The Privacy Implications of Whatsapp's Exposure to Brute Force Scraping

Introduction

Mobile messaging applications provide a free alternative for mobile texting on smartphones. WhatsApp is one of the leading mobile messaging applications these days. In the application, each user is characterized by a unique MSISDN number and may personalize their profile by defining a personal photo and an "About" text. The default privacy settings for these details is "Everyone", meaning that any other user of the application has access to these fields when querying for said user, raising the concern that companies and governments may exploit this easily for commercial purposes and nation-state espionage.

In this study, we collected MSISDN, profile photo and profile "About" text from users located in several locations around the world. The purpose of the study is to research the possible threats that WhatsApp users who use the default settings are exposed to. In the study, we collected and analyzed data of about 68k users from Europe and 34k users from the USA (total: 102k). The phone prefixes which were chosen to collect the study's users were selected at random. With more time and resources, the following method can be done on a much larger scale.

Related work

Over the last few years, researchers have begun to examine Whatsapp, and other mobile messaging applications, including research about privacy risks. For example, Andreas Buchenscheit et al. detected that sharing of presence information alone (e.g., being "Online" or not) is sufficient to identify daily routines and conversation partners [1]. Paspatis et al. examined a method to identify unknown Viber users by only processing self-exposed public data. Their method included using public and free databases and Google's reverse image lookup and revealed users' personal information such as their real name, address and other personal data - with a success rate of 75 percent [2]. The research by Malekhosseini et al. focused on the extraction of patterns in users' "About" text content. In their research, they found ten themes, including Emotional, Engagement moment/connection options and Personal information. In addition, their research found that users who have used themes with neutral polarity have the highest average privacy settings [3].

Methodology

In order to collect and analyze data about Whatsapp users, the work was divided into three main stages:

- Preparation/Offline Stage - We found telephone number ranges that are used for mobile communications services, for example, 376-3-XXXXX and 376-6-XXXXX

in Andorra [4] [5]. These telephone numbers ranges were converted to vCard files and thus transferred to our lab smartphone and loaded onto it (see Figure 1). In the Whatsapp application, the simulated contacts from the imported vCard file will be treated as MSISDN identifiers. It's essential to note that the mobile app allows a massive import of contacts and this feature was exploited in our research.

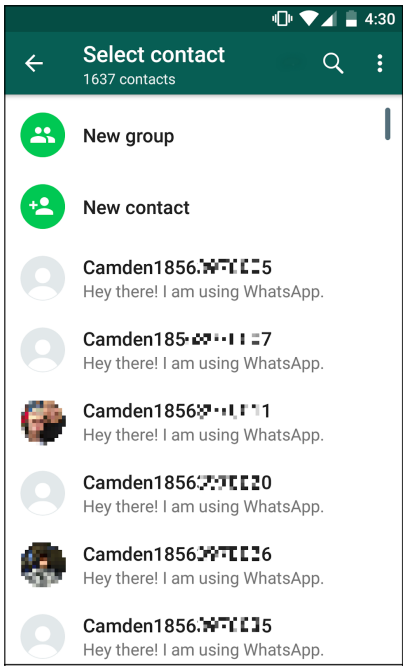

Figure 1: *The output of Preparation/Offline stage.*

- Collection/Online stage - We used a regular laptop as a server. From the server, we connected to WhatsApp's web platform, also known as Whatsapp Web (see Figure 2). After the first stage, when contacts were added to the lab smartphone, the Whatsapp app on it had access to the publicly available user data of said contacts. A Python program was written to connect through Whatsapp Web and collect this data (profile photo and user "About" text) correlated to each telephone number.

