Received 20 December 2022, accepted 22 February 2023, date of publication 9 March 2023, date of current version 16 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3254913

RESEARCH ARTICLE

Share and Multiply: Modeling Communication and Generated Traffic in Private WhatsApp Groups

ANIKA SEUFERT, FABIAN POIGNÉE, MICHAEL SEUFERT, (Member, IEEE), AND TOBIAS HOßFELD, (Senior Member, IEEE)
Chair of Communication Networks, University of Würzburg, 97070 Würzburg, Germany

Corresponding author: Anika Seufert (anika.seufert@uni-wuerzburg.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) under Grant HO 4770/7-1 and Grant SE 3163/1-1, and in part by the Open Access Publication Fund of the University of Wuerzburg.

ABSTRACT Group-based communication is a highly popular communication paradigm, which is especially prominent in mobile instant messaging (MIM) applications, such as WhatsApp. Chat groups in MIM applications facilitate the sharing of various types of messages (e.g., text, voice, image, video) among a large number of participants. As each message has to be transmitted to every other member of the group, which multiplies the traffic, this has a massive impact on the underlying communication networks. However, most chat groups are private and network operators cannot obtain deep insights into MIM communication via network measurements due to end-to-end encryption. Thus, the generation of traffic is not well understood, given that it depends on sizes of communication groups, speed of communication, and exchanged message types. In this work, we provide a huge data set of 5,956 private WhatsApp chat histories, which contains over 76 million messages from more than 117,000 users. We describe and model the properties of chat groups and users, and the communication within these chat groups, which gives unprecedented insights into private MIM communication. In addition, we conduct exemplary measurements for the most popular message types, which empower the provided models to estimate the traffic over time in a chat group.

INDEX TERMS Communication models, group-based communication, mobile instant messaging, mobile messaging application, private chat groups, WhatsApp.

I. INTRODUCTION

Today's ubiquitous Internet shows a complex interplay of Internet technology and human behavior. On the one hand, the Internet changes our daily lives, society, business, and industry. At the same time, also the Internet technology is driven by the adoption of end users and stakeholders in the ecosystem. The video streaming service YouTube is a prominent example for these interplays. It allows to upload and stream user-generated videos, which changed the Internet usage behavior of end users and led to an unprecedented increase of global Internet traffic. Then again, the evolved user behavior required to change the underlying Internet technology. In particular, the increased video demand on YouTube pushed the need for content delivery networks (CDNs) to place and cache videos closer to end users taking into account regional or social interests. This new Internet technology subsequently spread, and today many different types of services rely on CDNs.

Another drastic but often overlooked change of users' communication behavior is happening with the highly popular mobile instant messaging (MIM) applications, sometimes also referred to as mobile messaging applications (MMAs), such as WhatsApp, Facebook Messenger, or WeChat. These apps facilitate to communicate in fixed groups, which are created spontaneously, or which exist over a longer period. The popularity of this feature shows that users are not anymore solely relying on traditional communication paradigms, such

The associate editor coordinating the review of this manuscript and approving it for publication was Luca Bedogni.

as one-to-one (e.g., e-mail or SMS) or one-to-many (e.g., social networks). In contrast, group communication, which can be considered as many-to-many communication between a fixed number of participants, has gained a huge momentum.

Although often referred to as chat groups, typically, various types of media, such as text, voice, image, or video, can be exchanged among the members of a group. Note also that, in contrast to videoconferencing, group-based communication in MIM apps is still asynchronous. This means, no real-time communication channel is established and maintained, but MIM apps typically utilize a delay-tolerant publish-subscribe paradigm to transmit the text or media messages from the sender to other members of the chat group. Thus, the sharing of messages in chat groups multiplies the generated traffic on the network according to the number of receivers, which can have a massive impact on the underlying communication networks, especially when large media messages are shared in big chat groups.

In addition, due to omnipresent network connectivity and smartphone notifications, the activities of users mutually influence each other, for example, when posted messages trigger phases of increased messaging activity in these chat groups. Together with the large number of participants and high availability of (user-generated) media content, the time scales and data volumes of such communication are different from traditional communication paradigms and might have disruptive implications for the future Internet.

Today, mobile messaging applications utilize a publish-subscribe paradigm on application layer, which could be efficiently implemented on the network layer in the future. Furthermore, they rely on media compression to reduce the network demands. In the future, when quality demands increase, user-generated content (UGC) could be cached close to the edge and the social groups, or could be transmitted directly via device-to-device communication. This could also foster the implementation of current research proposals like information-centric networking. Moreover, the increasing privacy awareness of users and end-to-end encrypted data transfer will be a big challenge for network management. Thus, it might not be obvious yet, which Internet technology will be employed to cope with the new challenges and demands of group communication.

For the considered MIM applications, network operators still lack a thorough understanding of group communication and the underlying traffic generating processes for two reasons. First, as MIM applications like WhatsApp typically employ end-to-end encryption, network measurements can only provide aggregate traffic statistics of the status quo, e.g., for a given link or host, but they cannot not provide deeper insights into MIM communication, which would be helpful to model and predict future network loads. Second, given that the generation of traffic and its multiplication when sharing content depend on sizes of communication groups, the speed of communication, and the types of exchanged messages, it would be necessary to analyze and model chat groups and the communication within. However, these chat groups and the communication within are private, and thus, can only be accessed by group members.

In this paper, we tackle these issues and investigate group-based communication in private WhatsApp groups. For this, we rely on the export feature of WhatsApp to collect histories of chat groups using a web-based tool. As all chat histories are completely anonymized before storage and analysis, the privacy of all group members and of all messages is preserved. During the last four years, users voluntarily sent in their chats to contribute to this research, which resulted in a data set containing the communication metadata of 5,956 private chat histories, which contain 76,720,159 messages from 117,695 chat members, and cover time spans of up to 8.7 years. This is – to the best of our knowledge – by far the largest data set of WhatsApp chat groups. We not only publish the complete data set as open data, but also use the data to describe and model the properties of the communication within private WhatsApp chat groups. Furthermore, we conduct exemplary network measurements for the most popular types of messages, which empower the provided models to estimate the overall traffic in a chat group. Thus, our paper gives unprecedented insights into private MIM communication, which allows to better understand group-based communication in MIM applications, as well as the underlying traffic generating processes.

The contributions of this work can be summarized as follows:

- Publication of a large data set of anonymized communication in private WhatsApp chat groups
- Models for communication behavior in MIM applications
- Models for traffic generation depending on message type

II. RELATED WORK

Several works exist on how people communicate with each other and how this communication has been changing in the last decade due to MIM applications. However, most of this work focused on the social aspects of MIM and was conducted using surveys. In contrast, there are only few works, which investigated technical aspects of MIM or presented models for group-based communication. In the following, these related works are briefly summarized.

In [1], the temporal aspects and energy consumption of WhatsApp are investigated by analyzing message patterns of 51 users. 59% of messages were in single chats, and 41% were in group chats. While every day showed a similar trend, a gradual increase of messages over the course of day was observed with a single peak in the evening. Based on these data, they proposed a message aggregation technique that reduces the energy consumption by trading off against latency. The traffic behavior of WhatsApp was analyzed by [2] from passive measurements within a large cellular network. They found that WhatsApp is mainly used as a text

messaging service, with more than 93% of the transmitted flows containing text. However, 36% of the exchanged volume in uplink and downlink was caused by video sharing, and 38% by photo sharing and audio messages. However, these measurements are based on the old transport protocol (before 2016), which is why previous traffic measurements are no longer comparable with the current traffic. Reference [3] investigated user behavior patterns and traffic characteristics of the MIM application WeChat based on traffic measurements within a large cellular network. Thereby, they modeled the distributions of inter-arrival time of messages and message length and integrated them into an on/off-model to account for the keep-alive mechanisms of MIM apps. The resulting model was used to evaluate the impact of MIM on cellular network performance. The impact of data transfer time in contact-based messaging applications, i.e., MIM apps that exclusively rely on device-to-device communication, was analyzed by [4]. Their results show that if the message communication time is high, the overall message diffusion is bounded by this time, resulting in slightly increasing diffusion time when the number of nodes increases. Reference [5] monitored the network traffic of 3 million users on average for one month and identified WeChat traffic and video chat user behavior based on traffic analysis. They presented daily and weekly usage patterns and found that 20% of users contribute 95% of the video call traffic. Moreover, they observed that the calling times distribution follows a power-law distribution for which 96.5% of conversations are less than 5 minutes. Also [6] monitored traffic of 603,000 WeChat users to extract temporal patterns, flow characteristics, and message intervals. They identified nine usage patterns and analyzed the performance of media flows. As WhatsApp has been using end-to-end encryption since 2016, there are now only limited possibilities to investigate user behavior and traffic patterns when using WhatsApp. Reference [7] analyzed the semantics of encrypted network traffic generated by the WhatsApp application. By looking at the generated traffic, they were able to detect certain app functions used like call termination, missed/rejected calls, and blocked calls. In [8], a blind traffic detection technique is presented which is able to differentiate unique WhatsApp calls from encrypted traffic while the authors of [9] used wiretap data to identify possibilities to determine if someone is sending or receiving WhatsApp messages at a given time.

