WhatsApp Usage Patterns and Prediction Models*

Avi Rosenfeld[1], Sigal Sina[2], David Sarne[2], Or Avidov[2], Sarit Kraus[2,3]
[1]Department of Information Systems, Jerusalem College of Technology, Israel
[2]Department of Computer Science, Bar-Ilan University, Israel
[3] Institute for Advanced Computer Studies, University of Maryland, USA
rosenfa@jct.ac.il, sigal.sina@gmail.com, sarne@cs.biu.ac.il, oravi90@gmail.com, sarit@cs.biu.ac.il

Abstract

This paper presents an extensive study of the usage of the WhatsApp social network, an Internet messaging application that is quickly replacing SMS messaging. It is based on the analysis of over 4 million messages from nearly 100 users that we collected in order to understand people's use of the network. We believe that this is the first in-depth study of the properties of WhatsApp messages with an emphasis of noting differences across different age and gender demographic groups. It is also the first to use statistical and data analytic tools in this analysis. We found that different genders and age demographics had significantly different usage habits in almost all message and group attributes. These differences facilitate the development of *user* prediction models based on data mining tools. We illustrate this by developing several prediction models such as for a person's gender or approximate age. We also noted differences in users' *group* behavior. We created group behaviorial models including the likelihood a given group would have more file attachments, if a group would contain a larger number of participants, a higher frequency of activity, quicker response times and shorter messages. We present a detailed discussion about the specific attributes that were contained in all predictive models and suggest possible applications based on these results.

Introduction

Internet social networks have become a ubiquitous application allowing people to easily share text, pictures, audio and video files. Popular networks include Facebook, Reddit and LinkedIn which all maintain websites which serve as hubs facilitating people's information sharing. In contrast, the relatively new WhatsApp application is a smartphone application that enables people to share information directly via their phones. Since its introduction in 2009, its growth has steadily increased, and as of September 2015, it numbers over 900 million users[1]. While many alternatives to WhatsApp are currently available in different online application stores (e.g., Kik, Telegram, Line Messenger, BBM, WeChat), WhatsApp is currently the most popular messaging application with the largest name recognition, by far the largest user base, and the strongest corporate backing since its acquisition by Facebook in 2014. Given the emerging importance of this network it is not surprising that there is a growing interest in researching it, including user studies about people's WhatsApp use and possible applications (Fiadino, Schiavone, and Casas 2014; Church and de Oliveira 2013; Pielot et al. 2014; O'Hara et al. 2014; Bouhnik and Deshen 2014; Mudliar and Rangaswamy 2015; Montag et al. 2015; Johnston et al. 2015).

This paper contains two main contributions. First, it is the first work to study WhatsApp usage at the message level, basing itself on all message information short of the actual content of the message. While the messages' content was not stored due to privacy concerns, we did have general message information such as the messages' length, the size of the conversation group it was sent to and temporal properties such as the time it was sent and how much time elapsed between this message and the previous one. Prior WhatsApp work, as discussed in more detail in the following section, typically excludes message information and instead uses surveys and questionnaires constructed by the authors to test specific hypotheses. Second, we successfully created models that predict usage patterns between different types of users and groups. While such tools have been used to study other social networks, including Facebook (Wang, Burke, and Kraut 2013; Xiang, Neville, and Rogati 2010; Bakshy et al. 2012) and MySpace (Thelwall, Wilkinson, and Uppal 2010), applying these tools to the WhatsApp network is significantly more complicated because no public dataset currently exists, in contrast to these other networks. This is likely because of the medium involved – while other social networks are primarily web-based and thus given to compiling data through web crawling, the WhatsApp network is based on individuals' private phone use and thus not publicly available. Furthermore, these studies typically use the messages' actual content, something we did not have access to.

*This research is based on work supported in part by MAFAT and ISRAEL SCIENCE FOUNDATION (grant #1488/14).

[1]`http://fortune.com/2015/09/04/WhatsApp-900-million-users/`

As we further describe in the following sections, we performed an in-depth study based on WhatsApp messages and conversation groups through collecting over 4 million WhatsApp messages from 92 users. Through analyzing this data, our study revealed several key insights. First, we did in fact find significant differences in WhatsApp usage across people of different genders and ages. Second, we inputted the data into the Weka data mining package (Witten and Frank 2005) and studied the output from decision tree and Bayesian network algorithms, mainly as a proof of concept for the kind of results one may extract by applying machine learning and data mining tools on WhatsApp data when collected in the message level, even without getting exposed to the content itself. These algorithms were successful in building models that can accurately predict a person's gender and approximate age and are able to predict which WhatsApp groups have certain qualities, such as higher percentages of file attachments, quicker responses, larger discussion groups, and shorter messages. One key advantage in analyzing the results from the decision tree algorithm is that it outputs an unbiased assessment about which attributes and logical rules were important in building these prediction models, thereby providing additional insights. Last, we noted the importance of these results with possible future directions and applications.

Related Work

The WhatsApp social network is unique in several ways relative to other social networks. This application was developed to allow users to privately and freely send messages to each other through their smartphones. It provides a free alternative to SMS (Short Message Services) which is often still a metered (pay per use) service. Not only is WhatsApp often more cost effective than SMS, but it facilitates large group conversations, something that is difficult, if not impossible, through SMS. While freely sharing information over the Internet is common to many social networks, and other public messaging services, such as Twitter, exist, the private nature of the WhatsApp network makes it rather unique. A similar difference between WhatsApp and other social networks is that membership is created and updated directly via people's smartphones. Not only is registration done exclusively through one's phone number, but the smartphone is the primary interface for sending and receiving messages.[2] Third, WhatsApp conversation groups are the network's only communication medium and are formed by adding people's telephone numbers to that group. In contrast, other social networks are based on user membership and primarily focus on public messages where these messages are sent to all connected users (i.e these messages are called *Posts* in Facebook and *Tweets* in Twitter), and not through private groups. Given these and other differences between WhatsApp and other social networks, we believe that existing research about other networks is not necessarily applicable and a new and thorough analysis of WhatsApp is warranted.

