WhatsAnalyzer: a Tool for Collecting and Analyzing WhatsApp Mobile Messaging Communication Data

Anika Schwind, Michael Seufert*
Insitute of Computer Science, University of Würzburg
Würzburg, Germany
anika.schwind@informatik.uni-wuerzburg.de, michael.seufert.fl@ait.ac.at

***Abstract*—WhatsAnalyzer is a web-based service, which collects and analyzes chat histories of the mobile messaging application WhatsApp. Thereby, it leverages the e-mail export feature of WhatsApp to obtain the chat histories, which cannot be accessed otherwise due to encrypted storage on the mobile device and end-to-end encrypted transmission over the Internet. Thus, the major asset of the service is that real communication data can be collected without the bias introduced by observing or surveying participants. The collected communication data can be analyzed and provides valuable insights into the communication in WhatsApp and the resulting network traffic. To incentivize users to send chat histories, the privacy of users is respected by anonymizing all communication data. Moreover, some analyses of each chat history can be accessed on a web page by the sender of the chats.**

***Index Terms*—Communication model; WhatsApp; Mobile messaging application; Mobile instant messaging; Mobile networks.**

I. Introduction and Related Work

Mobile messaging applications (MMAs) offer real-time message transmission over the Internet. These apps are a free or low-cost alternative to operator-based messaging via SMS or MMS, and thus, show a growing popularity. In 2017, 1.82 billion people used MMAs at least once a month, and an increase to 2.48 billion in 2021 is expected [1]. Thereby, WhatsApp and Facebook Messenger are the most popular apps with 1.2 billion monthly active users in 2017. They are followed by WeChat (938 million) and QQ Mobile (678 million). Other popular MMAs are Skype, Snapchat, Kik, Viber, Line, BlackBerry Messenger, Telegram, and KakaoTalk, which are reported to have 300 million and less monthly active users [1], [2]. In [3], it was predicted that messaging traffic will reach up to 100 trillion MMA messages in 2019, which is 62.5% of global message traffic including MMAs, SMS, MMS, e-mail, rich communications suite, and social messaging. Thereby, the revenue generated from each MMA message is forecast to be less than 1% of that from SMS and MMS.

These statistics show that the network traffic created by ubiquitous communication through MMAs increases and puts a lot of load on mobile networks. To efficiently handle this traffic and provide a proper management of the cellular resources, it is necessary to understand how MMAs are used. Nowadays, MMAs are not purely text-based anymore, but several MMAs allow the transmission of (media) files, such as images or videos, and some even feature voice calls or videoconferencing. Additionally, most apps are not limited to one-to-one communication, but the creation of chat groups allows many-to-many communication. In contrast to regular chatting, a post in a group has to be transmitted to multiple recipients, and thus, multiplies the traffic on the network. While compression of media content is the default procedure how MMAs cope with huge amounts of data, the users' demand for high quality content and the multiplication of recipients might require additional network traffic management to cope with the traffic load.

Thus, it might not be obvious yet, which Internet technology will be employed to cope with the new challenges and demands of (group-based) messaging communication on the Internet. Still, it is the user behavior that dictates the path of technology through service acceptance, adoption, and usage. Therefore, it is important to analyze the way people communicate with each other using MMAs in order to develop effective traffic management algorithms to efficient deliver the generated data.

There is some literature on how people communicate with each other and how this communication has been changed in the last decade due to mobile messaging applications. Here, social aspects as well as technical aspects regarding WhatsApp and other applications were investigated [4]–[6]. Only very few papers deal with the abstract modeling of the communication within MMAs [7]–[9]. However, a comprehensive analysis and modeling of communication in MMAs is still missing.

