Potential Traffic Savings by Leveraging Proximity of Communication Groups in Mobile Messaging

Michael Seufert*, Anika Schwind, Marco Waigand, Tobias Hoßfeld
Insitute of Computer Science, University of Würzburg, Würzburg, Germany
michael.seufert.fl@ait.ac.at, {anika.schwind|marco.waigand|tobias.hossfeld}@informatik.uni-wuerzburg.de

***Abstract*—Communication groups in mobile messaging applications (MMAs) multiply the data transmissions, because every message has to be delivered to all members of the communication group. Thereby, they put a high load on mobile networks. As the number of recipients is still comparably small, the data-intensive user-generated content cannot be handled efficiently in large content delivery networks. However, small communication groups, such as groups of friends or teams, might often be in close proximity, which can be leveraged to locally deliver messages by applying edge caching or device-to-device (D2D) communication. In this work, a simulation study is conducted to investigate these potential traffic savings in the mobile network. It is based on a realistic communication model of the MMA WhatsApp and utilizes different models for human mobility. The user mobility and MMA communication are simulated for a single day in a small city to obtain the ratio of messages, which could be potentially transmitted locally when utilizing edge caching and D2D communication.**

***Index Terms*—Traffic management; WhatsApp; Mobile messaging application; Mobile instant messaging; Mobile networks.**

I. Introduction

More and more Internet applications provide many possibilities to their users to communicate with other users, and not only exchange text and voice, but can also send data-intensive media like images and videos. Thereby, the messages do not necessarily have only a single receiver, but often they are sent to many users in a communication group. In this work, the focus is on explicit groups in a mobile messaging application (MMA), such as WhatsApp. The ubiquitous and fast-paced communication through MMAs and the multiplication of sent messages, which have to be transmitted to many receivers in a communication group, generate huge amounts of data and pose a high load on constrained mobile networks. Network operators face the challenge to efficiently transmit the data, while reaching a high user satisfaction at the same time. Therefore, they utilize different mechanisms to optimize the network traffic, e.g., caching, which speeds up the data delivery and reduces the data volume in the network behind the cache.

However, users often transmit high volume media data, which was generated by themselves, i.e., user-generated content (UGC). UGC cannot be effectively cached in large content delivery networks, because the content is not globally popular, i.e., only few friends of the content creator are interested in a particular content item. Nevertheless, communication groups have interesting properties, which show a potential for traffic savings in the mobile network. For example, groups of friends or teams might often meet at the same time at the same place for certain social events. Thus, it could be feasible to apply edge caching [1], [2], i.e., to cache the content for a short time at the local base station. Another approach to reduce the load on mobile networks could be to share the content locally among users without using the mobile network infrastructure for data transmissions. Instead, devices could establish a direct device-to-device (D2D) connection to exchange data [3].

In this paper, the research question is answered if edge caching and D2D communication are effective for group-based communication in MMAs. A simulation study was conducted based on real communication data from WhatsApp and different models for human mobility. The traffic savings are computed in terms of the ratio of locally delivered messages, and the effectiveness of the different traffic management mechanisms in different mobility scenarios are compared.

Therefore, this paper is structured as follows. Section II outlines related works on social and technical aspects of MMAs, as well as on the investigated traffic management approaches, i.e., edge caching and D2D communication. Section III presents the utilized models and simulation procedures to obtain the performance evaluation results, which are shown and discussed in Section IV. Finally, Section V concludes.

II. Related Work

A. Social and Technical Aspects regarding MMAs

The WhatsApp behavior of more than 2 400 users in Germany was recorded by [4] over four weeks. WhatsApp was used for more than 30 min per day (around 20% of smartphone usage time). Females used it for significantly longer periods than males, and younger people longer than older people.

An online study in [5] showed, among other things, that user satisfaction with MMAs is depending on user experience with the technology, richness of the technology, and the influence of their immediate friends. The usage behavior of the MMA Snapchat was evaluated in an online survey with 209 participants in England [6]. It was reported that the snaps of each participant were sent to a single person in 72%, and to a group in 27%. A study about KakaoTalk collected data from more than 350 users using a questionnaire [7]. They found that perceived service quality and usability significantly affect user satisfaction and the intention to continue using MMAs.

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

978-3-903176-14-0 © 2018 IFIP

The temporal aspects and energy consumption of WhatsApp was investigated by [8] considering the 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 message 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 for latency. The traffic behavior of WhatsApp was analyzed by [9] 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. [10] investigated user behavior patterns and traffic characteristics of 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 MMAs. The resulting model was used to evaluate the impact of MMAs on cellular network performance.