Much recent work has been dedicated to the study of how people use WhatsApp and the role of this new application in social communication. Most works to date have analyzed peoples' behavior through conducting surveys and targeted interviews. For example, work by Church and Oliveira conducted on online study asking targeting questions to users aimed at understanding differences between WhatsApp to SMS usage (Church and de Oliveira 2013). Pielot et. al (Pielot et al. 2014) created a survey focusing on the question of if people expected an answer to their WhatsApp and SMS messages within several minutes. and O'Hara et. al interviewed 20 WhatsApp users for nearly an hour each, asking them semi-structured questions aimed at determining the nature of relationships forged with the people they communicated with (O'Hara et al. 2014). Mudliar and Rangaswamy (Mudliar and Rangaswamy 2015) spent over 350 hours observing 109 students, as well as conducted surveys to understand gender differences within Indian students' use of WhatsApp. All of these studies can be characterized as being formed through a desire to answer specific questions through conducting targeted surveys and interviews to answer that question.

This work is unique in that it uses statistical and data mining methods to study WhatsApp usage at the message level even without knowing the content of these messages. Our study contains the same motivation of previous WhatsApp research in that we also analyze differences between genders, the time that elapses between a message is answered, and the characteristics of larger and smaller discussion groups. However, our study is fundamentally different in that we are based solely on actual WhatsApp message data to perform our analysis without any possible human bias. The issue of human bias within smartphone usage analysis was recently studied and one of the study's conclusions was that people poorly reported their own usage in questionnaires (Lin et al. 2015). To our knowledge, only one other study, performed by Montag et. al (Montag et al. 2015), studied WhatsApp usage through logging data from nearly 2500 participants. While the number of participants in this study is impressive, the actual data logged was significantly less robust than in this study as they only collected general meta-data about use, only limited information about WhatsApp messages and no information about users' group activity.

In theory, even more accurate models may have been constructed had we analyzed messages' content. In fact, many such models that have been previously developed for other domains to successfully predicted an author's gender, age, native language or personality (Wang, Burke, and Kraut 2013; Argamon et al. 2009). For example, work by Argamon (Argamon et al. 2009) focused on creating models that identified word us-

[2] While we note that a computer interface for WhatsApp exists, it is exclusively an interface for people's smartphones and offers no additional functionality.

age differences between men and women on Internet blogs. Similarly, Wagner et. al (Wagner et al. 2015) focused on content differences between men and women in Wikipedia and Wang, Burke and Kraut performed a study of content differences between genders in Facebook (Wang, Burke, and Kraut 2013). However, as the WhatsApp network is inherently private, such approaches could not be applied due to privacy concerns. As we now detail, even despite this information we were indeed similarly successful in predicting a user's demographic and group behavior.

Dataset Creation and Description

Given the private nature of the WhatApp network, this study's first challenge was to create a WhatsApp message dataset while still insuring users' privacy. To do so, we developed software that integrated with the Android Debug Bridge (the ADB is an external tool which is able to backup an Android application). This enabled taking a "snapshot" of a person's groups and messages as appear in her phone. In order to make the data anonymous, the software encrypts the data that was pulled directly from the participant's smartphone by using the HMAC hash function. The entire process of obtaining a participant's data lasted approximately 15 minutes and we compensated each participant $12 for their time and temporary inability to use their phones. We also collected the participants' general demographic information including their age, gender, place of residence and educational background. In addition, we also asked them to self-rate their sociability and WhatsApp usage on a five-point Likert scale (Low to High), and four Boolean questions if they use WhatsApp for communication with work, family, friends or others. An IRB was obtained for ethical approval prior to beginning data collection.

We found it challenging to recruit participants, as people were quite reluctant to provide information about their WhatsApp messages, even when we emphasized that all content sent was encrypted, and that no non-encrypted content data was ever sent. While we attempted to recruit participants from all age groups, we found that students participants, found through advertisements on campus, were the demographic most willing to participate. Nonetheless we did make a concerted effort to find people in other demographics through word-of-mouth. Overall, we logged messages from 92 participants, of which 55 were female and 37 were male. 62 of the 92 participants were students and the remaining non-students. The participants were between 12 and 64 years of age with a median of 25. See Figure 1 for the complete age distribution. While this distribution is skewed in the sense of over-representing the demographic group of people younger than 35 and does not necessarily represent the true distribution of WhatsApp users in the general population[3], this is primarily due to the challenging data collection process as discussed above. Still, this does not produce a substantial bias in the statistics provided, as our analysis, and particularly the prediction models provided in the paper, still was overall successful in differentiating between these two segments of participants. The 92 participants sent and received a total of 4,355,783 messages, over an average period of approximately 15 months (average 14.96 months).[4]

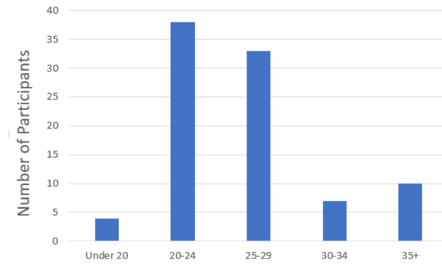

Figure 1: The age distribution of participants in the study

The defining characteristic for the logged data is that it intentionally contains no textual content. All types of textual content are unavailable, including if special characters or emojis exist in the messages. Similarly, we stress that we have no information about the message recipients as all data is anonymous.

