Can WhatsApp Counter Misinformation by Limiting Message Forwarding?

Philipe de Freitas Melo[1], Carolina Coimbra Vieira[1], Kiran Garimella[2],
Pedro O. S. Vaz de Melo[1], and Fabrício Benevenuto[1]

[1] Departamento de Ciência da Computação - UFMG, Belo Horizonte MG, Brazil,
`philipe, carolcoimbra, olmo, fabricio@dcc.ufmg.br`
[2] Institute for Data, Society and Systems - MIT, USA,
`garimell@mit.edu`

Abstract. WhatsApp is the most popular messaging app in the world. The closed nature of the app, in addition to the ease of transferring multimedia and sharing information to large-scale groups make WhatsApp unique among other platforms, where an anonymous encrypted messages can become viral, reaching multiple users in a short period of time. The personal feeling and immediacy of messages directly delivered to the user's phone on WhatsApp was extensively abused to spread unfounded rumors and create misinformation campaigns during recent elections in Brazil and India. WhatsApp has been deploying measures to mitigate this problem, such as reducing the limit for forwarding a message to at most five users at once. Despite the welcomed effort to counter the problem, there is no evidence so far on the real effectiveness of such restrictions. In this work, we propose a methodology to evaluate the effectiveness of such measures on the spreading of misinformation circulating on WhatsApp. We use an epidemiological model and real data gathered from WhatsApp in Brazil, India and Indonesia to assess the impact of limiting virality features in this kind of network. Our results suggest that the current efforts deployed by WhatsApp can offer delays on the information spread, but are ineffective in blocking the propagation of misinformation campaigns in public groups.

Keywords: WhatsApp, Misinformation, Fake news, Virality

1 Introduction

Messaging applications, such as WhatsApp, Facebook messenger, Telegram and Viber have gained a significant role in the daily lives of smartphone users. WhatsApp is the most popular app, with over 1 billion active users[3]. Besides being widely used to keep in touch with friends & family, run businesses, read news & get informed, WhatsApp has become an important platform for information dissemination and social mobilization, especially in Brazil and India [1].

There are a few key features that make WhatsApp unique among other platforms. First, WhatsApp allows the connection among like-minded individuals through chat groups. These chat groups have a limit of 256 users and can be

---

[3] https://blog.whatsapp.com/10000631/Connectinganuser-users-all-days

private or public. In the case of private groups, new members must be added by a member who assumes the role of group administrator. For public groups, the access is by invitation links that could be shared to anyone or be available on the Web. These public groups often come up to discuss hobbies and passions, but also specific topics such as health, education, and politics. Although the majority of groups are private, set up among people who share a social relationship (e.g., family, friends, workmates) public groups have been a catalyzing feature for the purpose of information diffusion: most of their members are strangers to each other. This is evident in countries like Brazil, where a survey reported that 76% of WhatsApp users are part of groups, 58% participate in groups with people they do not know, and 18% of these groups discuss politics [2]. For this reason, public groups can act as a shortcut for information to directly traverse distant parts of the underlying social network structure via a clique of weak ties, broadening and accelerating information dissemination [3].

Furthermore, the app has two sharing functions: broadcast, in which a contact list can be created to send messages to up to 256 contacts (users or groups) at once and forward, that a single message received can be forwarded to other 5 contacts (users or groups). Those characteristics allow the message to travel long distances by the network, but the end-to-end encryption makes it difficult to identify the source and track the spread of the messages. Because of these peculiarities, WhatsApp generated a controversy related to its anonymity and virality characteristics. This conflict is due the fact that we can view WhatsApp in two different ways, such as a technology company, or as a media platform. As a technology platform, it ensures user anonymity and security by encrypting your data. As a media platform, it transmits information and disseminates content in large-scale. Thus, messages sent anonymously reach thousands of people quickly and without any ethical or legal regulation of this disseminated content, promoting, for example, disinformation campaigns. The massive spread of (mis)information and rumors [4] led to requests from both the national governments[4] towards altering features that allow the platform to be abused to spread (mis)information at scale. This resulted in WhatsApp implementing restrictions on the way messages are forwarded[5] by reducing the limit for forwarding content to at most 5 users/groups at a time. However, there are no studies that investigate the impact of these limitations or whether the numbers chosen are sufficient to deal with the spread of viral content.