B. Edge Caching and Device-to-Device Communication

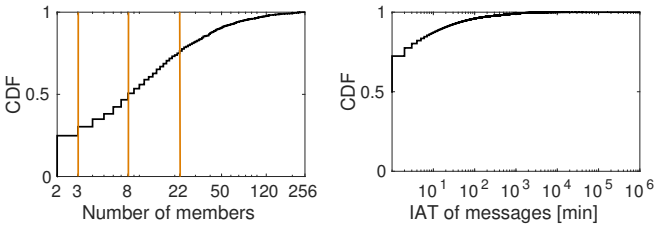
Edge caching refers to caching data at resources of network operators close to the end user, such as mobile base stations. Many works have considered algorithms for caching, including also edge caching [11]–[13]. D2D communication allows devices to transmit data directly, i.e., without transmitting data through the base station of a mobile cell [3], [14], [15]. Thereby, some approaches operate in the licensed spectrum, e.g., LTE Direct, and might even rely on using the infrastructure, e.g., for link discovery or user authentication [16]. Moreover, D2D communication can be implemented without any infrastructure in the unlicensed spectrum, e.g., Bluetooth or Wi-Fi Direct. Different D2D transmission techniques are compared in [17]. For offloading traffic from mobile base stations, the utilized techniques should show a high maximum transmission range, which can range up to $100\,\mathrm{m}$ for Bluetooth, $200\,\mathrm{m}$ for Wi-Fi Direct, and $1000\,\mathrm{m}$ for LTE Direct.

Several related works were focused on traffic management studies. The work in [18] investigated proactive caching at base stations and on mobile devices. The caching was supported by D2D connections between mobile devices. An architecture for caching popular video content to enable D2D communication was presented in [19]. In [20], random encounters of users are investigated and a framework for a distributed cache management with D2D communication was presented.

III. Simulation Concept

A. Models of WhatsApp Communication

In order to use realistic communication models for the simulation study, WhatsApp chat histories were collected and analyzed using the web-based service WhatsAnalyzer [21]. It can receive exported chat protocols from WhatsApp by email, anonymizes all data, and extracts communication data, which can be evaluated scietifically [22], [23]. These data include the size of the communication group, as well as for each message the timestamp, an anonymized user ID of the sender, and the type, i.e., text or media message. In this work, a WhatsAnalyzer data set of 2598 anonymized WhatsApp chats with $45\,729$ users was used to obtain communication models.



(a) Number of members per group (b) Inter-arrival time of messages

Fig. 1: Models for group-based communication

Figure 1a shows the empirical CDF of the number of members of all WhatsAnalyzer chats on a logarithmic x-axis. The orange lines indicate the categorization of group sizes from left to right: dyadic (two group members), small (three to seven group members), medium (eight to 21 group members), and large (22 or more group members). This categorization was chosen arbitrarily to almost equally divide the set of chats by chat size into four parts. The CDF shows that WhatsAnalyzer is mainly used to analyze groups with more than two people. The curve flattens with an increasing number of members, which means that communication in smaller groups is more prevalent compared to the maximum size of 256 members.

Next, the inter-arrival time (IAT) of messages, i.e., the time between two consecutive messages, was analyzed in Figure 1b. Although the x-axis represents the IAT in minutes on a logarithmic scale, the CDF shows a very steep increase, which means that IATs are generally short. In fact, 72.41% of all messages are answered within $1\,\mathrm{min}$, and 90.53% within $20\,\mathrm{min}$. Although only 3.61% of all messages are answered after more than $120\,\mathrm{min}$ ($2\,\mathrm{h}$), the mean IAT is $51.26\,\mathrm{min}$, which indicates the long tail characteristic of this distribution. This model shows that WhatsApp communication is mostly very fast, but also very long communication pauses can occur.

The data set also reveals that conversation in WhatsApp is mostly text-based. About three quarters of the collected chats have a media ratio of 10% or less. In about 10% of the communication groups no media has been sent at all.Only about 1% of all chats have a ratio above 50%, i.e., more media posts than text posts. Although the ratio of sent media messages in all sent messages is relatively low, the transmissions of compressed media files are responsible for the most network traffic generated by WhatsApp [9].