While we did not have messages' content or recipient information, we were nonetheless still able to glean a great deal of usage information regarding message and group statistics. The first type of information focused on general information surrounding the messages' characteristics such as when they were sent, their length and the messages' response time. Once we had all messages, we discretized their time into categories based on the percentages of messages sent over each hour interval– e.g. messages between 12 and 1 A.M. Similarly, we aggregated information about all messages' length and created categories with the total number of messages with 1 or 2, 3–5, 6–10, 11–20, and 20+ words. We then aggregated messages according to the time that elapsed between them – under 1, 1–2, 3–5, 6–15, 16–30, and 31–60 minutes. Messages that appear within a relatively short time interval, within the same group, may be related to the same conversation. We emphasize that by all means this does not imply that a message that appear more than an hour after the last message in a given group is not related to former messages, except that with no other supporting data (e.g., the content itself) it is impossible to make a concrete connection to prior messages. Hence, the time elapsed is the only possible, even though not a perfect, indication for relevance. We also aggregated messages according to their file at-

[3]E.g., as in http://www.statista.com/statistics/290447/age-distribution-of-us-WhatsApp-users/.

[4]The software we used collected all the data on the phone, hence the time period over which data was collected varied according to when users started using WhatsApp and their habit of deleting old messages (if at all).

tachments and created Boolean categories of messages with and without files. The second type of information focused on logging information about WhatsApp conversation groups. Note that groups with two participants are similar to a typical SMS conversation, and thus through logging this data we could test the degree to which WhatsApp has replaced traditional SMS messaging. However, groups might also be formed around a general topic, such as a discussion about work, leisure or family issue with many more than two participants. We logged information about the group size of all of the messages and categorized this information into the percentage of messages in trivially small groups of 2 people, groups of 3–5 participants, and those with 5 or more participants. We also collected group statistics that subsume those within the message analysis, but refer to the percentage of messages within a group having a certain attribute – e.g. the percentage of messages sent at a certain time, of a certain length, contain a file, etc.

Given the processed data, we created 3 types of datasets: information related to the users' WhatsApp messages, their groups, and their overall average usage. The processed average usage information from the 92 users and their 4,355,783 messages were the sizes of these datasets. These users were part of 6,185 groups which formed the size of this dataset. In this work we analyzed all 3 datasets and found statistically significant differences in usage for people of different genders and ages across all datasets.

Dataset Analysis and Statistics

We analyzed the basic distribution of messages, focusing on the statistical distributions across different genders, ages, and types of use. We found that over 75% (76.11%) of WhatsApp groups had only two participants (4,678 out of the 6,185), confirming previous assertions that WhatsApp is replacing SMS messaging (Church and de Oliveira 2013). On the flipside, over 50% of all messages were not in groups of two (2,073,096 out of 4,355,783), indicating that larger groups typically were fruitful grounds for larger discussions–something that SMS typically does not support. To better understand this point, please note these differences using the graphical distributions of the number of groups of each size in Figure 2 and the distribution of all messages in those same groups in Figure 3. We note that the number of groups of size two is overwhelmingly large (76.11%), but the number of messages in these groups is significantly smaller (49.97%). Please also note that while the number of groups with over 51 members is only about 0.5%, these groups have a disproportionably large number of messages (5.72%). We believe that the reason for this is clear– larger groups tend to have larger numbers of messages in each group. Thus, we find that a large percentage of WhatsApp activity is in fact taking the place of traditional SMS messages between two people. However, group messaging among large numbers of users, another key use of WhatsApp

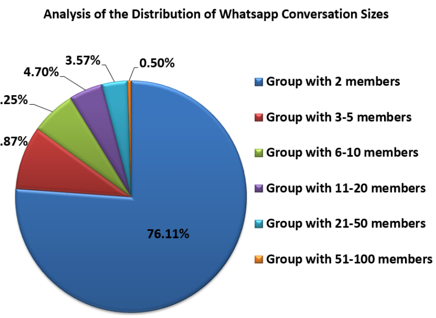

Figure 2: The distribution of the number of **groups** of each size

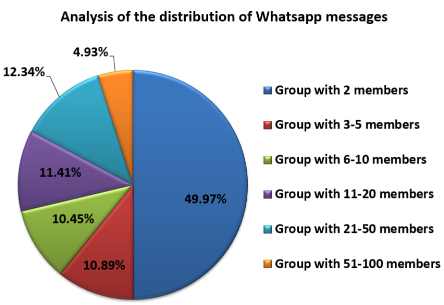

Figure 3: The distribution of the number of **messages** in each group size

which SMS is less successful in supporting, also constitutes a large percentage of the WhatsApp messages we collected.

We then studied the statistical distribution of the messages' attributes starting with the average response time (time elapsed between a message and a consecutive one when in conversation), found in Figure 4. Please note that the average response time is quite short. Nearly 1/3 (31.39%) of all messages are responses that were composed within 1 minute! This finding again confirms previous claims that WhatsApp has become a replacement to traditional SMS messaging, as most participants answer their messages quite quickly, something that is expected within SMS messaging (Church and de Oliveira 2013).

Next, we studied the distribution of the messages throughout the day, which is visually represented in Figure 5. As expected, very few messages were sent overnight with under 5% (4.36%) being being sent between midnight and 4:00 A.M. and only 2.37% being sent between sent between 4 and 8 A.M. Note that fewer number of messages were sent between 8:00 and noon (18.04%) compared to approximately 25% of all messages being sent in the other 4 hour intervals. In fact, we note no significant difference in the number of messages being sent in these three intervals (p-score > 0.1) while a significantly smaller number messages were sent between 8:00 and 12 P.M. (p-score $<< 0.01$).

