

Simon Chauchard

Kiran Garimella

Final Draft-September 2023

Collecting WhatsApp Data for Social Science Research: Challenges and a Proposed Solution.

Keywords: WhatsApp - Data donation – Ethics – Anonymization – Data Collection

1. Introduction

Over the past few years, alarming press reports assigning blame to WhatsApp usage for a variety of events have proliferated. In countries like Brazil and India, analyses have repeatedly suggested that group-based interactions on WhatsApp distort beliefs among the electorate (see, for instance Perrigo 2019; Bengani 2019; Benevenuto et al 2018), and beyond, that they impact various outcomes, including (but not limited to) individuals' propensity to engage in hostile, radical or violent behaviors (Chopra 2019, Magenta et al 2018; Ozawa et al 2023).

While academic research – including the contributions in this volume-, has recently started to examine these dramatic narratives, much research admittedly remains to be done to evaluate their validity, and to disentangle the mechanisms through which WhatsApp may or may not be

associated to these outcomes. Social scientists interested in hate speech or misinformation in the Global South will accordingly need access to WhatsApp data, potentially on a large scale, in years to come. Specifically, in order to ascertain the platform's role in the dissemination of problematic content, as well as the consequences of this dissemination, researchers still need to better understand (1) the type and the style of hateful content and/or disinformation that circulates on the platform, (2) the overall volumes of such contents, (3) their degree of virality, (4) the networks through which such contents are most likely to circulate and (5) the political, social and contextual factors in which these contents emerge and have real-world consequences. What is more, they will need to gain access to such data in a way that is practical, legal, and respectful of users' privacy.

Yet, access to WhatsApp data for research remains difficult and rare. While high-quality evidence about Facebook and Twitter users' "information diets" has existed for some time (Barbera et al 2015, Guess et al 2019), no comparable systematic evidence so far exists regarding WhatsApp, despite researchers' long-standing awareness of the fact that the platform is used to disseminate this type of content in much of the Global South (Tucker et al 2018). Besides, when researchers do have access to *some* WhatsApp data, they most likely access samples of data that are too limited in scope to answer all the aforementioned questions, or they do obtain such data in ways that may be suboptimal from an ethical standpoint.

How can researchers thus collect sufficiently interesting data in a way that minimizes these ethical concerns? To answer this question, this chapter presents a possible procedure (and the adjoining tool) to collect vast amounts of WhatsApp data. The *data donation* strategy we

introduce minimizes the practical aspects of WhatsApp data collection, while conforming to dominant norms about privacy. We detail our general strategy and propose a protocol in section 3. In sections 4 and 5, we discuss the pros and cons of that strategy. To set the stage for this, the next section starts by reviewing the challenges associated to WhatsApp Data Collection.

2. Challenges of WhatsApp Data Collection

Researchers eager to engage in WhatsApp Data collection may face technical, legal, privacy-related, and practical challenges.

Some of these challenges however strike us as being harder to overcome than others, and hence worthy of more attention. Technical challenges are, for one, relatively limited: extracting data from private WhatsApp threads is *technically* easy once a thread participant (whether or not they are an admin) consents to extract it: contrary to other platforms, WhatsApp makes it *very* easy for its users to archive the content of the conversations they are part of.¹ In our view, the most serious challenge is equally unlikely to be legal. *How* the platform will react if research we outline in the rest of this chapter becomes common remains to be completely seen. Nonetheless, in our experience, and based on our admittedly limited discussions with WhatsApp representatives, the platform may however be sympathetic to research that allows for the

¹ Concretely, in a matter of seconds, any user can go into a thread, press “export chat” and save the exported data in a variety of formats.

detection of problematic behaviors or for research on the causes of these. This is especially likely to be the case since WhatsApp's commercial promotion of its own encryption would make it difficult for the company to simultaneously research content circulating on the platform. In that sense, we hope that the delegation of this task to external researchers may not only be legally unproblematic but encouraged and supported by the platform.

These legal and technical hurdles notwithstanding, it remains that WhatsApp data collection on a large scale presents serious *privacy-related* and *practical* challenges. We detail these in the subsections that follow.

a. **Privacy-related Challenges**

Since users may easily export data from the threads they are on, and since researchers cannot access private WhatsApp threads without going through a thread's participant, any WhatsApp data *collection* effort needs to be a data *donation* effort. That is, one or several users need to *consent* to give away some data from the threads included in their account, and to engage in a series of actions to export these.

