Understanding the Limits of Information Dissemination on Whatsapp

Philipe de Freitas Melo
Departamento de Ciência da Computação - UFMG
philipe@dcc.ufmg.br

Carolina Coimbra Vieira
Departamento de Ciência da Computação - UFMG
carolcoimbra@dcc.ufmg.br

Kiran Garimella
Institute for Data, Society and Systems - MIT
garimell@mit.edu

Fabrício Benevenuto
Departamento de Ciência da Computação - UFMG
fabricio@dcc.ufmg.br

Pedro O. S. Vaz de Melo
Departamento de Ciência da Computação - UFMG
olmo@dcc.ufmg.br

WhatsApp has become an important platform for information dissemination and social mobilization, especially in Brazil, India and Southeast Asia, where this app is very popular and was massively exploited to disseminate political campaigns and spread false news [RMS+19]. These two countries went through general elections between 2018 and 2019 and were severely affected by the use of WhatsApp during the campaign period. In Brazil, 120 million people (approximately 48% of the population) use WhatsApp and most of them used it as primary source of information during elections[1]. This has put WhatsApp as the protagonist in the alarming problem of misinformation, enabling actors with extremist views that create ideological divisions in the country[2]. In India, 94% of all Android smartphones have WhatsApp installed and the volume of fake news on the app was also astonishing, where rumors and misinformation has lead to real world unrest in parts of the country[3].

The main features that have helped WhatsApp occupy such a prominent place in these countries are the possibility to send and forward messages to multiple users and create group chats. WhatsApp messages are encrypted and hence anonymous, and so tracing back the origin of a message that has spread across the entire network is an extremely difficult task. Because of these peculiarities, WhatsApp raises a controversy related to its characteristics of anonymity and virality[4]. The problem arises because, in practice, WhatsApp works as a communication and as a media platform. As a communication platform, it ensures user anonymity and security by encrypting our data. As a media platform, just like a radio or television channel, it transmits information and disseminates content and enable virality through its sharing functions: *broadcast* and *forward*. Using the *broadcast* feature, a contact list can be created to send messages to up to 256 contacts (or groups) at once. Using the *forward* feature, a message can be promptly sent to 5 distinct people or groups, a limit that was already imposed by WhatsApp to control the spread of information within the app[5]. These features can be used to spread anonymous messages to thousands of people without any ethical or legal regulation of their content.

Recent tools developed by Garimella and Tyson [GT18] enable collecting and studying WhatsApp data at scale. Their proposed method monitors public WhatsApp groups and exports the data from these groups. Resende et al. [RMS+19] made use of these tools to collect data from political groups in Brazil. They analyzed the network formed by members of these groups and the propagation of misinformation through images. Similarly, Resende et al. [RMR+19] analyzed the content of what is shared on popular public WhatsApp groups and found that messages with misinformation are much more viral, i.e., they are shared more times, by a larger number of users and in more groups.

Although we know messages in WhatsApp spread rapidly and broadly, we do not know if (and how) their dissemination can be controlled by limitations in the *broadcast* and *forward* functions. This is particularly true because of the very distinct WhatsApp network topology, which is composed by traditional social links and also very large cliques, i.e., the WhatsApp groups, where every user is connected to, and is exposed to messages sent by every other member of the group. Because these groups connect all their members simultaneously, they can work as a mass media platform. Public groups play a crucial role in disseminating messages through the network, as they easily connect people who are far apart in the natural (or organic) social structure (e.g. friends, family, coworkers) of the system.

In this work, we evaluate the dynamics of the spread of (mis) information on a network of public WhatsApp groups. We focus on the mass communication features of public chat groups and the forwarding/broadcasting of messages. More specifically, we study the anatomy of this emerging social network and comprehend its peculiarities to answer the question of how the forwarding tools contribute to the virality of (mis)information and whether system limitations are capable of preventing the spread of content. We also propose some hints on how the problem of large-scale dissemination can be countered.

Since chat groups on WhatsApp are mostly private, they are much harder to monitor than Facebook or Twitter discussions. Because of that, we use recent tools developed by Garimella and Tyson [GT18] to get access to messages posted on WhatsApp public groups. Given a set of invitation links to public groups, we automatically join these groups and save all data coming from them.