For this reason, this paper presents the web-based service WhatsAnalyzer, which can receive WhatsApp chat histories by e-mail and analyze the communication within the chat. Thereby, it leverages the e-mail export feature of WhatsApp to obtain a text-based version of the chat histories. As chat histories are stored in an encrypted database on the mobile device and messages are transmitted over the Internet with end-to-end encryption, this is currently the only option to access the chat data. Thus, the major asset of the WhatsAnalyzer service is that real communication data can be collected without

* Michael Seufert is now at AIT Austrian Institute of Technology GmbH, Vienna, Austria

978-0-9883045-5-0/18/$31.00 ©2018 ITC
DOI 10.1109/ITC30.2018.00020

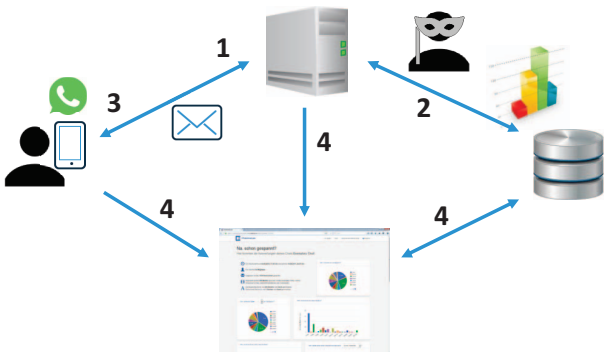


Fig. 1: Processing procedure of WhatsAnalyzer

the bias introduced by observing or surveying participants. When collecting the chat histories, the privacy of the users is respected, such that only timestamps, anonymized user names, message types, and message lengths are extracted from the chat history. These communication data can be analyzed to understand the communication in WhatsApp and the resulting network traffic, both in terms of frequency and volume. To give an incentive to use the service, some analyses of each chat history can be accessed on a web page with an individual link, which is given to the sender of the chat history by e-mail together with a mapping of real and anonymized user names. Thereby, also the senders of the chat history can get interesting insights into their own communication.

II. WhatsAnalyzer

This section describes the *WhatsAnalyzer* application. Figure 1 illustrates the general procedure. First, the user sends his or her WhatsApp chat history via e-mail to WhatsAnalyzer (1). The incoming mail is noticed by WhatsAnalyzer, which then automatically starts the processing procedure. The chat history is anonymized and statistically evaluated, and the anonymized chat and the statistics are stored in a database (2). Afterwards, WhatsAnalyzer sends an e-mail to the user (3) including a link to his or her analysis visualized on a webpage (4). In the following, the individual components of WhatsAnalyzer are described in detail.

A. Mail Handling

For mail handling, WhatsAnalyzer uses the mail server of the University of Würzburg. The communication with the server is done via standard protocols: The incoming e-mails are fetched from the server via the Internet Message Access Protocol (IMAP). To send e-mails, the Simple Mail Transfer Protocol (SMTP) is used.

To trigger an analysis of a WhatsApp chat, the user has to utilize a WhatsApp feature, which allows to send a copy of the chat history by e-mail. Internally, WhatsApp then generates a text document with the chat history, which is attached to an e-mail and can be sent via the device's e-mail application. The user has to send this text document of an individual chat or group chat without any other media files to the specific e-mail address of WhatsAnalyzer. In the following, this user will be referred to as chat owner. As soon as an e-mail arrives at WhatsAnalyzer's inbox, it is noticed by the mail handling module and the e-mail is being processed. As a first step, the module checks if the e-mail contains a valid WhatsApp chat history. In this case, the file is parsed, anonymized, evaluated, and stored as described below. Afterwards, the received e-mail is deleted automatically. The mail system also handles outbound e-mails to users and the administrator. In the main use case, WhatsAnalyzer replies to chat owners after their chat was analyzed. They will receive an e-mail containing a link to the web page on which the evaluation of their chat is displayed. In addition, this e-mail contains an assignment of the chat members' real names to their anonymized names.

B. Anonymization

While bringing the chat history into a standardized format, WhatsAnalyzer anonymises the chat and analyzes it afterwards. In this context, it should be noted that after the anonymization, the original chat history is deleted and only the anonymized version of it is used for further investigations.