We also analyzed the messages' types and length. Most of the messages (over 98%) are exclusively text

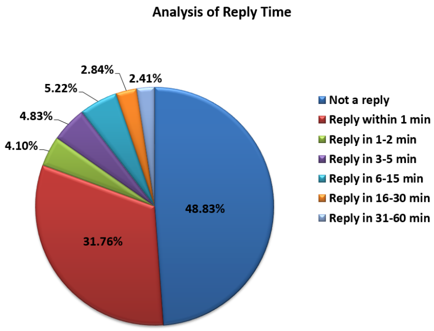

Figure 4: Analysis of reply time

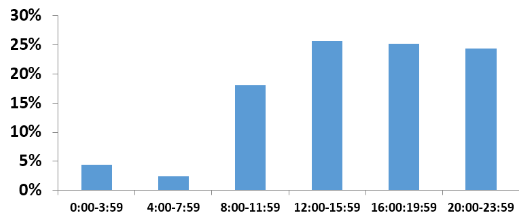

Figure 5: The distribution of messages per time of day

Table 1: WhatsApp statistics per gender

	Total	Male	Female	Higher Value	Ratio M/F
Participants	92	37	55	Female	0.67
AvgSent/Day	42.97	33.42	49.4	Female	0.68
AvgRcv/Day	103.15	97.18	107.17	Female	0.91
AvgMsg/Day	146.13	130.6	156.57	Female	0.83
RatioSend/Rcv	0.52	0.48	0.54	Female	0.89
AvgNumGroups	73.62	75.24	72.53	Male	1.04
% of Groups of 2	73.41%	72.53%	74.01%	Female	0.98
% of Groups of 3-5	9.64%	9.96%	9.43%	Male	1.06
% of Groups of 6-10	6.36%	5.85%	6.71%	Female	0.87
% of Groups of 11-20	5.27%	5.86%	4.87%	Male	1.2
% of Groups of 21-50	4.52%	4.57%	4.49%	Male	1.02
% of Groups of 51+	0.80%	1.23%	0.50%	Male	2.46

messages while only 2% included file attachments. Figure 6 shows the average length of all of the messages and also the average length of the participant's sent messages according to their gender. As expected in instant mobile messaging, most of the messages are short. About 33% of the messages contain 1 or 2 words, another 34% contain between 3 to 5 words, while less than 4% of the messages contain more than 20 words.

Recall that we only possess demographic information in all datasets for the 92 known users. Nonetheless, we can isolate messages these people sent to allow us to understand different message profiles according to these people's gender and age. For example, when we analyzed the length of messages sent by men and women, we found that women send longer messages than men. On average, women's messages include 6.5 words while men's messages include only 5.2 words. 32.86% men's sent messages include 1 or 2 words compared to 30.53% for women. At the other extreme of the distribution, 14.48% of women sent messages that included more than 10 words while 10.11% of the men sent messages of that length. While other works have noted that women often spend more time on WhatsApp than men (Montag et al. 2015), these results are the first to quantify these differences at the message level.

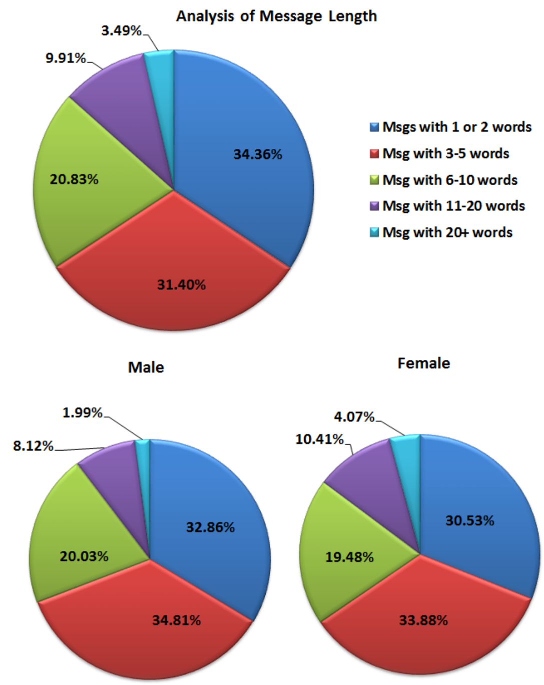

Figure 6: Distribution of message length overall (above), for men (lower left) and women (lower right)

Table 1 contains several additional gender related insights. First, we found that women on average sent and received more messages than men. Women sent and received nearly 150 messages a day while men sent and received approximately 130 messages (row 4), a difference of 15%. Of these messages, women sent on average almost 50 messages a day and received over 107 messages while men sent on average 33.4 messages a day and received about 97 (rows 2-3). Both genders have a similar 1:2 ratio between sent and received messages (row 5). Second, while on average both genders participate in a similar number of conversations (73-75 groups), the distribution of the various group sizes between the genders is different. Women are active in more smaller conversation groups (74.01% versus 72.53% in groups with two participants), but men are more active in larger groups (5.86% for men versus 4.87% for women in groups of 11-20, 4.57% versus 4.49% in groups of 21-50 and 1.23% versus 0.5% in groups bigger than 50) (rows 6-12).

Overall, we found that there were significantly different WhatsApp usage patterns between different genders and age groups. Table 2 provides details to support this claim where we present the general statistics of two different demographic groups: 1) men and women and 2) WhatsApp users younger than 25 (the median age) and older than 35 (significantly above the median age). Note the differences between the average number of total messages per day (AvgMsgDay), groups (AvgGroup/Usr) per user and differences from the users' responses to the questionnaire items where they self-rated their sociality (SocialLevel), overall usage (UsageLevel), and differences in the Boolean values (averaged based on values of 0 and 1), usage in communicating with friends (UsageFriend), family (UsageFamily), and work (UsageWork). In fact, we tested all pairs of numbers for statistical significance (2-tailed t-test) and found that **all** differences were significant (p-score << 0.05) except where noted with an "#" at the end of each pair, as is the case of the UsageWork numbers in the pair of people older than 35 and younger than 25. Additionally, we found significant differences in the usage patterns across group usage with people who were members of these different demographics. Note the differences in the average number of minutes a user took to respond to a message (AvgResponse), the percentage of their messages which were short – 5 words or less (Msgs5orLessWrd), the percentage of their messages which were quick responses within 5 minutes (%RespUnder5), the average message length (AvgTextLength), and the distributions of messages across different times (midnight until 4:00 A.M., 8:00 A.M. to 12:00 P.M., and 8:00 P.M. until midnight). We also found that usage styles were different in regard to the percentage of files found in users' groups of different genders and ages (UseFile) and the percentage of groups they were members of with 5 or more total users (isGrp5+).