In this work, we evaluate the dynamics of the spread of (mis)information in a network of public WhatsApp groups. 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.

---

[4] https://www.latimes.com/world/la-fg-india-whatsapp-2019-story.html
[5] blog.whatsapp.com/10000647/More-changes-to-forwarding

The rest of the paper is organized as follows. Section 2 describes related work. In Section 3 describes our WhatsApp data. A characterization of the data is presented in Section 4. In Section 5, we analyze the network structure of public groups in WhatsApp. In Section 6, we simulate the spread of misinformation within WhatsApp under different circumstances via a Susceptible-Exposed-Infected (SEI) epidemiological model [5]. Finally, Section 7 concludes the paper.

2   Related Work

Recently, there have been numerous research studies reporting misinformation campaigns on social networks [6, 7]. This includes popular platforms like Facebook, where, Ribeiro *et al.* [8] 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. There are also reports of attempts to manipulate political discourse with the use of social bots and even state-sponsored trolls [6, 9].

However, only recently, social message applications, such as WhatsApp were reportedly a means of abuse by misinformation campaigns [1, 10–12]. Particularly, Resende *et al.* [1] analyzed the dissemination of different kinds of content on WhatsApp, such as images, audio and videos, finding a large amount of misinformation in the form of memes and fake images. Resende *et al.* [10] provide an in-depth characterization of textual messages, showing that misinformation tends to be more viral, i.e., these messages are shared more times, by a larger number of users, and in more public groups. Bursztyn *et al.* showed that right-wing WhatsApp groups in Brazil were more active and engaged in spreading political content in WhatsApp along the 2018 Brazilian elections, in comparison with left-wing groups. Melo *et al.* [11] developed a system to help fact checkers, providing them a sample of the most popular images, messages, URLs, audios and videos shared hundreds of public groups in Brazil and India.

On the side of viral spreading, Rushkoff *et al.* [13] studies dissemination in digital media by comparing characteristics of biological and computational viruses, explaining this phenomenon on web and covering many efforts using epidemiological models to analysis viral spreading of disinformation on social networks. Our work is complementary to the above efforts as we investigate how limitations on virality features, such as limits on message forwarding recently deployed by WhatsApp, are effective in mitigating misinformation campaigns.

3   Datasets

Since chat groups on WhatsApp are mostly private, we use a recently proposed data gathering strategy [14, 1] to access messages posted on public groups. Given a set of invitation links to public groups, we automatically join these groups and save all data coming from them. 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 analyse and understand how a single piece of content flows across the network. To calculate a fingerprint for every image, we follow the same strategy of [1], using the Perceptual Hashing (pHash) algorithm to group together sets of images with the same content. Since similar images have the same hash value, 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, 60 days before and 15 after. 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, Brazil and India 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.

Data Limitations: Our methodology gathers a large dataset from public groups, but it is known that most of WhatsApp conversations occurs in private channels. A key limitation of our work is that our results reflect only users and content that circulate on the public layer of WhatsApp. We note, however, that there is evidence that suggests that public groups make up the key backbone of the misinformation campaigns on WhatsApp.[6] First, they are focused on political activism, where most of the shared content contain misinformation. For example, a fact checking agency in Brazil checked the top 61 images shared in these groups, finding that only 10% of them are true [1]. There is also evidence of the use of automatic tools to flood WhatsApp public groups with political content[7]. Then, the users in those groups would be responsible to amplify the misinformation campaign and propagate it to the private part of the network.[8]

Nevertheless, this project brings a considerable amount of data that can help to elucidate how WhatsApp is being abused for mass communication and the amplification backbone composed by public groups that distribute messages in bulk for thousands of users. At the least, our results provide a 'lower bound' on the ability of messages to spread on WhatsApp, since the network we consider is a subset of the entire WhatsApp network.