Donating data from one's own account may be problematic from a privacy-protection point of view insofar as it may contradict guidelines, norms or laws protecting individuals' privacy or limiting the processing of individuals' personal data. The European Union's General Data Protection's Regulation (GDPR) currently constitutes the main example of such regulation, though privacy laws around the world – such as India's Personal Data Protection Bill, Brazil's

LGDP or Canada's DCPI - echo most of the principles at the heart of the ruling when they exist. Besides, when it comes to the handling of users' personal data, principles of user consent, anonymization and limitations on the use that can be made of such data will likely deserve a discussion, whether a local equivalent to the GDPR exists or not.

So why and in what ways might donating data from one's own WhatsApp account violate the privacy of users, as defined in these norms? Though we acknowledge that norms and regulations will differ depending on the case chosen by researchers, we generally see five potential issues with a *WhatsApp* data donation program that relies on donations by what we will hereafter refer to as a consenting "gateway user":

- (1) the issue of consent (or the lack thereof) of third-party group participants (that is, users who are not our gateway into a group but whose data -- phone numbers, profile pic, messages, among other data -- nonetheless feature on threads). While gateway users consent to give away their data, they cannot speak for other users whose data will enter the dataset through *their* data donation. This suggests that consent should be obtained from these third-party group participants, which may be impossible or undesirable from a methodological standpoint, or that the data of third-party users should be credibly anonymized so as to make their data unidentifiable.
- (2) the issue of *how much* data should a research project be allowed to collect. Within the GDPR framework, this would for instance be related to the "*data minimization*" principle -- that is, the idea that personal data collected should be limited in time and scope to what

is directly relevant and necessary. But more generally, in the context of WhatsApp, this may indirectly raise the question of the type of threads (one-on-one vs. group, private vs. public etc...) that may be collected for research.

- (3) The issue of the anonymization strategy – that is, the strategy used to credibly anonymize stored data and minimize the potential for re-identification, whether this is regarding gateway users or third-party users. This is because researchers may want to or need to conceal the identity of discussion participants to protect their privacy. While protecting privacy is in and of itself important, this issue is especially likely to become an important issue if the data are later made available to others, as will tend to be the case under increasingly frequent open science agreements.
- (4) the issue of data protocols *pre-anonymization* – that is, what strategies will the researchers rely on when they need to transfer and store data in its pre-anonymization form. Concretely: how will researchers ensure that these data cannot be accessed by unauthorized actors or lost *before* they are anonymized?
- (5) the issue of “unexpected findings” and findings that are subject to legal disclosure obligations under international or local law. This raises the question of what protocols will researchers have in place if/when they stumble on data that are subject to legal disclosure obligations.

b. Practical challenges

We add to these ethical and privacy-related concerns several *practical* challenges that researchers would face to make such a WhatsApp data donation program sufficiently useful from a research standpoint, in line with the balancing principle enumerated by Ohme and Araujo (2022):

- (1) Firstly, there are various technical issues to ensure the smooth, rapid and private donation of such data. Concretely, any effort to obtain data from gateway users is likely to fail (low participation rates, for instance) if this is a tedious, expensive and/or time-consuming process. Similarly, it is also more likely to fail if the donation protocol requires that an enumerator or another associate of the research team present during the data collection scrolls through the data or accesses it in any way in its pre-anonymization form.
- (2) A second and related challenge relates to the ability of researchers to convince a broad (and ideally representative) sample of gateway users to donate some of their data. To put it simply, the ambitious research goals enumerated above would require us to obtain data from a sufficiently diverse cross-section of the population in order to be able to reach any scientifically valid conclusion about the populations targeted. This may require costly efforts by the research team to ensure that the sample of donors is sufficiently interesting.

3. A Possible Strategy

How can researchers overcome these *many* challenges?

As mentioned above, given end-to-end encryption, any ethical WhatsApp data *collection* effort by design needs to be a data *donation* effort. Considering the challenges we listed, this will additionally need to be a data donation strategy that facilitates a relatively effortless donation, in a privacy-preserving manner, and in a way that inspires confidence among a diverse group of donors.

As part of an ongoing research project requiring the collection of large amounts of private WhatsApp threads (the ERC POLARCHATS project²), we have spent much time developing such a solution over the past few years. As part of this process, we have developed a dedicated web interface called *WhatsApp Explorer*. While further tests will be needed to validate our strategy, these efforts will arguably allow us to minimize all these challenges while allowing to amass data that is interesting enough for research.

a. General Principles

We detail the broad principles of this strategy in this section, before getting into the weeds of the data protocol in the next section. Our general strategy is to contact users and ask them to donate

² The ERC-funded POLARCHATS project (2022-27) documents the extent of the misinformation crisis in India and Brazil and explores the causes and consequences of exposure to this misinformation. The project relies on qualitative insights, quantitative descriptions, and experimental methods to achieve these objectives.

some of their WhatsApp data for social science research. The key technical innovation of the project is to make this donation process relatively seamless for consenting users, so that they may donate the data they wish to donate with minimal effort, with the assistance of a research associates who will assist in the donation, but never access the data.