We find some of the differences in Table 2 intuitive and others surprising. We are not surprised to find that younger people are more likely than older ones to send messages late at night and thus older people send a higher percentage of their messages during the day. One could find support for gender differences found in people's self-rating of how much they use WhatsApp to communicate with family versus work based on previously observed differences in gender expressions (Kring and Gordon 1998). However, we could not find a clear explanation as to why men seem to send more files in their groups than women or why older people participate in larger groups more often than younger people. These differences might point to new directions that might be confirmed with further research and questionnaires. For example, a possible hypothesis for the differences in group sizes across different ages is that younger people have more thoroughly adopted WhatsApp as a replacement for SMS messaging and consequently a larger percentage of their communication can be found in these smaller groups. We believe that these results show differences that can spawn future research and discussion.

Predictive Models and Hypotheses

As we demonstrated in the previous section, significant differences do in fact exist between different types of WhatsApp users and groups. However, even statistically significant differences do not necessarily allow us to predict usage patterns. For example, the previous section demonstrated that men typically send shorter messages and women send more messages per day. However, these differences do not necessarily allow us to make a prediction about a specific user – something that data mining algorithms do in fact allow, as we now present.

In order to illustrate the potential of using the data collected for prediction purposes we created several predictive models for the user and group datasets, which we describe in this section. In general, we created predictive models for *user* and *group* usage. User models were based on the 92 users in this dataset and were built to identify whether the author of a given set of WhatsApp posts is of a given **gender** or **age**.

Our first hypothesis is that differences between WhatsApp authors can be predicted by exclusively using general statistics about usage, even without specific user content. In accordance with the results reported in the previous section, we posit that such differences will likely use attributes such as message length and response time as such attributes may be impacted by known gender differences (Kring and Gordon 1998). As such, one might find that women write more to better express their ideas or emotions, while men write more curtly. Similarly, one might find that differences in response time or average conversation size are reflective of emotional difference – e.g. women may prefer discussion in small groups while men prefer less personal, larger discussions. In a similar vein, one might find differences between ages, even within one gender. Such differences may be somewhat trivial, such as the time a message is sent – e.g. people of certain ages might be more or less likely to work and thus be less likely to send messages at certain times, but non-trivial differences might exist too, such as differences in message length.

Our second hypothesis, based on the differences reported in accordance to the different statistics reported in the former section, is that different types of group usage can be predicted based on general group attributes, again even without considering the messages' content. Specifically, we develop models that predict which groups will have a certain type of content such as *file attachments* or *shorter messages*. We also develop group models that predict which groups will have certain user activity such as *more participants*, a *larger quantity or more frequent messages*, and *quicker response times*. In theory other usage questions could have been studied such as if a message contained certain text – e.g. inappropriate or flagged for a certain type of content. However, as we have no access to message content, these issues cannot be evaluated. Similarly, it may be possible that certain messages are inherently different and thus likely to be more popular or important. Along these lines models might be created to predict which messages are apt to have certain characteristics, such as being forwarded – something that was previously studied within the Twitter network (Naveed et al. 2011). However, once again that study focused on the message content, which is often infeasible to rely on in real-life settings, either due to privacy or availability.

	Participants	AvgMsgDay	AveGroup/Usr	AveAge	SocialLevel	UsageLevel	UsageFriend	UsageFamily	UsageWork
Overall	92	146.12	73.62	26.98	3.99	3.82	0.9	0.57	0.18
Male	37	130.6	75.24	25.3	3.96	3.64	0.93	0.34	0.21
Female	55	156.57	72.53	28.11	4	3.94	0.88	0.72	0.17
LessThan25	42	206.65	80.29	21.98	4.06	4.03	0.91	0.54	0.11#
GreaterThan35	10	54.75	55.8	48.4	3.52	3.06	0.76	1	0.11#
	AveResponse	Msgs5orLessWrd	%RespUnder5	AvgTextLength	Hours0-4	Hours8-12	Hours20-24	UseFile	isGrp5+
Overall	5.95	0.41	0.4	5.97	0.04	0.18	0.24	0.15	0.17
Male	5.29	0.43	0.42	5.2	0.05	0.17	0.24#	0.17	0.18#
Female	6.42	0.39	0.39	6.5	0.03	0.19	0.24#	0.14	0.17#
LessThan25	5.87	0.39	0.4	5.66	0.05	0.17	0.26	0.10#	0.17
GreaterThan35	8.36	0.32	0.33	10.59	0.02	0.22	0.22	0.09#	0.24

Table 2: Results across different genders and ages in WhatsApp Dataset

Model Name	Size	Baseline	Accuracy	Recall Larger	Recall Smaller	AUC	Tree Attributes
User-Ages Women (D.T.)	38	80.85	80.85	1	0	0.43	Not Applicable
User-Ages Women (Bayes)	38	80.85	80.85	0.81	0.78	0.86	Not Applicable
User-Ages (D.T.)	68	85.29	82.35	0.86 (Younger)	0.6 (Older)	0.69	AvgMsgSent, MsgLen
User-Ages (D.T.) w/o Questions	68	85.29	82.35	0.86 (Younger)	0.6 (Older)	0.69	AvgMsgSent, MsgLen
User-Ages (Bayes)	68	85.29	85.29	0.85 (Younger)	0.9 (Older)	0.85	Not Applicable
Group-Ages (D.T.)	3090	82.25	89.44	0.95 (Younger)	0.63 (Older)	0.86	UsageLevel, UsageFamily, TextLength
Group-Ages (D.T.) w/o Questions	3090	82.25	82.74	0.92 (Younger)	0.37 (Older)	0.75	FileUsage, EarlyMorningUsage, ResponseTime
Group-Ages (Bayes)	3090	82.25	90.22	0.94 (Younger)	0.73 (Older)	0.87	Not Applicable
User-Gender (D.T.)	71	59.15	54.92	0.38 (Female)	0.68 (Male)	0.48	UsageFamily, AverageSent
User-Gender (D.T.) w/o Questions	71	59.15	52.11	0.6 (Female)	0.41 (Male)	0.48	GroupSize, TextSize, TimeofDay
User-Gender (Bayes)	71	59.15	60.56	0.6 (Female)	0.62 (Male)	0.5	Not Applicable
Group-Gender (D.T.)	6185	56.1	70.73	0.72 (Female)	0.69 (Male)	0.75	UsageLevel, UsageFamily, TextLength
Group-Gender (D.T.) w/o Questions	6185	56.1	62.16	0.82 (Female)	0.36 (Male)	0.62	TextLength, ResponseTime, ActivityLevel
Group-Gender (Bayes)	6185	56.29	69.01	0.71 (Female)	0.65 (Male)	0.74	Not Applicable

