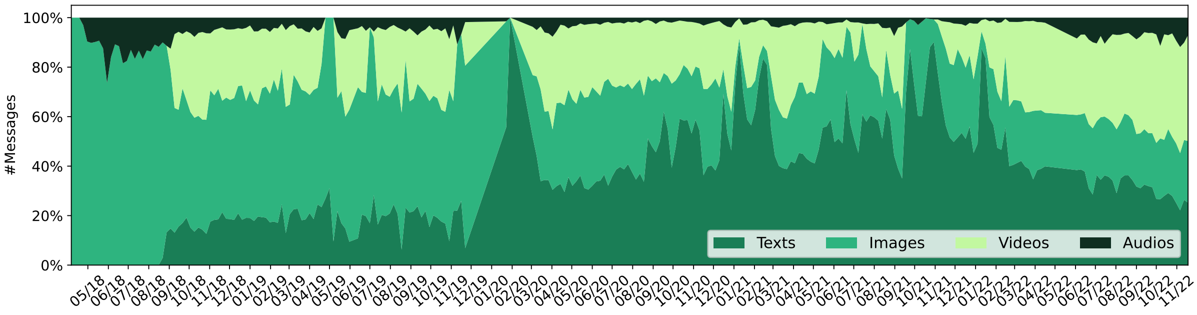
Before dive in the analysis of the content related to that period, there are some observations about on the timeline of data. Firstly, we need to discuss the impact of the limitations previously discussed in this chapter related to this long-term collection methodology. One is that it is possible to observe few small gaps, in which there is a much lower amount of collected data compared to the rest. Those, particularly at the beginning of 2020 and in May 2022, were periods when it was needed to make major adjustments in the infrastructure and codes of collection caused by limitations in resources for the research and also changes in how WhatsApp system works, causing the data gathering

Figure 8.8: Volume total of data collection from *WhatsApp Monitor* per week from 2018 to 2022.



Source: The Author.

Figure 8.9: Percentage of each media type on *WhatsApp Monitor* per week from 2018 to 2022.



Source: The Author.

to stop during those time intervals. In addition, while there is a notable growth in the volume of data between 2018 and 2022, and indeed events such as COVID-19 and Brazilian Presidential Elections populated the WhatsApp with an avalanche of new messages, it is worth to mention that, during this period, given the changes in this research by receiving new resources of smartphones and updates in the whole infrastructure, including searching and joining a large set of new public groups, that may also have influenced to the increasing of the data collected.

With those points in mind, while looking at this data, Figure 8.8 shows dataset collected by *WhatsApp Monitor*. There are some major events captured by our dataset: the 2018 Brazilian Elections, the COVID-19 pandemic and 2022 Brazilian Elections. We can relate the first to a spike between August and November 2018, which represent the period of Brazilian elections, this growth represents three times more messages than the periods just before and after the elections, reaching more than 50K messages a week near the actual voting day, which elected far-right candidate Bolsonaro as the new Brazilian President. This event is specially referred in Brazil by the abuse of fake news on WhatsApp and severe attacks against many politicians in the country. Another relevant event captured, after 2020 with a pick of 150K messages in March 2020, when COVID-19 pandemic emerged in Brazil. By this graphic of *WhatsApp Monitor* dataset, we can observe the huge volume of data shared on WhatsApp during the pandemic, collecting more than 100k messages per week during several weeks. Finally, between August and November 2022 there were another Brazilian Presidential Elections in which WhatsApp was again pointed as a key social media for sharing content regarding the candidates and spreading misinformation.

8.1.5 Final Considerations and Next Steps of *WhatsApp Monitor*

In this section, we presented our first approaching in combating misinformation issue on WhatsApp: the *WhatsApp Monitor*, a web support system for fact-checking in the fight against misinformation in the WhatsApp environment. This tool has been online since 2018, being updated daily with data from more than 1000 WhatsApp public groups on politics content and four year online. Our system is used as a data source by several researchers and journalists, including as a source for three *fact-checking* agencies.

The architecture of the system is able to collect, process, rank and display a large volume of multimedia content from WhatsApp groups on a web system. It is accessible by online browser only through username and password. Our methodology not only employs an innovative way to exhibit data from WhatsApp, but it has also proved to be effective in helping to combat misinformation.

With more transparency of WhatsApp data and ease of use of navigation by date showing the most popular content shared everyday, users are able to point out patterns and movements that emerge within the platform in a way that would be impossible without the system. This gives us valuable support for the laborious fact-checking task, reducing the effort needed to find the news and misinformation that goes viral on the Web.

Due to the enclosed and ephemeral nature of WhatsApp, our system acts as a kind of historical record of events that occurred on the platform, as this data would hardly be accessible otherwise, as the company does not store the content for a long time and even users involved in these chats may no longer have a register of sent messages anymore. Given the prominence and influence WhatsApp performed in society during the period f 2018 to 2022, specially in the context of political misinformation campaigns in Brazil, *WhatsApp Monitor* is a significant tool, even after these events have occurred, that allows people to look back and understand what that period was like within the universe of WhatsApp groups and have something that they can explore and search for.

Besides the critical usage of *WhatsApp Monitor* during elections, the system can be further expanded in different directions. It is possible to expand the system with more groups, but it can also be replicated for different contexts within WhatsApp and even for different instant messaging services. In relation to this last one, a copy of the system was built for Telegram. The *Telegram Monitor* was developed based in this system, also monitoring Brazilian political groups and channels, but in context of Telegram (JúNIOR et al., 2022). This system also enables a broader understanding of Telegram ecosystem in Brazil, supporting the fight against misinformation also in this network JúNIOR et al., 2021. These examples show how these kinds of monitoring tool for messaging services are beneficial to the society in investigating digital information disseminating on Web.

Moreover, we are also implementing periodic semi-automatic reports of specific dates and popular events occurring on WhatsApp to facilitate the interpretation of data in the system and provide information even to those without full access to the *WhatsApp Monitor*[6]. With all those steps, we hope that the system can continue to bring positive impact to our society and be a strong weapon in the fight against misinformation .

[6] <http://www.monitor-de-whatsapp.dcc.ufmg.br/brazil/reports.php>

8.2 Detecting Misinformation without Violating Users' Privacy

The popularity of IMPs such as WhatsApp revolutionized how users communicate and interact with the internet. Also, social communication around news is becoming more private as messaging apps continue to grow around the world. With so many users, WhatsApp plays an important role in this conjecture as it has become a primary network for discussing and sharing news in countries like Brazil and India where smartphones' use for news access is already much higher than other devices, including desktop computers and tablets (NEWMAN et al., 2019). Characteristics such as the immediacy of messages directly delivered to the user's phone and secure communication through end-to-end encryption have made this tool unique, but also allowed it to be extensively abused to create and spread misinformation. Due to the private encrypted nature of the messages, it is hard to track the dissemination of misinformation at scale or even propose means to combat this issue.

On other online social networks such as Twitter and Facebook, there is a constant moderation over the content shared there, not only for misinformation but also for hate speech, child exploitation and many other extraordinarily sensitive topics that go against the policy terms of these companies. Unlike such social platforms, which can enforce moderation, the end-to-end encrypted (E2EE) structure of WhatsApp creates a very different scenario where this is not possible. Only the users involved in the conversation have access to the content shared, shielding abusive content from being removed. The key challenge is to fight misinformation in WhatsApp, keeping it as a secure communication channel based on end-to-end encryption.

This format anonymizes messages and makes it difficult to track them, since only users involved in the conversation have access to the content shared, shielding abusive content from being removed or blocked but still allows for mass dissemination. Recently, this end-to-end encryption adopted by messaging platforms created a discussion concerning privacy and public safety in an open letter from the UK Home Secretary, the US attorney General, the US Acting Secretary of Homeland Security, and the Australian Minister for Home Affairs to Facebook (PATEL et al., 2019; MAYER, 2019), in which they contrast needs for enhancements in virtual security and vulnerability in the physical world.

The dichotomy between an encrypted anonymous communication and the ability of WhatsApp to act as a viral tool to spread rumors/misinformation is being widely debated in discussion around the world. Although the E2EE adopted by WhatsApp can help protect free speech, improve the safety of political dissidents and prevent censorship by both government and tech platforms, it also can make detecting crime more difficult, exacerbating the harm caused to victims. These discussions point to an incongruity between E2EE and content moderation, and lead to solutions being proposed that are forms of content moderation compatible with end-to-end encrypted messaging, without compromising important security principles or undermining policy values MAYER, 2019; GUPTA; TANEJA, 2018. In many cases, these discussions of content moderation for end-to-end encrypted messaging posed as a way of creating a backdoor into messaging apps that would allow access to the content of private communications (WONG, 2019). However, enabling content moderation for end-to-end encrypted messaging is a different problem from enabling law enforcement access to message content. The problems involve different technical properties, different spaces of possible designs, and different information security and public policy implications.

The idea is based on on-device checking, in which WhatsApp can detect when a user shares multimedia content, which has been previously labeled as misinformation by fact-checkers, without violating the privacy of the users. In the proposed a moderation methodology, WhatsApp could automatically detect when a user shares images and videos which have previously been labeled as misinformation, similar to how Facebook would flag content for fake news (HUNT, 2017) without violating the privacy of the user and compromising the E2EE within the messaging service. The solution is based on having hashes of previously fact-checked content on the device of the user, which can be quickly checked before the content is encrypted.

We evaluate the potential of this strategy in combating misinformation using data collected from both fact-checking agencies and WhatsApp during recent elections in Brazil and India. Using a large sample of labeled data from WhatsApp, we assess the potential of this strategy in combating misinformation and characterize the misinformation shared on WhatsApp using two distinct datasets of WhatsApp messages and fact-checkers from Brazil and India. Our results show that our approach has the potential to detect a considerable amount of images containing misinformation, reducing 40.7% and 82.2% of their shares in Brazil and India, respectively.

Currently, the only way of moderation of abusive messages is if users report the users/groups who post such messages directly to WhatsApp, even group administrators are restricted only to ban members while they are not allowed to remove the messages of other users within the chat. However, with our proposed approach, WhatsApp could have the content moderation executed on user messages directly on a user's phone, before encryption through an automatic architecture that checks the content against a dataset of fact-checked images. In this way, it would help WhatsApp prevent users from sharing content that violates its terms of service, without compromising the E2EE within the messaging service.

8.2.1 Background on WhatsApp and Security

The emergence of WhatsApp has made a new communication media available to smartphone users, seen as significantly more private and secure than other social media like Facebook (SIMON et al., 2016). The E2EE implemented on the platform is an important element that ensures user privacy and security, mainly during critical and crisis events. WhatsApp, in its technical explanation (WhatsApp Messenger, 2021), describes that all communication (including chats, group, images, videos, voice messages and files, and WhatsApp calls) between WhatsApp clients and WhatsApp servers is layered within a separate E2E encrypted channel, using Noise Pipes with Curve25519, AES-GCM, and SHA256 from the Noise Protocol Framework for long-running interactive connections.

MALKA; ARIEL; AVIDAR, 2015 demonstrate the importance of WhatsApp privacy in the lives of its users during the wartime in Israel, where users take advantage of the app's security to communicate with relatives on the battlefield and quickly circulate information of interest to the entire local community. They show the importance of the multi-functional role of the messaging app, functioning as a mass as well as an interpersonal communication channel. In this context, WhatsApp became the subject of public, media, and political discourse, especially within the context of protecting users' information. The security of smartphone messaging applications, though, is not a new concern. Earlier versions of WhatsApp (2.6.4) along together other messaging apps were evaluated in SCHRITTWIESER et al., 2012, prior to its acquisition by Facebook. They discovered several flaws in WhatsApp security at that date, more notable that the phone number verification process of WhatsApp at that was fatally broken, as the verification SMS message was generated on the phone itself and then sent to the server via a HTTPS connection. An attacker, then, could exploit this mechanism to hijack any WhatsApp account. All those security problems were fixed in recent versions, but privacy remains an issue for many researches.

RASTOGI; HENDLER, 2017 also analyze the security architecture of WhatsApp version (2.16.2) with a focus on its privacy preservation. They criticize WhatsApp's decision not to encrypt metadata sent to the service provider and raise privacy concerns as it can reveal just enough information to show connections between people, their patterns, and personal information. Currently, there is a broad discussion concerning multiple messaging apps and their security features (WILLIAMS, 2021).

The massive spread of misinformation and rumors on WhatsApp has led to public requests to curb virality features as a way to undermine the spread of misinformation at scale (TARDAGUILA; BENEVENUTO; ORTELLADO, 2018). Concerns about fake news dissemination have lead WhatsApp and encryption to become the subject of public, media, and political discourse, with requests from India government (BENGALI, 2019) and other governments arguing that WhatsApp needs to provide legal backdoors for security and law enforcement purposes (BARNES; ISAAC, 2019) or enable tracing the source of problematic messages (PHARTIYAL, 2018). On the other hand, privacy advocates and WhatsApp argue that such a move would completely break E2EE and hence a threat to protecting users' information. They argue that because of E2EE, moderating content is a challenging task on WhatsApp since any attempt to look at the content would compromise the security of the communication.

Although WhatsApp has deployed limits for forwarding, our study on forwarding dissemination showed that such limits can only offer delays on the information spread, but they are ineffective in countering the propagation of misinformation campaigns that are highly viral

However, broadly speaking, most of the efforts on WhatsApp involve studies that aim at better comprehending the phenomenon of misinformation spreading on WhatsApp, focusing on structural patterns of them or the characteristics of propagation dynamics (BURSZTYN; BIRNBAUM, 2019; MACHADO et al., 2019; CAETANO et al., 2019) and not in proposing mechanisms to combat it specially in the context of E2EE security of the app, and little was done in direction of countermeasures to minimize the impact of misinformation spreading there. The effort we proposed here is complementary to the existing ones, being the first of its kind to provide a practical proposition that if implemented by WhatsApp has the potential to significantly decrease the amount of misinformation *in real time*.

8.2.2 Balancing between Encryption and Moderation

In this proposed approach, we show that there is a simple way to find harmful messages without violating user privacy or creating a threatening surveillance backdoor. Automatic classification through machine learning, user reporting messages and repositories of popular content are forms to stop misinformation that are compatible with E2EE, as pointed by recent reports (GUPTA; TANEJA, 2018; MAYER, 2019).

Our proposed solution was designed with three main considerations: (a) It should be easy to implement, by allowing WhatsApp to port our solution to their existing infrastructure without much changes; (b) it must be flexible, be able to detect as much misinformation as possible, at scale, and adapt to the ever-changing trends in misinformation creation; and, (c) it must not compromise the end-to-end encryption services that WhatsApp provides.

Based on these considerations, we propose our solution, which covers part of the spectrum, basing on two key ingredients:

1. A database of previously fact-checked content: This approach requires a set of images labeled as misinformation by specialized fact-checking agencies and a technique to automatically match the images shared on WhatsApp to those checked. Since Facebook already has partnerships with several fact-checking agencies around the world, such a database is not hard to obtain. Moreover, Facebook also collects media items reported as problematic images (misinformation, hateful, etc) through its internal review processes. Fact-checking agencies could provide the images they have checked, creating a human labeled misinformation database. However, it is also possible to get the content checked directly from agencies channels via web crawling.