Consequently, in order to show the general potential for traffic savings, the traffic savings will not be expressed in number of bytes, which might soon become obsolete when file sizes of transmitted media change, e.g., when considering other MMAs, different compression algorithms, or source

media files with higher quality. Instead, the traffic savings will be shown only in terms of the share of messages, which can potentially be transmitted locally. This means, these messages do not have to be transmitted through the mobile network either because they are already cached at the mobile base station or because they can be directly transmitted to the recipient via D2D communication.

B. Mobility Models

The human mobility is the key influence factor for the effectiveness of edge caching and D2D communication. It is affected by many physical and social effects, e.g., age, fitness, social situation, profession, interests, relationships, or friendships. Currently, no models exist, which take all these effects into account and reach a high accuracy, but most models only focus on some aspects, e.g., [24]. To exclude the bias of a specific model on the results, this work attempts a parameter study of mobility based on the randomness included in the mobility. This means that four different mobility models for pedestrians are considered, which start from highly random movement, and decrease the underlying randomness of human mobility towards a very regulated scenario. Thereby, it is assumed that the real randomness of human mobility lies somewhere in between these models.

First, a random waypoint (RW) model is considered. In this simple model, pedestrians start at a random point and draw a random destination, which is uniformly distributed over the whole simulation area. Additionally, a random walking speed is drawn uniformly between $1\,\text{and}\ 1.4\,\mathrm{m/s}$ $(3.6-5\,\mathrm{km/h})$. The maximum distance of a trip is limited to $2\,\mathrm{km}$, i.e., a new destination is drawn until the walking distance is below this threshold, and the pedestrians follow a direct path (straight line) towards the destination. When the destination is reached, immediately a new destination and a new walking speed are drawn, and the walk is continued. Note that the mobility described by this model shows a very high randomness. This very high randomness is expected to make it a kind of worst case scenario in terms of potential traffic savings.

The next model resembles the RW model, but additionally introduces pauses when a pedestrian reaches a destination before starting a new walk. Therefore, this model will be referred to as random waypoint with pauses (RWP). The duration of the pauses are uniformly random between $0\,\text{and}\ 43\,200\,\mathrm{s}$ $(12\,\mathrm{h})$. As pedestrians are not moving during the pauses, the temporal randomness of the mobility is significantly reduced.

Another extension of the RW and RWP model is the random waypoint model with pauses on streets (RWPS). It additionally decreases the spatial randomness of the mobility by allowing the pedestrians to only walk on streets. Thus, pedestrians cannot follow the direct path (straight line) from source to destination anymore, which anyway will mostly be blocked, especially in urban environments. To make pedestrians walk on streets, mobility traces were generated with the Simulator of Urban MObility (SUMO) [25], [26]. SUMO allows to import OpenStreetMap data to produce a road network and can simulate the mobility of vehicles or pedestrians on these streets. After drawing a random waypoint, the closest street of the SUMO road network is selected as the start or end of the trip. Thereby, the street must be within $30\,\mathrm{m}$ of the random waypoint, otherwise, a new random waypoint is chosen. To compute the walking path, SUMO performs a fastest-path routing to determine the intermediate edges. Note that this model will lead to higher concentration of pedestrians in areas with many streets, such as city centers. The walking speed and pauses are implemented in the same way as for RWP.

Finally, the social randomness is decreased by allowing only meaningful walks. In this model, a pedestrian will only walk from home to work and back, and might then join a leisure activity with friends. Thus, this model is called home-/work-/leisure-mobility on streets (HWLS). First, user locations are randomly selected according to a simple model [27], which was designed to resemble the distribution of the locations of public Wi-Fi routers in cities. As this distribution should be correlated with population density, the model will be utilized to compute home, work, and leisure locations. The model is adjusted such that 95% of the home locations (work/leisure locations) are within half (quarter) of the diameter of the simulation area. This reflects the fact that work and leisure locations are often closer to the city center than home locations. Note that a new location has to be computed if any location is outside of the simulation area or if the maximum walking threshold of $2\,\mathrm{km}$ is violated. Now, a highly regulated schedule is implemented by HWLS to further decrease the randomness of the mobility. In the morning, the pedestrians are at home and start walking to work between $6\,\mathrm{am}$ and $12\,\mathrm{am}$ (uniformly random). They stay at their work location for a uniformly random duration between $4\,\text{and}\ 10\,\mathrm{h}$ before returning home. 10% of all pedestrians can now initiate a leisure activity. In case an initiator is at home at $6\,\mathrm{pm}$, a leisure location is randomly selected just like the work location, a starting time is drawn uniformly between $6\,\mathrm{pm}$ and $10\,\mathrm{pm}$, and an activity duration is selected randomly between $1\,\text{and}\ 4\,\mathrm{h}$. A communication group of the pedestrian is randomly selected and it is checked if the other members of the group can participate, i.e., if they are at home, their walking distance to the leisure location is below $2\,\mathrm{km}$, and they can reach the meeting on time. All participants then walk to the leisure location to meet at the starting time, stay there for the activity duration, and afterwards return home. Note that the meeting for the leisure activity reflects the proximity of groups of friends or teams, which can especially be exploited for transmitting messages locally. Together with the significantly reduced randomness, the HWLS model will describe a kind of best case in terms of potential traffic savings.