4   Spreading Coverage and Dynamics

Since we are able to track all occurrence of a given image, we can see the coverage and dynamics of spreading of these images in our data. To evaluate spreading metrics regarding time and coverage, we only consider the images that were

---

[6] https://www.bbc.com/news/world-asia-india-47797151

[7] https://www.bbc.com/news/technology-45956557

[8] https://time.com/5512032/whatsapp-india-election-2019/

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

(a) Total shares.   (b) Total groups   (c) Lifetime   (d) Inter-event
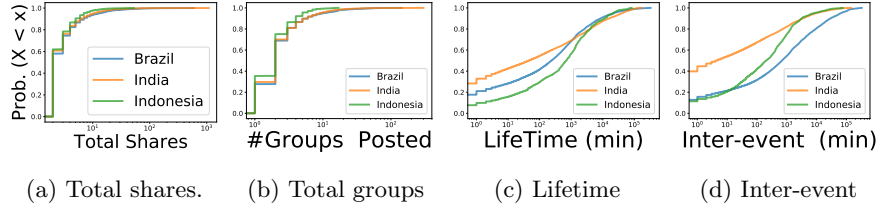
Fig. 1: CDF of sharing coverage and time dynamics metrics of images shared at least twice on WhatsApp.

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 the images for each country.

First, we calculate the total number of shares of each image and how many groups they have appeared. Figures 1a,1b 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 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 few times, this shows that WhatsApp can be used not only to particular conversations but also as a mass communication media with a potential virality of its content.

Time Analysis for WhatsApp Data: 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). We can also see that the majority (60%) of the images are posted before 1000 minutes after their first appearance. Moreover, in Brazil and India, around 40% of the shares were done after a day of their first appearance and 20% after a week. 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 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.

5   Network structure

In this section, we investigate the network structure of public WhatsApp groups and compare its characteristics with other real and synthetic social networks. To create a network from the WhatsApp groups, we connected two groups if they share a common user. Although WhatsApp is an encrypted personal chat application, 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. Although the WhatsApp group network resembles many other social networks, little is known about the differences in information dissemination.

In Figure 2, we show the distribution of groups per user and users per group. In order to compare WhatsApp peculiarities with another popular platforms, we also use data from Reddit network, modeling the subreddits as groups and users as members. Note that even though Reddit has the same group characteristic, we want to evaluate specific features of WhatsApp that lead to very different network structures. The maximum of 256 members in groups is a determining element in the network, capable of limiting group size, mainly in India (Figure 2a), where there are over 300k users and more than 5k groups.[9] 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, what creates big hubs of users. As both platforms have no limit on the number of groups users can join, we expected to see no differences in the total number of groups users participate. However, note that in Reddit, the distribution has a exponential decay, with a limit on ≈ 100 groups. On the other hand, all WhatsApp curves are similar with a well behaved power law curve, which naturally yields a larger variance. Note that in India we have users who participated in more than 300 groups.

In Figure 3 we show these networks for all three countries. 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 [15]. Observe that in all graphs there is an evident largest connected component and some other group clusters. Also, note that some groups position themselves as bridges and hubs, connecting different communities of the network structure.