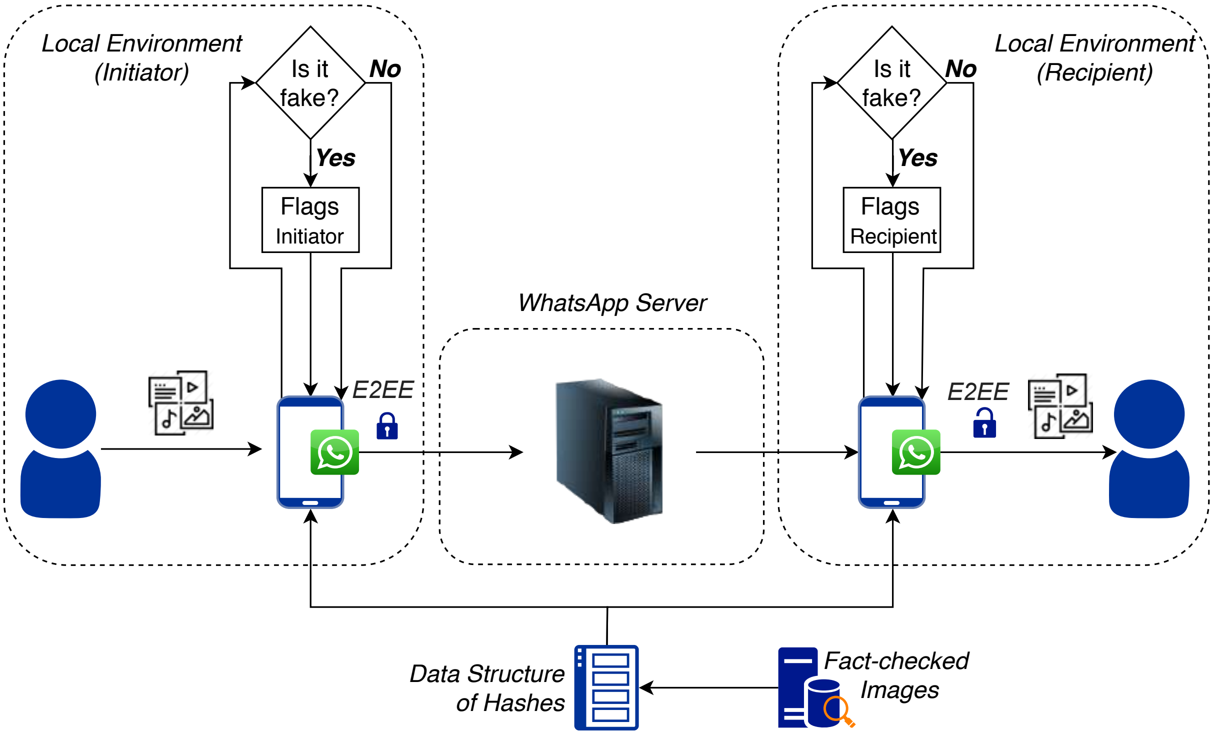
2. Algorithms for hashing and matching similar media content: A hashing algorithm provides a signature to represent an image or video. Given the exact same content, the hashing algorithm produces the same hash. Multiple types of hash functions exist to achieve this goal. In this work, we are primarily interested in two types of hash functions: a. Cryptographic hash, b. Perceptual hash. A cryptographic hash is a one-way hash function based on techniques like MD5 or SHA, and produces a string hash given an image. However, even changing a single pixel in the image changes the hash completely. Hence, cryptographic hashes can be used to only detect *exact* matches. On the other hand, perceptual hashing takes care of the drawbacks of a cryptographic hash and produces a hash that can be used to compare *similar* images. Even if the image is slightly rotated, cropped or has text added, a good perceptual hashing technique can produce a hash that is similar to the original image. There are multiple algorithms to produce perceptual hashes such as Facebook PDQ Hashing, pHash, Microsoft PhotoDNA[7], etc. Perceptual hashing is already widely used today for detecting known harmful content (FARID, 2018) and authentication of images (Swaminathan; Yinian Mao; Min Wu, 2006).

8.2.3 Architecture

An overview of the proposed architecture is shown in Figure 8.10 can be explained in the following steps: (i) WhatsApp maintains a set of hashes of images which have been previously fact-checked, either from publicly available sources or through internal review processes. (ii) These hashes are shipped with the WhatsApp app, storing it on a user's phone. This step can be periodically updated based on images that Facebook's moderators have been fact-checking on Facebook, which is much more openly accessible. This set

[7] <https://github.com/facebook/ThreatExchange/tree/master/hashing/tmk>, <https://www.phash.org>, and <https://www.microsoft.com/en-us/photodna>

Figure 8.10: Proposed architecture.



Source: The Author.

could be condensed and efficiently stored using existing probabilistic data structures like Bloom Filters (SONG et al., 2005). (iii) Once a user intends to send an image, WhatsApp checks whether it already exists in the hashed set on the user's device. If so, a warning confirmation is displayed, asking if the user really wants to share this content. (iv) The message is encrypted and transferred through the usual E2EE method. (v) When the recipient user receives the message, WhatsApp decrypts the image on the phone, obtains a perceptual hash, and checks it on a hashed set on the receiver's end. (vi) If it already exists, the content is flagged, and a warning is shown to the user indicating that the image could be a potential misinformation. Also, providing information about where the image was fact-checked; and in addition, also prevent the image from being forwarded further.

This architecture requires changes in WhatsApp, as it introduces a new component containing hashes stored on the phone and checking images. It provides high flexibility and the ability to detect near similar images, hence increasing the coverage and effectiveness in countering misinformation. This architecture also fully abides by the current E2EE pipeline WhatsApp has, where WhatsApp does not have access to any content information.[8] All the matching and intervention is done on the device without the need for any aggregate metadata in the message. Facebook could optionally keep statistics of how many times a match occurred to establish the prevalence and virality of different types of misinformation and to collect stats about users who repeatedly send such content. Note that similar designs have been proposed recently for informing policy decisions in light

[8] <https://www.whatsapp.com/security/WhatsApp-Security-Whitepaper.pdf>

of governments requesting a backdoor in the system (GUPTA; TANEJA, 2018; MAYER, 2019).

It is important to mention that while WhatsApp messages are secure in transit, the endpoint devices, such as smartphones and computers, do not offer security. In this sense, our architecture adds new components to the client, adding also more potential for security breaches.

We say that our solution is practical and deployable because it is an industry-standard to detect unlawful behavior in social media platforms (FARID, 2018). For example, WhatsApp scans all unencrypted information on its network such as user/group profile photos, and group metadata for unlawful content such as child pornography, drugs, etc. If flagged, these are manually verified (CONSTINE, 2018) and the abusing accounts are banned. Our proposal extends the same methodology to the user's device in order to enable private detection.

The method works to prevent coordinated disinformation campaigns that are particularly important during elections (ISAAC; ROOSE, 2018) and other high profile national events (ELLIS-PETERSEN, 2019), but also stops basic misinformation, where a lack of awareness leads to spreading. For instance, while manually labeling the fact-checked misinformation images, we observed that roughly 15% of the images in our data were related to false health information. These are forwarded mostly with the assumption that they might help someone in case they are true. In some cases (e.g. the child kidnapping rumors (ARUN, 2019)), such benign forwarding of misinformation lead to violence and killing (DIXIT; MAC, 2018).

8.2.4 Labeling a Dataset for Misinformation on WhatsApp

With the proposal presented in the previous section, it requires that a database of fact-checked images be matched with WhatsApp content to flag misinformation. In order to evaluate the practical potential of the proposed architecture, we built a dataset of WhatsApp messages containing misinformation and a dataset from real fact-checkers identifying which content is fake or not to simulate this process. To reach to this large WhatsApp misinformation dataset, we gathered data from public WhatsApp groups discussing politics from Brazil and India, as already extensively explained in the data collection of this thesis (Chapter 5). Complementary, we also collected a dataset of fact-checked misinformation images from well-known and publicly available fact-checking websites.

Table 8.2: WhatsApp collection.

	#Users	#Groups	Unique Images	Total images	Time Span
Brazil	17,465	414	4,524	34,109	2018/08 - 2018/11
India	63,500	4,250	509k	810k	2019/02 - 2019/06

8.2.4.1 WhatsApp Data.

To gather the data explored in this work we use available collection methodology to get access to messages posted on public WhatsApp groups. We selected over 400 and 4,200 groups from Brazil and India, respectively, dedicated to political discussions. The period of data collection for both countries includes the respective national elections in these countries. For this part of the work, we choose to filter only messages containing images. To evaluate our architecture, we selected here only images because this kind of media content is easier to track and keep immutable, while text messages tend to present slightly changes as it is disseminated on the network (e.g. add or remove a emoji, or a URL in the message), which is harder to check if it represent the same piece of misinformation. The dataset overview and the total number of users, groups, and distinct images are described in Table 8.2. Note that the volume of content in India is ten times bigger than Brazil.

8.2.4.2 Fact-checking Agencies Dataset.

As our methodology relies on the fact-checking task performed by specialized agencies, we need to build a wide dataset of previous labeled content for misinformation in order to compare it to WhatsApp. This would be the same process that the company can adopt in the design of the application, but it also can be replaced by a bigger partnership between WhatsApp and fact-checking agencies in which the agencies provide the labeled content direct to WhatsApp as they already do for Facebook (HUNT, 2017).

First, we create a list of the main and well-known fact-checking agencies in Brazil ("Folha–Lupa" <piaui.folha.uol.com.br/lupa/>, "Aos Fatos" <aosfatos.org>, G1–"É ou Não É?" <g1.globo.com/e-ou-nao-e/>, "e-Farsas" <www.e-farsas.com>, Veja–"Me Engana que eu Posto" <veja.abril.com.br/blog/me-engana-que-eu-posto/>, and "Boatos.org" <www.boatos.org>) and also some fact-checkers from India (<altnews.in>, <boomlive.in>, <smhoaxslayer.com>, <factchecker.in>, <factly.in>, <fakenewscounter.com>, and <check4spam.com>).

Then, for each agency, we developed a web crawler that navigates through the content of the website of the and collects all news page which have fact-checked content. Furthermore, for each of piece of news labeled, we collect the images shared along the page, as well we also obtained the label given by the fact-checker and the date when they were fact-checked. In total, we collected over 100k fact-checked images from Brazil and about 20k images from India.

Moreover, for this work specifically, we used the state-of-the-art perceptual hashing based image matching technique, PDQ hashing, to look for occurrences of the fact-checked images in our data from public groups. The PDQ hashing algorithm is an improvement over the commonly used pHash and produces a 256 bit hash using a discrete cosine transformation algorithm. PDQ is currently the method used by Facebook to detect similar content, and it is one of the best known state-of-the-art approach for clustering together similar images. The hashing algorithm can detect near similar images, even if they were cropped differently or they have small amounts of text overlaid on them, which is better to detect more duplicate images compared to pHash. By comparing the hashes from dataset from fact-checking agencies and WhatsApp images, we can match those from WhatsApp that is fake. By this, we can also compare the dates between when it was shared on WhatsApp and when it was fact-checked. As the perceptual hash allows comparison of values and not only exact matches, we use a threshold similarity of more than 0.8 to match two images as the same.

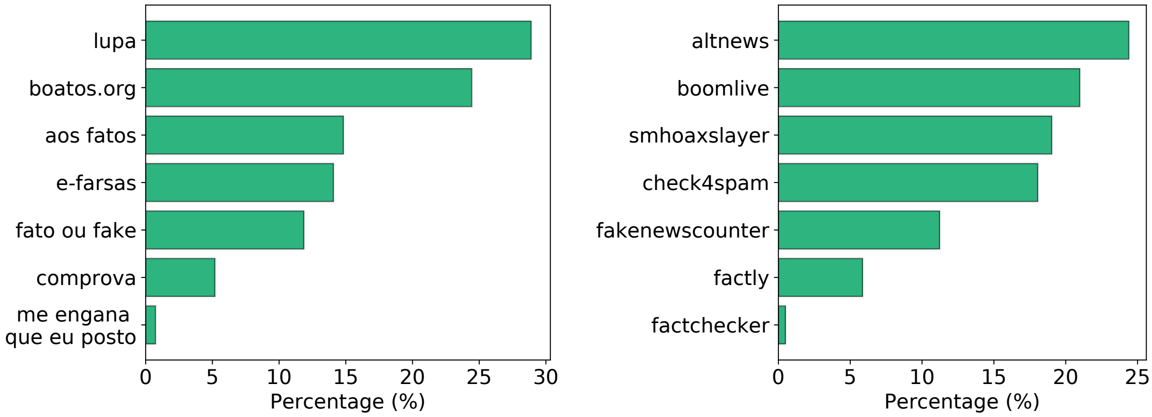
We also used a second strategy to expand and validate our fact-checking dataset of labeled misinformation images. For each WhatsApp Image from our dataset, we used Google reverse image search to check whether one of the main fact-checking domains were returned when searching for an image in our database. If so, we parsed the fact-checking page and automatically labeled the image depending on how the image was tagged on the fact-checking page. Finally, to make sure our dataset was accurately built, we manually verified each image that appears in both the fact-checking websites and in the WhatsApp data.

As shown in Table 8.3, this dataset of images previously fact-checked contains 135 images from Brazil and 205 images from India, which were shown to contain misinformation. It is important to highlight that many checking agencies do not post the actual image that has been disseminated. Often only altered versions of the image are posted and other versions of the false story are omitted to avoid contributing to the spreading of misinformation. This leads to us to have a small number of matches compared to the total number of fact-checked images we obtained, but that is sufficient to properly investigate the feasibility of the proposed architecture. Direct contact with the fact-checking agencies, like Facebook already does, certainly increases the size of the fact-checked set much more. Note that even though the set of fact-checked images is small, the fact that these images have been fact-checked means that they were popular and spread widely. Table 8.3 shows a summary of the fact-checked images and their activity in our dataset. It also shows the

Table 8.3: Amount of misinformation image shared on WhatsApp and comparison of shares before and after the checking date of fact-checking agencies.

	Misinformation Images found	100% Exact Matches	Total Shares	%Shares After Checking	Max Shares After Checking
Brazil	135	7	2,209	40.7	96
India	205	83	2,944	82.2	1,089

Figure 8.11: Distributions of WhatsApp misinformation images labeled per fact-checking.



(a) Brazil

(b) India

Source: The Author.

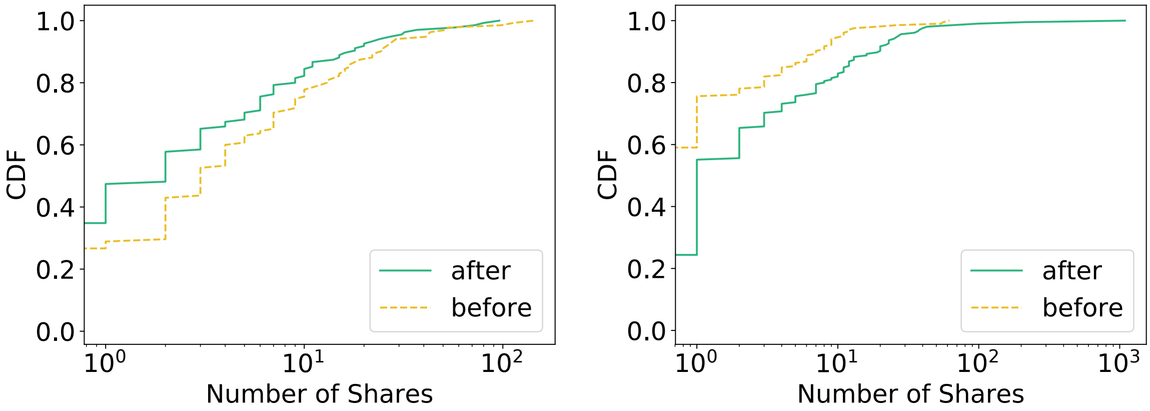
drawbacks of using a 100% exact match for hashing comparison. While similar perceptual hashes are able to identify more than a hundred images in both countries, using just the exactly same hash to find misinformation, only 5,1% of checked images from Brazil were retrieved and 40% of Indian images.

Figure 8.11 shows the breakdown of fact-checking agencies used and the amount of image retrieved from WhatsApp found in each of them. We find that Lupa is the fact-checking agency that most matches images containing misinformation in Brazil (28.9%) whereas, in India, Alt News has the highest number of matches (24.4%).

8.2.5 Potential Prevention of Misinformation

Furthermore, we evaluate the potential prevention of misinformation in case our architecture was implemented, and the spreading of these images were totally blocked immediately after the fact-checking happens. For this, we computed the timestamp of all the fact-checked images and the occurrence of these images in our WhatsApp data. This way, we are able to measure how many posts on WhatsApp were shared for each misinformation image before and after the first fact check of this image by any agency.

Figure 8.12: Total of shares of all misinformation images before and after they were fact-checked on each country.



(a) Brazil

(b) India

Source: The Author.

Figure 8.12 shows the cumulative distribution function (CDF) of the number of shares done before and after the checking of the misinformation images. We can observe in both countries that for the most broadly shared images there are as many posts before as after the checking date. Moreover, in India, there are more shares after checking than before and there are even images with up to 1,000 shares after fact-checking while the maximum shares before do not exceed 100.