[1] https://g1.globo.com/politica/eleicoes/2018/eleicao-em-numeros/noticia/2018/10/03/datafolha-quantos-eleitores-de-cada-candidato-usam-redes-sociais-leem-e-compartilham-noticias-sobre-politica.ghtml
[2] https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html
[3] https://www.bbc.com/news/world-asia-india-44435127
[4] https://www.wsj.com/articles/india-wants-access-to-encrypted-whatsapp-messages-11547551428
[5] https://blog.whatsapp.com/10000647/More-changes-to-forwarding

Table 1: Overview of the datasets.

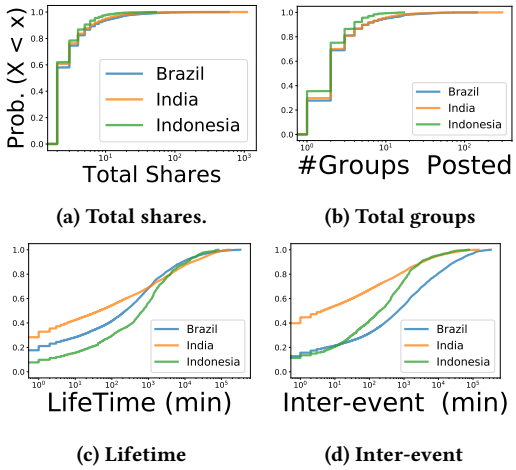
	#Users	#Groups	Unique Images	Total Images	Period ~2,5 months
Brazil	17,465	414	258k	416k	2018/08/15 - 2018/11/01
India	362,739	5,839	509k	810k	2019/03/15 - 2019/06/01
Indonesia	8,388	217	16k	21k	15/03/2019 - 2019/06/01

We selected groups from Brazil, India and Indonesia dedicated to political discussions. These groups have a large flow of content and are mostly operated by individuals affiliated with political parties, or local community leaders. We monitored the groups during the electoral campaign period and, for each message, we extracted the following information: (i) the country where the message was posted, (ii) name of the group the message was posted, (iii) user ID, (iv) timestamp and, when available, (v) the attached multimedia files (e.g. images, audio and videos).

As images usually flow unaltered across the network, they are easier to track than text messages. Thus, we choose to use the images posted on WhatsApp to analyze and understand how a single piece of content circulates across the network. To calculate a fingerprint for images, we follow the same strategy of [RMS+19], using the Perceptual Hashing (pHash) algorithm to group together sets of images with the same content. Then, we can count its popularity and track its spreading across the network. In total for all three countries, 784k unique image objects were tracked. For all three countries, we analyzed the data around the election day. We kept the same time span for the three countries to ease the comparison among them. The dataset overview and the total number of distinct images are described in Table 1. As expected, India and Brazil have a much larger volume of data shared on WhatsApp compared to Indonesia, as they have much more groups and users registered in our data collection system.

We calculate coverage and dynamics of spreading of all these images in our data. To evaluate spreading metrics regarding time and coverage, we only consider the images that were posted in at least two groups, since we cannot see the effect of spreading of images only shared 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 for each country. First, we calculate the total number of shares of each image and how many groups they have appeared using the Cumulative Distribution Function (CDF) as shown in Figures 1a,1b. Even though nearly 80% images on WhatsApp were posted only once, there are some very popular images broadly shared over 100 times that reached multiple groups. This shows that WhatsApp can be used as a mass communication media and the potential of virality of content.

Besides looking at the spread of images on WhatsApp, we also analyze their "lifetimes" in Figure 1c. The lifetime is given by the difference between the last and first occurrence of the image in our dataset. In short, while most of the images (80%) last no more than 2 days, there are images in Brazil and in India that continued to appear even after 2 months of the first appearance (10^5 minutes). Also, 60% of the images are posted before 1000 minutes after their first appearance. Further analysis, in Figure 1d shows the distribution of the "inter-event times" between posts of the same image. 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



(a) Total shares. **(b) Total groups**

(c) Lifetime **(d) Inter-event**

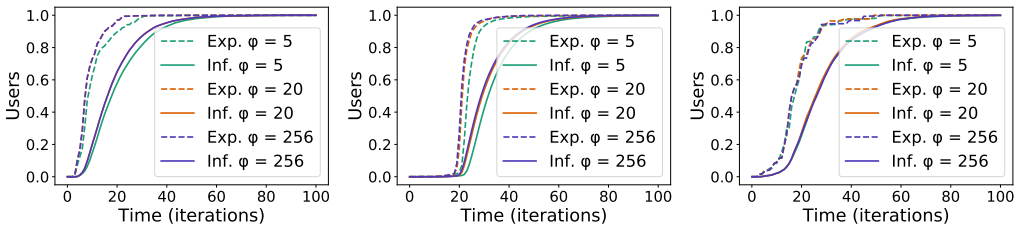
Figure 1: CDF of sharing coverage and time dynamics metrics of images shared at least twice on WhatsApp.