C. Traffic Management Simulation

The simulation area is based on the city of Würzburg, Germany, with a population of around $125\,000$ inhabitants. A map of Würzburg of size $10.349\,\mathrm{km}$ x $7.324\,\mathrm{km}$ was obtained from OpenStreetMap and imported into SUMO as a road network. The center of the map is located at $49.790\,615°\mathrm{N}$, $9.944\,470°\mathrm{E}$. In this area, the mobility of $70\,000$ pedestrians is

simulated with Java (in case of RW and RWP) or SUMO (in case of RWPS and HWLS), which roughly approximates the share of 55% of Germans reported to use WhatsApp on a daily basis in 2017 [28]. The simulation duration is a single day with a time granularity of 1 s, and ten different mobility traces were generated per mobility model. The abstract traffic management simulation is implemented in Java and uses the output mobility traces as well as the presented WhatsApp communication models as follows. Thereby, one mobility trace corresponds to one simulation run. In each run, 100 000 WhatsApp groups of different sizes are created according to the distribution presented in Figure 1a. The members of the groups are selected uniformly random from the set of pedestrians. In every group, the inter-arrival time of the next message is drawn randomly according to the distribution shown in Figure 1b. For each message, one of the group members is randomly selected as the sender of the message, the other members are the recipients. This process is repeated until the simulation time is exceeded. The granularity of the communication is also 1 s.

Three approaches are considered for saving traffic in the mobile network by locally transmitting messages. The first approach is edge caching, i.e., the caching of messages directly at the base station. Therefore, the simulation area is divided into a grid of square cells. The length of each square cell is a parameter and varies from 100 m to 500 m. Every time a message passes through a mobile base station (i.e., when the sender uploads the message to the MMA server, or when a recipient downloads the message from the MMA server), the message is cached at the base station. If other recipients are in the same cell, the message can be locally delivered to them from the edge cache. This implies that every message is transmitted through the mobile network to any cell at most once. Second, D2D communication allows close devices to transmit data directly, i.e., without transmitting data through the base station of a mobile cell. The local transmission area of a device is implemented as a circle with a parametrized radius ranging from 25 m to 100 m. To effectively utilize D2D communication, a message delay had to be introduced in order to allow users to move into the transmission circles of other users and relay the message. Therefore, after the sender uploaded the message to the MMA server, a local relay threshold of 5 s is introduced for the sender to locally transmit the message via D2D communication. If the message cannot be transmitted to any user within 5 s, the message is sent from the MMA server to a single, random recipient via the mobile network. Then, the process repeats, i.e., every recipient of the message has again 5 s to locally transmit the message via D2D communication to another user before the message is sent from the MMA server to another, random recipient. After a global relay threshold of 30 s is reached, the message is transmitted to all remaining recipients via the mobile network in order to limit the message delivery delay. Note that only members of the group can receive and relay the messages. Finally, a combination of edge caching and D2D communication is investigated. It resembles the D2D communication approach, but whenever a message is up- or downloaded through the mobile network, the message is also cached at the base station, and can be delivered locally to all recipients in that cell.

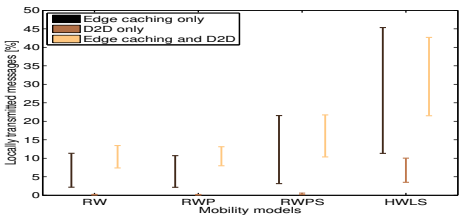


Fig. 2: Traffic saving potential of edge caching and D2D communication

Note that this simulation is abstract of any actual network technologies and does not consider degradations on the wireless transmission path, such as shielding by obstacles or interference by other devices, as well as constraints in the mobile network in terms of capacity or cache size at the mobile base station. This is why the results, which are presented in the following section, could be seen as an optimistic estimation. However, considering the maximum transmission ranges of widely used transmission techniques (cf. Section II), the chosen cell sizes and transmission ranges were sufficiently reduced. In the considered ranges, the impact of degradations on the wireless transmission path should not be very severe, such that the presented results still exhibit practical relevance.