In addition to these network measurement-based studies, few works have explicitly studied and modeled group communication using chat histories. In [10], a survey was conducted and private chat histories of 243 users were evaluated to obtain first statistics of chat groups, which allowed for a characterization of groups and the derivation of a simple communication model. Reference [11] further analyzed the same data set to come up with a refined model for the active participation of users in group chats and the resulting network traffic. Reference [12] collected and analyzed a data set of public WhatsApp group communication consisting of 178 public groups, which contained around 45,000 users and 454,000 messages. They evaluated the number of messages per group and per user, the location of users, the content and language of messages, as well as, for the most active groups, the number of messages per day. Their methodology to collect public WhatsApp group chats was also utilized by [13], which analyzed the communication in 141 and 364 public political groups. For this, they focused on the number of messages per content type, category and propagation of images, and the network structure among groups and users. The study in [14], which is the closest to our work, investigated WhatsApp usage patterns of 100 users to predict demographic characteristics. In contrast to our work, they analyzed all chats of the users, but the data set is much smaller than the one used in this work, and the analyses mostly focus on different demographics, such as gender, age, or education.

The presented related works already gave interesting insights in group-based communication in MIM applications. However, the major limitation of previous works was a lack of depth of evaluations due to encrypted traffic or the availability of only small data sets, which explicitly allow a comprehensive analysis and modeling of group communication. To overcome this issue, in this work, we focused on the popular messaging application WhatsApp, and collected a huge data set of 5,956 private chat histories. This is – to the best of our knowledge – the largest data set of private WhatsApp chat histories, and it allows to present more refined and reliable models of group communication in MIM applications, as well as models for traffic generation in private chat groups.

III. WhatsApp DATA SET

In this section, the collection of chat histories is described. Moreover, the limitations of the data set are discussed and general characteristics of the data set are presented.

A. DATA COLLECTION METHODOLOGY

To obtain insights into the communication behavior in private WhatsApp groups, a large data set of real chat histories is necessary. To collect these chat histories, our web-based tool *WhatsAnalyzer* [15] was used. WhatsAnalyzer is a free-to-use application for collecting and automatically evaluating WhatsApp chats. It is based on a built-in feature in WhatsApp (email chat), which allows to send an email with a text file containing the chat history. With this feature, users can send a chat to our tool, where a unique identifier is generated, system messages are detected, and the remaining chat is anonymized to protect the users' privacy. The unique identifier is based on a hash of the chat name and the email address of the sender, which allows to easily detect whether a chat was sent multiple times. To reliably detect system messages, we supported only four popular languages. However, users could change the system language of their phone to one of these languages to obtain a compliant export format.

As can be seen in Figure 1, anonymization means, that user names, phone numbers, and message content are removed,

```
05/02/22, 17:29:12 - Alice created group "IMC'22"
05/02/22, 17:30:58 - Alice added you
05/02/22, 17:32:43 - Bob: Hi Carol
05/02/22, 17:57:08 - +12 1234 12345678: 😊 Hi!
05/03/22, 08:29:51 - Alice: <video omitted>
```

```
2022-05-02 17:29:12, User1, <created group>
2022-05-02 17:30:58, You, <added>
2022-05-02 17:32:43, User2, 8 chars, 0 emojis
2022-05-02 17:57:08, User3, 4 chars, 1 emojis
2022-05-03 08:29:51, User1, <video>
```

FIGURE 1. Chat history as exported by WhatsApp (left), and corresponding anonymized, normalized version (right).

and only relevant metadata about the communication behavior are kept. Moreover, a normalized version of the chat is stored for further analysis because export formats of WhatsApp greatly vary due to system language, operating system and/or WhatsApp version. Finally, some basic statistic evaluations of the communication within that chat are calculated. As an incentive to use our tool, these evaluations are sent back to the users, after they sent their chat, to also provide them some analyses of their own communication behavior.

All collected chat histories were sent voluntarily by people who were interested in the resulting evaluations. No targeted study or recruitment was made. Our tool was only advertised using a web page and social media and can easily be found in popular search engines by people who want to get insights into the communication in their WhatsApp chats.

In total, our tool collected 7,400 group chats from 3,978 different senders between October 2017 and December 2021. To avoid bias from chats, which were sent multiple times, 1,444 chats were filtered out based on the unique chat identifier, such that only the newest version of each duplicate chat was kept. Note that a higher number of group members than the limit of 256 members is possible due to the fact that people can leave and new people can join the conversation during the lifetime of the group. Furthermore, if people use an invitation link to join a public group chat, there is no limitation of the group's size. As this work focuses on the communication in private WhatsApp group, we apply a second filter, which removes all groups with more than 256 group members. This further excludes 88 groups from the data set, which are so large that they are clearly public chat groups, although some smaller public groups might still reside in the data set. However, we expect that they behave rather similar to private groups, and thus, consider their impact on the evaluations as marginal.

Finally, after filtering duplicates and obviously public groups, 5,956 chats remain for our analyses, which were sent from 3,899 different persons. The data set contains 76,720,159 messages from 117,695 users, which means that each user contributed on average about 652 messages. With respect to the 5,956 collected chat histories, this means that there are about 12,881 messages in each chat on average. The average covered time span of a chat is 1.32 years with a standard deviation of 1.55 years and a maximum of 8.73 years.

We provide our filtered data set, i.e., the data set on which all presented evaluations in this work are based, as open data in a repository[1] The repository also contains a `README` file explaining the structure of the data set. Before we describe the general characteristics of the data set, and present the detailed analyses of group communication in WhatsApp, we first outline the limitations of our data set.

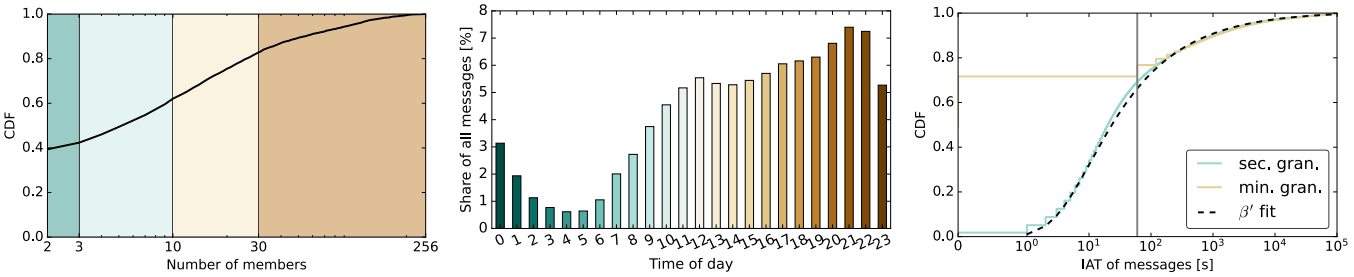
[1]https://figshare.com/articles/dataset/WhatsApp_Data_Set/19785193

B. LIMITATIONS OF THE DATA SET

As the data was collected using a WhatsApp feature, which only allows to export and send single chats per email, the data set only contains single, selected chats of each user. This makes it possible to analyze the sending behavior of the users, what we do in this work, but the data do not contain any information about the users' overall WhatsApp usage, e.g., times when a user is passively reading messages. Moreover, as there is a strict anonymization per group, the data do not provide any insights into the content of sent messages, and users and contents cannot be identified over different groups. This prohibits the usage of the data set for the analysis of linguistic aspects, such as the topics discussed within group chats, or the mood or emotions of participants.