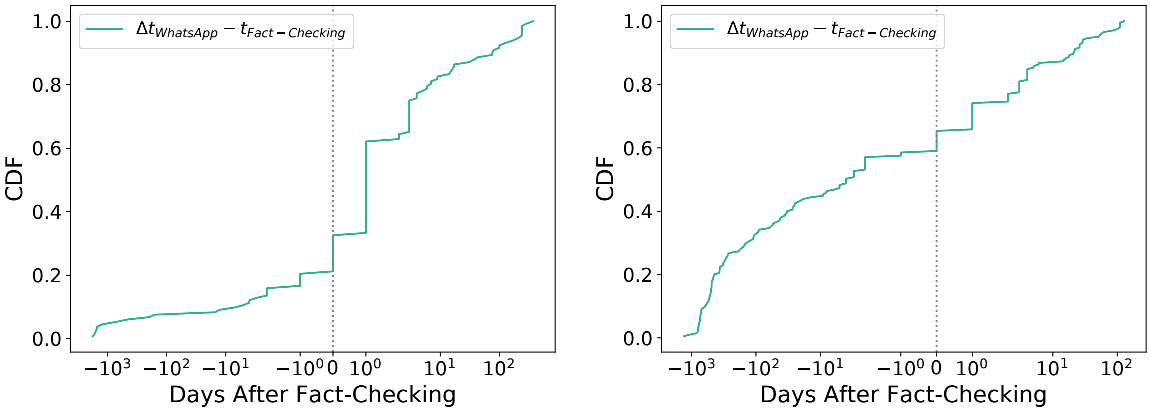
Summing all shares, we find that **40.7% of the misinformation image shares in Brazil and 82.2%**[9] **of the shares in India could have been avoided** by flagging the image and preventing it from being forwarded after being fact-checked. This demonstrates the importance of using fact-checking agencies to combat misinformation on WhatsApp highlighting the potential of our proposed approach.

8.2.5.1 When are Images Fact-checked?

As mentioned, for each fact-checked image, we know the exact time when it was fact-checked. In Figure 8.13 we evaluate the time difference between an image to appear on one of our WhatsApp groups, and to be fact-checked. Precisely, we compute the difference between the date of the first share on WhatsApp ($t_{whatsapp}$) and date of the fact-checking ($t_{fact-checking}$). For Brazil, we see that 20% of the images had already been checked even before their first appearance in our WhatsApp dataset, 12% of them were checked on the same day that they appeared in our dataset, and, 68% of the images were

[9] This number drops to 71.7% if we remove the outlier image with the maximum number of shares, as it was shared over 1000 times.

Figure 8.13: CDF between the date of the first share on WhatsApp and date of the fact-checking .



(a) Brazil

(b) India

Source: The Author.

checked after they had already appeared in our WhatsApp collection. For India, 58% of the images had already been checked before they appeared in our dataset, 7% of them were checked in the same day while only 35% of the images were checked after their first occurrence on WhatsApp. Therefore, using our approach has the potential to completely avoid the emergence on WhatsApp of roughly 60% of the misinformation images in India and one-fifth of the images in Brazil.

From Figure 8.13, we also can see that there are images that have been fact-checked more than 3 years ago by the fact-checking agencies that are still spreading (Negative values represent that checking occurred before the image first appeared on WhatsApp while positive values were those checked by some agency after appearing on WhatsApp). Although our collection represents only a portion of WhatsApp, we can see that some fake content continues circulating long after it has been checked and could be prevented before reaching the platform. On the other hand, there are also images that take over 100 days after its appearance on WhatsApp to be fact-checked by some agency, showing that there are also kinds of content that require fast attention to be checked.

Those results evidence the importance of fact-checking context of fake news and the potential of our approach in the combat of misinformation. Firstly, as some misinformation took a lot of time to enter in sight of the fact-checking agencies, as we find images that were checked only 100 days after we first appeared on WhatsApp, the WhatsApp Monitor can make it easier for those journalists labeling misinformation to find relevant content to be checked. On the other hand, as there are checked content that was labeled years before reach the WhatsApp, an moderation approach based on the checked content would avoid this content to still circulate in the platform after so long.

8.2.5.2 Characterizing Misinformation on WhatsApp

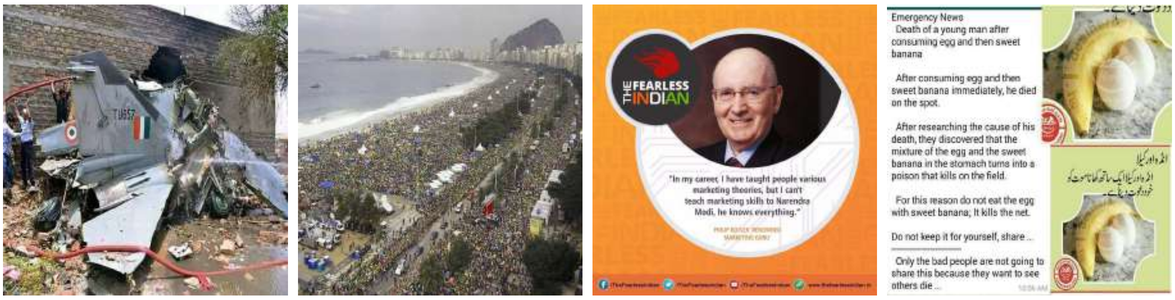
Since we only have a few hundred examples of fact-checked misinformation images, we manually explored these examples to obtain a loose typology of the various types of image based misinformation and coded the images into multiple categories.

Figure 8.14 show examples of the most common types of misinformation images we find in our dataset, which are old existing images re-shared out of context and images with false textual messages with the image. In example of Figure 8.14(a) we have an image of a plane crash taken out of its real context. This is an old image of an Indian plane crash, but the image spread widely during the India-Pakistan conflict in March 2019, that this is a Pakistani plane downed by India; Figure 8.14(b) represents an image of World Youth Day, which brought together nearly one million faithful to accompany Pope Francis on the Copacabana Beach in Brazil. This image was spread during the 2018 Brazilian presidential elections as if related to a popular manifestation in favor of a presidential candidate Jair Bolsonaro. In Figure 8.14(c) there is a false quote praising the Indian prime minister, Figure 8.14(d) is about a false health scare. Health misinformation is very prevalent in our dataset, even though we specifically monitored political groups.

Such a typology could help us understand the different types of deception used. Upon independent manual coding of 135 images from Brazil and 205 images from India, we came up with four main categories of the type of image based misinformation: (i) Images taken out of context: These represent old images that are taken out of context and re-shared (e.g. Figures 8.14(a) and 8.14(b)), making up roughly 30-40% of the images; (ii) Photoshopped images: images which are manipulated/edited (20%); (iii)False statistics and quotes: for funny memes, yet misleading images, usually containing incorrect stats or quotes (10%); (iv) Other: containing all other types of fake images, consisting of health scares, fake alerts, etc.

For both India and Brazil, we found that old images that are taken out of context,

Figure 8.14: Examples of the most common types of misinformation images from our dataset — old images re-shared out of context.



"In my career, I have taught people various marketing theories, but I can't teach marketing skills to Narendra Modi, he knows everything."

Emergency News
Death of a young man after consuming egg and then sweet banana

After consuming egg and then sweet banana immediately, he died on the spot.

After researching the cause of his death, they discovered that the mixture of the egg and the sweet banana in the stomach turns into a poison that kills on the field.

For this reason do not eat the egg with sweet banana; It kills the net.

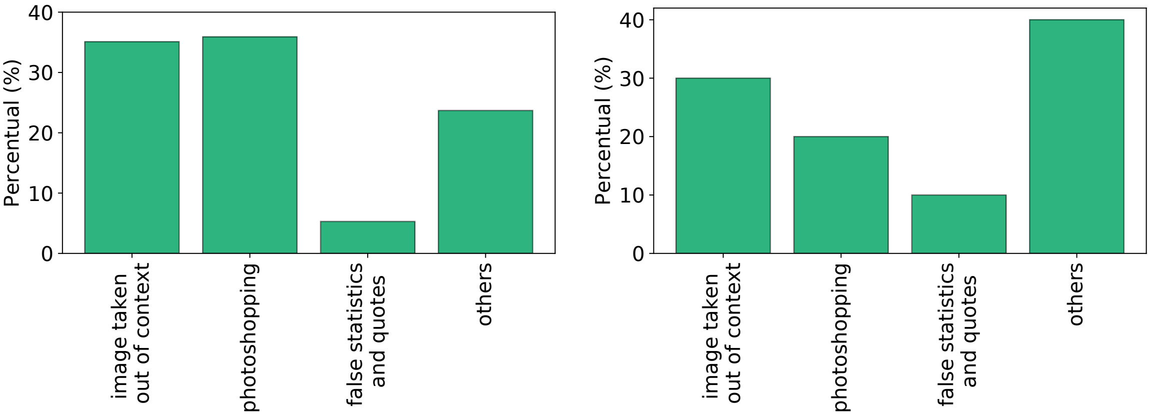
Do not keep it for yourself, share ...

Only the bad people are not going to share this because they want to see others die ...

(a) (b) (c) (d)

Source: The Author.

Figure 8.15: Distributions of image categories of misinformation.



(a) Brazil

(b) India

Source: The Author.

also termed cheap or shallow fakes (PARIS; DONOVAN, 2019) make up roughly 30% of our misinformation image dataset. This finding shows that even without any new fact-checking efforts, if WhatsApp creates just a database of old, already shared images, it could detect roughly 30% of the misinformation images that are being shared.

Figure 8.15 presents the fraction of images that fall into each of these categories. The next popular category is the simple doctored/photoshopped images, which make up over 20% of the images, followed by fake quotes/stats which make up 10% of the images. Note that just the top three categories of misinformation constitute over 60% of the fake images.

8.2.6 Discussion on Results

In this proposed approach to combat misinformation on WhatsApp, we propose a practical solution that WhatsApp could implement to prevent misinformation from spreading while ensuring user's privacy. The solution is based on having a set of already fact-checked image hashes on the user's device and matching these images with the content being shared. We would expect that by the time fact-checking organizations receive and fact-check a piece of content, most of its spread would be done, thus defeating the purpose of fact checking. However, as our results show, in part because of the closed nature of the platform, and the lack of a central authority to stop the spread, there are images that keep spreading even long after being fact checked.

Looking at the actual sharing of these images in our data, we show that over 40% of the spreading of misinformation detected in Brazil and 82% in India could have been prevented by implementing these measures in the public groups we monitor. Besides that, our results show that there are images that took over 100 days to be checked. Hence, a still open problem is the proposal of solutions to help fact-checking agencies in identifying early images containing misinformation. Our manual categorization of the misinformation images reveals that over 30% of the misinformation images are just old images that were re-shared out of context. Such images could be prevented with our proposal. Apart from presenting a simple, practical and deployable solution to the problem, our paper presents a counter-voice to strong claims by governments to allow backdoors in encryption for law and order purposes. Although our strategy is applied specifically to WhatsApp, it could be easily adapted for other contexts that also make usage of E2EE technology such as Telegram, Viber, Line, Signal, and others. Finally, this approach is also in line with WhatsApp's efforts to limit forwarding. As discussed in Chapter 7, this approach can impose delays in the content dissemination, which represent an extra time for fact-checking and more effectiveness for our approach.

Limitations. Labeling and implementing forwarding restrictions on already known fake images can only help to a certain degree. This proposal has a few limitations: (i) Does labeling actually make a difference? Firstly, careful considerations must be taken to prevent backfire effect (NYHAN; REIFLER, 2010; LEVIN, 2017); (ii) The dataset from WhatsApp is not representative, since it comes from public groups which are a small fraction of all groups. However, this is the largest available sample of WhatsApp data to test such an architecture. (iii) The amount of misinformation that could be prevented could be an overestimate because these fact-checked images are already popular. Even though our approach does not remove all misinformation, it can help remove popular, viral misinformation that has already been fact-checked. Given that only a small amount of content gets viral on WhatsApp, such efforts are helpful to prevent lethal mis/disinformation campaigns and rumors.

Chapter 9

Conclusion

In this chapter, we present the general outline of this thesis, summarizing the main results achieved, the final discussions and describing the future steps for this work.

9.1 General Outline

Given the open research questions that drive this thesis, we have obtained substantial results for each of them. In the following lines, we present the general outline of accomplished results of our study for those questions.

RG1 – Collect and build a consistent dataset of WhatsApp that can be explored to understand real-world events.

We developed a systematically and solid methodology to collect WhatsApp data, our methodology differs from previous efforts in collecting WhatsApp data, as it does not require the rooting process of smartphones to access the data, and then does no break any encryption security of the phone. Furthermore, the proposed approach includes means to discover public groups, and it can also track and merge duplicate piece of information circulating within the groups monitored and, then, provide aggregated metadata from each content collected. Moreover, our methodology to find relevant public groups on WhatsApp and collect data from them is novel approach to study IMPs environment, being an effective way to analysis WhatsApp data at scale not only in the context of this thesis but also replicated in several different scenarios (BURSZTYN; BIRNBAUM, 2019; CHANG, 2020; YADAV et al., 2020; JAVED et al., 2020; CABRAL et al., 2021; CARDOSO et al., 2020).

We also expand the approach to specific scenarios as our methodology to collect WhatsApp data covers since the groups discovery, profile creation until the steps of processing, storing and displaying the WhatsApp data in an organized way. This methodology is extensively described in Chapter 5 in order to anyone could reproduce while all scripts were made publicly available[1].The data collected was also available to researches and journalists through the web system WhatsApp Monitor (Chapter 8.1) by restricted login and password access, and a considerable portion of this content was labeled of misinformation and was released to other researches in a published dataset paper (REIS et al., 2020).

By developing a solid methodology of collection of large-scale of WhatsApp, this study provided consistent information relating some events such as the 2018 Brazilian presidential elections (RESENDE et al., 2019b, 2019a; MAROS et al., 2020) and COVID-19 pandemic (VASCONCELOS et al., 2020)) that played a big role in the recent years of our society and how they unfolded through the lenses of WhatsApp. We also made contributions by publishing a dataset paper with content labeled for disinformation (REIS et al., 2020). Our investigation on the process of discovery of public groups from IMPs resulted in a wide study that provided a better understanding in the ecosystem of groups from messaging platforms (HOSEINI et al., 2020).

RG2 – Understanding how the structure of the WhatsApp impacts on the dissemination of information within its network.

We used the large-scale data of WhatsApp public groups from Brazil, India and Indonesia to perform a series of experiments simulating the propagation of information in WhatsApp, comparing it to other relevant social medias that have similar group-based communication: Reddit and Telegram. We observed that WhatsApp network has similar characteristics of other social networks, with well-connected groups in which messages can circulate quickly and in bulk.

The experiments show that a message in WhatsApp can be viral and spread to the whole network, even it being encrypted (MELO et al., 2019b). We assess how the "Forwarding" mechanic of WhatsApp application impact on the dissemination of information through its network, concluding that it slows down the propagation, but is not enough to block the viral content to spread in WhatsApp. These limits may be further restricted to reduce viral spread in situations of misinformation or any other harmful content, such as pedophilia (CONSTINE, 2018).

RG3 – Finding measures that can be taken to combat misinformation circulating in WhatsApp After collecting data and observing how fake news spread within IMPs network, we evaluated means to combat misinformation in these platforms. Our efforts went in direction of proposing a solid methodology that can be adopted by instant messaging services to prevent the dissemination of misinformation. We analyzed the encrypted structure of WhatsApp and we also idealized an approach that could be adopted that preservers the privacy and security of the users in EE2E, while providing means to combat misinformation by storing popular hashes of pre-labeled fake news within the application. We proceed to evaluate our methodology with a labeled dataset

[1] <https://github.com/Phlop/WhatsApp_Crawler>

of misinformation on WhatsApp finding that more than 40% of the shares containing misinformation could be avoided if our approach was used. This metric was published in (REIS et al., 2020).

Additionally, and more important, we seek to implement other measures in direction to aid the hard task of fact-checking, showing that it is possible to offer techniques that seek to diminish the consequences of misinformation shared on WhatsApp. we deployed the WhatsApp Monitor, publishing the whole structure of the system in (MELO et al., 2019a).

As part of the project "Eleições Sem Fake" ("Elections Without Fake")[2] and inspired by this work, we implemented the "WhatsApp Monitor" – a Web tool to navigate through the content of the WhatsApp public groups monitored in this study available at <http://www.monitor-de-whatsapp.dcc.ufmg.br/>. This system allows the user to check what was the most popular images, videos, audios, and text message shared each day on WhatsApp groups. They can choose a day or even an entire period and the system retrieves the medias posted that time ranked by popularity. This system is restricted, accessed only by login and password as can contain sensitive data, but we provide access to use to more than a hundred journalists and to three fact-checking agencies who explicitly mentioned our system as a data source. More importantly, dozens of pieces of news have referred to our system or used its data during the Brazilian elections, suggesting it was useful to better understand the political campaigns and discussions within WhatsApp.