intervals of 10 minutes or less, while just 20% of shares were done in this same time interval in Brazil and Indonesia. We manually looked for reasons behind the short period of time between posts and found that in the data from India, there is more automated, spam-like behavior compared to in Brazil and Indonesia. In conclusion, these results suggest that WhatsApp is a very dynamic network and most of its image content is ephemeral, i.e., the images usually appear and vanish quickly. The linear structure of chats make it difficult for an old content to be revisited, yet there are some that linger on the network longer, disseminating over weeks or even months.

Furthermore, we construct the network of WhatsApp groups with the groups as nodes and edges representing number of users they share in common. This network has a 65%, 92% and 55% of groups in the largest connected component for Brazil, India and Indonesia respectively, indicating that they are very well connected allowing easy information dissemination.

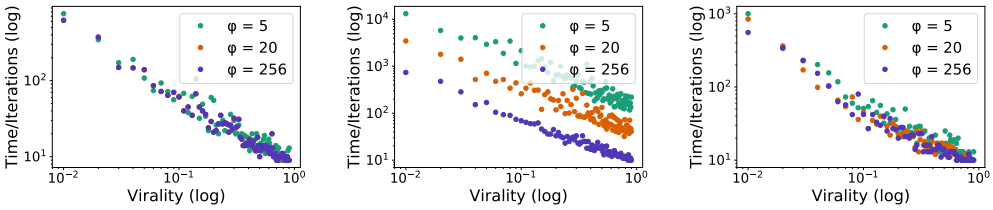
We use these networks to run simulations of the epidemiological model of Susceptible-Exposed-Infected (SEI) [GZ04] and perform several experiments to compare and estimate the the dissemination of malicious messages in WhatsApp groups by assuming misinformation as an infection that spreads to users through the group network. In our scenario, nodes are useres, which are members of groups and 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 didn't share it; *Infected* (I) is the final stage in which a user effectively shares this infection. We consider that users are infected when they forward or broadcast this content, as it indicates a degree of belief in the shared message. This model has two basic parameters: *virality* (α) and *exposition* (β). We also implemented a third parameter *forward limit* (φ) to test the restrictions on sharing by WhatsApp.

The **virality** (α) of malicious content is a parameter that controls the rate of infected users. This parameter indicates the probability of infected users to share the content that they had contact with. The **exposition** parameter (β) refers to the rate at which exposed



(a) Brazil **(b) India** **(c) Indonesia**

Figure 2: SEI model varying the forward limit (φ). $\alpha = \beta = 0.1$.



(a) Brazil **(b) India** **(c) Indonesia**

Figure 3: Time to infect all user in the network on simulations of SEI model varying the virality (α) from 0.001 up to 1.0 .

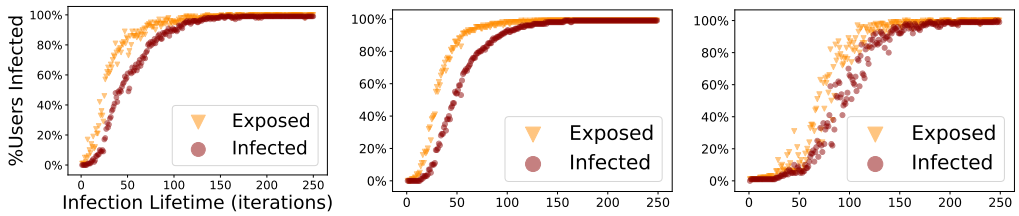
users become infected. It represents the probability of an exposed user to transform in an infected one. Lastly, the **forward limit** (φ) of infection is a specific parameter we use to restrict the spread of the infection to simulate the actual conditions on WhatsApp. This parameter indicates the maximum amount of groups an infected node can spread the infection to.

Figure 2 represents results of simulation showing the fraction of users infected over WhatsApp networks for each country varying the forward limit (φ). We considered the limit of forwarding to 5 groups (the actual scenario), 20 groups (the previous limit), and 256 groups (the current limit for broadcasting). Notice that the rate of users exposed in the network grows very fast, regardless of forwarding limits, showing that a message can infect the entire network in 60 iterations. Also, observe that limitations on forwarding slightly diminish the velocity of spreading, but does not stop it completely, especially for exposed users.