Table 3: Authorship identification prediction of Gender and Age based on average WhatsApp user data

Model Name	Size	Baseline	Accuracy	Recall Larger	Recall Smaller	AUC	Tree Attributes
10% Files (Bayes)	6185	87.37	87.84	0.93 (No File)	0.55 (Older)	0.9	Not Applicable
10% Files (D.T.)	6185	87.37	88.89	0.95 (No File)	0.53 (Older)	0.74	SocialLevel, Education Level, Age, MsgLength
Time 8-12 (Bayes)	6185	81.27	73.04	0.83 (Not 8-12)	0.3 (8-12)	0.62	Not Applicable
Time 8-12 (D.T.)	6185	81.27	81.27	1.0 (Not 8-12)	0 (8-12)	0.5	None (Majority Class)
Group Size 5+ (Bayes)	6185	83.04	87.51	0.9 (Not 5+)	0.76 (Yes 5+)	0.91	Not Applicable
Group Size 5+ (D.T.)	6185	83.04	88.83	0.94 (Not 5+)	0.62 (Yes 5+)	0.77	Message Frequency, MsgLength
5+ Msg/day in Group (Bayes)	6185	76.32	78.36	0.84 (Not 5+)	0.6 (Yes 5+)	0.79	Not Applicable
5+ Msg/day in Group (D.T.)	6185	76.32	79.53	0.9 (Not 5+)	0.48 (Yes 5+)	0.66	GroupSize, MsgLength
Short Messages (Bayes)	6185	58.41	71.75	0.74 (Not Short)	0.68 (Short)	0.8	Not Applicable
Short Messages (D.T.)	6185	58.41	73.62	0.8 (Not Short)	0.65 (Short)	0.71	Age, Message Frequency,
Quick Responses 0.25 (Bayes)	6185	69.16	74.14	0.86 (Not Quick)	0.48 (Quick)	0.75	Not Applicable
Quick Responses (0.25) (D.T.)	6185	69.16	74.03	0.9 (Not Quick)	0.38 (Quick)	0.68	MsgLength

Table 4: Predicting group activity in WhatsApp Dataset

The advantage to using data mining algorithms to test these hypotheses is the objectivity of the outputted results. On a technical level, we built models from decision trees, as implemented in the accepted C4.5 algorithm (Quinlan 1996) to create classifiers between two choices (Boolean). The C4.5 algorithm was chosen because of two main advantages. First, C4.5 identifies which attributes are most important for accurate prediction by using the InfoGain measure to rank the predictive ability of all attributes. This allows us to objectively identify which factors are most important for accurate prediction. Second, the if-then rules outputted by these algorithms allow us to observe and analyze the exact range of values within the selected attributes that form the prediction model. Furthermore, we consider many tasks, such as if a user is male / female or above / below a certain age, which are inherently Boolean decisions and are thus well suited for C4.5. In theory, other models, such as regression analysis, are better suited for numeric prediction, such as predicting the time that will elapse before a message is answered, but lack the attribute analysis provided by C4.5. To overcome this limitation for these types of prediction tasks, we transformed continuous target variables into two categories through binning according to preset cutoff thresholds. For example, in creating the quick response time model, we chose a response threshold of 1 minute. We then created a Boolean classifier and assumed anyone who answered within 1 minute answered quickly and those who answered after 1 minute, even if they answered only seconds after 1 minute, did not. More specifics of the models and their findings are in the next section.

Data Analytic Results

In general, we built two types of models using the popular open source Weka data mining package (Witten and Frank 2005)– C4.5 decision trees and probabilistic models based on Bayesian networks. We did consider other models, and as we found that probabilistic Bayesian network models were at times more accurate than the C4.5 algorithm, we present results from this algorithm for comparison. All models were constructed using the accepted method of 10-fold cross validation to ensure that the models were valid and created and tested from two different sets of data.

Overall, we were successful in predicting an author's gender and approximate age based on users' general data as can be seen in Table 3. We present three types of models: a decision tree model built from all of collected information including the average usage statistics and the answers from the 6 questions in the questionnaire, a decision tree model that is exclusively based on average usage statistics and a Bayesian network model with all of the information. We also considered two training and testing datasets: the user dataset and the group dataset. The first column in the table presents how many records were in each dataset, the next two columns present the accuracy of the baseline model and the created one, and the next four columns present the

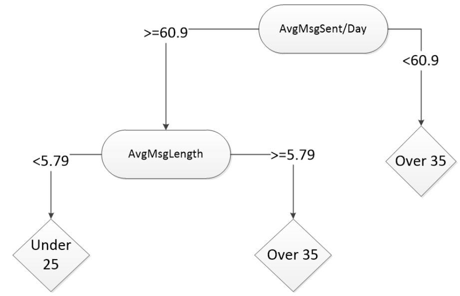

Figure 7: Predicting author age from user data

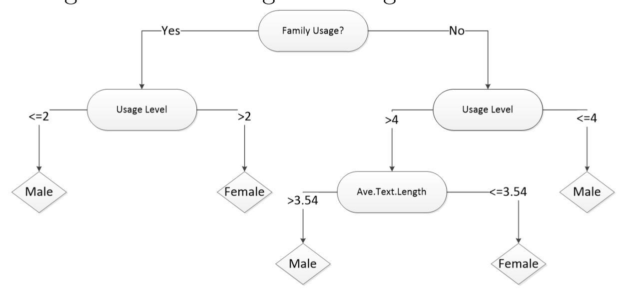

Figure 8: Predicting Male / Female from all collected data

recall of both the majority and minority classes, the Area under the ROC Curve (AUC) of the model and the attributes found in the highest levels of the tree for these models. We define the baseline as the accuracy of the model that predicted all instances which were in the larger category. Minimally, a successful model should at least be more accurate than this value while still achieving relatively high recall for both categories.