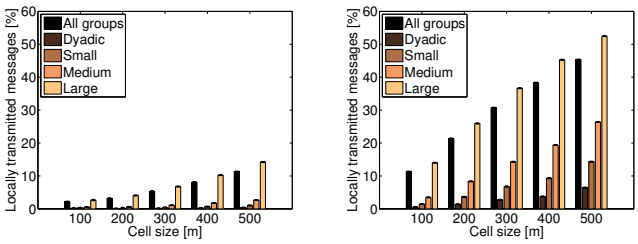
IV. Evaluation

In this section, the simulative performance evaluation results are presented. Figure 2 summarizes the simulation results for the four mobility models. It can be seen that the underlying randomness of the mobility models significantly affect the traffic saving potential. The models are depicted on the x-axis, and the y-axis shows the ranges of the mean ratio of locally transmitted messages over all groups for edge caching only (black), D2D communication only (brown), and the combination of edge caching and D2D communication (yellow). The results show that depending on the mobility model, traffic savings up to 13.43% (RW), 13.13% (RWP), 21.72% (RWPS), and 45.34% (HWLS) are possible when the largest cell size of 500 m and/or D2D transmission range of 100 m are considered.

In this work, it is assumed that the real randomness of human mobility lies somewhere in between the selected models, which are bounded by the “worst case” RW and the “best case” HWLS model. Thus, the worst case (RW) and best case (HWLS) situation will be further studied in detail. First, the scenario, in which only edge caching is implemented. This corresponds to the black whiskers in Figure 2. Then, the situation with only D2D communication is investigated (brown whiskers). The yellow whiskers describe scenarios with both edge caching and D2D, which will be explored last.

A. Edge Caching

In this section, the impact of pure edge caching on locally transmitted messages is analyzed. Figure 3 shows two bar plots



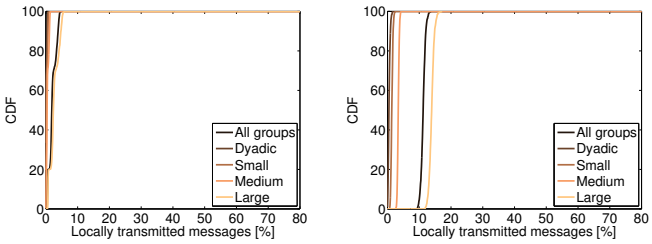
(a) RW (b) HWLS

Fig. 3: Impact of edge caching on locally transmitted messages

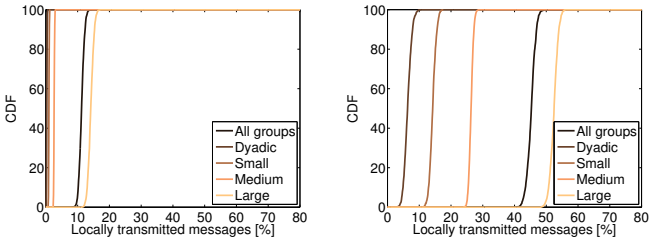
for RW (worst case) and HWLS (best case), which indicate the ratio of locally transmitted messages (y-axes) for different cell sizes (x-axes) ranging from 100 m to 500 m. Not only the overall savings for all communication groups are shown in the black bars, but also the savings for different group sizes. Thereby, the dark brown bars represent dyadic groups with only two members, light brown bars represent small groups with three to seven members, orange bars represent medium-sized groups with eight to 21 members, and yellow bars represent large groups with 22 or more members. All bars indicate the mean ratio of locally transmitted messages in the respective groups over all simulation runs, and the 95% confidence intervals are plotted on top. The confidence intervals are very small, which is due to the high number of simulated communication groups.

It can be seen for both the RW model in Figure 3a and the HWLS model in Figure 3b that the traffic savings increase with the cell size. This is an expected result as an increasing cell size leads to a higher probability that members of the same group are in the same cell. Also the ratio of locally transmitted messages increases with the group size, because there is a higher probability that some recipients are in the same cell when there are more recipients. Interestingly, it can be seen that the high traffic saving ratio for large groups (yellow bar) has a big impact on the overall traffic saving ratio (black bar), although only roughly a quarter of all groups are large groups. This is due to the fact that large groups have many recipients, so the communication in these groups accounts for a much higher share of the transmitted messages. The absolute numbers of the mean of locally transmitted messages clearly show the differences between the worst case (RW) and the best case (HWLS) mobility model. For RW, the overall traffic savings for all groups range from 2.17% for a cell size of 100 m to 11.34% for 500 m. For HWLS, the savings range from 11.32% for 100 m to 45.34% for 500 m.