Figure 2 shows an original chat history as it can be sent from WhatsApp and its anonymized version. The anonymization step not only protects users' private data (real names of communication partners and content of messages), but also transforms the chat histories into a standardized format. This is necessary because WhatsApp chat histories occur in a variety of formats (e.g., date, separator, system messages) depending on operating system, system language, and WhatsApp version, which makes parsing the histories challenging. Each line of a chat history represents a message of a user. Despite the various formats, each post starts with a timestamp followed by the name of the author and the content of the message.

The format of the timestamp varies considerably depending on the system language. To cope with this, the timestamps are parsed and normalized for post-processing in the following format: `dd.mm.yyyy, HH:MM`.

The second part of a line contains the name of the author of the post as stored in the contact list of the device. Each of these names is replaced by a unique user ID in order to be able to keep track of the individual behavior of different users. A list of the original names of the users and their IDs is temporarily stored to be sent to the chat owner in the response e-mail and will be deleted afterwards. With this list, the chat owner, i.e., the user who sent the chat history, is able to identify all participants, while nobody else is.

The last part of a line contains the actual message that has been sent. This can be some text written by a participant, a placeholder indicating that a media file was sent, or a system message (e.g., informing the participants that somebody has changed his or her phone number or changed the chats' name). To protect the users' privacy, the content of text messages is discarded but only the number of written characters is saved.

```
02/08/17, 17:29 – Michael created group "Test"
02/08/17, 17:30 – Michael added you
02/08/17, 17:32 – Marco: Hi Anika
02/08/17, 17:57 – +49 1234 12345678: Hey
02/08/17, 18:43 – Anika: Hello everybody
03/08/17, 08:29 – Michael: <Media omitted>
```

```
02.08.2017, 17:29 – User1: created group
02.08.2017, 17:30 – You: were added
02.08.2017, 17:32 – User2: 8 chars
02.08.2017, 17:57 – User3: 3 chars
02.08.2017, 18:43 – User4: 15 chars
03.08.2017, 08:29 – User1: <media>
```

Fig. 2: Original WhatsApp chat history and its anonymized version

After the anonymization, the original chat history is deleted. For all further analysis, only the anonymized version of the chat is used.

C. Standard Evaluation

In the next step, the anonymized chat history is statistically evaluated and stored in a database. First, a unique ID for the respective chat is generated. This ID is provided to the chat owner via the response e-mail so that he or she can access the associated visualization. Next, the anonymized chat history is analyzed and different types of statistics are produced:

Temporal Characteristics: In this analysis, the date of the first post and the last post of the chat is saved. Note that this timespan does not necessarily cover the complete conversation of the chat. Parts of the conversation can be lost when the chat owner changed or reset his or her device, or when the chat owner was added or removed from the chat. In addition, it is investigated how much posts were sent in specific time intervals, i.e., the number of messages per day, per weekday, and per daytime.

Chat Characteristics: For the whole chat conversation, WhatsAnalyzer counts the number of members, the number of sent messages, and the number of sent media files like photos or videos. Moreover, every post is analyzed with respect to its length. In particular, the number of characters in the shortest and in the longest text message are counted and the corresponding users are identified.

User Characteristics: For each user, the number of sent posts is calculated. Thereby, text posts and media posts are differentiated. A communication matrix is generated, which counts how often each user answers to any other user. In this context, each message is considered to be an answer to the previous message. WhatsAnalyzer also determines the frequency of starting a new session per user. A session is defined via a fixed pause threshold t, i.e., a session is a sequence of posts, such that any two consecutive posts have not been sent more than t minutes apart. This analysis is repeated with t set to 30, 60, and 1 440 (24 hours). For every participant in the chat, it is counted how often he had the final say. A message was counted as final say if it was the last message before a discussion break, which is at least one hour.

After the statistical analyses, the data are stored in a database for later visualization and further evaluations. Once the data are stored, the response e-mail containing the link to the visualized statistics of the chat is sent as described above.