Although it is possible to analyze the communication behavior in the sent chat group, this means that it is not possible to analyze whether the communication behavior of a user differs in different groups. Consequently, also no information can be obtained about a user's communication behavior over all chats, although the same user could potentially be part of several collected chats. Further, there is no possibility to analyze the social network of a user or groups, e.g., with respect to the sharing of content in multiple chats. Note that this limitation was deliberately introduced during the collection of the WhatsApp chats in order to protect the users' privacy and further incentivize users to send their chats to our tool.

Regarding the collection of WhatsApp chats, it has to be noted that there was no systematic recruitment of participants. As mentioned above, our tool was only advertised using a web page and social media as a "tool to get insights into communication behavior in WhatsApp chats". Thus, the collection is purely based on voluntary participation. This leads to a potential bias that chats were sent to our tool mainly by persons who are interested in the evaluation of the communication behavior within their WhatsApp chats, such that they searched and found our webpage. Consequently, the data set is by no means representative for all group chats in WhatsApp. Nevertheless, it can be assumed that, most likely, chats were collected and analyzed by our tool that are meaningful to or valued by the sender. Moreover, the collected chats cover a huge diversity of WhatsApp chats,



(a) Number of members of all chats. (b) Messages per time of day. (c) Interarrival time of messages.

FIGURE 2. Group sizes and temporal aspects of conversations in the collected data set.

as chats were sent by 3,899 different persons from different countries. Although no information was collected from the sender, which includes the country of origin, this observation follows from the different system languages and timestamp formats that were found in the collected WhatsApp chats, as well as from support requests of potential users.

Finally, the analysis of the communication behavior within a single group is additionally limited by the different export formats of WhatsApp, which depend on system language, operating system, and WhatsApp version. In particular, some export formats just report that a media message was sent, but do not provide any information about the type of media (e.g., image, video, location). Moreover, some export formats provide timestamps with seconds granularity, i.e., using the format hh:mm:ss, while some export formats only report on minutes granularity, i.e., using the format hh:mm. Thus, some analyses have to be based on a reduced data set, which will always be reported throughout this work. In addition, the time range of exported messages might be limited, if the user, who sent the chat, was added later to the chat group, or due to the WhatsApp export limit of 40,000 messages [16]. This means that an exported chat history does not necessarily contain the complete conversation from the creation of the chat group until the time of the export. Lastly, it should also be considered that users, who did not join or leave the conversation, and never sent a message during the analyzed excerpt of the chat history, do not appear in the exported chat history, and thus, could not be counted nor analyzed.

Nevertheless, although some limitations exist, the collected data set allows to gain unprecedented, detailed insights into the communication behavior in private WhatsApp chat groups. In the following, the data set is characterized, before the insights into the communication behavior of groups and users are presented.

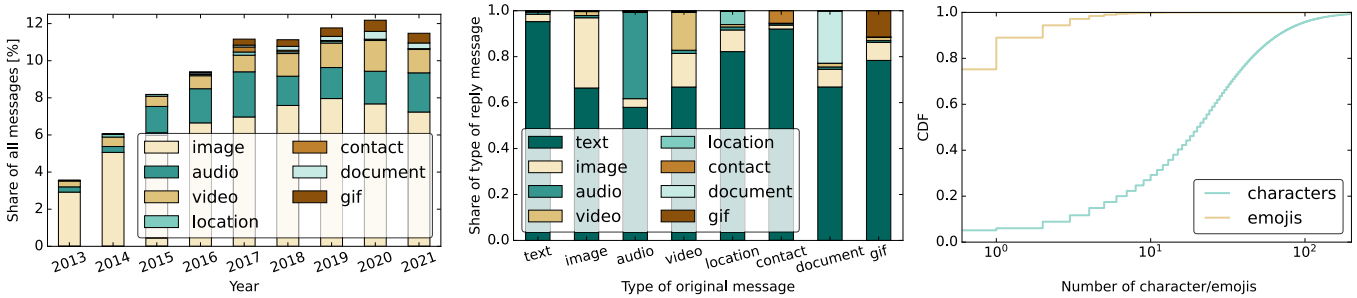
C. GENERAL CHARACTERISTICS OF DATA SET

In Figure 2a, we look at the CDF of the number of members, i.e., the **group size**, in all collected chats. Here, due to the wide range of observed group sizes, a logarithmic scale is used on the x-axis, starting from two members. The colored background indicates a categorization of group sizes from left to right: dyadic (two group members), small (three to ten group members), medium-sized (eleven to 30 group members), and large (31 or more group members). This categorization was chosen arbitrarily to almost equally divide the set of actual group chats, i.e., chats with three or more members, into three parts. In fact, the collected data set contains 42.34% dyadic groups, 19.73% small groups, 20.77% medium-sized groups, and 17.16% large groups. This shows that our web-based tool is mainly used to analyze groups with more than two people. 95% of all chats have 97 members or less, and the maximum observed group size is 252.

The **temporal distribution of the sent messages** is depicted in Figure 2b. The x-axis shows the hour of the day from 0 to 23, and the height of the bars represents the share of messages that were sent in the corresponding hour. A typical diurnal pattern can be observed, which has its minimum at 4 am with 0.61% of messages. In the morning a steep rise can be observed until noon, where the share of messages surpasses 5%. Around noon, from 11 am to 4 pm, the activity of users almost plateaus, before the increase in the evening reaches the maximum at 9 pm (7.40%). During the night, the share of messages quickly decreases.

Afterwards, the **interarrival time (IAT)** of messages, i.e., the time between two consecutive messages, is analyzed and the corresponding CDF is depicted in Figure 2c. Here, this analysis considers all pairs of consecutive messages over all chats. Note that the export format of WhatsApp can be either on seconds granularity, i.e., timestamps are exported as hh:mm:ss, or on minutes granularity (hh:mm). In the collected data set 1,598 chats have seconds granularity, the remaining chats have minutes granularity. Both CDFs are shown in the plot. It can be seen that the yellow CDF of minutes granularity is well aligned with the green CDF of seconds granularity, so, in the following, IATs will be investigated only on seconds granularity.

Although the x-axis represents the interarrival time in minutes on a logarithmic scale, the green CDF shows a very steep increase, which means that IATs are generally short. In fact, 69.29% of all messages are replied within the same minute. Considering a maximum IAT of 2 minutes, 76.49% of all messages are replied within this time period. Although only



(a) Evolution of the average distribution of media types per group.

(b) Distribution of consecutive message types.

(c) Message lengths of text posts.

FIGURE 3. **Message content characteristics observed in the collected data set.**

5.03% of all messages have a reply after more than 60 minutes (1 hours), the mean IAT is 42.47 minutes, which indicates a long tail characteristic. This shows that WhatsApp is mainly used as very fast communication channel, but also very long communication pauses can occur.

We fit the distribution of IATs starting from 1 second using a beta prime distribution ($\beta`(x) = I_{\frac{x}{1+x}}(\alpha, \beta)$, where I is the regularized incomplete beta function), which can be seen as the black dashed CDF in Figure 2c. For all curve fittings in this work, we investigated different functions and show the one with the best fit. The fitted distribution has parameters $\alpha = 4.7970$ and $\beta = 0.4513$, and nicely overlaps the green empirical CDF with only marginal deviations in the range from 30 to 120 seconds. The high goodness of fit is confirmed in terms of the coefficient of determination R^2, which reaches a very high score of $R^2 = 0.9921$. Thus, this model can be used to accurately generate very realistic interarrival times for messages in private WhatsApp groups.

After investigating general characteristics of the communication within the analyzed WhatsApp chat groups, we describe the characteristics of the collected messages.