Figure 4 depicts the CDF for the two mobility models and two cell sizes of 100 m and 500 m. The CDFs different group sizes are plotted in the same colors as above. For RW with a cell size of 100 m in Figure 4a, members of the same group are almost never in the same cell, which results in very low traffic savings. For around 20% of the groups, almost no traffic can be saved at all. The mean traffic savings are 0.22% for dyadic, 0.33% for small, 0.58% for medium, and 2.70% for large groups. For the remaining subplots, all CDFs

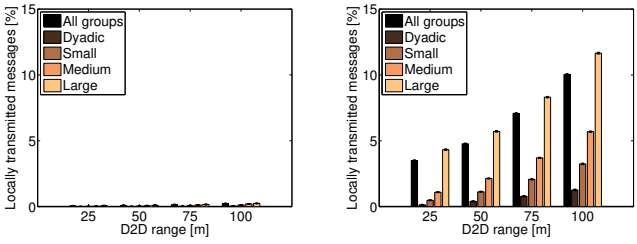


(a) RW, cell size 100 m (b) HWLS, cell size 100 m



(c) RW, cell size 500 m (d) HWLS, cell size 500 m

Fig. 4: Impact of edge caching for different group sizes



(a) RW (b) HWLS

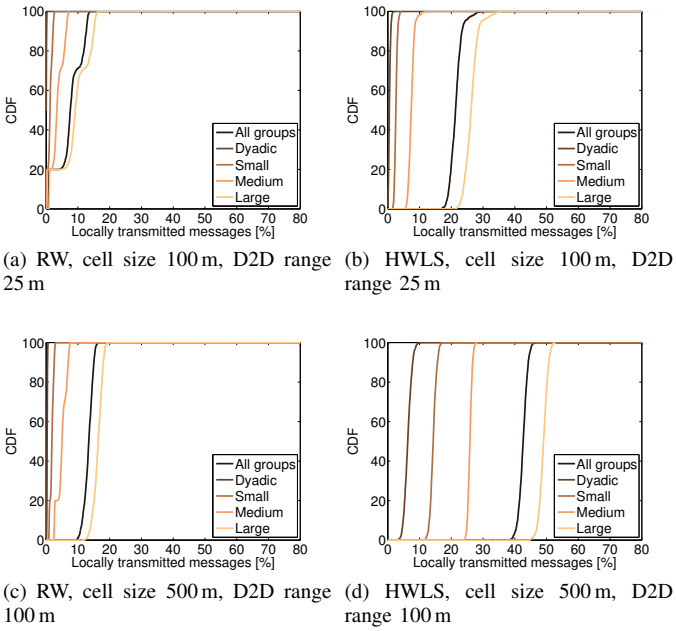
Fig. 5: Impact of D2D communication on locally transmitted messages

are close to vertical, which means that all groups have more or less the same ratio of locally transmitted messages. The plots for HWLS with 100 m (Figure 4b) and for RW with cell size 500 m (Figure 4c) are very similar and show only some considerable savings for large groups, on average 14.25% (RW, 500 m) and 14.01% (HWLS, 100 m), respectively. Dyadic, small, or medium groups stay below 5%. Only in HWLS with cell size 500 m (Figure 4d), caching is more efficient and reaches savings of 6.46%/14.34%/26.42%/52.50% for dyadic/small/medium/large groups.

B. D2D Communication

Now scenarios with only D2D communication are considered. Figure 5 indicates the ratio of locally transmitted messages (y-axes) for different D2D transmission ranges (x-axes) ranging from 25 m to 100 m and the two considered models RW and HWLS. The color coding is the same as above, and the bars again indicate the mean ratio of locally transmitted messages in the respective groups over all simulation runs with the 95% confidence intervals on top.