To reward users for their time and contribution to research, we provide them with small amounts in phone credits. We also provide them with extensive guarantees regarding privacy and anonymization and highlight that their (anonymized-at-the-source) data will *at no point* be shared beyond the main members of the research team. Importantly, in the design we outline below, no field staff (enumerators and local partners) will have access to the data collected. That is, while field staff make the data collection possible, the data never transit through their own devices (they are instead instantly encrypted and uploaded to a secured server that only we PIs have access to).³

Importantly, we also refrain from asking users to donate one-on-one threads, and concentrate on *group threads*, to limit privacy concerns (details in the full protocol below)

To overcome privacy-related challenges, we have devised an extensive strategy to anonymize data *as it is uploaded to our servers*. We do not store any raw de-anonymized data.

With regards to text data, we anonymize any personally identifiable information like the names, phone numbers, and emails from the dataset. The anonymization is done through a state-of-the-

³ Nor do field staff subsequently have access to the server on which the data are securely stored.

art privacy preserving algorithms which are a well-established and widely used library provided by *Google* called the *Google Data Loss prevention API*.⁴

Regarding visual content, we proceed to irreversibly anonymize most images we store as we upload it, with the exception of images/videos which are shared by at least k groups/threads (say $k = 5$) in our data. This ensures that we do not access the vast majority of un-anonymized visual content. Importantly, the viral content we keep and analyze is extremely unlikely to be personal or private content, as it will be, by definition, content that is shared on many online communities. To anonymize visual content, we proceed in several steps which we detail below. We first use automated tools to systematically blur faces and a few additional identifying features of images/videos (for instance, car plates). We also implement a second, human-supervised stage of anonymization BEFORE analyzing the data, to strengthen an already thorough anonymization strategy.

b. Illustration: a Possible Data Protocol

Here are how these general principles may translate into a concrete data collection protocol.

While a fully-online process may be possible⁵, we think an in-person process may be more adapted to most global south contexts, in which most users will not own the hardware necessary (concretely, two screen-equipped devices, whether these are phones, tablets, laptop or desktop

⁴ Full technical specifications are below.

⁵ We are currently exploring this strategy in one of our study sites to complement the in-person strategy we outline here.

computers) to complete the process themselves online, or will not have the skills to do so. In addition, in the global south or elsewhere, we note that an in-person process may be necessary to efficiently deliver guarantees about privacy and to generate trust among respondents.

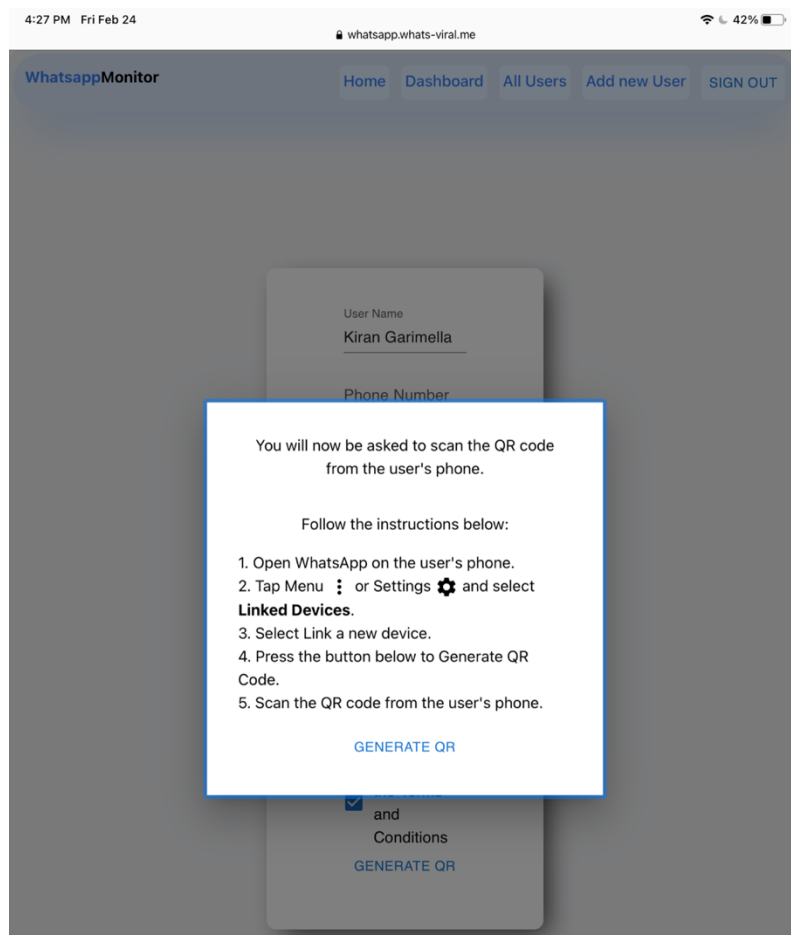
For this reason, we focus in what follows on describing a protocol for in-person collection. How would this look like? Concretely, a research associates (a trained enumerator from a partner survey firm) would visit *randomly* selected citizens face to-face (at their residence) and offer them to participate in a research study about their social media activity, particularly with regards to discussion groups they are part of and about the content that circulates on these groups.