First, the **evolution of the share of media messages over the years** is analyzed in our data set. Note that the export format of some WhatsApp versions just reports that a media message was sent but does not report the type of media. Thus, for this analysis, we rely on 1,664 chats, for which media types could be differentiated. For each year from 2013 to 2021, we calculated the share of each type of media per group. The resulting average shares of each type of media per group are displayed as a stacked bar plot in Figure 3a. The averages are based on large sample sizes of at least 124 active chats per year, or even at least 650 active chats when considering 2015 to 2021. Thus, confidence intervals are small and were omitted from the plot to increase the legibility. In addition, the height of the stacked bar plot indicates the overall share of media messages in all messages in that year, which is 11.49% in 2021. At the bottom of the bars, there are images, which are the most sent media in WhatsApp having a share of 7.24% in 2021. Next, audio and video messages are depicted, which also have become more popular since 2013, and currently reach a share of 2.11% (audio) and 1.26% (video) of all WhatsApp messages in 2021, respectively. Finally, the plot shows that location and contact are very rarely used with shares below 0.1% in 2021, while document (0.30% in 2021) and gif (0.52% in 2021) have a small share of all messages.

Regarding the evolution of the share of media messages over the course of the last eight years, it can be seen that media messages have increased from 3.56% in 2013 to 11.48% in 2021. This, in turn, implies that the share of text messages has decreased from 96.44% in 2013 down to 88.52% in 2021. Note that the increase of the share of media messages has to be considered relative to the overall number of messages. This means that to obtain an estimate for the absolute volume of media messages, the overall growth of WhatsApp usage has to be considered accordingly. As the number of messages per day has been growing from 1 billion messages per day in 2011 [17] to 100 billion messages in 2020 [18], and is probably still growing, it can be followed that the absolute number of media messages is growing at an even higher rate.

In Figure 3b, we look at the **usage of different message types**. In particular, we investigate what transitions between message types appear in a conversation, e.g., which message types are mixed together and whether long sequences of the same message type occur. For this, first, each sent message is considered together with the directly following message, which we call a ‘‘reply’’ to the original message. Note that a ‘‘reply’’ does not need to be sent by another person, but even if users send two consecutive messages, we also consider their second message as a reply to their first message.

The x-axis of Figure 3b shows the different message types of the original message. The stacked bars with different colors indicate the share of message types among the corresponding replies. First of all, it can be seen that for all message types, the most frequent reply is a text message. This comes as no surprise since text messages comprise the vast majority of all message types. For example, it can be seen that 95.26% of text messages, 66.36% of images, and 57.94% of audio messages are followed by a text message.

Another observation is that a considerable number of media messages is followed by another media message of the same type. This can be seen, for example, for 30.56% of images, 37.54% of audio messages, 16.59% of videos, or 22.71% of documents. Other media types are rarely used, except for images, which account for 14.61% of replies to videos, 9.42% of replies to locations, or 7.86% of replies to gifs. The reason for this observation could be that recipients of a message with a certain type often choose to keep the same modality. This means, they reply using a message of the same type, for example, when recording and sending an audio message in reply to an audio message. Moreover, based on the above definition of reply, another reason could be that users often choose to send a batch of messages, which have the same type, e.g., choosing to share a set of images with the members of the chat group at the same time.

However, analyzing the sequence lengths per message type, i.e., the number of subsequent messages in each group with the same message type, we see that text messages have an average sequence length of 20.16, but the sequences of media messages are generally short. Audio messages have the longest average sequence length of 1.60, before images (1.44), documents (1.29), and video (1.20). The other media types have shorter sequence lengths, and thus, are often sent as a single, standalone message. Note that we did not differentiate between users in each sequence, so the average lengths of batches, which are sequences of messages from the same user, will be even shorter.

Next, we look in more detail at text messages. For this, the cumulative distribution functions (CDFs) of the **number of characters** (green) **and emojis** (yellow) per text message is shown in Figure 3c. Since we observed a wide range of message lengths, the CDFs are depicted on a logarithmic x-axis. It can be seen that the green curve initially is growing fast. It reaches the median value at 19 characters, which means that 50% of all text message have 19 characters or less. Afterwards, the text message lengths only slowly increase. Although the longest observed message has 83,161 characters, the 95th percentile resides at 93 characters.

Considering the usage of emojis, it can be seen from the yellow CDF that 75.22% of all text message do not contain any emojis. In contrast, the green CDF shows that 5.14% of all text messages contain no characters, but only emojis. Although also messages with many emojis exist (maximum 65,536 emojis), 95% of all texts contain only 2 or less emojis.

The general characteristics of the data set show that WhatsApp contains a very fast paced communication with a small number of persons. The communication is heavy on text and consists of rather short messages than long stories.

IV. MODELING GROUP COMMUNICATION

After the general characteristics of the collected data set have been described, we will now focus on detailed insights into chat communication in private WhatsApp groups. We begin by characterizing private groups, before also the communication behavior of the single users within the groups is investigated and modeled.

A. CHARACTERIZATION OF PRIVATE GROUPS

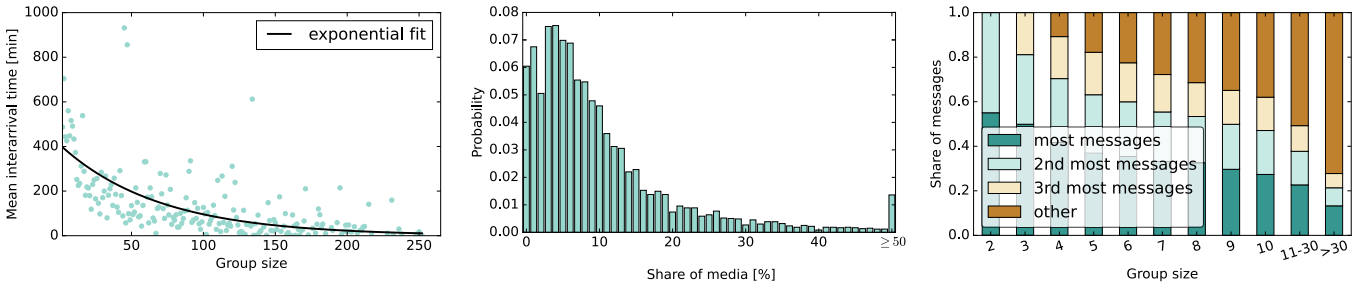
First, we take a look at how the **communication speed** differs between different private groups depending on their group size. For this, Figure 4a depicts the different group sizes on the x-axis, and the y-axis shows the corresponding mean interarrival time (IAT) between two consecutive messages that was observed in groups of that size. It can be seen that small groups below 10 participants start out with a mean IAT of around 8 hours. Considering the few participants, who can reply to a message, this means already a fast communication speed for a mainly text-based communication. As group sizes increase, the mean IATs generally decrease, which means that the communication further speeds up. This comes as no surprise, since with more participants in larger groups, it is more likely to receive a reply quickly.

The plot shows that the mean IAT often decreases to only a few minutes for large groups above 125 members. Note that some group sizes exist with higher mean IATs, which could be due to the fact that only few messages were observed in groups of these sizes. Nevertheless, despite the low mean IATs, the long tail characteristic of the distribution of IATs (see above) has to be considered. This means that these groups show both very active phases and long breaks between the sessions of a conversation.

To account also for these outliers, an exponential function $a \cdot b^x$ was fit to the data, which is plotted as the black line in Figure 4a. The parameters of the best fit are $a = 409.4868$ and $b = 0.9856$, reaching a rather low $R^2 = 0.4282$ due to the high number of outliers. However, it confirms the observed trends that communication speeds are very high, and that they are generally increasing for an increasing group size.

Next, we differentiate groups based on the overall **ratio of media messages**, which are used in the conversation. Here, interestingly only a small effect of the group size on the usage of media could be found. This became evident when investigating the Spearman's rank order correlation coefficient ρ, which is only 0.23. Thus, regarding the different media usage among groups, we can look at the corresponding histogram in Figure 4b. It can be seen that in only 6.04% of all groups no media has been sent, but all messages were text messages. This low share shows that media messages are an important means of communication in WhatsApp. However, communication is not solely media-based. Only 1.36% of all groups have a ratio of 50% or higher, i.e., equal or more media messages than text messages. Instead, it can be observed that most groups can be found at rather low levels of media usage. For example, 67.12% of all chats have a media ratio of 10% or less.

This analysis confirms the above observation that conversation in WhatsApp groups is mostly text-based. The reasons



(a) Average interarrival time per group size. (b) Share of media in group chat messages. (c) Participation of members per group size.