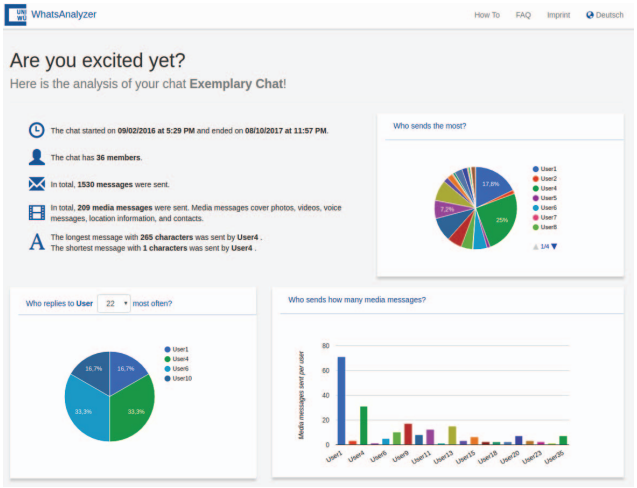


Fig. 3: Screenshot of the evaluation website

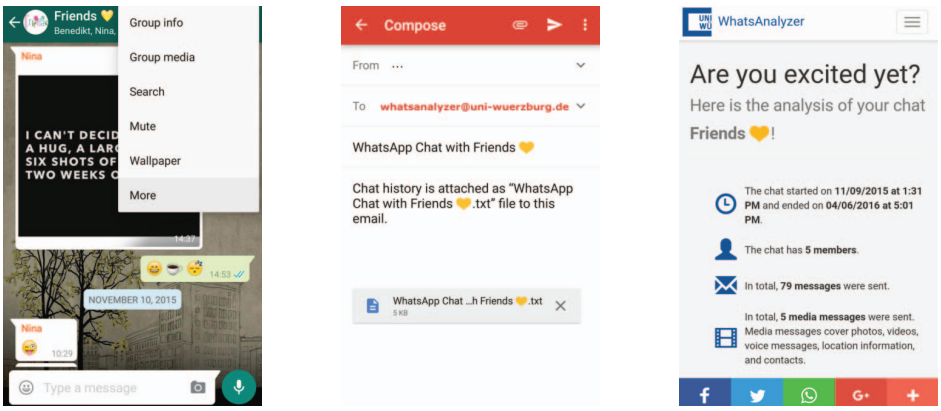
D. Visualization

The last step in the WhatsAnalyzer process is the visualization of the chat analysis. This web page only shows statistics using the anonymized user IDs. However, the chat owner is able to identify all participants by using the list he or she also received in the response e-mail.

Figure 3 shows a screenshot of an exemplary evaluation webpage. In the upper left part, a listing of important statistics can bee seen. Here, the date of the first and the last post of the chat, the number of members of the chat, and the number of sent text and media messages is shown. Additionally, the member who wrote the longest and the member who wrote the shortest message are displayed. On the right of the screenshot, a pie chart shows the percentage of messages each member posted during the conversation. In the lower part, the left pie chart shows how often a user answered to every other member, while the right chart indicates the number of sent media messages per person as bar plot. Note that the color assigned to each user is the same for every chart on the web page, so that the identification of particular chat members is simplified. The whole exemplary evaluation webpage can be found at `https://goo.gl/hvV8eF`.

III. Demostration

The demonstration shows the capabilities of WhatsAnalyzer. Here, the procedure presented in Figure 1 is shown. For the demonstration, WhatsAnalyzer will run on a local server and



(a) Exporting chat (b) Sending chat (c) Getting the analysis

Fig. 4: Three steps of how to get a WhatsApp chat analyzed using WhatsAnalyzer

give insights into each step of the procedure. The users can test the tool by sending either a personal WhatsApp chat from their smartphone or by sending an exemplary chat from a given demo smartphone.