9.2 Real World Impact of the Research

Besides all the findings discovered through our data collection, experiments and analysis, it is important to evaluate further contributions achieved with the execution of this work. Given the novel approach we propose to investigate WhatsApp network of public groups and insights about this kind of platform, it has influenced other researches and studies on the platform. Also, with our long-term and large-scale data collection, we provide a rich resources support for many journalists and news outlets in the coverage of important topics happening on WhatsApp and which led to the publication of articles and news that helped to bring more information to the public about the content of our research. Next, we list some of them.

[2] <www.eleicoessemfake.dcc.ufmg.br>

9.2.1 Scientific Contributions

In addition to the results mentioned above, several studies, carried out outside our research group at UFMG, also used the *WhatsApp Monitor* as a methodology and data source to advance research in the areas of WhatsApp and fake news. The system helps these researchers to develop their work, providing transparency and ease to navigate past periods of WhatsApp (SOARES et al., 2021b; SILVA, 2021; SOARES et al., 2021a; GOMES; NAKAGAWA; CARDOSO, 2020; RECUERO; SOARES; VINHAS, 2021).

MüZELL, 2020, in his master's thesis, he drew on system data during the 2018 elections, identifying strategies and patterns about how the political campaign took place on WhatsApp and how they impacted in the election. With the help of *WhatsApp Monitor*, he was also able to point out various popular misinformation circulating at that time. ALMEIDA et al., 2019, to find out what contents were aired in the 2018 elections, also used the *WhatsApp Monitor* to explore characteristics of the circulation of this platform, considering the closed space for exchanging information. OLIVEIRA; CASALECCHI; BACHINI, 2020 also investigates the *WhatsApp Monitor* data from the elections to the COVID-19 pandemic as the platform makes it difficult for the actors involved in the exchange of messages to be aware of the information and subjects that circulates there and for these to be questioned, thus causing a perverse effect of skewed and little-diverse circulation of information on WhatsApp public groups. SOARES et al., 2021a, with the help of our tool, also studied the misinformation about COVID-19 on WhatsApp, observing how the pandemic is framed as a political debate. In another direction, the study by TOMáS; TOMáS; ANDREATTA, 2020 investigated the fake news against public universities in Brazil, using our system to analyze a series of images that went viral on WhatsApp that attack public universities as a place of depravity and waste of resources.

All these studies rely on data collected and exhibited by *WhatsApp Monitor* to investigating misinformation on WhatsApp environment and fully explored the potential of our tool to perform their researches. These results found that the system is able not only to help journalists, but can also be useful to researchers as they can concentrate their time analyzing the data instead to worry in gathering or extracting data from WhatsApp.

Moreover, as a result of the analysis and experiments performed by our research, some papers and articles containing partial or related findings of this work were published in high-impact venues as listed in Appendix A.

9.2.2 Social Contributions

During all years of our study, our system of *WhatsApp Monitor* has been used for hundreds of journalists, researchers and *fact-checking* agencies that explicitly mentioned our system as a data source during their checks. Additionally, dozens of news have made reference to this study or have used our data during their investigation of events within WhatsApp ecosystem. Major events such as truck's driver in 2018, presidential elections from 2018 and 2022, and the COVID-19 pandemic were widely covered by our WhatsApp data besides many other minor elements taking place within our monitored WhatsApp groups and chats which have helped to provide a big picture of WhatsApp for the society. More specifically, news articles from BBC, The Guardian, El País, The Intercept, O Globo, Estadão, Folha, Uol, among several others that investigate WhatsApp. A more detailed list of news articles referencing our research is provided in Appendix B

In 2018 and 2022, UFMG partnered with the Superior Electoral Court (TSE) through our monitoring system to contain the dissemination of fake news[3].

Moreover, during 2020, we developed a technology transfer project, implementing the WhatsApp collection tool from our system to the Public Ministry of Minas Gerais (MPMG) as one of the projects from the Analytical Capabilities Program, which aims to provide more transparency on public data online[4].

9.3 Future Works

Instant Messaging Platforms as WhatsApp become an important tool for communication as it is free and easy to use application that only requires Internet access for smartphone users to be in contact to millions of other users around the world. This popularity also transformed the app in a channel for information among its users, creating a distinctive environment that is both used for personal private encrypted chats and also for massive, viral and public messages. This is only possible because key features present in WhatsApp system of public groups and forwarding messages.

This particular kind of network also need specific forms of analysis to better comprehend is structure and the impact of its usage in our society, given their unique char-

[3] <https://www.tse.jus.br/imprensa/noticias-tse/2018/October/tse-studa-possibility-of-signing-a-partnership-with-university-to-inhibit-fake-news-no-whatsapp>

[4] <https://www.mpmg.mp.br/comunicacao/noticias/mpmg-inicia-trabalhos-de-convenio-com-ufmg-para-ampliar-capacidade-de-analyse-of-data.htm>

acteristics and challenges. Thinking about that, this study has collected WhatsApp at scale and provides valuable insights into the content within the platform, highlighting the harmful misinformation present there, investigates how the messages circulate and get viral in this network and explore an effective means to combat fake news in such peculiar environment.

Our results propose an effective methodology to observe and monitor events within WhatsApp public groups, working as a large-scale and long-term data collection for this social media. Since the data from WhatsApp have an ephemeral nature, in which a large volume of messages, users and groups appears and vanishes in short period of time, this gathering of events happening within WhatsApp public groups context may represent a crucial record of this platform. This collection methodology is able to register how this platform was used and what content circulate through the network even though it does not last longer in original sources of smartphone of the users nor even in WhatsApp server.

Furthermore, we show the potential virality of messages to disseminate to thousands of people around the network and, moreover, that some measures taken by WhatsApp to block the spread of misinformation (i.e. reducing the number of simultaneous forwards), even though help to slow down the spreading of messages, are not enough to avoid them to get viral. Also, the forwarding structure is an easy tool to circumvent and difficult to trace, as many duplicates of messages are being created and shared undetectable by WhatsApp.

Finally, we explore viable and effective means to tackle misinformation issue on WhatsApp. We propose an architecture that could be implemented to prevent users from sharing messages containing misinformation, whereas it still protects their privacy. We also build a real world online system, named WhatsApp Monitor, that monitors and daily updates the most popular messages, images, videos, audios shared on WhatsApp public groups we observed. That system proved to be useful in supporting researches and journalists to investigate general trending on WhatsApp and quickly understand how users are articulated with this platform.

However, as WhatsApp is a relatively new social media, there is plenty of space for further analysis and novel studies. A clear direction is to continue collecting messages from selected groups during other important events on WhatsApp as well as further expand our collection universe to another contexts and different groups (e.g. other countries or other topics besides politics). Moreover, we want to explore deeply the content nature of the data collected. We plan to investigate the specific contexts and events that occurred within the WhatsApp public groups we monitor, specially in terms of COVID-19 pandemic and 2022 Brazilian election, in which those events were widely abused by misinformation campaigns and introduced new challenges for the research on WhatsApp.

We also want to dive in analysis of the senders of the messages, investigating the massive distribution of messages on WhatsApp, seeking evidence of automation in the process of sending messages, unauthentic behavior patterns, and use of bots within the platform. Finally, by providing support to WhatsApp Monitor to keep working, we expect that it could still be used by journalists and researchers in effective combat of misinformation on WhatsApp.

References

ABDIN, L. Bots and fake news: the role of whatsapp in the 2018 brazilian presidential election. *Casey Robertson*, v. 41, n. 1, 2019.

ABILOV, A. et al. VoterFraud2020: a Multi-modal Dataset of Election Fraud Claims on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 15, n. 1, p. 901–912, May 2021.

ABU-SALMA, R. et al. The Security Blanket of the Chat World: An Analytic Evaluation and a User Study of Telegram. In: INTERNET SOCIETY. *Proceedings of the European Symposium on Usable Security*. [S.l.], 2017. (EuroUSEC '17).

AGARWAL, P. et al. Characterising user content on a multi-lingual social network. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 14, n. 1, p. 2–11, May 2020.

AKBARI, A.; GABDULHAKOV, R. Platform surveillance and resistance in iran and russia: The case of telegram. *Surveillance & Society*, v. 17, n. 1/2, p. 223–231, 2019.

ALIAPOULIOS, M. et al. An Early Look at the Parler Online Social Network. In: *ICWSM*. [S.l.: s.n.], 2021.

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, v. 31, n. 2, p. 211–36, May 2017.

ALLEN, A. *Hashing and Encryption: A Likely Pair*. 2013. *Innovative Routines International (IRI) Total Data Management*. [Online; Updated on 11-Jun-2017]. Available at: <https://www.iri.com/blog/data-protection/hashing-tables-encryption>.

ALLYN, B. *Telegram is the app of choice in the war in Ukraine despite experts' privacy concerns*. 2022. *Nevada Public Radio*. [Online; Updated on 14-March-2022]. Available at: <https://knpr.org/npr/2022-03/telegram-app-choice-war-ukraine-despite-experts-privacy-concerns>.

ALMEIDA, S. L. de et al. Whatsapp: a desordem da informação na eleição presidencial brasileira de 2018. In: *Anais do VI Simpósio Internacional LAVITS 2019*. Salvador, BA: [s.n.], 2019. (LAVITS19, VI).

ANGLANO, C.; CANONICO, M.; GUAZZONE, M. Forensic analysis of telegram messenger on android smartphones. *Digital Investigation*, Elsevier, v. 23, p. 31–49, 2017.

Anti-Defamation League. *Telegram: The Latest Safe Haven for White Supremacists*. 2019. *ADL Blog*. [Online. Posted on 02-Dec-2019]. Available at: <https://www.adl.org/blog/telegram-the-latest-safe-haven-for-white-supremacists>.

ARUN, C. On WhatsApp, Rumours, and Lynchings. *Economic & Political Weekly*, v. 54, n. 6, p. 30–35, 2019.

ASNAFI, A. R. et al. Using mobile-based social networks by iranian libraries: The case of telegram messenger. *Libr. Philos. Pract*, v. 2017, n. 1, 2017.

BADRINATHAN, S. Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review*, Cambridge University Press, p. 1–17, 2020.

BAKARE, A. S.; ABDURRAHAMAN, D. T.; OWUSU, A. Forwarding of Messages Via WhatsApp: The Mediating Role of Emotional Evocativeness. *Howard Journal of Communications*, Routledge, v. 33, n. 3, p. 265–280, 2022.

BANAJI, S. et al. Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india. Department of Media and Communications, London School of Economics and Political Science, 2019.

BARNES, . B. J. E.; ISAAC, M. *Barr Pushes Facebook for Access to WhatsApp Messages*. 2019. *The New York Times*. [Online; Posted on 03-October-2019]. Available at: <https://www.nytimes.com/2019/10/03/us/politics/barr-whatsapp-facebook-encryption.html>.

BAUMGARTNER, J. et al. The pushshift reddit dataset. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2020. v. 14, p. 830–839.

BAUMGARTNER, J. et al. The Pushshift Telegram Dataset. In: *ICWSM*. [S.l.: s.n.], 2020.

BENGALI, S. *How WhatsApp is battling misinformation in India, where 'fake news is part of our culture'*. 2019. *Los Angeles Times*. [Online; Posted on 04-February-2019]. Available at: <https://www.latimes.com/world/la-fg-india-whatsapp-2019-story.html>.

BERNSTEIN, M. et al. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. 2011.

BESSI, A.; FERRARA, E. Social bots distort the 2016 us presidential election online discussion. *First Monday*, v. 21, n. 11-7, 2016.

BHATTACHARJEE, S. D.; TALUKDER, A.; BALANTRAPU, B. V. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In: *Big Data*. [S.l.: s.n.], 2017.

BIANCHI, A. et al. Exploitation and mitigation of authentication schemes based on device-public information. In: *Proceedings of the 33rd Annual Computer Security Applications Conference*. NY, USA: ACM, 2017. (ACSAC '17), p. 16—27. ISBN 9781450353458.

BLABST, N.; DIEFENBACH, S. Whatsapp and wellbeing: A study on whatsapp usage, communication quality and stress. In: *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*. Swindon, GBR: BCS Learning & Development Ltd., 2017. (HCI '17).

BLANCO-HERRERO, D.; AMORES, J. J.; SáNCHEZ-HOLGADO, P. Citizen perceptions of fake news in spain: Socioeconomic, demographic, and ideological differences. *Publications*, v. 9, n. 3, 2021. ISSN 2304-6775.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003.

BLONDEL, V. D. et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2008, n. 10, p. P10008, 2008.

BONEVA, B. S. et al. Teenage communication in the instant messaging era. In: _____. [S.l.]: Oxford University Press, 2006. (Computers, Phones, and the Internet). ISBN 9780195312805.

BORGOLTE, K.; KRUEGEL, C.; VIGNA, G. Meerkat: Detecting website defacements through image-based object recognition. In: *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, 2015. p. 595–610. ISBN 978-1-939133-11-3.

BOWLES, J.; LARREGUY, H.; LIU, S. Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe. *PloS one*, Public Library of Science San Francisco, CA USA, v. 15, n. 10, p. e0240005, 2020.

BURDOVA, C. *What Is Jailbreaking and Is It Safe?* 2021. *Avast.* [Online; Updated on 23-May-2022]. Available at: <https://www.avast.com/c-jailbreaking>.

BURSZTEIN, E. et al. Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. In: *The World Wide Web Conference*. NY, USA: ACM, 2019. (WWW '19), p. 2601—2607. ISBN 9781450366748.

BURSZTYN, V. S.; BIRNBAUM, L. Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups. In: *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. [S.l.: s.n.], 2019. (ASONAM '19).

CABRAL, L. et al. FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages. In: INSTICC. *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS,*. [S.l.]: SciTePress, 2021. p. 63–74. ISBN 978-989-758-509-8.

CAETANO, J. A. et al. Characterizing attention cascades in whatsapp groups. In: *Proceedings of the 10th ACM Conference on Web Science*. [S.l.]: ACM, 2019. (WebSci'19), p. 27–36.

CAMERON, A. F.; WEBSTER, J. Unintended consequences of emerging communication technologies: Instant messaging in the workplace. *Computers in Human behavior*, Elsevier, v. 21, n. 1, p. 85–103, 2005.

CARDOSO, G. et al. Social media disinformation in the pre-electoral period in portugal. CIES e-Working Papers, Portugal, 2020.

CASSITA, D. *Agora é ilegal vender curtidas e seguidores em redes sociais*. 2019. *Tecmundo*. [Online; Posted on 09-Fev-2019]. Available at: <https://www.tecmundo.com.br/redes-sociais/138405-ilegal-vender-curtidas-seguidores-redes-sociais.htm>.

CECI, L. *Forecast of the number of Whatsapp users in Brazil from 2019 to 2028*. 2022. *Statista*. [Online; Updated on 01-Dec-2022. Accessed on 31-Dec-2022]. Available at: <https://www.statista.com/forecasts/1145210/whatsapp-users-in-brazil>.

CESARINO, L. Como vencer uma eleição sem sair de casa: a ascensão do populismo digital no Brasil. *Internet & Sociedade*, v. 1, n. 1, p. 91–120, Feb. 2020.

CHA, M. et al. Measuring user influence in twitter: The million follower fallacy. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 4, n. 1, p. 10–17, May 2010.

CHA, M. et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. [S.l.: s.n.], 2007. p. 1–14.

CHAGAS, V. Meu malvado favorito: os memes bolsonaristas de whatsapp e os acontecimentos políticos no brasil. *Estudos Históricos (Rio de Janeiro)*, SciELO Brasil, v. 34, p. 169–196, 2021.

CHAMOSO, P. et al. A hash based image matching algorithm for social networks. In: PRIETA, F. De la et al. (Ed.). *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017*. Cham: Springer International Publishing, 2018. p. 183–190. ISBN 978-3-319-61578-3.

CHANDRA, Y. U. et al. Higher education student behaviors in spreading fake news on social media: A case of LINE group. In: IEEE. *2017 International Conference on Information Management and Technology (ICIMTech)*. [S.l.], 2017. p. 54–59.