Concretely, this is how we envision the data collection will look like, step-by-step:

- Field enumerators explain the goal of the study to the individuals contacted and ask for consent, using a standard consent procedure. At this time, they also provide respondents with a printed flyer explaining who the researchers are, what their goals are, and how they can be contacted. This includes logos of all partner organizations, a link to the relevant registry of data processing activities, detailing the research plans and the legal basis for it, a hotline phone number for asking questions and extensive details on our anonymization strategy in relatively untechnical language. Finally, this document also contains clear technical instructions on how to *end* participation in the project.⁶

⁶ A login code is stored to obtain data for 2 months after which it is automatically deleted. The users however have a chance to logout any time before that on their own phones, using the instructions detailed here: https://faq.whatsapp.com/539218963354346/?locale=en_US

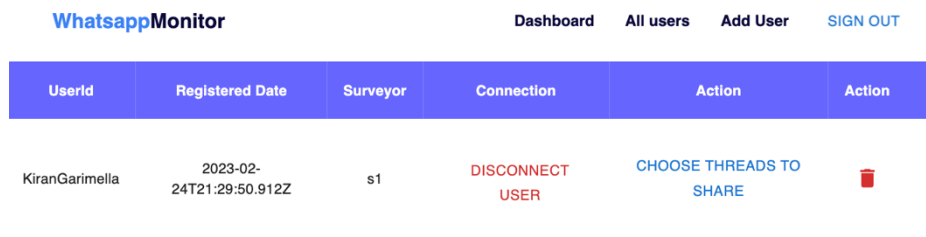
- IF they consent to participate, the enumerator requests that the respondent scans our generated QR code through their *WhatsApp* App on their own smartphone. Concretely, the research associate – using the web interface we designed (<https://whatsapp.whats-viral.me/>) – generates a QR code and asks the respondent to scan it with their phone (this is easily done within *WhatsApp* through the “linked device” function anyone can use to connect their WhatsApp, for instance, on a computer).




Note: the early version of the interface was named WhatsApp Monitor, as shown here.

Importantly, throughout the process we describe here, the enumerators at no point need to handle the respondents' devices nor to see their content.

- Once this is done, the enumerator connects with the user's WhatsApp, by pressing "connect user" on this screen.