FIGURE 4. **Temporal and content-related characteristics of group chats.**

could be that media creation, selection, and upload consume significantly more time than text messages, and that the expressiveness of text messages for communication purposes can be much higher. Thus, most groups range at low levels of media usage, and an increasing group size does not lead to substantially more usage of media messages.

Finally, the chat groups are characterized by the contribution of their members. Figure 4c shows a stacked bar plot for different group sizes, which indicates the **average shares of the members' contributions in terms of sent messages**. Here, we differentiate the top-3-contributors, which refers to the members who sent the most messages (green bars), the second most messages (light green bars), and the third most messages (yellow bars), respectively. All messages of the remaining users are summed up in brown bars.

The plot shows that dyadic conversations (chats with two members) as well as group conversations are rarely balanced. For dyadic conversations, the top contributor sends on average 55.03% of the messages, which already shows a significant deviation from an expected balanced conversation, i.e., 50%. As group sizes increase, the share of the top contributor decreases, but the observed discrepancy even grows. For example, in groups of size 10, the top contributors account for roughly one fourth of all messages on average (27.35%), whereas they would only be expected to contribute one tenth in a balanced conversation. Even in large groups with more than 30 participants, the top contributor still sends on average 13.25% of all messages. This shows that WhatsApp conversations cannot be considered balanced conversations.

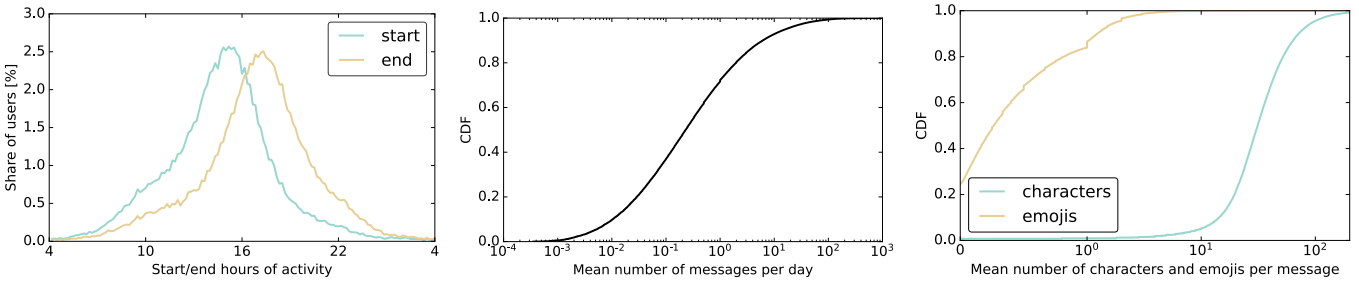
When further inspecting the contribution of the members, who send the second and third most messages, we see similar effects of over-represented contributions, but on a smaller scale. Nevertheless, the top-3-contributors clearly dominate the conversations for small and medium-sized groups. As can be seen in the plot, on average they send at least 77.09% of messages in small groups, almost half of messages for medium-sized groups of 11-30 members (49.24%), and still account for 27.76% of messages in large groups of more than 30 members. This leads to the conclusion that most conversations in group chats are dominated by few active members. It follows that a large share of passive members only read or even ignore the messages. In the next section, we will characterize the members of private chat groups in detail.

B. CHARACTERIZATION OF WhatsApp USERS

In this section, we gain further insights into the characteristics of users and their WhatsApp behavior. The evaluated data set consists of 117,695 chat members.

First, we investigate the activity of different WhatsApp users. For this, we compute the **average time at which users send their first and last message** every day, respectively. We break the day at 4 am, which is the time with the lowest overall communication activity, and we aggregate each timestamp into bins of 10 minutes. Figure 5a depicts the resulting probability density functions (PDF) on the x-axis, which shows the hours of the day from 4 am to 4 am. Here, the green curve represents the average start time of daily activity. It can be seen that only few users send the first message on average before 10:00 am (8.45%). Moreover, only half of the users regularly send messages before 2:50 pm, which is the median for average daily start time over all users. Around this time, at 3:10 pm, there is also the peak of the PDF, where 2.57% of all users, i.e., 3,023 users, have their average activity start. Note that there are also 8.75% of users, which regularly send their first message of the day after 7:00 pm.

Considering the yellow curve, we can see that 9.77% of the users have already ended their activity at noon. In contrast, 6.22% of users send their last message on average after 10:00 pm. The peak of the green PDF can be seen at 5:20 pm at similar height to the green curve. In general, it can be seen that both the yellow and the green PDF resemble each other in shape, but they are shifted by around 2 hours. This shift indicates the average length of the daily activity, which turns out to have an average of 132 minutes when considering all users. Note that the investigated daily activity only considers the message sending behavior. The passive message reading behavior might substantially deviate from these times, however, the collected data unfortunately cannot provide any insights into this aspect.



(a) Average start and end of daily activity.

(b) Number of messages per day per user.

(c) Mean number of characters and emojis per text message per user.

FIGURE 5. **User behavior during message sending.**

Next, we look at the active contributions of different users. For this, we investigate the **number of messages, which each user sends per day** on average in a single group. Figure 5b depicts the corresponding distribution on a logarithmic x-scale. It can be seen that the majority of users, i.e., 71.52%, send less than 1 message per day on average in the observed chat group, 62.37% of users send even less than 0.5 messages per day, i.e., 1 message in 2 days. On the other side, there are also few power users, which send a significantly higher number of messages in a single chat group. These users might correspond to the dominating users, which were identified above. For example, the top 5% of users send on average at least 16.11 messages per day. Here, the maximum observed value is an average of 927.27 messages per day, which corresponds to 1 message every 100 seconds, excluding any sleep times. This maximum mean value was observed for a user in a dyadic chat group over a chat duration of 22 days. The other member of this group also reached a very high average value of 890.86 messages per day, which is the second largest mean value overall. Considering the relative daily contribution with respect to the particular group of each user, i.e., we divide the average number of sent messages of users by the average amount of messages of their whole group, the observed trend is confirmed. 75% of the users contribute on average less than 3.53% of the daily messages in their group. On the other side, the top 5% of users account on average for more than 31.71% of their group's messages per day.

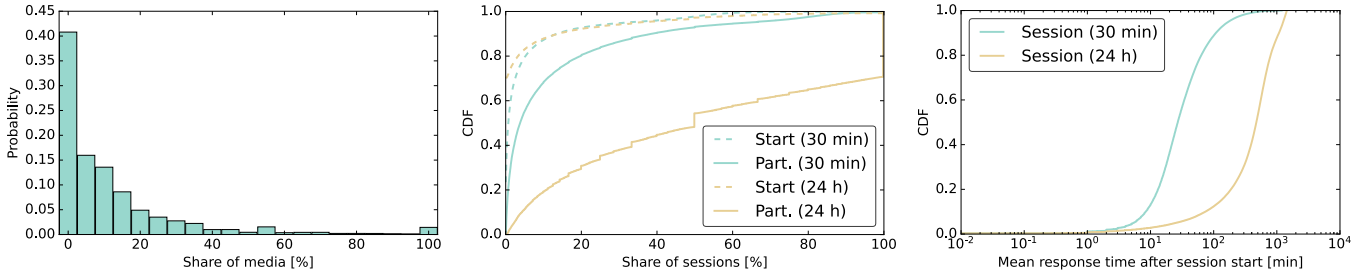
Thus, we can clearly see two different types of WhatsApp users. While most users only sporadically communicate within a private chat group, there are a few users, who send messages very frequently. However, the differences in the contributions of users might not only depend on the personality of the user, but also on the group size, the interest in the conversation topic, or the perceived importance of that particular group chat, which was collected in our data set.

To evaluate the quantitative number of individual contributions, Figure 5c analyzes the **mean text message length per user** on a logarithmic x-axis. We see that the green curve, which indicates the CDF of the mean number of characters per message, increases slowly up to 20 characters (21.43% of users). Afterwards, it increases strongly until it flattens for a mean number of more than 95 characters (94.99% of users). The median of the mean text message length is 31.21 characters, the average is 40.13 characters. Considering that the average word length in English is 4.79 characters [19], this translates to roughly 8 words (including whitespaces) in an average message of an average user. This shows that the quantitative content of WhatsApp message is rather low. Potential reasons might be the informal character of many conversations in private groups, or the high speeds of conversations, which require to shorten individual messages.