Next, we compare the characteristics of the WhatsApp group network and other social network graphs: (i) random generated graphs using the Barabasi-Albert scale free model, the Erdős–Rényi model, the small world model [16] and the Forest Fire network model [17], for which we used the same number of nodes

---

[9] 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.
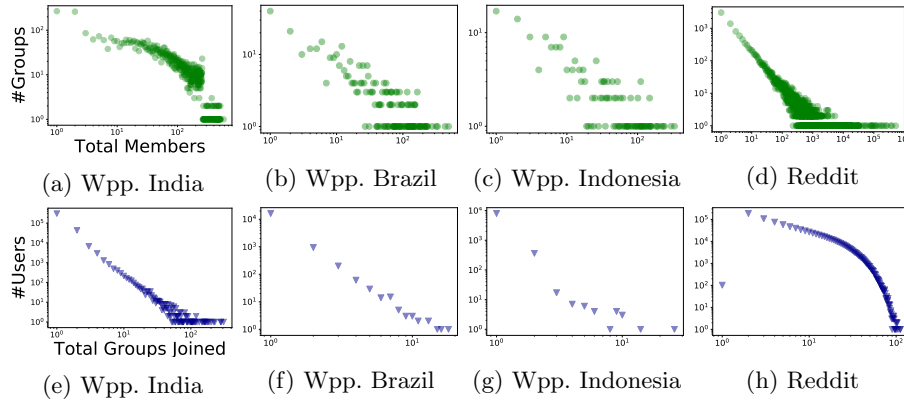
(a) Wpp. India   (b) Wpp. Brazil   (c) Wpp. Indonesia   (d) Reddit

(e) Wpp. India   (f) Wpp. Brazil   (g) Wpp. Indonesia   (h) Reddit

Fig. 2: Distributions of the number of members per group and total groups joined per user in WhatsApp (Wpp) and in Reddit.
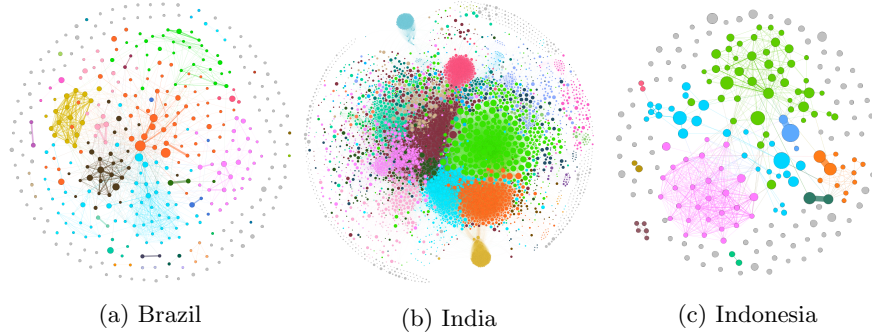

(a) Brazil   (b) India   (c) Indonesia

Fig. 3: WhatsApp Public Groups Network for each country. Each node is a group and edges represents members in common.

in the Indian dataset in order to create a comparable network; (ii) the network of subreddits from Reddit [18], and, (iii) the Flickr network [19], which, different from the WhatsApp and Reddit group networks, the Flickr graph represents the network of images shared by users on the platform. The results are shown in Table 2. We 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 a social network. WhatsApp also shows a higher Pearson coefficient, in which nodes tend to be connected with other nodes with similar degree values. In epidemic analyses, it can help 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.

6   Impact of forwarding limitations on information spread

We use the epidemiological model of Susceptible-Exposed-Infected (SEI) [5] to estimate the virality of malicious messages in WhatsApp groups by assuming

Table 2: Network metrics for WhatsApp compared to other networks.

|	#Nodes	#Edges	Mean Degree	Clustering Coefficient	Diameter	APL*	Density	LCC**	Pearson Coefficient
Wapp. India	5,839	407,081	69.71	0.59	11	3.17	0.0239	92.6%	0.295
Wapp. Brazil	414	1,400	6.76	0.32	8	3.19	0.0164	65.2%	0.346
Wapp. Indonesia	217	699	6.44	0.38	9	3.09	0.0298	55.3%	0.290
Bar.-Albert	5,839	792,300	271.38	0.10	3	1.95	0.0465	100%	0.008
Erdos-Renyi	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
ForestFire	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	4.8	0.0004	99.8%	0.247

*Average Path Length
**Largest Connected Component

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 didn't 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). We also implemented a third parameter *forward limit* (φ) to test the restrictions on sharing by WhatsApp.