Figure 4 shows how WhatsAnalyzer can be used in this demonstration. First, as can bee seen in the left and the middle part of the Figure, the user has to select a WhatsApp chat on the smartphone and send it to WhatsAnalyzer via e-mail. Therefore, the user has to open a chat and click the menu button at the top right. Then he has to select 'More', choose 'E-mail chat' and click 'Without media'. Next, an e-mail app is opened and the e-mail and the attached chat history has to be send to `whatsanalyzer@uni-wuerzburg.de`. As soon as the e-mail arrives at WhatsAnalyzer's mail server, the chat is automatically anonymized and evaluated. Afterwards, it replies to the user via e-mail, sending a link and an assignment of the anonymized names to the real names. The link leads to a web page that shows several statistics of the sent chat as can bee seen in the right part of the Figure.

IV. Conclusion and Outlook

Due to increasingly popular mobile messaging applications, the way people communicate has been evolved in the last years. To analyze this development, we presented WhatsAnalyzer, a web-based tool to collect and analyze chat histories of the mobile messaging application WhatsApp. Chat histories can be extracted in WhatsApp and sent by e-mail to WhatsAnalyzer, which is currently the only option to access chat data. Users are encouraged to send their chat histories by emphasizing the protection of the users' privacy. This means, all analyzed and stored data is completely anonymized, keeping only timestamps, anonymized user names, message types, and message lengths. These communication data suffice to analyze the properties of WhatsApp chats, the users, and the communication within. Moreover, some data are visualized and presented to the chat senders to show them some basic insights into their communication.

In future work, it is planned to extend WhatsAnalyzer to add the possibility of analyzing chat histories without anonymization. If the users give their approval, an evaluation of the plain text of the chat can be done. Thus, linguistic analyses could be carried out to get more insights into the communication within WhatsApp. For example, it could be evaluated how the used language evolves during chatting in mobile messaging applications. Also other user and group characteristics could be extracted, such as the ratio of emoticons usage or the mood of conversations.

References

[1] C. Boyle, "Messaging App Usage Worldwide: eMarketer's Updated Forecast, Leaderboard and Behavioral Analysis," eMarketer, Tech. Rep., 2017. [Online]. Available: http://www.emarketer.com/Chart/Mobile-Phone-Messaging-App-Users-Worldwide-2016-2021-billions-change/209369, http://www.emarketer.com/Chart/Users-of-Select-Mobile-Messaging-Apps-Worldwide-2016-2017-millions/209534

[2] Statista, "Most popular mobile messaging apps worldwide as of January 2017, based on number of monthly active users (in millions)," 2017. [Online]. Available: https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/

[3] L. Foye, "Mobile & Online Messaging: SMS, RCS & IM Markets 2015-2019," Juniper Research, Tech. Rep., 2015. [Online]. Available: https://www.juniperresearch.com/press/press-releases/messaging-revenues-down-600m-traffic-up-200pc, https://www.juniperresearch.com/press/press-releases/traffic-from-messaging-reach-438bn-per-day-by-2019

[4] L. Piwek and A. Joinson, ""What do they Snapchat about?" Patterns of Use in Time-limited Instant Messaging Service"," *Computers in Human Behavior*, vol. 54, pp. 358–367, 2016.

[5] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting WhatsApp through Large-scale Measurements in Mobile Networks," in *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4. ACM, 2014, pp. 133–134.

[6] K. Church and R. de Oliveira, "What's up with WhatsApp? Comparing Mobile Instant Messaging Behaviors with Traditional SMS," in *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services (MOBILE HCI)*, Munich, Germany, 2013.

[7] A. Rosenfeld, S. Sina, D. Sarne, O. Avidov, and S. Kraus, "A study of whatsapp usage patterns and prediction models without message content," *arXiv preprint arXiv:1802.03393*, 2018.

[8] M. Seufert, T. Hoßfeld, A. Schwind, V. Burger, and P. Tran-Gia, "Group-based communication in whatsapp," in *IFIP Networking Conference (IFIP Networking) and Workshops, 2016*. IEEE, 2016, pp. 536–541.

[9] M. Seufert, A. Schwind, T. Hoßfeld, and P. Tran-Gia, "Analysis of group-based communication in whatsapp," in *International Conference on Mobile Networks and Management*. Springer, 2015, pp. 225–238.