We first created an author prediction model for age to differentiate between authors under 25 and over 35. Please note that this dataset size is only 68 of the 92 users as the other users were not in this age range. Also note that the vast majority of people were below 25 (85.29%), making this task even more challenging. In the first row, we present the results from the decision tree model (C4.5) built with all available data. While for most other learning tasks we found that the C4.5 model built with all of the data outperforms the model without the questionnaire data, here this is not the case. The reason for this is that the Infogain metric found that the most significant attributes for prediction were the average number of sent messages (AvgMsgSent) and Message Length (MsgLen), while the information added from the questionnaire was not as important. Thus, the algorithm performed equally well without having the added data from the questionnaires. Also, please note that the Bayesian model performs as well or better than the C4.5 models across all performance metrics. However, it does not provide the same insight as to which attributes were most helpful. For example, observe the resultant decision tree for predicting age in Figure 7. We note that the model identifies older authors as those who sent less than an average of 60.9 messages a day or

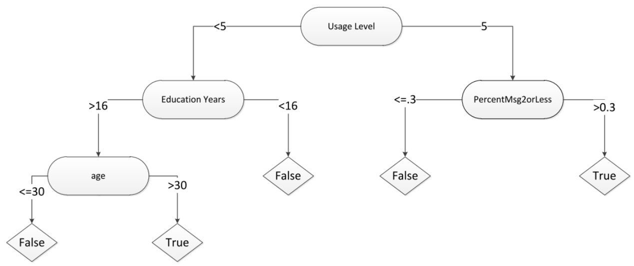

Figure 9: Predicting file usage within groups

have messages with at least an average of 5.79 words. This model is consistent with the findings from Table 2 that older people write less frequently but longer messages. But the decision tree also provides insight as to the exact cutoff for making these prediction models.

We also considered an age prediction task with the group dataset. As we overall have much more group data (6,185 records) than average user data (92), it is not surprising that these models performed better relative to the baseline. We again eliminated all groups where the known author was not in the age range for the two groups we are classifying, leaving only 3,090 of the 6,185 groups. While we did not have any information about other participants, it seems that the added value of the group information was still helpful. Also, the questionnaire data was helpful here, with the results in line 4 outperforming those in line 5. Again, the decision trees provide insight as to which attributes are most helpful: here in the first decision tree we found that older people had higher values for the overall usage and were more likely to use WhatsApp for family communication as noted in the questionnaire, and in the second model older people were found to have longer messages, send and receive more messages overall and in the mornings, and had slower response times.

We also created a prediction model for gender using both user and group data. As we now identified that the data is influenced by age, we wished to eliminate this bias in this model. As such, we intentionally used data from participants close in age – between the ages of 20–29. Please note from lines 6–9 of Table 3 that we used only 71 participants of the total group of 92 as a result. We again found that the models built with the decision trees did not perform better than the baseline model, while the Bayesian network model did. Again, this is likely due to the relatively small size of the dataset, as models built with group data did in fact perform better. We present the decision tree built with all group data (line 10 of Table 3) in Figure 8. Here we found that users who self-rated themselves as using WhatsApp for family communication and rated their usage level above the second level in the Likert scale were women. The model also predicted that those who did not use WhatsApp for family communication, but still rated themselves at the 5th level (above 4) on the Likert scale for usage and wrote messages with an average of less than 3.54 words were also women. Men were predicted for the other conditions.

We also built models that predicted group usage characteristics, the results of which are found in Table 4. Specifically, we built models to predict which groups will contain files attachments in at least 10% of all messages, will have more than 25 percent of their messages sent between 8:00 and 12:00, (rows 3 and 4), which messages are characteristic of groups with 5 or more users (rows 5 and 6), will average at least 5 messages per day (rows 7 and 8), will on average contain short texts with five words or less (rows 9 and 10), and will have at least one quarter of all messages responded to within one minute (rows 11 and 12). Please note that Baseline and Recall Larger columns help give insight into each of these tasks. For example, for the 10% file task, note that the baseline is 87.37% meaning that 87.37% of the groups are in the larger category. Also note that the larger category is the No File one as per the Recall Larger column. Thus, one may deduce that under 13% of all groups have file attachments in at least 10% of the messages.

For all models we present both the results from the models built through Bayesian networks and decision trees. Once again, the decision trees perform slightly worse overall, but facilitated the ability to understand which attributes were most influential in predicting the group's behavior. The decision tree from the file model, found in Figure 9, indicates that the attributes of how people rated their social level, their educational level, age and average message length were most important. Specifically, groups with more files will be those that have a participant we identified as having a high usage level (5) and had participants who sent over 30% of their messages with only two words or less. Additionally, groups with users who didn't rate themselves with high usage levels, but had high educational levels (above 16 years) and were above 30 still typically sent more file attachments. Observe that the time prediction task was not successful, with the decision tree finding no important attributes, and simply adopting the majority case that messages were not sent in this timeframe. As one might intuitively believe, and as Figure 3 demonstrates, larger groups have more messages and thus typically have messages sent with a higher frequency, but shorter messages are also indicators that the group is large. Similarly, groups with a higher frequency of messages are typically larger, but also have typically smaller messages (rows 7–10). Younger people typically send shorter messages, while older people typically send longer messages less frequently (row 10). Similarly, we found that shorter messages also indicated that people responded quickly (row 12).