Although the same trends can be observed as with edge caching, the traffic savings are much smaller, which is due



(a) RW, cell size 100 m, D2D range 25 m

(b) HWLS, cell size 100 m, D2D range 25 m

(c) RW, cell size 500 m, D2D range 100 m

(d) HWLS, cell size 500 m, D2D range 100 m

Fig. 6: Combined impact of edge caching and D2D for different group sizes.

to the much lower D2D transmission range compared to the cell size. For RW in Figure 5a, almost no traffic savings can be seen. The reason is that with such high randomness of the mobility, it is very unlikely that any members of the same communication group come close enough to establish a D2D communication. The mean ratio of locally transmitted messages for all groups range from 0.06% for a D2D range of 25 m to 0.23% for 100 m. The corresponding mean traffic savings per group type are 0.004%/0.01%/0.03%/0.07% for dyadic/small/medium/large groups for a D2D range of 25 m, and 0.05%/0.13%/0.21%/0.24% for 100 m. For HWLS in Figure 5b, the results indicate a higher potential for traffic savings, as it is more likely that people meet in the city center. Here, the overall savings range from 3.50% for 25 m to 10.03% for 100 m. Again, for 25 m, large groups stand out and reach a mean traffic saving of 4.32%. In case of a D2D range of 100 m, a considerable ratio of messages can be transmitted locally, i.e., 1.26%/3.24%/5.69%/11.65% for dyadic/small-/medium/large groups. Similar to edge caching, the simulation results show homogenous savings for all groups.

C. Edge Caching and D2D Communication

Finally, the combination of edge caching and D2D communication is investigated. Figure 6 shows the detailed CDFs for different group sizes for both RW and HWLS in two scenarios. Thereby, Figures 6a and 6b consider a cell size of 100 m and D2D range of 25 m, and Figures 6c and 6d assume a cell size of 500 m and D2D range of 100 m. The CDFs show basically the same trends as discussed above. However, it can be seen that, in the RW (cell size 100 m, D2D range 25 m), the traffic savings of the combined edge caching and D2D communication are higher than the sum of the separate strategies with only edge caching or only D2D communication (cf. Figures 4a, 5a) by around 9%. This is due to an increased traffic saving potential of edge caching. As the proximity of group members in the same cell is very unlikely in this scenario due to small cell sizes and high randomness of mobility, edge caching benefits from the relay thresholds introduced by D2D communication in the combined scenario. This means, cached messages can still be delivered to group members, when they walk into the cell within the relay thresholds, which results in a slightly increased traffic savings of the combined approach. This can also be observed with smaller effect size of around 5% for HWLS (cell size 100 m, D2D range 25 m), and for RW (cell size 500 m, D2D range 100 m) with around 3%.

In contrast, for HWLS (cell size 500 m, D2D range 100 m), the traffic savings of the combined approach are smaller than the traffic savings of edge caching only (cf. Figure 4d) by around 3%. The reason is that the proximity of users is generally higher and the large cells result in a high ratio of locally delivered messages for edge caching only. In the combined scenario, several messages might already be delivered via D2D communication. Less recipients remain and the delay caused by the relay thresholds could negatively influence the proximity, which makes caching less efficient.

V. Conclusion

This work presented a first simulative estimation of the potential traffic savings for group communication in MMAs. The simulation considered edge caching and device-to-device (D2D) communication to locally transmit messages, and thereby reduce the load on the mobile network. The results showed that the ratio of locally transmitted messages depends heavily on the assumed randomness of the human mobility and the resulting proximity of members of a communication group. Nevertheless, a plausible range of potential traffic savings was obtained by focusing on worst case and best case scenarios for mobility, cell sizes, and D2D transmission ranges.

Overall, edge caching was the most promising traffic management approach. It could save between 2.17% and 11.34% of all messages in the mobile network in the worst case and between 11.32% and 45.34% in the best case. Thereby, high savings were possible especially for large communication groups. Additional D2D communication was not effective per se, but introduced message delivery delays, which raised the lower bounds, especially when proximity of group members is less likely. However, it could also slightly reduce the upper bound, and might generally not be desired in fast-paced MMA communication. Although the lower bounds of potential traffic savings look small, it has to be remembered that they are based on very high randomness of mobility and/or very pessimistic assumptions on cell size and D2D transmission range. In reality, the actual traffic savings should be closer to the upper bounds of the presented results. Thus, the results show that local transmissions of group communication messages can significantly reduce the MMA traffic in mobile networks.

REFERENCES

[1] A. Dabirmoghaddam, M. M. Barijough, and J. Garcia-Luna-Aceves, "Understanding optimal caching and opportunistic caching at the edge of information-centric networks," in *Proceedings of the 1st ACM Conference on Information-centric Networking (ICN)*, 2014.

[2] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the Wireless Edge: Design Aspects, Challenges, and Future Directions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 22–28, 2016.