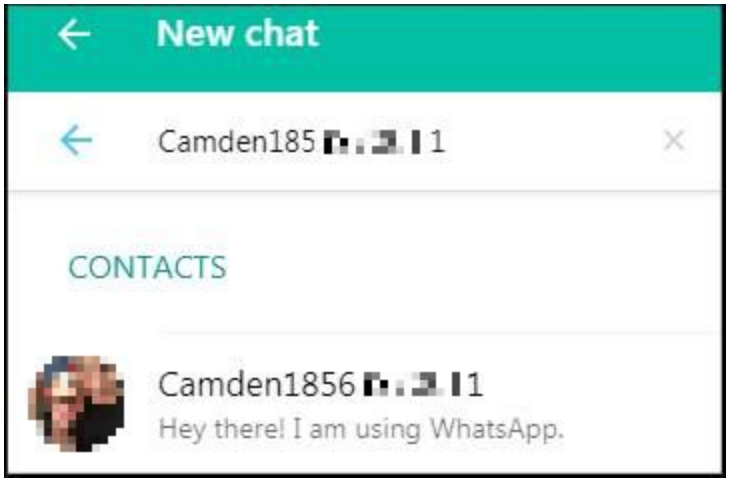

Figure 2: *Search in WhatsApp Web platform.*

- Analyze/Processing stage - Google has publicly available services to help characterize different digital content types. A python program was written to utilize Google's AI and machine learning's products, Google Vision and Google, in order to analyze the collected users' "About" texts and images (see Figure 3).

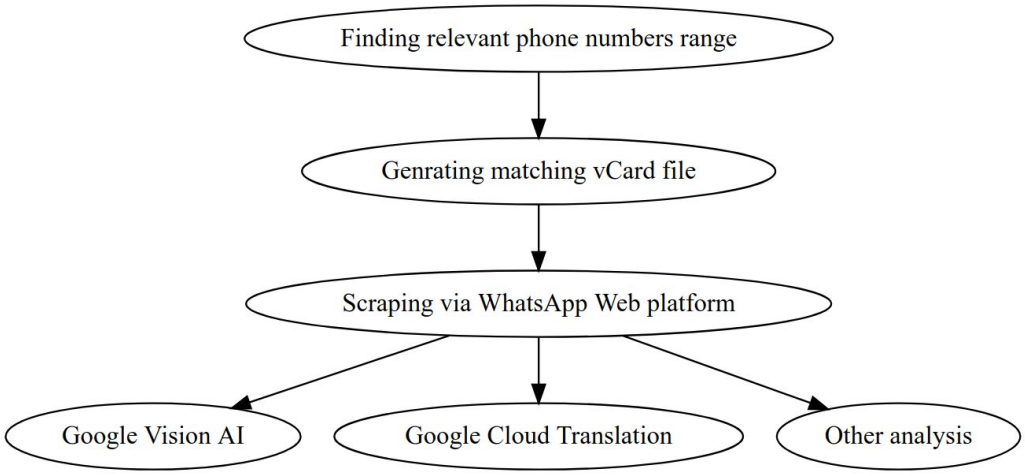

Figure 3: *The general process.*

Data collected per telephone numbers prefixes:
Andorra prefixes: 376-{3,6}-XXXXX
Liechtenstein prefixes: 423-{770, 777, 780-789, 791-794}-XXXX
Naples, Florida prefixes: 1239-{206, 207, 227, 248, 272, 293, 315, 331, 398, 404, 595, 682, 776, 844, 877}-XXXX
Camden, New Jersey prefixes: 1609-{504, 636, 841, 876, 932}-XXXX, 1856-{289, 397, 408, 445, 564, 571, 613, 676, 883, 952, 986}-XXXX

Amount of collected users per prefix:

Prefix	Amount of collected users
Andorra	57,635
Liechtenstein	10,647
Naples, Florida	24,069
Camden, New Jersey	10,255

Discussion

It's essential to note that what has been described can and might be implemented on a much larger scale. Methods like the one described may be exploited by various threat actors. The collected data is useful for both commercial purposes and nation-state espionage.

Type of collected data per prefix:

	Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
Image and “About” text	46%	44%	43%	44%
Image only	21%	17%	7%	4%
“About” text only	9%	16%	35%	38%
None	24%	23%	15%	14%

- None (not image and not “About” text) - Only the fact that there is a Whatsapp user identified with an MSISDN number may indicate that the said number is active. These phone numbers might be used as a target for spam messages, phishing messages and the like.
- “About” text - It’s possible to use the “About” text for social engineering purposes. Even if the “About” text is a default one - it’s possible to learn from its language the device’s or the user’s language.

To demonstrate, consider the following table of most popular language for each prefix:

Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
English (43%)	German (45%)	English (88%)	English (91%)
Spanish (27%)	English (44%)	Spanish (8%)	Spanish (6%)
French (13%)	Italian (9%)	French (2%)	French (<1%)
Catalan (7%)	Spanish (6%)	Portuguese (<1%)	Portuguese (<1%)

An issue that we observed during the research is how many users chose to share a name in their “About” text. We intersected the names in the SSA (The United States Social Security Administration) database [6] with “About” text, and this is the rate of the users who shared a name in their “About” text:

Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
0.8%	0.6%	0.1%	0.1%

- Images - It’s interesting to look at which characteristics of images are shared in similar rates in the Europan and American samples, and which are shared mainly only in one of them.