CHANDRASEKHARAN, E. et al. The bag of communities: Identifying abusive behavior online with preexisting internet data. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. [S.l.: s.n.], 2017. p. 3175–3187.

CHANG, A. Networks in a world unknown: Public whatsapp groups in the venezuelan refugee crisis. 2020.

CHEESEMAN, N. et al. Social media disruption: Nigeria's whatsapp politics. *Journal of Democracy*, Johns Hopkins University Press, v. 31, n. 3, p. 145–159, 2020.

CHEN, Y.; LIANG, C.; CAI, D. Understanding wechat users' behavior of sharing social crisis information. *International Journal of Human–Computer Interaction*, Taylor & Francis, v. 34, n. 4, p. 356–366, 2018.

CHURCH, K.; OLIVEIRA, R. de. What's up with whatsapp? comparing mobile instant messaging behaviors with traditional sms. In: *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*. NY, USA: ACM, 2013. (MobileHCI '13), p. 352–361. ISBN 9781450322737.

COCKERELL, I. *Inside China's Massive Surveillance Operation*. 2019. *Wired Backchannel*. [Online; Posted on 9-May-2019]. Available at: <https://www.wired.com/story/inside-chinas-massive-surveillance-operation/>.

CONDE, M. Á. et al. Whatsapp or telegram. which is the best instant messaging tool for the interaction in teamwork? In: ZAPHIRIS, P.; IOANNOU, A. (Ed.). *Learning and Collaboration Technologies: New Challenges and Learning Experiences*. Cham: Springer, 2021. p. 239–249.

CONROY, N. J.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. In: *ASIS&T*. [S.l.: s.n.], 2015.

CONSTINE, J. *WhatsApp has an encrypted child abuse problem*. 2018. *Techcrunch*. [Online; Posted on 02-December-2018]. Available at: <https://techcrunch.com/2018/12/20/whatsapp-pornography/>.

COX, J. *The Gaming Site Discord Is the New Front of Revenge Porn*. 2018. *The Daily Beast*. [Online; Posted on 17-Jan-2018]. Available at: <https://www.thedailybeast.com/the-gaming-site-discord-is-the-new-front-of-revenge-porn>.

DAVIDSON, T. et al. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 11, n. 1, p. 512–515, May 2017.

DELAM, H.; EIDI, A. WhatsApp Messenger role in Coronavirus Disease 2019 (COVID 19) Pandemic. *Journal of Health Sciences amp; Surveillance System*, Shiraz University Of Medical Sciences, v. 8, n. 4, p. 183–184, 2020. ISSN 2345-2218.

DIETRICH, C. J.; ROSSOW, C.; POHLMANN, N. Exploiting visual appearance to cluster and detect rogue software. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. NY, USA: ACM, 2013. (SAC '13), p. 1776–1783. ISBN 9781450316569.

DISCORD. *Discord API*. 2020. [Online. Accessed on 20-Nov-2022]. Available at: <https://discord.com/developers/docs/resources/guild>.

DISCORD. *Discord OAuth API*. 2020. [Online. Accessed on 20-Nov-2022]. Available at: <https://discord.com/developers/docs/topics/oauth2>.

DIXIT, P.; MAC, R. *How WhatsApp Destroyed A Village*. 2018. *BuzzFeed News*. [Online; Posted on 09-September-2018]. Available at: <https://www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india>.

DIXON, S. *Leading countries based on number of Twitter users as of January 2022*. 2022. *Statista*. [Online. Posted on 22-Nov-2022, Accessed on 10-Dec-2022]. Available at: <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>.

DOU, E. *Inside China's Massive Surveillance Operation*. 2017. *The Wall Street Journal*. [Online; Posted on 8-Dec-2017]. Available at: <https://www.wsj.com/articles/jailed-for-a-text-chinas-censors-are-spying-on-mobile-chat-groups-1512665007>.

ELLIS-PETERSEN, H. *Social media shut down in Sri Lanka in bid to stem misinformation*. 2019. *The Guardian*. [Online; Posted on 21-April-2019]. Available at: <https://www.theguardian.com/world/2019/apr/21/social-media-shut-down-in-sri-lanka-in-bid-to-stem-misinformation>.

ELíAS, C.; CATALAN-MATAMOROS, D. Coronavirus in spain: Fear of 'official' fake news boosts whatsapp and alternative sources. *Media and Communication*, v. 8, n. 2, p. 462–466, 2020. ISSN 2183-2439.

ESPINOZA, A. M. et al. Alice and bob, who the FOCI are they?: Analysis of end-to-end encryption in the LINE messaging application. In: *7th USENIX Workshop on Free and Open Communications on the Internet*. [S.l.]: USENIX Association, 2017. (FOCI'17).

EVANGELISTA, R.; BRUNO, F. Whatsapp and political instability in brazil: targeted messages and political radicalisation. *Internet Policy Review*, Alexander von Humboldt Institute for Internet and Society, Berlin, v. 8, n. 4, p. 1–23, 2019. ISSN 2197-6775.

FARID, H. Reining in online abuses. *Technology & Innovation*, v. 19, n. 3, p. 593–599, 2018. ISSN 1949-8241.

FAROOQ, G. Politics of fake news: how whatsapp became a potent propaganda tool in india. *Media Watch*, Centre for Academic Social Action, v. 9, n. 1, p. 106–117, 2017.

FERRARA, E. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday*, v. 22, n. 8, 2017.

FIGUEIREDO, F.; BENEVENUTO, F.; ALMEIDA, J. M. The tube over time: characterizing popularity growth of youtube videos. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. [S.l.: s.n.], 2011. p. 745–754.

Freitas, C. et al. Reverse Engineering Socialbot Infiltration Strategies in Twitter. In: *2015 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining*. [S.l.: s.n.], 2015. (ASONAM'15), p. 25–32.

FUNKE, D. *How misinformation spreads on Line – one of the most popular messaging apps in Southeast Asia*. 2018. *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; Posted on 21-Aug-2018].

GAGLANI, J. et al. Unsupervised whatsapp fake news detection using semantic search. In: *4th International Conference on Intelligent Computing and Control Systems*. [S.l.: s.n.], 2020. (ICICCS'20), p. 285–289.

GALLAGHER, K. *THE SOCIAL MEDIA DEMOGRAPHICS REPORT: Differences in age, gender, and income at the top platforms*. 2017. *Insider*. [Online; Posted on 07-Jul-2021]. Available at: <https://www.businessinsider.com/the-social-media-demographics-report-2017-8>.

GARIMELLA, K.; ECKLES, D. Images and misinformation in political groups: Evidence from whatsapp in india. *Harvard Kennedy School Misinformation Review*, 2020.

GARIMELLA, K. et al. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In: *WWW*. [S.l.: s.n.], 2018.

GARIMELLA, K.; TYSON, G. Whatapp doc? A first look at whatsapp public group data. In: *Twelfth International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2018. (ICWSM '18).

GILBERT, E. Widespread underprovision on reddit. In: *Proceedings of the 2013 conference on Computer supported cooperative work*. [S.l.: s.n.], 2013. p. 803–808.

GIMENEZ, M.; ZIRPOLI, R. Trastornos psicológicos vinculados al uso de whatsapp. In: FACULTAD DE PSICOLOGÍA-UNIVERSIDAD DE BUENOS AIRES. *VII Congreso Internacional de Investigación y Práctica Profesional en Psicología XXII Jornadas de Investigación XI Encuentro de Investigadores en Psicología del MERCOSUR*. [S.l.], 2015.

GOETHEM, T. V. et al. Purchased fame: Exploring the ecosystem of private blog networks. In: *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. NY, USA: ACM, 2019. (Asia CCS '19), p. 366—378. ISBN 9781450367523.

GOGA, O.; VENKATADRI, G.; GUMMADI, K. P. The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks. In: *Proceedings of the 2015 Internet Measurement Conference*. NY, USA: ACM, 2015. (IMC '15), p. 141—153. ISBN 9781450338486.

GOMES, R. C. L. F.; NAKAGAWA, R. M. de O.; CARDOSO, T. de S. Epistemologias mutiladas e a exploração política de vieses cognitivos: o negacionismo engendrado pela retórica bolsonarista. *Revista Mídia e Cotidiano*, v. 14, n. 3, p. 31–52, 2020.

GUO, L.; ZHANG, Y. 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*, Taylor & Francis, v. 21, n. 15, p. 2176–2195, 2020.

GUPTA, H.; TANEJA, H. *WhatsApp has a fake news problem – that can be fixed without breaking encryption*. 2018. *CJR – Columbia Journalism Review*. [Online; Posted on 23-Aug-2018]. Available at: <https://www.cjr.org/tow_center/whatsapp-doesnt-have-to-break-encryption-to-beat-fake-news.php>.

HAMRICK, J. et al. An examination of the cryptocurrency pump and dump ecosystem. *Available at SSRN 3303365*, 2018.

HAO, Q. et al. It's Not What It Looks Like: Manipulating Perceptual Hashing Based Applications. In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. NY, USA: ACM, 2021. (CCS '21), p. 69—85. ISBN 9781450384544.

HARVEY, C. G.; STEWART, D. B.; EWING, M. T. Forward or delete: What drives peer-to-peer message propagation across social networks? *Journal of Consumer Behaviour*, Wiley Online Library, v. 10, n. 6, p. 365–372, 2011.

HASE, M. *WhatsApp Wrapper*. 2018. *GitHub*. [Online. Accessed on 20-Nov-2022]. Available at: <https://github.com/mukulhase/WebWhatsapp-Wrapper>.

HASHEMI, A.; CHAHOOKI, M. A. Z. Telegram group quality measurement by user behavior analysis. *Social Network Analysis and Mining*, Springer, v. 9, n. 1, p. 33, 2019.

HERRERO-DIZ, P.; CONDE-JIMéNEZ, J.; CóZAR, S. R. de. Teens' motivations to spread fake news on whatsapp. *Social Media + Society*, v. 6, n. 3, p. 2056305120942879, 2020.

HINE, G. E. et al. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *arXiv preprint arXiv:1610.03452*, 2016.

HITCHEN, J. et al. Whatsapp and nigeria's 2019 elections: Mobilising the people, protecting the vote. *Portal Africa*, Centre for Democracy and Development (CDD), Jul 2019. Available at: <https://www.africaportal.org/publications/whatsapp-and-nigerias-2019-elections-mobilising-people-protecting-vote/>.

HOSEINI, M. et al. Demystifying the Messaging Platforms' Ecosystem Through the Lens of Twitter. In: *Proceedings of the 2020 Conference on Internet Measurement Conference*. [S.l.: s.n.], 2020.

HOU, A. et al. The effects of push-pull-mooring on the switching model for social network sites migration. In: *Proceeding of the 19th Pacific Asia Conference on Information Systems*. [S.l.: s.n.], 2014. (PACIS 2014). ISBN 978-988-8353-22-4.

HOWARD, P. N.; HUSSAIN, M. M. The Upheavals in Egypt and Tunisia: The Role of Digital Media. *Journal of Democracy*, Johns Hopkins University Press, v. 22, n. 3, p. 35–48, 2011.

HUCKLE, S.; WHITE, M. Fake News: A Technological Approach to Proving the Origins of Content, Using Blockchains. *Big Data*, v. 5, n. 4, p. 356–371, 2017.

HUNT, E. *'Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news*. 2017. *The Guardian*. [Online; Posted on 22-March-2017]. Available at: <https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news>.

ISAAC, M.; ROOSE, K. *Disinformation Spreads on WhatsApp Ahead of Brazilian Election*. 2018. *The New York Times*. [Online; Posted on 19-October-2018]. Available at: <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html>.

JACCARD, P. The distribution of the flora in the alpine zone. *New Phytologist*, v. 11, n. 2, p. 37–50, 1912.

JAIN, S. *Indians most active on WhatsApp with 390.1 million monthly active users in 2020*. 2021. *Forbes India*. [Online; Posted on 27-August-2021]. Available at: <https://www.forbesindia.com/article/news-by-numbers/indians-most-active-on-whatsapp-with-3901-million-monthly-active-users-in-2020/70059/1>.

JAO, N. *WeChat now has over 1 billion active monthly users worldwide*. 2018. *Technode*. [Online; Posted on 5-Mar-2018]. Available at: <https://technode.com/2018/03/05/wechat-1-billion-users/>.

JAVED, R. T. et al. A First Look at COVID-19 Messages on WhatsApp in Pakistan. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.: s.n.], 2020. p. 118–125.

JIANG, J. A. et al. Moderation Challenges in Voice-based Online Communities on Discord. *Proceedings of the ACM on Human-Computer Interaction*, ACM NY, USA, v. 3, n. CSCW, p. 1–23, 2019.

JOHNSON, N. et al. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, Nature Publishing Group, v. 573, n. 7773, p. 261–265, 2019.

JúNIOR, M. et al. Telegram Monitor: Monitoring Brazilian Political Groups and Channels on Telegram. In: *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*. NY, USA: ACM, 2022. (HT '22), p. 228—231. ISBN 9781450392334.

JúNIOR, M. et al. Towards Understanding the Use of Telegram by Political Groups in Brazil. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. NY, USA: ACM, 2021. (WebMedia '21), p. 237—244. ISBN 9781450386098.

KAFRUNI, S. *Bloqueio do WhatsApp deixa rastro de prejuizos pelo pais*. 2016. *Correio Braziliense*. [Online; Posted on 16-May-2016]. Available at: <https://www.correiobraziliense.com.br/app/noticia/economia/2016/05/04/internas_economia,530305/bloqueio-do-whatsapp-deixa-rastro-de-prejuizos-pelo-pais.shtml>.

KALOGEROPOULOS, A. *Groups and Private Networks –- Time Well Spent?* 2019. *Reuters Institute*. [Online. Accessed on 21-Aug-2021]. Available at: <https://www.digitalnewsreport.org/survey/2019/groups-and-private-networks-time-well-spent>.

KAZEMI, A. et al. Tiplines to combat misinformation on encrypted platforms: A case study of the 2019 indian election on whatsapp. *arXiv preprint arXiv:2106.04726*, 2021.

KHARRAZ, A.; ROBERTSON, W.; KIRDA, E. Surveylance: Automatically detecting online survey scams. In: *2018 IEEE Symposium on Security and Privacy (SP)*. [S.l.: s.n.], 2018. p. 70–86.

KHURANA, P.; KUMAR, D. Sir Model for Fake News Spreading Through Whatsapp. In: *Proc. of 3rd Int'l Conf. on Internet of Things and Connected Technologies*. India: [s.n.], 2018. (ICIoTCT).

KIENE, C.; HILL, B. M. Who uses bots? a statistical analysis of bot usage in moderation teams. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts*. [S.l.: s.n.], 2020. p. 1–8.

KULSOLKOOKIET, R. et al. Using social media to change smoking behavior: Line instant messaging application perspectives. *The Journal of Applied Business and Economics*, North American Business Press, v. 20, n. 1, p. 106–119, 2018.

KWAK, H. et al. What is twitter, a social network or a news media? In: ACM. *Proc. of the 19th Int'l Conf. on World Wide Web (WWW'10)*. [S.l.], 2010.

KWANDA, F. A.; LIN, T. T. C. Fake news practices in indonesian newsrooms during and after the palu earthquake: a hierarchy-of-influences approach. *Information, Communication & Society*, Routledge, v. 23, n. 6, p. 849–866, 2020.

LACHER, L.; BIEHL, C. Using discord to understand and moderate collaboration and teamwork. In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. [S.l.: s.n.], 2018. p. 1107–1107.

LAMBTON-HOWARD, D. et al. Whatfutures: Designing large-scale engagements on whatsapp. In: *Proceedings of the International Conference on Human Factors in Computing Systems*. New York, USA: ACM, 2019. (CHI'19), p. 1–14. ISBN 9781450359702.

LAZER, D. M. J. et al. The science of fake news. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018.

LEE, J. Y. et al. Smiley face: Why we use emoticon stickers in mobile messaging. In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. NY, USA: ACM, 2016. (MobileHCI '16), p. 760–766. ISBN 9781450344135.

Legal Information Institute. *17 U.S. Code §1201 - Circumvention of copyright protection systems*. 2021. *Cornell Law School*. [Online; Accessed on 18-Jun-2021]. Available at: <https://www.law.cornell.edu/uscode/text/17/1201#a_3>.