Considering the usage of emojis in the yellow CDF, it can be seen that 24.48% of users never use any emoji in their text messages. This also means that emojis are well adopted by a vast majority of users. Overall, an average user inserts 0.52 emojis per message, which means that 1 emoji is used in roughly every 2 messages. However, the usage of emojis is generally limited, considering that 95.59% of the users use 2 or less emojis per message on average. Compared to the mean number of characters, it can be followed that emojis are far from replacing characters in text messages. Still, many users frequently equip messages with emojis.

One could imagine that users, which frequently and burstily send messages, write shorter messages due to reduced typing time or the splitting of a contribution over several messages. However, the assumption that the mean number of characters of text messages is correlated to the burstiness of the message sending behavior cannot be confirmed, as the mean message length has no correlation to the mean IAT of a user ($\rho = -0.0209$), and only a small correlation of $\rho = 0.2025$ to the mean response time of a user, which is the time a user needs to answer to the previous message. This means that short interarrival times and long message lengths do not exclude each other, probably due to fast typing skills or advanced text input methods on mobile devices (e.g., predictive text, autocorrection, swiping).

Next, the **sharing of media** is investigated for the different users. In Figure 6a, we can see that there are many users (40.80%) who never send media. While the mean ratio of media posts per user is 10.31%, there are also several users



(a) Share of media in own messages. (b) Ratio of started and participated sessions. (c) Mean response time for participation.

FIGURE 6. **Characteristics of media usage, session participation, and response times.**

who send more media than text messages (3.12%). Some users seem to send media instead of long texts, e.g., a picture is worth a thousand words, as the correlation ρ to the mean length of the text messages of the user is -0.1843 and thus can be considered as slightly negative correlated. Thus, it can be concluded that for some people media usage replaces text usage, but as the correlation is only very low, for most users it rather supplements the mainly text-based conversation.

Lastly, the communication behavior of users is analyzed with respect to **sessions**, i.e., phases of active communication within the group. Thereby, a session is defined via a pause threshold, i.e., a timeout value for the session end. This means, if no message is replied within the defined pause threshold after the last message, the session is considered to be ended, and the next sent message is considered to start a new session. Two fixed pause thresholds are considered, namely, 30 minutes and 24 hours. A threshold of 30 minutes considers the fast-paced communication in WhatsApp and the ''always on'' phenomenon of smartphone users, which can be informed about new messages via smartphone push notifications and can instantly join and continue the conversation due to ubiquitous Internet connectivity. Moreover, 30 minutes are a typical think-time, which is also often utilized to split web sessions, e.g., as proposed in [20]. The threshold of 24 hours will leave sessions open for a much longer time to account for users who are using WhatsApp less frequently and for slow paced conversations, such as event planning or discussions about business topics.

We begin by investigating how often users start a session of the conversation within the private group. Figure 6b shows the CDF of the ratio of sessions, which were started by users (dashed) and in which they actively participated by sending a message (solid). The green curves represent the short threshold of 30 minutes, while the yellow curves correspond to the long threshold of 24 hours. Note that each start of a session with threshold 24 hours is also a start of a session with threshold 30 minutes. The mean ratio of started sessions is 5.06% (30 minutes) and also 5.06% (24 hours). However, it can be seen that 70.15% of the users never started a session after a pause of 24 hours. This confirms that only few users dominate the communication in groups, such as reactivating the communication by starting new sessions after a long pause, while the remaining users rarely or never contribute to a conversation. For shorter pauses of 30 minutes, this ratio is much smaller (23.51% of the users), which can also be traced back to users that only sporadically follow a conversation, and thus, their responses to the last message are sent later than 30 minutes. On the other side, the 99th percentiles of started sessions are only 56.70% (30 minutes) and 80.00% (24 hours), respectively. This means, in most cases there are no single users, who dominate the chat activity being the only users who start sessions within the conversation.

When considering the participation in sessions (solid curves), on average, a user participates in 12.68% of short sessions, and 52.77% of long sessions. It is clear that a higher activity can be observed, as session starts are a subset of session participations. It can be seen that both CDFs increase faster for lower ratios of sessions, which means that still many users only participate in a small share of the sessions. However, for a session threshold of 24 hours, we see a share of 29.33% of the users, which participated in every session of their group chat. Note that the high share can also be partially attributed to large groups, in which messages are sent every day, since these groups only consist of a single session with respect to the session timeout of 24 hours. The same cannot be observed for a threshold of 30 minutes, since much more sessions exist, and almost no user could participate in every session due to the brevity of resulting sessions. Nevertheless, as the green CDF indicates, 0.70% of the users manage to participate in at least 90% of the sessions. This shows that there is a substantial amount of users, which show a high motivation to engage in ongoing conversations in their group.

Now, we investigate how quickly users react after a new session has been started. Thereby, we consider the **response time**, which is the time difference from the start of a session until the first message of the users, if they participate in the corresponding session. Note that this case considers only sessions, which were started by a different chat member. Figure 6c depicts the CDFs of mean response time per user for both session thresholds (30 minutes in green, 24 hours in yellow) on a logarithmic x-axis. 0.55% of the users, who participate in a short session, are responding on average

within the same minute. 12.91% of the users respond on average within 10 minutes to the session starter, and 54.47% within 30 minutes on average. The mean of the mean response time for this threshold is 50.18 minutes. This shows that users, who participate in short sessions, actively follow and join the conversations and respond quickly. However, since the mean response time is longer than the session threshold, many users wait for other messages before they reply themselves.

For the longer session threshold of 24 hours, the mean response times are generally longer as the sessions contain more messages and have an equal or higher duration. Only 8.65% of users respond on average within 1 hour, but many users respond to session starters rather slowly considering that the mean of the mean response time for this session threshold is 8.67 hours. As the mean is very high, this shows an even more reluctant response behavior. Thus, we found that many users take a lot of time to respond, either due to using WhatsApp and following conversations less frequently, or due to high passivity during conversations, which includes waiting for other users to reply before themselves.

To sum up, it could be seen that most members of group conversations are rather passive and rarely start sessions. The ratio of participated sessions is low for most users and also mean response times can take large values for some users, which both might indicate high passivity during the conversations. Still, a significant amount of users respond in every session, especially for longer session thresholds. The response times with respect to the session start are short for most users compared to the session thresholds, which indicates that many participating users are very responsive and reply shortly. However, there is also a large share of users, which are reluctant to participate, such that they consistently reply late, if they reply at all.

V. TOWARDS SOURCE TRAFFIC MODELS FOR WhatsApp CHAT GROUPS

In order to understand and investigate the impact of WhatsApp chat group traffic on networks, source traffic models are extremely useful. They are mathematical or programmatic models, which are able to describe or imitate the traffic generated in MIM group communication. Thus, they can be used as input to simulation studies of communication networks.

However, it is obvious that different research questions require different models, for example, when investigating the traffic of a single chat group vs. many chat groups, or when being interested in traffic on time scales of hours vs. seconds. Thus, in this work, we refrain from providing a single source traffic model for WhatsApp group communication, but instead, describe how the published data set and the presented analyses in this work can be used to generate and validate source traffic models, which perfectly fit the research question(s) at hand. For this, we focus on the use case of modeling the generated traffic in access networks.

First, it is important to **model the communication** in MIM applications. This includes to decide on the desired number of chat groups, which shall be considered, and select their sizes according to the group size distribution, cf. Figure 2a. When simulating the shared messages in each group, different options are available. For example, the simple, general IAT distribution in Figure 2c can be used, when message types are not differentiated. As with all CDFs, inverse transform sampling [21] can be easily applied to draw random numbers, which exactly follow the desired distribution.