We also evaluate the time needed by (mis)information to infect all users with different potential viralities. Figure 3 shows the time needed 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 the propagation, mainly in larger networks, as in India. In short, **limits on forwarding and broadcasting can reduce velocity of dissemination by one order of magnitude for any of α virality**.

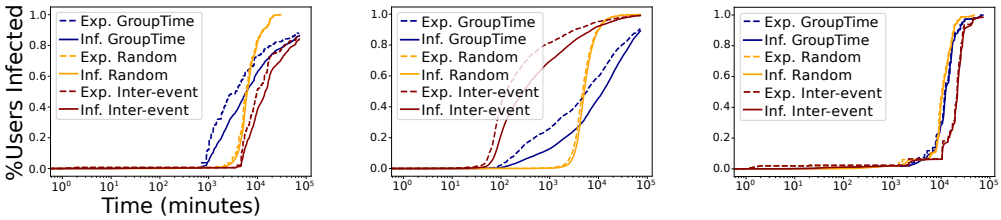
In reality, users may lose interest in some topics through time, so it is natural for a time limit on the content spread. We add this time limit to our model through a "lifetime", which denotes the maximum duration of an infection in the simulation before it is entirely extinguished. Figure 4 shows the percentage of users infected by increasing 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 an infection could last. We observe that for all three countries, an infectious content that lasts 100 iterations or more is powerful enough to expose more than half population. When this content persists in the network for at least 150 iterations, it usually infects almost 100% of the users. 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 and nullify its virality (e.g. disabling forwarding on that piece of content), thus preventing further contagion.

Finally, in the previous simulations using the SEI model, the spread of information was measured in terms of the *number of iterations*. Now, we use real data to adapt the SEI model and measure the spread in terms 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 1d; (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 on initial steps smaller than in the subsequent steps. During the simulation, we track the number of times 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 5 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 user in the first 2 day (3000



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

Figure 4: Users infected by time in simulations of the SEI model using max lifetime for infections. $(\alpha) = (\beta) = 0.1$. Forward limit $(\varphi) = 5$.



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

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

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.

The closed nature of WhatsApp and the ease of transferring multimedia and sharing information to large-scale groups makes WhatsApp an extremely hard environment for the deployment of countermeasures to combat misinformation. WhatsApp opens a paradoxical use of its platform, allowing at the same time the viral spread of a content and encrypted personal chat. Together those two features can be widely abused by misinformation campaigns. Our results show that a content can spread quite fast through the network structure of public groups in WhatsApp, reaching later the private groups and individual users. Our empirical observations about the network of WhatsApp public groups in three different countries provides a means of inferring the information velocity in terms of minutes related to real-world scenarios[6]. We verified that most of the images (80%) last no more than 2 days in WhatsApp which, in India, 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.

Using a SEI model we investigate a set of what-if questions about the limits that WhatsApp can impose in the information propagation. While the limit on the number of users per groups can prohibit the creation of giant hubs to spread information through the network, this limit, however, is not able to prevent a content to reach a large portion of entire platform. More important, our analysis show that low limits imposed on message forwarding and broadcasting (e.g. up to five forwards) offer a delay in the message propagation of up to two orders of magnitude in comparison with the original limit of 256 used in the first version of WhatsApp. We 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 [RMS+19]. Thus, as a counter-measurement, WhatsApp could implement a quarantine approach to limit infected users to spread misinformation. This could be done by temporarily restricting the virality features of suspect users and content, especially during elections, preventing coordinated campaigns to flood the system with misinformation.

[6]Please, see a full version of this work at [dFMVG+19] with more details and further discussions of our findings.

REFERENCES

[dFMVG+19] Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro O. S. Vaz de Melo, and Fabrício Benevenuto. Can WhatsApp Counter Misinformation by Limiting Message Forwarding?, 2019.

[GT18] Kiran Garimella and Gareth Tyson. Whatapp doc? a first look at whatsapp public group data. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[GZ04] Li Guihua and Jin Zhen. Global stability of an sei epidemic model. *Chaos, Solitons & Fractals*, 21(4):925–931, 2004.

[RMR+19] Gustavo Resende, Philipe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In *Proceedings of the ACM Conference on Web Science*, WebSci'19, 2019.

[RMS+19] 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 *Proceedings of The Web Conference*, WWW'19, 2019.