[3] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-device Communication in Cellular Networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[4] C. Montag, K. Błaszkiewicz, R. Sariyska, B. Lachmann, I. Andone, B. Trendafilov, M. Eibes, and A. Markowetz, "Smartphone Usage in the 21st Century: Who is Active on WhatsApp?" *BMC Research Notes*, vol. 8, no. 1, p. 331, 2015.

[5] S. O. Ogara, C. E. Koh, and V. R. Prybutok, "Investigating factors affecting social presence and user satisfaction with mobile instant messaging," *Computers in Human Behavior*, vol. 36, pp. 453–459, 2014.

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

[7] A. P. Oghuma, C. F. Libaque-Saenz, S. F. Wong, and Y. Chang, "An expectation-confirmation model of continuance intention to use mobile instant messaging," *Telematics and Informatics*, vol. 33, no. 1, pp. 34–47, 2016.

[8] E. J. Vergara, S. Andersson, and S. Nadjm-Tehrani, "When Mice Consume like Elephants: Instant Messaging Applications," in *Proceedings of the 5th International Conference on Future Energy Systems (ACM e-Energy)*, Cambridge, UK, 2014.

[9] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting Whatsapp Through Large-scale Measurements in Mobile Networks," in *ACM Conference on SIGCOMM*, Chicago, IL, USA, 2014.

[10] 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 Communications Letters*, vol. 18, no. 3, pp. 389–392, 2014.

[11] C. Wang, Y. He, F. R. Yu, Q. Chen, and L. Tang, "Integration of Networking, Caching and Computing in Wireless Systems: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys & Tutorials*, 2017.

[12] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[13] L. Li, G. Zhao, and R. S. Blum, "A Survey of Caching Techniques in Cellular Networks: Research Issues and Challenges in Content Placement and Delivery Strategies," *IEEE Communications Surveys & Tutorials*, 2018.

[14] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device Communication in LTE-Advanced Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 1923–1940, 2015.

[15] P. Gandotra and R. K. Jha, "Device-to-device Communication in Cellular Networks: A Survey," *Journal of Network and Computer Applications*, vol. 71, pp. 99–117, 2016.

[16] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device Communication as an Underlay to LTE-Advanced Networks," *IEEE Communications Magazine*, vol. 47, no. 12, 2009.

[17] D. Feng, L. Lu, Y. Yuan-Wu, G. Li, S. Li, and G. Feng, "Device-to-device Communications in Cellular Networks," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 49–55, 2014.

[18] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[19] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling Behavior for Device-to-device Communications with Distributed Caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286–4298, 2014.

[20] K. Machado, A. Boukerche, E. Cerqueira, and A. A. Loureiro, "A Socially-Aware In-Network Caching Framework for the Next Generation of Wireless Networks," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 38–43, 2017.

[21] A. Schwind and M. Seufert, "WhatsAnalyzer: a Tool for Collecting and Analyzing WhatsApp Mobile Messaging Communication Data," in *Proceedings of the 30th International Teletraffic Congress (ITC30)*, Vienna, Austria, 2018.

[22] M. Seufert, A. Schwind, T. Hoßfeld, and P. Tran-Gia, "Analysis of Group-based Communication in WhatsApp," in *Proceedings of the 7th EAI International Conference on Mobile Networks and Management (MONAMI)*, Santander, Spain, 2015.

[23] M. Seufert, T. Hoßfeld, A. Schwind, V. Burger, and P. Tran-Gia, "Group-based Communication in WhatsApp," in *Proceedings of the 1st IFIP Internet of People Workshop (IoP)*, Vienna, Austria, 2016.

[24] T. Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 483–502, 2002.

[25] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent Development and Applications of SUMO-Simulation of Urban MObility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3&4, 2012.

[26] Deutsches Zentrum für Luft- und Raumfahrt, "Simulation of Urban MObility - Wiki," 2018. [Online]. Available: http://sumo.dlr.de/wiki

[27] M. Seufert, C. Moldovan, V. Burger, and T. Hoßfeld, "Applicability and Limitations of a Simple WiFi Hotspot Model for Cities," in *Proceedings of the 13th International Conference on Network and Service Management (CNSM)*, 2017.

[28] W. Koch and B. Frees, "ARD/ZDF-Onlinestudie 2017: Neun von zehn Deutschen online," *Media Perspektiven*, no. 9/2017, pp. 434–446, 2017. [Online]. Available: http://www.ard-zdf-onlinestudie.de/files/2017/Artikel/917_Koch_Frees.pdf