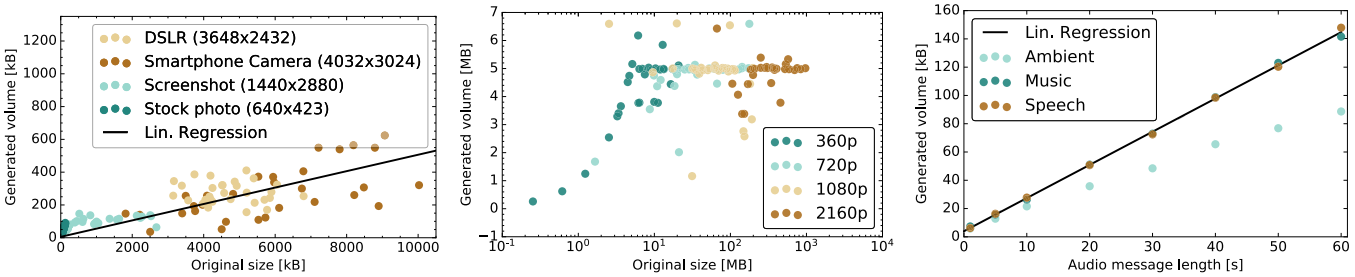
However, improved accuracy can be reached by considering different message types, e.g., using a semi-Markov model. Here, it is required to obtain a matrix of transition probabilities, which describes the transitions between different media types as depicted in Figure 3b, as well as separate IAT distributions per transition, which were not presented in this paper but can be obtained from the published data set.

Of course, the granularity can become even more fine-grained, e.g., by following non-Markovian approaches or by additionally considering properties of messages, such as the distribution of text lengths, cf. Figure 3c. Also different user types could be considered, thus, composing the chat groups with members having different daily activity patterns, cf. Figure 5a, ranging from frequent posters to silent readers, cf. Figure 4c. The data needed to obtain such fine-granular models is available in the published data set, which, ultimately, also allows to replay communication in chat groups exactly as it happened.

Note that when implementing the communication models in a simulation, apart from selecting appropriate input distributions, it is important to distinguish transient and steady-state behavior and to follow best practices for building valid, credible, and appropriately detailed simulation models, such as [21]. Finally, the resulting communication models and simulated group chats can be validated using the presented analyses in this paper, which have not been used as an input to the model, for example, considering the resulting overall average IAT per group size (cf. Figure 4a) or the resulting share of media in group chat messages (cf. Figure 4b) when implementing the above described semi-Markov model for a large number of diverse chat groups.

While the communication model is crucial, in order to estimate the generated network load, a source traffic model needs to consider the resulting data transmissions when sharing messages in chat groups. In its simplest form, this includes to model the data volumes of the messages, however, more fine-granular models could additionally consider packet-level characteristics of the resulting traffic, such as packet sizes or bursts. To obtain the required data, network measurements can be conducted using a simple testbed.

Here, we exemplarily **model the data volumes of the most popular message types**, i.e., text, image, video, and audio, via measurements. Our tests are executed on a Google Pixel 2XL smartphone running Android 11 and the current WhatsApp version (2.22.9.78). It is connected via USB to a PC using reverse tethering, thus, relaying the traffic to the PC



(a) Resulting traffic for image sharing.

(b) Resulting traffic when sharing 20 s videos.

(c) Resulting traffic for audio messages.

FIGURE 7. **Resulting traffic for most popular media types.**

for *tshark* [22] monitoring and capture. We filter the captured traffic for packets from the smartphone to WhatsApp servers and consider the generated volume of a message as the sum of the payload of the uplink packets. Due to end-to-end encryption, the measurements could include uplink traffic, which is not related to message sending, e.g., key exchanges, online status updates, or typing notifications. However, we consider this a minor impact for this exemplary measurements. Note that, since we send all messages to a second phone in our lab, we are also able to inspect the shared media at the receiver.

We use four data sets, namely, (1) text from [23] and [24] in different lengths for measuring text message volumes, (2) smartphone pictures and screenshots, DSLR images [25], and low resolution stock photos [26] for measuring image message volumes, (3) a subset of the YouTube UGC data set [27] for measuring video message volumes, and (4) music, speech, or ambient noise recorded via WhatsApp's audio message feature for measuring audio message traffic.

For text messages, the generated traffic volume of different texts and lengths is used to fit a linear regression model which results in $y = 695\,\mathrm{B} + \sum_{c=0}^{l} 3.17 \cdot be_c$ where y is the generated volume and l is the text length. For each character c, be_c is the UTF-8 byte equivalent of c. For example, an ASCII character needs 1 B (byte), while a basic emoji requires 2-7 B [28]. The goodness of fit is confirmed with a high coefficient of determination of $R^2 = 0.9972$. The fixed size of 695 B is the message overhead most likely containing metadata, such as sender id, group chat id, and timestamp.

For image messages, Figure 7a shows the generated traffic volume based on the original file size for 30 images in each category. Although no public information about the compression mechanisms of WhatsApp are available, it can be seen that the image is compressed by the application before sending, such that even large images of up to 10 MB are generating traffic volumes of less than 700 kB in the network. On average, the generated volume per image in our data set is 174.5 kB and the compression ratio is 14.7, for which all images, except the already small stock photos, have a lower resolution at the reception side than the original images. In general, we again observe a linear trend of $y = 5.6\,\mathrm{kB} + s \cdot 0.05\,\mathrm{kB}$ ($R^2 = 0.6148$), i.e., the larger the original file size s, the higher the generated traffic volume, with an average message overhead of 5.6 kB.

Figure 7b depicts the resulting traffic based on our video data set, which consists of 30 videos per resolution (360p, 720p, 1080p, 2160p) with a length of 20 s each. Similar to image media, also video files are compressed before sending and have a resolution of 360p at the receiver. However, contrary to images, for videos of at least 720p, we can't see a strong relationship between original video size and generated volume. Instead, most of the videos larger than 5 MB are compressed to about 5 MB and generate similar traffic volume. On average, the generated traffic volume for the tested videos is 4.63 MB, the compression ratio is 27.65, and the overhead is 45.1 kB. We also tested different video lengths confirming that the resulting traffic volume is proportional to the length, i.e., a 40 s video generates twice the traffic of a 20 s video, and thus, hinting at a default target bitrate of around 2 Mbps. However, when the application estimates a resulting video size of more than 16 MB, the user can only send a specific part of the video [29].

For audio messages, the generated volume per message length is presented in Figure 7c. It can be seen that the resulting traffic for messages containing music and speech follows a linear trend and is very similar with overlapping points in the figure. The resulting function depending on the message length l in seconds is $y = 3.95\,\mathrm{kB} + l \cdot 2.35\,\mathrm{kbit\,s^{-1}}$, which gives a high $R^2 = 0.9971$. For audio messages with only ambient noise the resulting traffic volume is 33.7% lower. Finally, with audio messages, we made the interesting observation that they show a different behavior on the network. While image and video messages are sent as bulk after compression, audio chunks are sent periodically to the server during message recording. Thus, the resulting traffic shows similarities to streaming traffic.

To sum up, source traffic models for MIM group communication depend on two important ingredients. First, there are models for communication behavior, which could be derived from the published WhatsApp data set and the presented analyses in this paper. They are rather stable over time, thus our insights should stay useful in the next years, and might well be generizable to other Mim applications. Second, there are models for generated traffic, which we decided to keep separately. The reason is that the observed traffic may be

different for other MIM applications due to different compression methods and may likely change over time, which has happened often in the past due to technological advancements. For example, in the future, we might expect MIM applications to transmit lossless formats or even original files when exchanging media content, as our desire for improved quality increases. However, the presented methodology is general and can easily be applied to measure generated traffic with this or other applications now or in the future. Thus, the resulting traffic models can be easily combined with the communication models to describe and simulate the traffic generated from group chats. This will be of great value to network operators, which need to plan and manage their network accordingly to efficiently cope with multiplying MIM traffic while reaching a high user satisfaction.

VI. CONCLUSION AND OUTLOOK

As mobile instant messaging (MIM) applications allow users to communicate in groups with many people at any time of day and from anywhere in the world, the way people communicate evolved in recent years. The sharing of text and media messages in chat groups multiplies the generated traffic on the network, which can have a massive impact on the underlying communication networks. However, due to end-to-end encryption, measurements cannot bring valuable and highly needed insights for network operators into group communication and the underlying processes, which drive the generation of traffic in such groups.