The **virality** (α) of malicious content controls the rate of infected users. This indicates the probability of an infected user to share the content to 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 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. The simulation starts by selecting one user randomly to be the initial infected node to start the spreading. For each user exposed, they have a probability given by α to share the malicious message. When these infected nodes decide to forward, there is a limitation given by φ, the maximum number of groups they will send the content to. After that, each user in the groups that received the message are exposed. Then, each exposed user has also a probability β of becoming an infected node and sharing the content. This iteration repeatedly continues until all users are infected.

We perform several experiments using our SEI model comparing the dissemination in different scenarios by enforcing limits of broadcast and forward. Since it would not be possible to reach isolated nodes using the whole structure, only the largest connected component was considered. Figure 4 shows the fraction of users infected over time for all countries when the forward limit (φ) is varied, i.e., how the restrictions implemented by WhatsApp can interfere with the spread. We considered the limit of forwarding to 5 groups (the real scenario), 20

(a) Brazil (b) India (c) Indonesia
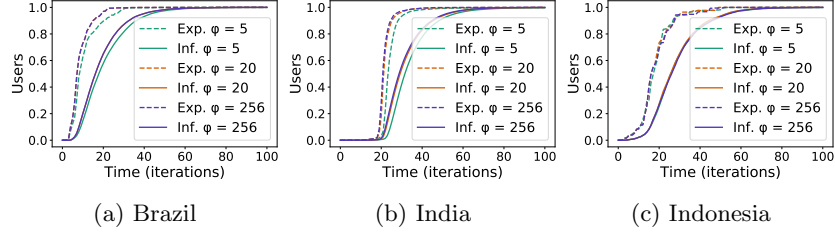
Fig. 4: SEI model varying the forward limit (φ). $\alpha = \beta = 0.1$.

(a) Brazil (b) India (c) Indonesia
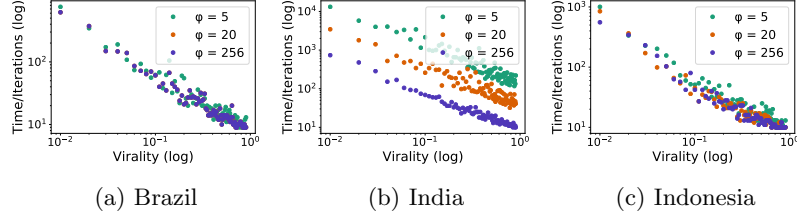
Fig. 5: Time to infect all user in the network on simulations of SEI model varying the virality (α) from 0.001 up to 1.0 .

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, 60 iterations is enough to infect the entire network. Also, 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 for (mis)information with different potential viralities to infect all users. Figure 5 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, i.e., content circulates until it loses attention and stagnates. We add this time limit to our SEI model, calling this period "lifetime", which denotes the maximum duration of an infection in the simulation before it is entirely extinguished. Figure 6 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
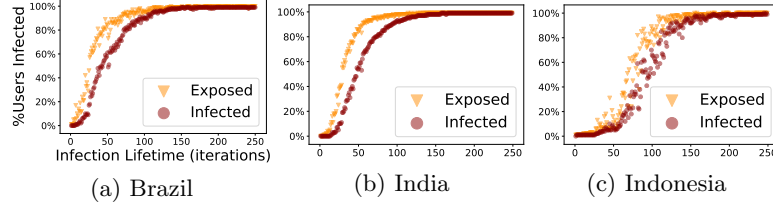
(a) Brazil   (b) India   (c) Indonesia

Fig. 6: Users infected by time in simulations of the SEI model using max lifetime for infections. $(\alpha) = (\beta) = 0.1$. Forward limit $(\varphi) = 5$.
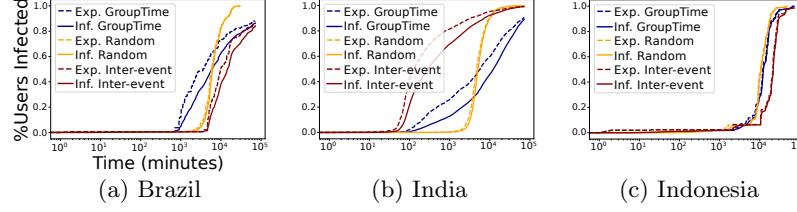