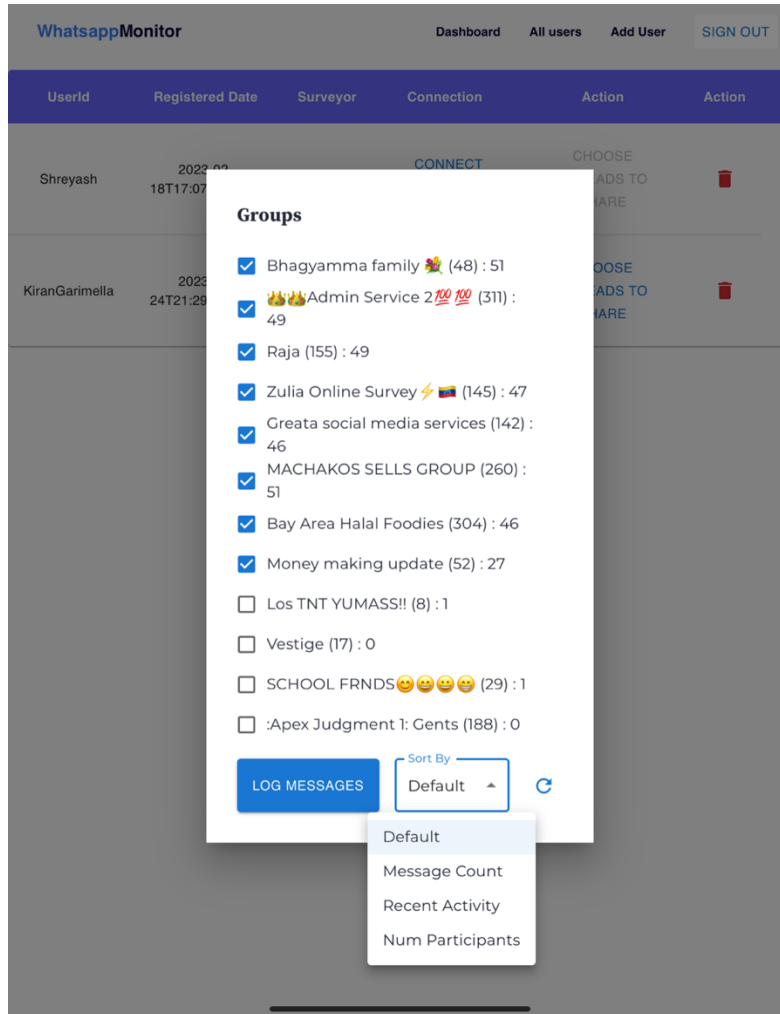


The screenshot shows a web interface for 'WhatsappMonitor'. At the top, there are navigation links: 'Dashboard', 'All users', 'Add User', and 'SIGN OUT'. Below this is a table with the following data:

Userid	Registered Date	Surveyor	Connection	Action	Action
KiranGarimella	2023-02-24T21:29:50.912Z	s1	DISCONNECT USER	CHOOSE THREADS TO SHARE	

Note: the early version of the interface was named WhatsApp Monitor, as shown here.

Once this is done, the enumerator presses "choose threads to share":



At this point the enumerator can show the respondent on the tablet they use for the survey the full list of WhatsApp group names that exist in the respondent's *WhatsApp*, without yet having access to their content or to any other information besides metadata.[1] Once more, throughout the whole process, the enumerators do NOT have access to the content of the threads, as these can be directly uploaded to a server that only the PIs will have access to, at the end of the process we describe in what follows.

Importantly we at this stage automatically *exclude* one-on-one threads to minimize the amount of data we manipulate, and to further protect privacy.⁷

We can order these group threads in three different ways: by date of most recent post, by number of people in a group and by total number of posts over the past two weeks.

On this screen, we by default present all threads with 6 or more participants and that count 10 or more messages over the past two months to the participants.

However, this interface allows participating users to select themselves the groups they are willing to donate by ticking the corresponding box in the right panel.

- We at this point ask participants to share data from the subset of these groups that they did select for the two months before they entered the program (i.e., two months before the date of their interview), and for two months going forward, though we readily note that they may themselves choose to restrict their donation to either part of this (as shown below). That is, provide us with *only* past or *only* future data, in addition to only sharing a subset of these groups.
- As this inclusion/exclusion process takes place, our interviewer asks which type of data they are willing to give:
 - a) *Historical* data (past two months)

⁷ Incidentally, we also believe this exclusion to constitute an important incentive for respondents to participate, as their most private data are likely to be in these one-on-one threads.

- b) Future data (in the next two months)
- c) Both historical and future data.

If the respondent indicated willingness to share data going forward, we explain to them what they ought to do to ensure that the collection happens over the next two months, as well as what they can do to ensure it does *not*, should they change their mind.

Importantly, they can take these steps *in seconds, on their own phones* should they want to disconnect. This is explained in detail in the draft flier we distribute when asking for users' consent. The enumerator may also demonstrate it in person.

- Once this inclusion/exclusion process is completed, the enumerator presses the “log messages” button.

Doing so leads to two things:

- An upload of the last two months of content of the selected threads onto our secured server (details below), in an anonymized manner (details on anonymization strategy below). That is, we never upload nor store non-anonymized content. The content is encrypted as it is transferred from the field staff's devices to our cloud server.⁸ This

⁸ Concretely, we rely on the following procedure: 1. The unencrypted data is exported and temporarily saved in a location that the research team cannot temporarily access as its access is temporarily protected. 2. An automated procedure encrypts it. 3. The encrypted copy is stored safely in the research team's server. 4. As soon as step 3 is completed, the unencrypted data that was exported and temporarily saved is permanently deleted.

guarantees that neither the PI nor research staff ever has access to such content, either now or in the future.