LESKOVEC, J.; KLEINBERG, J.; FALOUTSOS, C. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In: *Proc. of the 11th SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining*. [S.l.: s.n.], 2005. p. 177–187.

LEVIN, S. *Facebook promised to tackle fake news. But the evidence shows it's not working*. 2017. *The Guardian*. [Online; Posted on 16-May-2017]. Available at: <https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working>.

LI, G.; ZHEN, J. Global stability of an SEI epidemic model with general contact rate. *Chaos, Solitons & Fractals*, v. 23, n. 3, p. 997 – 1004, 2005.

LI, T. *Fake news proves to be a huge problem for Tencent as it blocks 1,000 articles each minute*. 2017. *g0vnews*. [Online; Posted on 21-Dec-2017]. Available at: <https://www.scmp.com/tech/china-tech/article/2125200/fake-news-proves-be-huge-problem-tencent-it-blocks-1000-articles>.

LIMA, L. et al. Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. [S.l.: s.n.], 2018. (ASONAM'18), p. 515–522.

LIU, Y. et al. Analyzing facebook privacy settings: user expectations vs. reality. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. [S.l.: s.n.], 2011. p. 61–70.

LORENZ-SPREEN, P. et al. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, Nature Research, v. 2020, jun. 2020. ISSN 2397-3374.

LU, Z. et al. The Government's Dividend: Complex Perceptions of Social Media Misinformation in China. In: ______. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. NY, USA: ACM, 2020. (CHI'20), p. 1–12. ISBN 9781450367080.

MACHADO, C. et al. A Study of Misinformation in WhatsApp Groups with a Focus on the Brazilian Presidential Elections. In: *Companion Proceedings of The 2019 World Wide Web Conference*. [S.l.]: ACM, 2019. (WWW '19), p. 1013–1019.

MAENG, W. et al. Can mobile instant messaging be a useful interviewing tool? a comparative analysis of phone use, instant messaging, and mobile instant messaging. In: *Proceedings of HCI Korea*. Seoul, KOR: Hanbit Media, Inc., 2016. (HCIK '16), p. 45–49. ISBN 9788968487910.

MAGENTA, J. G. M.; SOUZA, F. *How WhatsApp is being abused in Brazil's elections*. 2018. *BBC News*. [Online; Posted on 24-October-2018]. Available at: <https://www.bbc.com/news/technology-45956557>.

MAGENTA, M.; GRAGNANI, J.; SOUZA, F. *How WhatsApp is being abused in Brazil's elections*. 2018. *BBC News*. [Online; Posted on 24-Oct-2018]. Available at: <https://www.bbc.com/news/technology-45956557>.

MAHADEOKAR, J.; PESAVENTO, G. *Open Sourcing a Deep Learning Solution for Detecting NSFW Images*. 2016. *yahoo! Engineering*. [Online; Posted on 30-September-2016]. Available at: <https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for>.

MALKA, V.; ARIEL, Y.; AVIDAR, R. Fighting, worrying and sharing: Operation 'Protective Edge' as the first WhatsApp war. *Media, War & Conflict*, SAGE, v. 8, n. 3, p. 329–344, 2015.

MAROS, A. et al. Analyzing the use of audio messages in whatsapp groups. In: *Proceedings of The Web Conference 2020*. Taipei, Taiwan: ACM, 2020. (WWW'20), p. 20–24. Available at: <https://doi.org/10.1145/3366423.3380070>.

MAYER, J. *Content Moderation for End-to-End Encrypted Messaging*. 2019. *Princeton University*. [Open Letter. Online; Posted on 06-Oct-2019]. Available at: <https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf>.

MCAULEY, J.; LESKOVEC, J. Image Labeling on a Network: Using Social-Network Metadata for Image Classification. In: *12th European Conf. on Computer Vision (ECCV12)*. [S.l.: s.n.], 2012.

MCDONELL, S. *China social media: WeChat and the Surveillance State*. 2019. *BBC News*. [Online; Posted on 7-Jun-2019]. Available at: <https://www.bbc.com/news/blogs-china-blog-48552907>.

MEDVET, E.; KIRDA, E.; KRUEGEL, C. Visual-similarity-based phishing detection. In: *Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks*. NY, USA: ACM, 2008. (SecureComm '08). ISBN 9781605582412.

MELO, P. et al. WhatsApp Monitor: A Fact-Checking System for WhatsApp. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2019. (ICWSM '19, v. 13), p. 676–677.

MELO, P. et al. Can whatsapp counter misinformation by limiting message forwarding? In: *International Conference on Complex Networks and Their Applications*. [S.l.]: Springer, 2019. p. 372–384.

MILLWARD, S. *It's time for messaging apps to quit the bullshit numbers and tell us how many users are active*. 2014. *Tech In Asia*. [Online; Posted on 23-Jan-2014]. Available at: <https://www.techinasia.com/messaging-apps-should-reveal-monthly-active-users>.

MILLWARD, S. *Line just lost even more users. But that's apparently fine*. 2017. *Tech in Asia*. [Online; Posted on 16-May-2017]. Available at: <https://www.techinasia.com/line-loses-users-again>.

MISLOVE, A. et al. Growth of the flickr social network. In: *Proceedings of the first workshop on Online social networks*. [S.l.: s.n.], 2008. p. 25–30.

MISLOVE, A. et al. Understanding the demographics of twitter users. In: *ICWSM*. [S.l.: s.n.], 2011.

MISLOVE, A. et al. Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. [S.l.: s.n.], 2007. p. 29–42.

MITTOS, A. et al. "and we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 14, n. 1, p. 452–463, May 2020.

MONDAL, M.; SILVA, L. A.; BENEVENUTO, F. A measurement study of hate speech in social media. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. NY, USA: ACM, 2017. (HT '17), p. 85–94. ISBN 9781450347082.

MONTAG, C. et al. Smartphone usage in the 21st century: who is active on whatsapp? *BMC Research Notes*, BioMed Central, v. 8, n. 1, p. 331, Aug 2015. ISSN 1756-0500.

MORENO, A.; GARRISON, P.; BHAT, K. Whatsapp for monitoring and response during critical events: Aggie in the ghana 2016 election. In: *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*. [S.l.: s.n.], 2017. (ISCRAM'17), p. 645–655.

MüZELL, R. B. *Desinformação e Propagabilidade: uma Análise da Desordem Informacional em Grupos de Whatsapp*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Rio Grande do Sul, 2020.

NAPOLI, P.; CAPLAN, R. Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, v. 22, n. 5, May 2017. Available at: <https://journals.uic.edu/ojs/index.php/fm/article/view/7051>.

NAPPA, A.; RAFIQUE, M. Z.; CABALLERO, J. Driving in the cloud: An analysis of drive-by download operations and abuse reporting. In: RIECK, K.; STEWIN, P.; SEIFERT, J.-P. (Ed.). *Detection of Intrusions and Malware, and Vulnerability*

Assessment. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 1–20. ISBN 978-3-642-39235-1.

NASERI, M.; ZAMANI, H. Analyzing and Predicting News Popularity in an Instant Messaging Service. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. [S.l.: s.n.], 2019. p. 1053–1056.

NATH, G.; ADHI, G. An attempt to detect fake messages circulated on whatsapp. In: *Proceedings of 7th International Conference of Business Analytics and Intelligence*. [S.l.: s.n.], 2019.

NELLY. *How to use Discord for your classroom*. 2020. *Discord Blog*. [Online; Posted on 27-Mar-2020]. Available at: <https://blog.discord.com/how-to-use-discord-for-your-classroom-8587bf78e6c4>.

NEMER, D. *The three types of WhatsApp users getting Brazil's Jair Bolsonaro elected*. 2018. *The Guardian*. [Online. Posted on 25-Oct-2018]. Available at: <https://www.theguardian.com/world/2018/oct/25/brazil-president-jair-bolsonaro-whatsapp-fake-news>.

NEWMAN, N. *Executive Summary and Key Findings of the 2020 Report*. 2020. *Reuters Institute*. [Online; Accessed on 21-Aug-2021]. Available at: <https://www.digitalnewsreport.org/survey/2020/overview-key-findings-2020/>.

NEWMAN, N. et al. *Reuters Institute Digital News Report 2019*. 2019. Reuters Institute for the Study of Journalism.

NEWMAN, N. et al. *Reuters Institute Digital News Report 2021*. 2021. Reuters Institute for the Study of Journalism.

NEWMAN, N. et al. *Reuters Institute Digital News Report 2020*. 2020. Reuters Institute for the Study of Journalism.

NIKIFORAKIS, N. et al. Stranger Danger: Exploring the Ecosystem of Ad-Based URL Shortening Services. In: *Proceedings of the 23rd International Conference on World Wide Web*. NY, USA: ACM, 2014. (WWW '14), p. 51—62. ISBN 9781450327442.

NIKKAH, S.; MILLER, A. D.; YOUNG, A. L. Telegram as An Immigration Management Tool. ACM, 2018.

NOBARI, A. D.; RESHADATMAND, N.; NESHATI, M. Analysis of Telegram, An Instant Messaging Service. In: ACM. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. [S.l.], 2017. p. 2035–2038.

NYHAN, B.; REIFLER, J. When corrections fail: The persistence of political misperceptions. *Political Behavior*, Springer, v. 32, n. 2, p. 303–330, 2010.

OFUSORI, L. O.; KARIUKI, P. The power of whatsapp as a communication tool for elections observation and monitoring in kwazulu-natal: Ngo case study in south africa. In: *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance*. NY, USA: ACM, 2017. (ICEGOV '17), p. 536–537. ISBN 9781450348256.

OGHUMA, A. P. et al. An expectation-confirmation model of continuance intention to use mobile instant messaging. *Telematics and Informatics*, v. 33, n. 1, p. 34–47, 2016. ISSN 0736-5853.

O'HARA, K. P. et al. Everyday dwelling with whatsapp. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. NY, USA: ACM, 2014. (CSCW '14), p. 1131—1143. ISBN 9781450325400.

OHASHI, K.; KATO, F.; HJORTH, L. Digital genealogies: Understanding social mobile media LINE in the role of Japanese families. *Social Media+ Society*, SAGE, v. 3, n. 2, p. 2056305117703815, 2017.

OLIVEIRA, G.; CASALECCHI, G. .; BACHINI, N. Informação, voto e whatsapp na eleição presidencial brasileira de 2018. In: *Anais do 44º Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais*. Online: [s.n.], 2020. (ANPOCS).

OSENI, K.; DINGLEY, K.; HART, P. Instant Messaging and Social Networks – The Advantages in Online Research Methodology. *International Journal of Information and Education Technology*, International Journal of Information and Education Technology, v. 8, n. 1, p. 56–62, 2018.

PARIS, B.; DONOVAN, J. Deepfakes and cheap fakes: the manipulation of audio and visual evidence. *Data and Society*, 2019.

PASTRANA, S. et al. Measuring ewhoring. In: *Proceedings of the Internet Measurement Conference*. NY, USA: ACM, 2019. (IMC '19), p. 463—477. ISBN 9781450369480.

PATEL, P. et al. *Open Letter: Facebook's "Privacy First" Proposals*. 2019. *U.S. Department of Justice*. [Open Letter. Online; Posted on 04-Oct-2019]. Available at: <https://www.justice.gov/opa/press-release/file/1207081>.

PEREZ, S. *WhatsApp officially launches its app for businesses in select markets*. 2018. *Tech Crunch*. [Online; Posted on 18-Jan-2018]. Available at: <https://techcrunch.com/2018/01/18/whatsapp-officially-launches-its-app-for-businesses-in-select-markets/>.

PHARTIYAL, S. *India government meets with WhatsApp over tracing of fake news: source*. 2018. *Reuters*. [Online; Posted on 07-December-2018]. Available at: <https://www.reuters.com/article/us-india-whatsapp-government/india-government-meets-with-whatsapp-over-tracing-of-fake-news-source-idUSKBN1O60GO>.

PONNIAH, K. *WhatsApp: The 'black hole' of fake news in India's election*. 2019. *BBC News*. [Online; Posted on 06-April-2019]. Available at: <https://www.bbc.com/news/world-asia-india-47797151>.

PRUCHA, N. Is and the jihadist information highway–projecting influence and religious identity via telegram. *Perspectives on Terrorism*, v. 10, n. 6, 2016.

QIU, X. et al. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, Nature Publishing Group, v. 1, n. 7, p. 0132, 6 2017.

RAFIQUE, M. Z. et al. It's free for a reason: Exploring the ecosystem of free live streaming services. In: INTERNET SOCIETY. *Proceedings of the 23rd Network and Distributed System Security Symposium*. [S.l.], 2016. (NDSS'2016), p. 1–15.

RAJ, A. *India's Sandes takes on WhatsApp, citing security and misinformation concerns*. 2021. *TWA – Tech Wise Asia*. [Online. Posted on 03-Aug-2021]. Available at: <https://techwireasia.com/2021/08/indias-sandes-takes-on-whatsapp-citing-security-and-misinformation-concerns/>.

RAJASEGARAN, J. et al. A multi-modal neural embeddings approach for detecting mobile counterfeit apps. In: *The World Wide Web Conference*. NY, USA: ACM, 2019. (WWW '19), p. 3165–3171. ISBN 9781450366748.

RAMAN, A. et al. Challenges in the decentralised web: The mastodon case. In: *Proceedings of the Internet Measurement Conference*. [S.l.: s.n.], 2019. p. 217–229.

RASTOGI, N.; HENDLER, J. WhatsApp Security and Role of Metadata in Preserving Privacy. *Int'l Conference on Cyber Warfare and Security*, p. 269–XVI, 2017. Available at: <https://search.proquest.com/docview/1897684003?accountid=134127>.

RATHI, K. et al. Forensic analysis of encrypted instant messaging applications on android. In: *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*. [S.l.: s.n.], 2018. p. 1–6.

RECUERO, R.; SOARES, F.; VINHAS, O. Discursive strategies for disinformation on whatsapp and twitter during the 2018 brazilian presidential election. *First Monday*, 2021.

REIS, J. C. et al. Explainable machine learning for fake news detection. In: *WebScience*. [S.l.: s.n.], 2019.

REIS, J. C. et al. Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School (HKS) Misinformation Review*, 2020.

REIS, J. C. S. et al. Supervised learning for fake news detection. *IEEE Intelligent Systems*, IEEE, v. 34, n. 2, 2019.

REIS, J. C. S. et al. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and India Elections. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 14, n. 1, p. 903–908, May 2020.

REPNIKOVA, M. *China's Lessons for Fighting Fake News*. 2018. *Foreign Policy*. [Online; Posted on 6-September-2018]. Available at: <https://foreignpolicy.com/2018/09/06/chinas-lessons-for-fighting-fake-news/>.

RESENDE, G. et al. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In: *Proceedings of the 10th ACM Conference on Web Science*. [S.l.]: ACM, 2019. (WebSci '19), p. 225–234.

RESENDE, G. et al. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In: *The World Wide Web Conference*. [S.l.]: ACM, 2019. (WWW '19), p. 818–828.

RIBEIRO, F. N.; BENEVENUTO, F.; ZAGHENI, E. How biased is the population of facebook users? comparing the demographics of facebook users with census data to generate correction factors. In: *12th ACM Conference on Web Science*. NY, USA: ACM, 2020. (WebSci '20), p. 325—334. ISBN 9781450379892.

RIBEIRO, F. N. et al. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. NY, USA: ACM, 2019. (FAT* '19), p. 140—149. ISBN 9781450361255.

RIBEIRO, M. H. et al. Everything I Disagree With is#FakeNews": Correlating Political Polarization and Spread of Misinformation. In: *DSJM@KDD*. [S.l.: s.n.], 2017.

RIBEIRO, M. H. et al. Does platform migration compromise content moderation? evidence from r/the_donald and r/incels. In: *CSCW*. [S.l.: s.n.], 2021.

RICARD, J.; MEDEIROS, J. Using misinformation as a political weapon: Covid-19 and bolsonaro in brazil. *Harvard Kennedy School (HKS) Misinformation Review*, 2020.

RIVAS, A. et al. Image matching algorithm based on hashes extraction. In: *Progress in Artificial Intelligence: 18th EPIA Conference on Artificial Intelligence, EPIA 2017, Porto, Portugal, September 5-8, 2017, Proceedings*. Berlin, Heidelberg: Springer, 2017. p. 87—94. ISBN 978-3-319-65339-6.