Conclusions and Future Work

This work represents the first exhaustive analysis of WhatsApp messages. We collected over 4 million messages from nearly 100 participants, and differentiated between different types of *user* and *group* usage of the network. Through performing extensive statistical analysis, we found that many message and group characteristics significantly differed across users of different demographics such as gender and age. Additionally, we believe that one key novelty of this work is that we use data analytics to predict users' gender, age and group activity. As we base our findings exclusively on the algorithms' output, there is no possibility that an author bias exists in our results. This is one key advantage to using data analytics, and this difference is especially clear from the decision tree results presented in this paper.

Overall, our results provide several new insights into WhatsApp usage. We find that older people typically use this network less frequently, but when they do, write longer messages. We also find that more education and age are positive factors in predicting how frequently people will send file attachments. Overall, women use this network more often than men and they reported that they use it more often to both generally communicate and communicate with family. Men on the other hand are members overall of larger communication groups and send shorter messages. Additionally, larger groups are not only defined by their large number of users, or even the large numbers of messages that are frequently sent, but are also typically defined as having shorter messages than those in private one-to-one communications. Decision tree models were not only helpful in identifying these attributes, but were useful in providing the thresholds within the if-then rules for the models that predicted these results.

In building upon this work, we believe that two types of studies will likely lead to fruitful results. First, we believe that additional studies should be undertaken to improve upon and extend the study we present. While this study analyzed over 4 million messages, it is still somewhat limited in containing only 92 users. We believe that even more accurate models can be built through studying data from more users. Similarly, we did not study all group tasks and other tasks such as which messages will be forwarded remain unexplored. In a related matter, while we intentionally built models without analyzing user content in order to safeguard privacy, even more accurate models might be built in the future if user consent could be obtained for this information.

We believe a second type of direction should focus on applying the lessons learned from this paper's models. It may be wise to customize user interfaces for certain types of users and tasks based on the attributes found to be important in this paper. For example, users who are more educated or older might prefer a different WhatsApp interface from less educated or younger users as their usage patterns differ significantly. Similarly, as larger groups are characterized by shorter messages, it may be that the interface for these types of interactions should be customized with this information in mind as well. We hope that these and other issues will be explored in greater detail in future work.

References

[Argamon et al. 2009] Argamon, S.; Koppel, M.; Pennebaker, J. W.; and Schler, J. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52(2).

[Bakshy et al. 2012] Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, 519–528.

[Bouhnik and Deshen 2014] Bouhnik, D., and Deshen, M. 2014. Whatsapp goes to school: Mobile instant messaging between teachers and students. *Journal of Information Technology Education: Research* 13:217–231.

[Church and de Oliveira 2013] Church, K., and de Oliveira, R. 2013. 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*, MobileHCI '13, 352–361.

[Fiadino, Schiavone, and Casas 2014] Fiadino, P.; Schiavone, M.; and Casas, P. 2014. Vivisecting whatsapp through large-scale measurements in mobile networks. In *SIGCOMM 2014*.

[Johnston et al. 2015] Johnston, M.; King, D.; Arora, S.; Behar, N.; Athanasiou, T.; Sevdalis, N.; and Darzi, A. 2015. Smartphones let surgeons know whatsapp: an analysis of communication in emergency surgical teams. *American Journal of Surgery* 209:45–51.

[Kring and Gordon 1998] Kring, A. M., and Gordon, A. H. 1998. Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology* 686–703.

[Lin et al. 2015] Lin, Y.-H.; Lin, Y.-C.; Lee, Y.-H.; Lin, P.-H.; Lin, S.-H.; Chang, L.-R.; Tseng, H.-W.; Yen, L.-Y.; Yang, C. C. H.; and Kuo, T. B. J. 2015. Time distortion associated with smartphone addiction: Identifying smartphone addiction via a mobile application (app). *Journal of psychiatric research* 65:139–145.

[Montag et al. 2015] Montag, C.; Błaszkiewicz, K.; Sariyska, R.; Lachmann, B.; Andone, I.; Trendafilov, B.; Eibes, M.; and Markowetz, A. 2015. Smartphone usage in the 21st century: who is active on whatsapp? *BMC Research Notes* 8(1):1–6.

[Mudliar and Rangaswamy 2015] Mudliar, P., and Rangaswamy, N. 2015. Offline strangers, online friends: Bridging classroom gender segregation with whatsapp. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 3799–3808.

[Naveed et al. 2011] Naveed, N.; Gottron, T.; Kunegis, J.; and Alhadi, A. C. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11: Proceedings of the 3rd International Conference on Web Science*.

[O'Hara et al. 2014] O'Hara, K. P.; Massimi, M.; Harper, R.; Rubens, S.; and Morris, J. 2014. Everyday dwelling with whatsapp. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, 1131–1143.

[Pielot et al. 2014] Pielot, M.; de Oliveira, R.; Kwak, H.; and Oliver, N. 2014. Didn't you see my message?: Predicting attentiveness to mobile instant messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 3319–3328.

[Quinlan 1996] Quinlan, J. R. 1996. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research* 4:77–90.

[Thelwall, Wilkinson, and Uppal 2010] Thelwall, M.; Wilkinson, D.; and Uppal, S. 2010. Data mining emotion in social network communication: Gender differences in myspace. *J. Am. Soc. Inf. Sci. Technol.* 61(1):190–199.

[Wagner et al. 2015] Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *ICWSM2015*.

[Wang, Burke, and Kraut 2013] Wang, Y.-C.; Burke, M.; and Kraut, R. E. 2013. Gender, topic, and audience response: An analysis of user-generated content on facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 31–34.

[Witten and Frank 2005] Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann Series in Data Management Systems*.

[Xiang, Neville, and Rogati 2010] Xiang, R.; Neville, J.; and Rogati, M. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 981–990.