- An anonymized mirror copy of the selected threads, allowing us to collect data on these selected threads *going forward*. Importantly, we see it as important to set such a limit for data going forward, after which the process is automatically disabled, with the user being disconnected and their contact information deleted from our database. Importantly, we do not have access to the data until the data collection period is concluded and/or the user disconnects. At that time, the temporarily saved unencrypted data which is saved in a location we cannot access is encrypted and sent to our server, where it is stored (and the unencrypted data is permanently deleted, as described above).
- At this stage, once messages/threads are logged, respondents answer a brief series of questions about up to 20 of the threads whose data were harvested, and about their own demographic characteristics.
- Within a few weeks, they receive a reward (for instance, phone credits) directly on their phones. When they do, they once again receive contact information for the “hotline” and a link towards information about the data donation program.

4. Advantages

Why do we believe this strategy to constitute the right compromise for WhatsApp data collection, considering the ethical and practical challenges we listed above, and the need to strike the right balance between privacy and the need for research, as enumerated by Ohme and Araujo (2022)?

After several pre-tests, we are convinced that our strategy technically works, and that it is practical and rapid (hence solving the first practical challenge).

More importantly, we are confident that using *WhatsApp Explorer* (and its adjoining protocol) strikes the right balance between the imperious need to collect WhatsApp data and respect for privacy and ethical standards.⁹ This is because of 1. the extensive, multi-stage anonymization we put in place, 2. the development of clear procedures to handle “unexpected findings”, and 3. the procedure’s respect for the data minimization principle. We detail each of these points in the following subsections.

a. Strong Anonymization and Privacy Protection

⁹ We reach this conclusion after a year-long ethical review process that involved the ethical review experts of the European Research Council (ERC), the University Carlos 3 Madrid’s data protection officer (José Furones, whom we are especially grateful for, for his dedication to the project), the University Carlos 3 Madrid’s Ethical Review Committee, and the Ethical Review Committee of the Fundação Getulio Vargas in Brazil. Importantly, all three institutions have now officially given us a green light on the design and protocol, hopefully creating a useful precedent within the GDPR (or the LGDP, Brazil’s equivalent of GDPR) frameworks for research of this type.

To understand why and how our anonymization strategy arguably protects the privacy of users, it may first be useful to fully detail what data are collected through this strategy, as well as how and when it is anonymized.

As part of this process,, we collect information on:

- who users are connected to (i.e. users address book).
- who they chatted with.
- how many groups they are part of.
- the respective size of these groups.
- and most importantly, the content from chats the users consented to share. The information we get from the content of the chats includes messages, images and videos exchanged in the chat and the timestamps of when they were sent and who sent them.

These data however go through several stages of anonymization. A first, automated anonymization stage happens before it is stored on our servers.¹⁰ At that stage, we rely on automated procedures to anonymize any personally identifiable information like the names, phone numbers, and emails from the dataset. The anonymization is done through a state of the art privacy preserving algorithms which are well established and widely used library provided by *Google* called the *Google Data Loss prevention API* . Each bit of sensitive information is encoded and replaced with a unique identifier. While it can technically be used for re-identification, we delete the original key to make this impossible as soon as we have verified that

¹⁰ On servers: the current plan is to store these data on UC3M servers, for which the necessary guarantees in terms of privacy and data transfer are already in place.

the data is safely stored. We do not store group icon pictures (i.e., profile pictures), nor audio messages.

In addition, for pictures/videos included *within the threads* that we collect, we create the following pipeline:

1. We start by storing the pictures and videos securely on our servers but also create hashes of them.
2. We then use the aforementioned hashes to identify if the same image/video were in data shared by multiple users.
3. We proceed to irreversibly anonymize most images we store, with the exception of images/videos which are shared by at least 5 groups/threads in our data AND that do not contain personal data (for instance, a nude that was forwarded around). Since this eventually be limited to a small number of items, the PI will check each of these images individually to decide whether this is the case. Overall this procedure ensures that we do not access the vast majority of un-anonymized visual content. Importantly, the viral content we keep and analyze is extremely unlikely to be personal or private content.

To anonymize visual contents, we rely on tools such as [Brighter AI](https://brighter.ai/)¹¹ to blur out faces. Such tools provide us with an automated (and hence convenient) procedure (the data never leaves our servers during this process) to blur faces and a few additional identifying features of images/videos (for instance, car plates). We at that point replace the un-anonymized images and

¹¹ <https://brighter.ai/>

videos stored on our servers with these anonymized images and permanently delete the un-anonymized originals.