RIVEST, R. L. *The MD5 Message-Digest Algorithm*. RFC Editor, 1992. *RFC – Request for Comments*. (Request for Comments, 1321). Doi: 10.17487/RFC1321. [RFC 1321. Online. Posted on 01-Apr-1992]. Available at: <https://www.rfc-editor.org/info/rfc1321>.

ROOSE, K. *This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville.* 2017. *New York Times*. [Online; Posted on 15-Aug-2017]. Available at: <https://www.nytimes.com/2017/08/15/technology/discord-chat-app-alt-right.html>.

ROSENFELD, A. et al. Whatsapp usage patterns and prediction models. In: *ICWSM/IUSSP Workshop on Social Media and Demographic Research*. [S.l.: s.n.], 2016.

RUSSELL, J. *Stickers: From Japanese craze to global mobile messaging phenomenon.* 2013. *TNW – The Next Web*. [Online; Posted on 12-Jul-2013]. Available at: <https://thenextweb.com/news/stickers>.

SALGANIK, M. J.; DODDS, P. S.; WATTS, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, American Association for the Advancement of Science, v. 311, n. 5762, p. 854–856, 2006.

SAM. *Invites 101*. 2020. *Discord Support*. [Online. Updated on 31-May-2022]. Available at: <https://support.discord.com/hc/en-us/articles/208866998-Invites-101>.

SAMANTA, P.; JAIN, S. Analysis of Perceptual Hashing Algorithms in Image Manipulation Detection. *Procedia Computer Science*, v. 185, p. 203–212, 2021. ISSN 1877-0509. Big Data, IoT, and AI for a Smarter Future.

SATRYA, G. B.; DAELY, P. T.; NUGROHO, M. A. Digital forensic analysis of telegram messenger on android devices. In: IEEE. *2016 International Conference on Information & Communication Technology and Systems (ICTS)*. [S.l.], 2016. p. 1–7.

SCHRITTWIESER, S. et al. Guess who's texting you? evaluating the security of smartphone messaging applications. In: *Proc. of the 19th Annual Symposium on Network and Distributed System Security*. [S.l.: s.n.], 2012.

SEUFERT, M. et al. Group-based Communication in WhatsApp. In: *IFIP Networking Conf. and Workshops*. [S.l.: s.n.], 2016. (NETWORKING).

SEUFERT, M. et al. Analysis of Group-Based Communication in WhatsApp. In: *Mobile Networks and Management*. [S.l.]: Springer, 2015. p. 225–238.

SHEHABAT, A.; MITEW, T.; ALZOUBI, Y. Encrypted jihad: Investigating the role of telegram app in lone wolf attacks in the west. *Journal of Strategic Security*, JSTOR, v. 10, n. 3, p. 27–53, 2017.

SHI, S.; WANG, X.; LAU, W. C. Mossot: An automated blackbox tester for single sign-on vulnerabilities in mobile applications. In: *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. NY, USA: ACM, 2019. (Asia CCS '19), p. 269—282. ISBN 9781450367523.

SHU, K. et al. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, ACM, NY, USA, v. 19, n. 1, p. 22–36, sep. 2017. ISSN 1931-0145.

SILVA, G. G. da. Memes war:: The political use of pictures in brazil 2019. *Philósophos - Revista de Filosofia*, v. 25, n. 2, abr. 2021. Available at: <https://www.revistas.ufg.br/philosophos/article/view/64490>.

SILVA, L. et al. Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 10, n. 1, p. 687–690, 2016.

SIMON, T. et al. Kidnapping whatsapp–rumors during the search and rescue operation of three kidnapped youth. *Computers in Human Behavior*, Elsevier, v. 64, p. 183–190, 2016.

SIMONS, H. *WhatsApp bans 2 million accounts each month: Here's how they do it.* 2019. *Android Authority.* [Online; Updated on 7-February-2019]. Available at: <https://www.androidauthority.com/whatsapp-ban-accounts-951307/>.

SINGER, P. et al. Evolution of reddit: from the front page of the internet to a self-referential community? In: *Proceedings of the 23rd international conference on world wide web*. [S.l.: s.n.], 2014. p. 517–522.

SOARES, F. B. et al. Desinformação sobre o COVID-19 no WhatsApp: a pandemia enquadrada como debate político. *Ciência da Informação em Revista*, v. 8, n. 1, p. 74–94, 2021.

SOARES, F. B. et al. Research note: Bolsonaro's firehose: How covid-19 disinformation on whatsapp was used to fight a government political crisis in brazil. *Harvard Kennedy School (HKS) Misinformation Review*, 2021.

SONG, H. et al. Fast Hash Table Lookup Using Extended Bloom Filter: An Aid to Network Processing. In: *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. NY, USA: ACM, 2005. (SIGCOMM '05), p. 181—192. ISBN 1595930094.

SPEICHER, T. et al. Potential for Discrimination in Online Targeted Advertising. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. NY, USA: PMLR, 2018. (FAT'18, v. 81), p. 5–19.

STARBIRD, K. et al. Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains. In: *Proc. of 12th Int'l Conf. on Web and Social Media (ICWSM'18)*. [S.l.: s.n.], 2018.

STRUPPEK, L. et al. Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. NY, USA: ACM, 2022. (FAccT '22), p. 58—69. ISBN 9781450393522.

SUN, Y. et al. Understanding users' switching behavior of mobile instant messaging applications: An empirical study from the perspective of push-pull-mooring framework. *Computers in Human Behavior*, v. 75, p. 727–738, 2017. ISSN 0747-5632.

Swaminathan, A.; Yinian Mao; Min Wu. Robust and secure image hashing. *IEEE Transactions on Information Forensics and Security*, v. 1, n. 2, p. 215–230, 2006.

TAN, C.; LEE, L. All who wander: On the prevalence and characteristics of multi-community engagement. In: *Proceedings of the 24th International World Wide Web Conference*. [S.l.: s.n.], 2015. (WWW'15).

TAN, R. *Terrorists' love for Telegram, explained*. 2017. *Vox*. [Online; Posted on 30-Jun-2017]. Available at: <https://www.vox.com/world/2017/6/30/15886506/terrorism-isis-telegram-social-media-russia-pavel-durov-twitter>.

TANKOVSKA, H. *Most popular global mobile messenger apps as of April 2021, based on number of monthly active users*. 2021. *Statista*. [Online. Posted on 02-Aug-2021]. Available at: <https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/>.

TARAFDAR, A. et al. Spam detection using threshold method on whatsapp image data. In: BANERJEE, S.; MANDAL, J. K. (Ed.). *Advances in Smart Communication Technology and Information Processing*. Singapore: Springer Singapore, 2021. p. 317–325. ISBN 978-981-15-9433-5.

TARDAGUILA, C.; BENEVENUTO, F.; ORTELLADO, P. *Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It*. 2018. *New York Times*. [Online. Posted on 17-Oct-2018]. Available at: <https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html>.

TELEGRAM. *Telegram API*. 2020. [Online. Accessed on 20-Nov-2022]. Available at: <https://core.telegram.org/method/channels.joinChannel>.

THIVYA.G, S. Survey on vigilance of instant messages in social networks using text mining techniques and ontology. *International Journal of Innovative Research in Computer and Communication Engineering*, v. 03, n. 19, p. 734–739, 2015.

TOMáS, R.; TOMáS, L.; ANDREATTA, E. Da Depravação ao Desperdício de Recursos: Estratégias de Desconstrução da Universidade Pública em Redes de Fake News. *VERBUM. Cadernos de Pós-Graduação*, v. 9, n. 2, p. 141–167, 2020. ISSN 2316-3267.

TRAMèR, F. et al. Adversarial: Perceptual ad blocking meets adversarial machine learning. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. NY, USA: ACM, 2019. (CCS '19), p. 2005—2021. ISBN 9781450367479.

TRUJILLO, M. et al. What is bitchute? characterizing the. In: *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. NY, USA: ACM, 2020. (HT '20), p. 139—140. ISBN 9781450370981.

TWITTER. *Twitter's Search API*. 2020. *Twitter Developers*. [Online. Accessed on 20-Nov-2022]. Available at: <https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets>.

TWITTER. *Twitter's Streaming API*. 2020. *Twitter Developers*. [Online. Accessed on 20-Nov-2022]. Available at: <https://developer.twitter.com/en/docs/tweets/filter-realtime/overview>.

VADREVU, P.; PERDISCI, R. What you see is not what you get: Discovering and tracking social engineering attack campaigns. In: *Proceedings of the Internet Measurement Conference*. NY, USA: ACM, 2019. (IMC '19), p. 308—321. ISBN 9781450369480.

VALERIANI, A.; VACCARI, C. Political talk on mobile instant messaging services: a comparative analysis of Germany, Italy, and the UK. *Information, Communication & Society*, Routledge, v. 21, n. 11, p. 1715–1731, 2018.

VASCONCELOS, M. et al. Analyzing YouTube Videos Shared on Whatsapp in the Early COVID-19 Crisis. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. NY, USA: ACM, 2020. (WebMedia '20), p. 25—28. ISBN 9781450381963.

VIEIRA, C. et al. The Paradox of Encrypted Information Virality on WhatsApp. In: *Proceedings of the XXXVII Brazilian Symposium on Computer Networks and Distributed Systems*. Gramado, Brazil: SBC, 2019. (SBRC'19), p. 403–416.

VIJAYKUMAR, S. et al. 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*, Springer Nature, v. 8, n. 1, p. 1–12, 2021.

VISWANATH, B. et al. On the evolution of user interaction in facebook. In: *Proceedings of the 2nd ACM workshop on Online social networks*. [S.l.: s.n.], 2009. p. 37–42.

VOIDA, A.; NEWSTETTER, W. C.; MYNATT, E. D. When conventions collide: The tensions of instant messaging attributed. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. NY, USA: ACM, 2002. (CHI '02), p. 187–194. ISBN 1581134533.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1146–1151, 2018. ISSN 0036-8075.

WANG, S. S. More Than Words? The Effect of Line Character Sticker Use on Intimacy in the Mobile Communication Environment. *Social Science Computer Review*, v. 34, n. 4, p. 456–478, 2016.

WANG, Y. et al. Eann: Event adversarial neural networks for multi-modal fake news detection. In: *KDD*. [S.l.: s.n.], 2018.

WANG, Y. et al. Understanding the use of fauxtography on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 15, n. 1, p. 776–786, May 2021.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, Nature Publishing Group, v. 393, n. 6684, p. 440, 1998.

WENG, L. et al. Competition among memes in a world with limited attention. *Scientific reports*, Nature Publishing Group, v. 2, p. 335, 2012.

WHATSAPP. *Keeping WhatsApp Personal and Private*. 2020. *WhatsApp Blog*. [Online; Posted on 07-Apr-2020. Accessed on 10-Sep-2022]. Available at: <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>.

WHATSAPP. *Two Billion Users – Connecting the World Privately*. 2020. *WhatsApp Blog*. [Online; Posted on 12-Feb-2020. Accessed on 10-Sep-2022]. Available at: <https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately>.

WhatsApp Messenger. *WhatsApp Encryption Overview – Technical white paper*. 2021. *WhatsApp Blog*. [Online; Updated on 14-Jul-2021]. Available at: <https://www.whatsapp.com/security/>.

WILLIAMS, M. *Secure Messaging Apps Comparison*. 2021. *Securemessagingapps.com*. [Online; Updated on 13-Oct-2021. Accessed on 22-Nov-2022]. Available at: <https://www.securemessagingapps.com>.

WONG, J. C. *US, UK and Australia urge Facebook to create backdoor access to encrypted messages*. 2019. *The Guardian*. [Online; Posted on 04-October-2019]. Available at: <https://www.theguardian.com/technology/2019/oct/03/facebook-surveillance-us-uk-australia-backdoor-encryption>.

WYTZE, A. *Killing Fake News Dead on Taiwan's Most Popular Messaging App*. 2017. *g0vnews*. [Online; Posted on 15-Feb-2017]. Available at: <https://g0v.news/killing-fake-news-dead-on-taiwans-most-popular-messaging-app-c99d93582cbe>.

YADAV, A. et al. Understanding the Political Inclination of WhatsApp Chats. In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. NY, USA: ACM, 2020. (CoDS COMAD 2020), p. 361—362. ISBN 9781450377386.

YAYLA, A. S.; SPECKHARD, A. Telegram: The mighty application that isis loves. *International Center for the Study of Violent Extremism*, 2017.

YEE, A. Post-truth politics & fake news in asia. *Global Asia*, v. 12, n. 2, p. 66–71, 2017.

YOON, C.; JEONG, C.; ROLLAND, E. Understanding individual adoption of mobile instant messaging: A multiple perspectives approach. *Information Technology and Management*, Springer, v. 16, n. 2, p. 139–151, 2015.

YUAN, K. et al. Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion. In: *2019 IEEE Symposium on Security and Privacy*. [S.l.: s.n.], 2019. (SSP'19), p. 952–966.

YUSOFF, M.; DEHGHANTANHA, A.; MAHMOD, R. Forensic Investigation of Social Media and Instant Messaging Services in Firefox OS: Facebook, Twitter, Google+, Telegram, OpenWapp, and Line as Case Studies. In: CHOO, K.-K. R.; DEHGHANTANHA, A. (Ed.). *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*. [S.l.]: Syngress, 2017. p. 41–62. ISBN 978-0-12-805303-4.

ZANNETTOU, S. et al. What is gab: A bastion of free speech or an alt-right echo chamber. In: *The Web Conference 2018*. [S.l.: s.n.], 2018. p. 1007–1014.

ZANNETTOU, S. et al. On the origins of memes by means of fringe web communities. In: *Proceedings of the Internet Measurement Conference 2018*. NY, USA: ACM, 2018. (IMC'18), p. 188–202. ISBN 9781450356190.

ZANNETTOU, S. et al. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 14, n. 1, p. 774–785, May 2020.

ZANNETTOU, S. et al. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In: ACM. *Proceedings of the 2017 Internet Measurement Conference*. [S.l.], 2017. p. 405–417.

ZANNETTOU, S. et al. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In: *Companion Proceedings of The 2019 World Wide Web Conference*. [S.l.: s.n.], 2019. (WWW'19), p. 218–226.

ZANNETTOU, S. et al. Who let the trolls out? towards understanding state-sponsored trolls. In: *Proceedings of the 10th acm conference on web science*. [S.l.: s.n.], 2019. p. 353–362.

ZAUNER, C. Implementation and benchmarking of perceptual image hash functions. na, 2010.

ZHANG, C. *WeChatting American politics: Misinformation, polarization, and immigrant Chinese media*. 2018. *Tow Center for Digital Journalism*. [Online; Posted on 19-Apr-2018]. Available at: <https://www.cjr.org/tow_center_reports/wechatting-american-politics-misinformation-polarization-and-immigrant-chinese-media.php>.

Appendix A

Academic Contributions

During the course of this extended study, some results have been achieved and, as a result, some pieces of this work were published in relevant conferences in the field. Below, there is a list of some publications resulting directly from the steps taken by this thesis while investigating the Instant Messaging Platforms ecosystem.

- Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. **(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures**. In *The World Wide Web Conference*, WWW '19, pages 818–828. ACM, 2019b

- Gustavo Resende, Philipe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara M. Almeida, and Fabrício Benevenuto. **Analyzing Textual (Mis)Information Shared in WhatsApp Groups**. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, pages 225–234. ACM, 2019a

- Carolina Vieira, Philipe Melo, Pedro Olmo Vaz de Melo, and Fabrício Benevenuto. **The Paradox of Encrypted Information Virality on WhatsApp.** In *Proceedings of the XXXVII Brazilian Symposium on Computer Networks and Distributed Systems*, SBRC'19, pages 403–416, Gramado, Brazil, 2019. SBC

- Philipe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. **WhatsApp Monitor: A Fact-Checking System for WhatsApp**. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13 of ICWSM '19, pages 676–677, Jul 2019a

- Philipe Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro OS Vaz de Melo, and Fabrício Benevenuto. **Can WhatsApp counter misinformation by limiting message forwarding?** In *International Conference on Complex Networks and Their Applications*, pages 372–384. Springer, 2019b

- Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. **A Dataset of Fact-Checked Images Shared on**

WhatsApp During the Brazilian and India Elections. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):903–908, May 2020a. doi: doi.org/10. 5281/zenodo.3734805

- Julio CS Reis, Philipe Melo, Kiran Garimella, and Fabrício Benevenuto. **Can whatsapp benefit from debunked fact-checked stories to reduce misinformation?** *Harvard Kennedy School (HKS) Misinformation Review*, 2020b

- Mohamad Hoseini, Philipe Melo[1], Manoel Junior, Fabrício Benevenuto, Balakrishnan Chandrasekaran, Anja Feldmann, and Savvas Zannettou. **Demystifying the messaging platforms' ecosystem through the lens of twitter**. In *Proceedings of the 2020 Conference on Internet Measurement Conference*, 2020

- Daniel Pimentel Kansaon, Philipe De Freitas Melo, and Fabrício Benevenuto. **"Click Here to Join": A large-scale analysis of topics discussed by brazilian public groups on whatsapp**. In *Proceedings of the Brazilian Symposium on Multimedia and the Web, WebMedia* '22, page 55–65, NY, USA, 2022. ACM. ISBN 9781450394093

- Érica Verônica Pereira, Philipe Melo, Manoel Júnior, Vitor O. Mafra, Julio C. S. Reis, and Fabrício Benevenuto. **Analyzing youtube videos shared on whatsapp and telegram political public groups**. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 28–37, NY, USA, 2022. ACM. ISBN 9781450394093. doi: 10.1145/3539637.3556997

Furthermore, we have under review process two other submissions that expand this research. One about multimedia format of stickers and their impact on public groups on WhatsApp and a second regarding the Forwarding tools of WhatsApp and evaluating the labeling WhatsApp apply over messages to prevent they to get viral within the network.

[1] Mohamad Hoseini and Philipe Melo contributed equally and both are first authors of the work.

Appendix B

Media Coverage

We are pleased to mention that our study had some repercussions in the news media and impact on society, helping journalists and researches to combat the increasing misinformation on WhatsApp. We list some of the most relevant news that covered our study as follows:

- Alessandra Monnerat. **Limite de encaminhamento no WhatsApp não consegue frear desinformação na plataforma, aponta pesquisa**. *Estadão*, Sep 2019. Available at <https://politica.estadao.com.br/blogs/estadao-verifica/limite-de-encaminhamento-no-whatsappnao-consegue-frear-desinformacao-na-plataforma-aponta-pesquisa>. [Online; Posted on 29-September-2019]

- Daniela Flamini. **Whatsapp efforts to curb misinformation aren't entirely effective, research shows**. *Poynter*, Sep 2019. Available at <https://www.poynter.org/fact-checking/2019/whatsapp-efforts-to-curb-misinformation-arent-entirely-effective-research-shows/>.[Online; Posted on 27-September-2019]

- Angela Chen. **Limiting message forwarding on whatsapp helped slow disinformation**. *MIT Technology Review*, Sep 2019. Available at <https://www.technologyreview.com/2019/09/26/434/whatsapp-disinformation-message-forwarding-politics-technology-brazilindia-election/>. [Online; Posted on 26-September-2019]

- Isabel Rubio. **Limitar el reenvío de mensajes en WhatsApp ralentiza la difusión de noticias falsas, pero no la detiene**. *El País*, Oct 2019. Available at <https://elpais.com/tecnologia/2019/10/01/actualidad/1569938947_969288.html>. [Online; Posted on 05-October-2019]

- Donna Lu. **WhatsApp restrictions slow the spread of fake news -– but don't stop it**. *New Scientist*, Sep 2019. Available at <https://www.newscientist.com/article/2217937-whatsapp-restrictions-slow-the-spread-of-fake-news-but-dont-stop-it/>. [Online; Posted on 27-September-2019]

- Géssica Brandino. **Quais as falhas do WhatsApp no combate à desinformação**. *Nexo*, Oct 2019. Available at <https://www.nexojornal.com.br/expresso/

2019/10/11/Quais-asfalhas-do-WhatsApp-no-combate-à-desinformação>. [Online; Posted on 11-October-2019]

- Laura Hazard Owen. **WhatsApp's message forwarding limits do work (somewhat) to stop the spread of misinformation**. *NiemanLab*, Sep 2019. Available at <https://www.niemanlab.org/2019/09/whatsapps-message-forwarding-limits-do-work-somewhat-tostop-the-spread-of-misinformation/>. [Online; Posted on 27-September-2019]

- Daniel Avelar. **Whatsapp fake news during brazil election 'favoured bolsonaro'**. *The Guardian*, Oct 2019. Available at <https://www.theguardian.com/world/2019/oct/30/whatsapp-fake-news-brazil-election-favoured-jair-bolsonaro-analysis-suggests>. [Online; Posted on 30-October-2019]

- Guilherme Pavarin. **Como a milícia digital bolsonarista resgatou sua máquina de fake news para atacar universitários**. *The Intercept Brasil*, May 2019. Available at <https://theintercept.com/2019/05/14/milicia-digital-bolsonarista-contra-universidades>. [Online; Posted on 15-May-2019]

- Letícia Mori. **Por que convocação de ato pró-Bolsonaro está rachando a direita**. *BBC News Brasil*, May 2019. Available at <https://www.bbc.com/portuguese/brasil-48360324>. [Online; Accessed on 01-December-2022]

- Pedro Capetti. **Decisivos na campanha, grupos bolsonaristas no WhatsApp agora atuam para desfazer crises**. *O Globo*, May 2019. Available at <https://oglobo.globo.com/brasil/decisivos-na-campanha-grupos-bolsonaristas-no-whatsapp-agora-atuam-paradesfazer-crises-23676755>. [Online; Accessed on 01-December-2022]

- Matheus Roca. **Entre fake-news e correntes, Whatsapp dominou debate eleitoral**. *Época*, Oct 2018. Available at <https://oglobo.globo.com/epoca/entre-fake-news-correnteswhatsapp-dominou-debate-eleitoral-23149667>. [Online; Accessed on 01-December-2022]

- Bryan Harris, Carolina Unzelte, and Andres Schipani. **Brazilian students and teachers join marches against Bolsonaro**. *Financial Times*, May 2019. Available at <https://www.ft.com/content/144fe138-7716-11e9-be7d-6d846537acab>. [Online; Accessed on 01-December-2022]

- Beatriz Jucá. **Mobilização por educação confronta bolsonaristas nas redes e testa força nas ruas**. *El País Brasil*, May 2019. Available at <https://brasil.elpais.com/brasil/2019/05/> [Online; Accessed on 01-December-2022]

- Amanda Ribeiro. **Após cortes no MEC, envio de imagens de estudantes nus cresce 950% em grupos de WhatsApp em 24 horas**. *Aos Fatos*, May 2019. Available at <https://www.aosfatos.org/noticias/apos-cortes-no-mec-envio-de-imagens-de-estudantes-nuscresce-950-em-grupos-de-whatsapp-em-24-horas/>. [Online; Accessed on 01-December-2022]

- Cristina Tardáguila. **Fotos (velhas) de universitários nus inundam WhatsApp para 'provar' a 'balbúrdia' apontada por Weintraub**.*Piauí Folha – Lupa*, May 2019b. Available at <https://lupa.uol.com.br/jornalismo/2019/05/13/artigo-universidade-whatsapp/>. [Online; Accessed on 01-December-2022]

- Reinaldo José Lopes. **Research Reveals Sexual Content of Social Media Images that Mocked Universities**. *Folha de São Paulo*, May 2019b. Available at <https://www1.folha.uol.com.br/internacional/en/brazil/2019/05/research-reveals-sexual-content-of-social-mediaimages-that-mocked-universities.shtml>. [Online; Accessed on 01-December-2022]

- Reinaldo José Lopes. **Pesquisa revela teor sexual de imagens que ridicularizam universidades**. *Folha de São Paulo*, May 2019a. Available at <https://www1.folha.uol.com.br/ciencia/2019/05/pesquisa-revela-teor-sexual-de-imagens-que-ridicularizam-universidades.shtml>. [Online; Accessed on 01-December-2022]

- Madhumita Murgia, Stephanie Findlay, and Andres Schipani. **India: the WhatsApp election**. *Financial Times*, May 2019. Available at <https://www.ft.com/content/9fe88fba-6c0d-11e9-a9a5-351eeaef6d84>. [Online; Accessed on 01-December-2022]

- Estadão Conteúdo. **Como as imagens sobre o protesto do dia 26 estão circulando no WhatsApp**. *Revista Exame*, May 2019b. Available at <https://exame.abril.com.br/brasil/como-as-imagens-sobre-o-protesto-do-dia-26-estao-circulando-no-whatsapp/>. [Online; Accessed on 01-December-2022]

- Estadão Conteúdo. **Bolsonaro condena ataques a Congresso e STF: 'Está mais para Maduro'**. *Revista Veja*, May 2019a. Available at <https://veja.abril.com.br/politica/bolsonaro-condena-ataques-a-congresso-e-stf-esta-mais-para-maduro/>. [Online; Accessed on 01-December-2022]

- Reynaldo Turollo, Tulio Kruse, and Diogo Magri. **O perigoso vale-tudo no submundo dos grupos do Telegram**. *Revista Veja*, Apr 2022. Available at <https://veja.abril.com.br/brasil/o-perigoso-vale-tudo-no-submundo-dos-grupos-do-telegram/>. [Online; Accessed on 01-December-2022]

- José Benedito da Silva. **O tamanho do poder da família Bolsonaro no Telegram**. *Revista Veja*, Apr 2022. Available at <https://veja.abril.com.br/coluna/maquiavel/o-tamanhodo-poder-da-familia-bolsonaro-no-telegram/>. [Online; Accessed on 01-December-2022]

- Ryan Broderick. **This Election Offered A Window Into WhatsApp's Wild, Sometimes Fact-Free World**. *Buzzfeed News*, Oct 2018. Available at <https://www.buzzfeednews.com/article/ryanhatesthis/no-one-knows-how-bad-fake-news-is-on-whatsapp-but-if>. [Online; Accessed on 01-December-2022]

- UOL Confere. **Para inflar economia do Brasil, áudio inventa que Mercedes esgotou produção**. *UOL Notícias*, Set 2022. Available at <https://noticias.uol.com.br/comprova/ultimas-noticias/2019/11/14/audio-inventa-que-mercedes-esgotou-producao-para-inflareconomia-do-brasil.htm>. [Online; Accessed on 01-December-2022]

- Levy Teles, Samuel Lima, and Gustavo Queiroz. **Grupos bolsonaristas recorrem a pânico para mobilizar no WhatsApp e Telegram**. *UOL Notícias*, Set 2022. Available at <https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2022/09/04/grupos-recorrema-panico-para-mobilizar-no-whatsapp-e-telegram.htm>. [Online; Accessed on 01-December-2022]

- Cristina Tardáguila. **Marielle, Suzano e STF transformaram o WhatsApp num pântano de horror e ódio**. *Época*, Aug 2019a. URL *https://oglobo.globo.com/epoca/marielle-suzano-stf-transformaram-whatsapp-num-pantano-de-horror-odio-23531221*. [Online; Accessed on 01-December-2022]

- Projeto Comprova. **É falso texto que acusa Embaixada do Brasil nos EUA de ser "reduto do PT"**. *Revista Exame*, Aug 2019b. Available at <https://exame.com/brasil/efalso-texto-que-acusa-embaixada-do-brasil-nos-eua-de-ser-reduto-do-pt/>. [Online; Accessed on 01-December-2022]

- Redação Estadão. **Vídeo que contesta desmatamento citando chuvas traz dados errados; saiba como são feitas as medições do INPE**. *Estadão Verifica*, Jul 2019. Available at <https://politica.estadao.com.br/blogs/estadao-verifica/video-que-contesta-desmatamentocitando-chuvas-traz-dados-errados-saiba-como-sao-feitas-as-medicoes-do-inpe/>. [Online; Accessed on 01-December-2022]

- Projeto Comprova. **Chuvas não impedem desmatamento na Amazônia; veja como o Inpe monitora as florestas**. *GaúchaZH Ambiente*, Aug 2019a. Available at <https://gauchazh.clicrbs.com.br/ambiente/noticia/2019/08/chuvas-nao-impedem-desmatamentona-amazonia-veja-como-o-inpe-monitora-as-florestas-cjyrjs7sh00gu01 html>. [Online; Accessed on 01-December-2022]

- Bernardo Barbosa. **O que levou aos protestos em defesa da pauta de Bolsonaro**. *Uol Notícias*, May 2019b. Available at <https://noticias.uol.com.br/amp-stories/o-que-levou-aos-protestos-em-defesa-da-pauta-de-bolsonaro/>. [Online; Accessed on 01-December- 2022]

- Bernardo Barbosa. **Pauta radical é de "lobo solitário", diz organizador de ato pró- Bolsonaro**. *Uol Notícias*, May 2019a. Available at <https://noticias.uol.com.br/politica/ultimas-noticias/2019/05/25/pauta-radical-e-de-lobo-solitario-diz-organizador-de-atopro-bolsonaro.htm>. [Online; Accessed on 01-December-2022]

- Juliana Gragnani. **Como planos de bloqueio saíram de grupos de mensagem para as ruas**. *BBC News – World Service Disinformation Team*, Nov 2022b. Available at <https://www.bbc.com/portuguese/brasil-63479074>. [Online; Accessed on 01-December- 2022]

- Juliana Gragnani. **'Não vamos parar': a reação de grupos bolsonaristas nas redes ao discurso de Bolsonaro**. *BBC News – World Service Disinformation Team*, Nov 2022a. Available at <https://www.bbc.com/portuguese/brasil-63480237>. [Online; Accessed on 01-December-2022]

- Shin Suzuki. **O que WhatsApp, Telegram, TikTok, Facebook e YouTube prometem fazer contra fake news nas eleições**. *BBC News Brasil*, Mar 2022. Available at <https://www.bbc.com/portuguese/brasil-60896482>. [Online; Accessed on 01-December- 2022]

- Patrícia Campos Mello. **Menções ao 7 de Setembro explodem em grupos de mensagens e incluem teor golpista**. *Folha de São Paulo*, Aug 2022. Available at <https://www1.folha.uol.com.br/poder/2022/08/convocacoes-golpistas-para-o-7-de-setembro-explodemem-grupos-de-mensagens.shtml>. [Online; Accessed on 01-December-2022]

- Patrícia Campos Mello, Paula Soprana, and Renata Galf. **Fake news sobre urnas, pesquisas e TSE dominam eleição de 2022**. *Folha de São Paulo*, Set 2022. Available at <https://www1.folha.uol.com.br/poder/2022/09/fake-news-sobre-urnas-pesquisas-e-tse-dominameleicao-de-2022.shtml>. [Online; Accessed on 01-December-2022]

- Laura Scofield. **Grupo armamentista apontado como danoso no Facebook convoca para o 7 de Setembro**. *Agência Pública*, Aug 2022. Available at <https://apublica.org/sentinela/2022/08/grupo-armamentista-apontado-como-danoso-no-facebook-convocapara-o-7-de-setembro/>. [Online; Accessed on 01-December-2022]

- Laura Scofield and Nathallia Fonseca. **TikTok e Kwai levam desinformação sobre urnas e Forças Armadas ao WhatsApp**. *Agência Pública*, Set 2022. Available at <https://apublica.org/sentinela/2022/09/tiktok-e-kwai-levam-desinformacao-sobre-urnas-e-forcasarmadas-ao-whatsapp/>. [Online; Accessed on 01-December-2022]

- Laura Scofield and Matheus Santino. **Grupos bolsonaristas dominam redes com mentiras e críticas a pesquisas eleitorais**. *Agência Pública*, Set 2022. Available at <https://apublica.org/sentinela/2022/09/grupos-bolsonaristas-dominam-redes-com-mentirase-criticas-a-pesquisas-eleitorais/>. [Online; Accessed on 01-December-2022]

- Ana Flávia Gussen. **Pelo fim do vale-tudo, o TSE fecha o cerco às milícias digitais**. *Carta Capital*, Mar 2022. Available at <https://www.cartacapital.com.br/politica/o-tsefecha-o-cerco-as-milicias-digitais-e-pesquisadores-criam-tecnologia-que-detecta-fakenews/>. [Online; Accessed on 01-December-2022]

Some news cover specifically our dissemination analysis from MELO et al., 2019b, while others refer to *WhatsApp Monitor* online system or use some data or report provided by our extension project of our research group at UFMG "Eleições sem Fake". There were also some TV news making usage of this study, showing the importance of our results and impact for society.