To address this issue, we collected and published 5,956 WhatsApp private chat histories. Using this data set, we created models describing the properties of groups, the users, and their communication. In addition, we also conducted exemplary network traffic measurements to quantify the resulting data transmissions when sending different types of WhatsApp messages. Finally, we presented a guideline on how to use the given data set and the presented communication models in combination with network traffic measurements to develop source traffic models for describing and simulating WhatsApp communication behavior and the resulting traffic. The big advantage is that these models allow for different granularities, and thus, can be perfectly aligned with the particular research questions.

In future work, we plan to investigate several research questions related to the network load of MIM applications by implementing appropriate source traffic models in simulation environments. In particular, using these models, we will investigate how the increasing network load of MIM applications can be properly managed, e.g., using edge caching.

ACKNOWLEDGMENT

The authors alone are responsible for the content.

REFERENCES

[1] E. J. Vergara, S. Andersson, and S. Nadjm-Tehrani, "When mice consume like elephants: Instant messaging applications," in *Proc. 5th Int. Conf. Future Energy Syst. (ACM e-Energy)*, Cambridge, U.K., Jun. 2014, pp. 97–107.

[2] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting WhatsApp through large-scale measurements in mobile networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 133–134, Feb. 2015.

[3] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Understanding the nature of social mobile instant messaging in cellular networks," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 389–392, Mar. 2014.

[4] E. Hernández-Orallo, J. Herrera-Tapia, J.-C. Cano, C. T. Calafate, and P. Manzoni, "Evaluating the impact of data transfer time in contact-based messaging applications," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1814–1817, Oct. 2015.

[5] S. Lin, W. Zhou, and J. Liu, "Network traffic and user behavior analysis of internet-based mobile messaging applications: A case of WeChat," in *Proc. 8th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, Aug. 2016, pp. 567–572.

[6] Q. Deng, Z. Li, Q. Wu, C. Xu, and G. Xie, "An empirical study of the WeChat mobile instant messaging service," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 390–395.

[7] G. De Luca Fiscone, R. Pizzolante, A. Castiglione, and F. Palmieri, "Network forensics of WhatsApp: A practical approach based on side-channel analysis," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.* Cham, Switzerland: Springer, 2020, pp. 780–791.

[8] C. Shubha, S. A. Sushma, and K. H. Asha, "Traffic analysis of WhatsApp calls," in *Proc. 1st Int. Conf. Adv. Inf. Technol. (ICAIT)*, Jul. 2019, pp. 256–260.

[9] R. Cents and N.-A. Le-Khac, "Towards a new approach to identify WhatsApp messages," in *Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2020, pp. 1895–1902.

[10] M. Seufert, A. Schwind, T. Hoßfeld, and P. Tran-Gia, "Analysis of group-based communication in WhatsApp," in *Proc. Int. Conf. Mobile Netw. Manage.* Cham, Switzerland: Springer, 2015, pp. 225–238.

[11] M. Seufert, T. Hoßfeld, A. Schwind, V. Burger, and P. Tran-Gia, "Group-based communication in WhatsApp," in *Proc. IFIP Netw. Conf. (IFIP Networking) Workshops*, May 2016, pp. 536–541.

[12] K. Garimella and G. Tyson, "WhatsApp, doc? A first look at WhatsApp public group data," 2018, *arXiv:1804.01473*.

[13] 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 *Proc. World Wide Web Conf.*, May 2019, pp. 818–828.

[14] A. Rosenfeld, S. Sina, D. Sarne, O. Avidov, and S. Kraus, "WhatsApp usage patterns and prediction of demographic characteristics without access to message content," *Demographic Res.*, vol. 39, pp. 647–670, Sep. 2018.

[15] A. Schwind and M. Seufert, "WhatsAnalyzer: A tool for collecting and analyzing WhatsApp mobile messaging communication data," in *Proc. 30th Int. Teletraffic Congr. (ITC)*, vol. 1, Sep. 2018, pp. 85–88.

[16] WhatsApp. *WhatsApp FAQ—How to Save Your Chat History*. Accessed: Oct. 21, 2022. [Online]. Available: https://faq.whatsapp.com/android/chats/how-to-save-your-chat-history/

[17] (2011). *WhatsApp Blog—One Billion Messages*. Accessed: Oct. 21, 2022. [Online]. Available: https://blog.whatsapp.com/one-billion-messages

[18] M. Singh. (2020). *WhatsApp is Now Delivering Roughly 100 Billion Messages a Day*. Accessed: Oct. 21, 2022. [Online]. Available: https://tcrn.ch/3jHg1YC

[19] P. Norvig. (2012). *English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU*. Accessed: Oct. 21, 2022. [Online]. Available: http://norvig.com/mayzner.html

[20] L. Vassio, I. Drago, M. Mellia, Z. B. Houidi, and M. L. Lamali, "You, the web, and your device: Longitudinal characterization of browsing habits," *ACM Trans. Web*, vol. 12, no. 4, pp. 1–30, Nov. 2018.

[21] A. M. Law, *Simulation Modeling and Analysis*, 5th ed. New York, NY, USA: McGraw-Hill, 2015.

[22] G. Combs. (2022). *TShark—Dump and Analyze Network Traffic*. Accessed: Oct. 21, 2022. [Online]. Available: https://www.wireshark.org/docs/man-pages/tshark.html

[23] Wasai LLC. (2015). *Lorem Ipsum*. Accessed: Oct. 21, 2022. [Online]. Available: https://loremipsum.io/

[24] P. Tracy. (2015). *Office Ipsum*. Accessed: Oct. 21, 2022. [Online]. Available: http://officeipsum.com/

[25] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3277–3285.

[26] Pexels GmbH. (2014). *The Best Free Stock Photos, Royalty Free Images & Videos Shared by Creators*. Accessed: Oct. 21, 2022. [Online]. Available: http://pexels.com/
[27] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 1–5.
[28] The Unicode Consortium. (2021). *Full Emoji List, V14.0*. Accessed: Oct. 21, 2022. [Online]. Available: https://unicode.org/emoji/charts/full-emoji-list.html
[29] WhatsApp. *I Get a Message That My Video is Too Long and it Won't Send*. Accessed: Oct. 21, 2022. [Online]. Available: https://faq.whatsapp.com/general/i-get-a-message-that-my-video-is-too-long-and-it-wont-send/?lang=en

ANIKA SEUFERT received the master's degree in computer science from the University of Würzburg, Germany, in 2017, where she is currently pursuing the Ph.D. degree. She is also a Research Assistant with the Chair of Communication Networks, University of Würzburg. Her research interests include user-centric modeling of mobile applications, measuring internet experience from end-user perspective, and optimizing end users' quality of experience.

FABIAN POIGNÉE received the master's degree in computer science from the University of Würzburg, in 2020, where he is currently pursuing the Ph.D. degree. He is also a Research Assistant with the Chair of Communication Networks, University of Würzburg. His research interests include quality of experience of internet applications, artificial intelligence and machine learning for networks, and group-based communication systems.

MICHAEL SEUFERT (Member, IEEE) received the bachelor's degree in economathematics and the Diploma and Ph.D. degrees in computer science from the University of Würzburg. He is currently the Research Group Leader of the Chair of Communication Networks, University of Würzburg, Germany. He holds the first state examination degree in mathematics, computer science, and education for teaching in secondary schools. His research interests include user-centric communication networks, including QoE of internet applications, AI/ML for QoE-aware network management, and group-based communications.

TOBIAS HOßFELD (Senior Member, IEEE) received the Ph.D. degree, in 2009. His professorial thesis (Habilitation) "Modeling and Analysis of Internet Applications and Services," in 2013, with the University of Würzburg, Germany. He was heading the "Future Internet Applications & Overlays" research group. From 2014 to 2018, he was the Head of the Chair "Modeling of Adaptive Systems" with the University of Duisburg-Essen, Germany. He has been a Professor with the Chair of Communication Networks, University of Würzburg, since 2018. He has published more than 100 research papers in major conferences and journals, and received five best conference paper awards, three awards for his Ph.D. thesis, and the Fred W. Ellersick Prize 2013 (IEEE Communications Society) for one of his articles on QoE. He is also a member of the Advisory Board of the ITC Conference and the Editorial Board of IEEE Communications Surveys & Tutorials.