Even with this extensive protocol, we however must acknowledge that *perfect, foolproof* anonymization is never possible, nor that it can be left to automated procedures alone. Hence, we also implement a second stage of anonymization, this time human-driven. We implement this systematic anonymization audit (SAA) BEFORE analyzing the data, in order to potentially strengthen - and hopefully perfect - an already thorough anonymization strategy.

Concretely, the PI and close associates with knowledge of the context will systematically review all the text and visual content already anonymized using automated methods and evaluate the potential for re-identification of personal data.

Our strategy is to start by further anonymizing the text content. We systematically remove any mention of:

- A location (neighborhood, city, region etc..)
 - To do so we replace locations with the marker [LOCATION]
- identification numbers (for instance a passport number or another ID number, though we will add to the list.)
- Mentions of an individual *physical, physiological, genetic, mental, economic, cultural or social attribute*.

We also conditionally redact:

- Information relating to individuals' possessions, *if it is likely to enable identification* (because it is rare and hence distinctive).
- Information relating to individuals' company or social network, *if it is likely to enable identification, (again, because it is rare and hence distinctive).*

Once this is done, we proceed to further anonymize the visual content, where we see a need to do so. While almost all identifying content has been removed from the textual part of the WhatsApp thread, this should be a necessary additional step in some cases.

Concretely, we make sure to blur:

- Street signs or store/office fronts indicating or providing hints / location.
- Potentially distinctive or atypical landmarks in the background.
- Dress of individuals if it is distinctive or identifying.
- Distinctive body marks (scars, tattoos, etc. if not on face since face will be already blurred).

And pledge to add to this list as need occurs.

b. Provision for Unexpected Findings

How do we deal with unexpected findings?

While we see this probability as low ex-ante, we also recognize that either this data collection project or the analysis of said data might lead to some “unexpected findings”, that is, findings that fall outside of the scope of the principal research objectives but necessitate action on the part of the researchers, for instance disclosure of information to appropriate or designated authorities.

In the context of a project on social media and violence, these ethics issues may be considered as ‘serious and/or complex’ if the research yields, for example:

- social or behavioral unexpected or incidental findings that may require interventions to safeguard the well-being of research participants (e.g. signs of physical abuse, self-harm, or drug dependency or neglect in minors).

OR

- findings that are subject to positive disclosure obligations the national laws of countries in which the research takes place, requiring researchers to breach the confidence of research participants. Examples include criminal conduct such as crimes, child sexual exploitation, human trafficking or terrorism.

If we detect such content, the research team will within 3 days (from the time of discovery) consult with the POLARCHATS project ethics advisory board (which contains both Brazil and India specialists) to establish the best way forward on a case by case basis. We refrain from

establishing a blanket policy ex-ante considering the rapidly changing political landscapes in both countries and in light of the potential for political bias in the judicial system.

c. Restraint in Amount of Data Collected

As explained above, we *ask* participants to share *up to* four months of data (2 months prior, two months after) on a very specific subset of their threads (i.e. threads with 6 or more participants and that count 10 or more messages over the past two months – ALL OTHER THREADS, including one-on-one threads, are altogether excluded from the data collection), though we also provide them with an easy way to share only a subset of this subset.

That is, they may share either historical data or data going forward; they may also exclude any thread on the list we initially present them with, and as noted above, as many threads on that list as they wish. We also note that they may quit the program at any time after they consented.

This means, conversely, that we do not remain indefinitely connected and that we expressly restrain from collecting data from certain types of threads. We believe these parameters to constitute the right compromise between 1. the data minimization principle, 2. The feasibility of our anonymization-intensive strategy and 3. our ability to conduce meaningful scientific enquiry – and especially statistical analyses - in the public interest.

5. Remaining Challenges & Conclusion

In sum, the strategy we detail here should – once the tool is made available in the next few months – provide researchers with an efficient, privacy-protecting and secure methodology to collect WhatsApp data to answer a variety of research questions. In that sense, we hope our work will help develop research on the platform and its uses around the world.

Of course, we acknowledge that this strategy has many severe limitations. First, we recognize that the multiple, extensive stages of anonymization we implement eventually fall short of eliminating 100% of the possible risks of identification of the individuals involved. We however believe it comes extremely close to doing that, in practice, and note that researchers willing to undertake WhatsApp research must, in one way or another, be willing to deviate from the strictest guidelines about privacy protection in order to make research in the public interest possible, or must be willing to interpret these guidelines creatively.