(a) Brazil   (b) India   (c) Indonesia

Fig. 7: Real Time SEI model using "incubation time" before spread infection and each iteration equals 1 minute (log). $(\alpha) = (\beta) = 0.1$. Forward limit $(\varphi) = 5$.

content but were not infected and nullify its virality (e.g. disabling forwarding on that piece of content), thus preventing further contagion.

In the SEI model, the spread of information was measured in terms of the *number of iterations*. In this experiment, 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 7 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 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   Conclusions

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. 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 [1]. 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.

Acknowledgments

References

1. G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto, "(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures," in *The World Wide Web Conf. (WWW'19)*, 2019, pp. 818–828.
2. N. Newman, R. Fletcher, A. Kalogeropoulos, and R. K. Nielsen, "Reuters Institute Digital News Report 2019 ," Reuters Institute for the Study of Journalism, 2019.

3. E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The Role of Social Networks in Information Diffusion," in *The World Wide Web Conf. (WWW'12)*, 2012, pp. 519–528.

4. C. Arun, "On WhatsApp, Rumours, and Lynchings," *Economic & Political Weekly*, vol. 54, no. 6, pp. 30–35, 2019.

5. G. Li and J. Zhen, "Global stability of an SEI epidemic model with general contact rate," *Chaos, Solitons Fractals*, vol. 23, no. 3, pp. 997 – 1004, 2005.

6. A. Bessi and E. Ferrara, "Social bots distort the 2016 us presidential election online discussion," *First Monday*, vol. 21, no. 11-7, 2016.

7. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

8. F. N. Ribeiro, K. Saha, M. Babaei, L. Henrique, J. Messias, F. Benevenuto, O. Goga, K. P. Gummadi, and E. M. Redmiles, "On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook," in *Proc. of the Conf. on Fairness, Accountability, and Transparency*, 2019, pp. 140–149.

9. S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web," in *Companion Proc. of The 2019 World Wide Web Conf. (WWW'19)*, 2019, pp. 218–226.

10. G. Resende, P. Melo, J. C. S. Reis, M. Vasconcelos, J. M. Almeida, and F. Benevenuto, "Analyzing Textual (Mis)Information Shared in WhatsApp Groups," in *Proc. of the 10th Conf. on Web Science (WebSci19)*, 2019, pp. 225–234.

11. P. Melo, J. Messias, G. Resende, K. Garimella, J. Almeida, and F. Benevenuto, "WhatsApp Monitor: A Fact-Checking System for WhatsApp," in *Proc. of the Int'l AAAI Conf. on Web and Social Media*, vol. 13, no. 01, Jul 2019, pp. 676–677.

12. V. S. Bursztyn and L. Birnbaum, "Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups," in *Proc. of the IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019.

13. D. Rushkoff, D. Pescovitz, and J. Dunagan, "The Biology of Disinformation: Memes, media viruses, and cultural inoculation," 2018.

14. K. Garimella and G. Tyson, "Whatapp doc? A first look at whatsapp public group data," in *Int'l AAAI Conf. on Web and Social Media*, 2018.

15. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

16. D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.

17. J. Leskovec, J. Kleinberg, and C. Faloutsos, "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*, 2005, pp. 177–187.

18. R. S. Olson and Z. P. Neal, "Navigating the Massive World of Reddit: Using Backbone Networks to Map User Interests in Social Media," *PeerJ Computer Science*, vol. 1, p. e4, 2015.

19. J. McAuley and J. Leskovec, "Image Labeling on a Network: Using Social-Network Metadata for Image Classification," in *12th European Conf. on Computer Vision (ECCV12)*, 2012.