  Labels that Google vision classified in similar ratio in Europe and USA:

	Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
Headgear	2.2%	2.2%	2.2%	2.3%
Child	7.2%	6.4%	7.2%	6.6%
Toddler	3.3%	3.1%	3.7%	3.3%

  It is worthwhile to pay attention to the relatively high rate of users who chose to share images that Google vision classified as “Child”. These can be exploited by pedophiles for spying and etc.

Labels that Google vision classified more in Europe than in the USA:

	Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
Snow	5.2%	5.3%	0.2%	0.3%
Mountain	3.4%	5.0%	0.3%	0.3%
Wilderness	2.6%	2.7%	0.3%	0.4%

Labels that Google vision classified more in the USA than in Europe:

	Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
Lip	4.6%	3.1%	12.6%	10.6%
Black hair	2.0%	0.7%	6.4%	4.5%
Eyebrow	13.5%	10.5%	30.9%	27.6%
Hairstyle	7.4%	6.2%	16.7%	15.2%
Suit	1.0%	1.2%	2.4%	2.3%

Another point that is worth paying attention to is the high rate of users who chose to share images that Google vision classified as "Face":

Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
26.6%	22.6%	44.9%	42.0%

It’s impossible to decide based on the collected data which of the faces belongs to the users, we’d like to take a careful guess that most of the faces are the faces of the users or their family members. A possible threat is that these images might be used for face recognition, e.g search phone number by face, a tool which may be used for spying and commercial purposes.

An additional possible exploit is finding WhatsApp’s users in social networks based on the same image (users who uploaded the same image to more than one platform). A further study could try accomplishing this method by using Google's reverse image lookup.

Another possible exploit is targeting users by the images they shared in Whatsapp, for example:

	Andorra	Liechtenstein	Naples, Florida	Camden, New Jersey
Glasses	8.1%	9.7%	11.5%	11.0%
Child	7.2%	6.4%	7.2%	6.6%
Dog	4.0%	3.1%	1.8%	2.6%
Toddler	3.3%	3.1%	3.7%	3.3%
Muscle	1.9%	1.5%	2.8%	3.2%

So it's possible to advertise dog food for people who uploaded an image of a Dog, for example. The advertisement can be executed by direct marketing (Telemarketing, SMS messages, etc) and targeted advertising.

Future work

An interesting progression of this research may be work in the spirit of Andreas Buchenscheit at el.'s paper [7]: Developing a prediction mechanism based on the data users share voluntarily in WhatsApp to find highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, age, etc. Another direction may be analyzing the semantics behind the emoji characters in the user "About" text, which were not taken into consideration in this paper.

Conclusion

In this research, we showed some possible utilizations of collecting images and "About" text of Whatsapp's users who use the default "Everyone" settings.

The mobile app allows a massive import of contacts and this feature was exploited in our research. It's possible to fix this breach by changing the default privacy settings to "My contacts", limiting the number of contacts a user may hold in the application or limiting the number of contacts a user is able to see per day. The thought that various threat actors (commercial, nation-state espionage, private investigators, etc) may hold massive data about Whatsapp's user is worrying.

References

[1] Buchenscheit ,A., Könings B., Neubert A., Schaub F., Schneider M., Kargl F. (2014). Privacy implications of presence sharing in mobile messaging applications. MUM 2014: 20-29.
https://doi.org/10.1145/2677972.2677980

[2] Paspatis I., Tsohou A., Kokolakis S. (2017). Mobile Application Privacy Risks : Viber Users' De-Anonymization Using Public Data.

[3] Malekhosseini, R., Hosseinzadeh, M., Navi, K. (2018). Evaluation of users' privacy concerns by checking of their WhatsApp status. Software: Practice and Experience, 48(5), 1143-1164.

[4] Andorra's National Numbering Plan:
https://www.itu.int/oth/T0202000005/en
Liechtenstein's National Numbering Plan:
https://www.llv.li/inhalt/11098/amtsstellen/nummerierung

[5] Search area code exchange by ratecenter state, TelcoData.us.
https://www.telcodata.us/search-area-code-exchange-by-ratecenter-state

[6] The "national data" at this link:
https://www.ssa.gov/oact/babynames/limits.html

[7] Kosinski, M., Stillwell, D., Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.