Second, we still lack sufficient data to speak to the representativeness of the data we will eventually manage to extract. Until a larger study is run, we will remain unclear as to whether the strategy will for instance function among some demographics, and the extent to which respondents will be selective in terms of the groups they choose to donate.¹² There is in addition little doubt that researchers focusing on populations by nature difficult to investigate (say for instance, members of a rebel army or of a vigilante group, as several authors in this volume) will

¹² Note that this may be one further argument for an automated online self-administered procedure, which would likely allow us to access different demographics.

continue to struggle to obtain data to study the influence that WhatsApp networks may have in these processes. Our technology may not entirely change the reticence that many users may have when approached and asked to donate their smartphones' content.

Third and relatedly, our strategy is costly in labor, infrastructures, and resources, especially if researchers are going to provide rewards or incentives to potential donors. This implies that many researchers relying on it will not be able to collect large and/or representative datasets in the case of their choosing.

In spite of these important limitations, we believe the technology we present in this chapter, and which we will keep improving over the next few years, will dramatically improve current research opportunities and practice. Our early experiments in the field in India and Brazil on our own project (ERC Polarchats) suggest that we will be able to obtain large datasets from a diverse, if not representative, group of users. This is, in and of itself, an improvement over the status quo, and one that should allow us to answer some important research questions and monitor the virality of problematic contents on the app.

Further, while we acknowledge that most researchers will not be able to collect as much data as we plan to due to the rather costly nature of our strategy, we also hope it will help set the standards for *how* to collect WhatsApp data, regardless of the amount of data collected by specific researchers. Important discussions about consent, privacy and anonymization are at stakes and need to be balanced with the impervious need to access this data to document and analyze some pressing dangers. Even if researchers assemble datasets more limited in scope than

the ones we are planning to assemble, we believe their strategy should equally go through this balancing exercise and provide clear safeguards to users. In that sense, we hope this chapter will push researchers to reflect on what “fair” WhatsApp data collection should look like – a thorny issue we have tried to solve –, in addition to assisting their practical needs.

References

Barberá, Pablo, and Gonzalo Rivero. (2015). "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* 33 (6): 712-729.

Barbera, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. (2015.) *Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?* *Psychological Science*, 26 (10): 1531-1542.

Benevenuto, F., Ortellado, P., & Tardáguila, C. (2018), [Fake news is poisoning Brazilian politics](#). WhatsApp Can Stop It. *The New York Times*.

Bengani, P. (2019, October 16). India had its first 'WhatsApp election.' We have a million messages from it. *Columbia Journalism Review*. Retrieved May 9, 2023, from https://www.cjr.org/tow_center/india-whatsapp-analysis-election-security.php

Chopra, Rohit (2019). "In India, WhatsApp is a Weapon of Antisocial hatred." *The Conversation*, April 23 (2019).

Guess, Andrew M, and Benjamin A Lyons. 2020. "Misinformation, disinformation, and online propaganda." *Social media and democracy: The state of the field, prospects for reform* 10.

Guess, Andrew, Joshua Tucker and Jonathan Nagler (2019). "Less than you think: Prevalence and predictors of fake news dissemination on Facebook", *Science Advances*.

Magenta, Matheus, Juliana Gragnani, and Felipe Souza (2018), "How WhatsApp Is Being Abused in Brazil's Elections," October 24, sec. *Technology*. (<https://www.bbc.com/news/technology-45956557>).

Ohme, Jakob, and Theo Araujo. "Digital data donations: A quest for best practices." *Patterns* 3.4 (2022): 100467.

Ozawa, J. V. S., Woolley, S. C., Straubhaar, J., Riedl, M. J., Joseff, K., & Gursky, J. (2023). How Disinformation on WhatsApp Went From Campaign Weapon to Governmental Propaganda in Brazil. *Social Media + Society*, 9(1), 20563051231160630. <https://doi.org/10.1177/20563051231160632>

Perrigo, Billy (2019). How Volunteers for India's Ruling Party Are Using WhatsApp to Fuel Fake News Ahead of Elections. *Time*, January 25 (2019).

Tucker, Joshua A., Yannis Theocharis, Margaret E. Roberts, and Pablo Barberá. (2017) "From liberation to turmoil: social media and democracy." *Journal of democracy* 28 (4): 46-59.

Tucker, Joshua Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan (2018). *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*. Hewlett-Packard Foundation.