

AMERICAN UNIVERSITY OF BEIRUT

MODELING MOBILITY AND CONTENT  
SIMILARITY FOR EFFICIENT  
DEVICE-TO-DEVICE DATA SHARING

by  
LYNN WAJDI AOUDE

A thesis  
submitted in partial fulfillment of the requirements  
for the degree of Master of Engineering  
to the Department of Electrical and Computer Engineering  
of the Faculty of Engineering and Architecture  
at the American University of Beirut

Beirut, Lebanon  
May 2016

AMERICAN UNIVERSITY OF BEIRUT

MODELING MOBILITY AND CONTENT  
SIMILARITY FOR EFFICIENT  
DEVICE-TO-DEVICE DATA SHARING

by  
LYNN WAJDI AOUDE

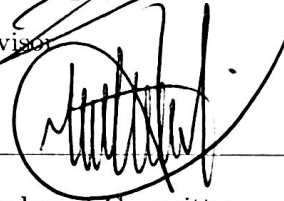
Approved by:

Dr. Zaher Dawy, Professor  
Electrical and Computer Engineering



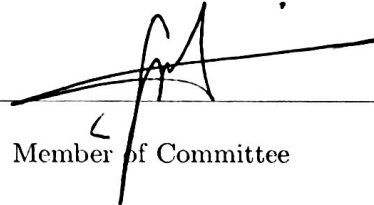
Advisor

Dr. Hassan Artail, Professor  
Electrical and Computer Engineering



Member of Committee

Dr. Fadi Karamah, Associate Professor  
Electrical and Computer Engineering



Member of Committee

Date of thesis defense: May 31, 2016



# Acknowledgements

This work was made possible by NPRP grant 7-1529-2-555 from the Qatar National Research Fund (a member of The Qatar Foundation).

I would like to thank my family and friends for their support and encouragement which helped me complete this thesis. A special thanks to my father who urged me to get my Masters degree, and has been my constant supporter during my studies.

I would like to express my special gratitude and thanks to my advisor, Professor Zaher Dawy, for his guidance and support throughout my years as a graduate student.

My thanks to the members of the committee, Drs. Artail and Karameh, for their endorsement and approval of my work.

I would like to thank Karim Frenn, former AUB student, for his help in developing the mobile application used in the experiment on which this work is based.

My appreciation goes to the students who participated in the experimental study on which this thesis is based. Their readiness to answer the proposed survey and use their smartphones for data collection made this work possible. Also, my thanks to the Institutional Review Board for the approving of the experimental procedures needed to make this work possible.

# An Abstract of the Thesis of

Lynn Wajdi Aoude for Master of Engineering  
Major: Electrical and Computer Engineering

Title: Modeling Mobility and Content Similarity for Efficient Device-to-Device Data Sharing

Creation and availability of user-generated content is facilitated by the widespread use of smartphones, which provide mobility and connectivity along with user friendly graphical interfaces. Considering that the majority of internet traffic is produced by end users, network offloading using device-to-device (D2D) communications is an attractive approach to enhance network performance. For any two users to share data via D2D links, they need to be in close proximity for a long enough period of time, and have similar content interests. In this work, we model mobility and content similarity between smartphones using an experimental study focused on wireless D2D content sharing applications. In our study, engineering students from the American University of Beirut filled a users content interest survey and downloaded an Android application on their smartphones in order to collect location and neighbor discovery data. The survey responses are used to perform content similarity analysis, while the data collected by the Android application is used to empirically model and analyze the number of contacts and inter-contacts between devices, and the contact and inter-contact durations, as well as to identify hub locations. The obtained models and insights can serve as key inputs to generate more accurate performance results for wireless network designs with D2D communications.

# Contents

<b>Acknowledgements</b>	<b>v</b>
<b>Abstract</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background</b>	<b>3</b>
2.1 Modeling Using Large Data Sets . . . . .	5
2.1.1 Mobility Related Modeling . . . . .	5
2.1.2 Content Related Modeling . . . . .	6
2.2 Modeling Using Small Data Sets . . . . .	7
2.2.1 Mobility Related Data Collection . . . . .	7
2.2.2 Content Related Data Collection . . . . .	8
2.2.3 Mobility Related Modeling . . . . .	8
2.2.4 Content Related Modeling . . . . .	10
2.2.5 Mobility Related Prediction and Validation . . . . .	11
2.3 Social-Aware Modeling . . . . .	12
2.4 Results Regeneration of Previous Work . . . . .	14
2.4.1 Data Sets . . . . .	15
2.4.2 Procedure and Results . . . . .	15
2.4.3 Summary . . . . .	18
<b>3 Problem Definition and Thesis Objectives</b>	<b>20</b>
3.1 Problem Definition . . . . .	20
3.2 Thesis Objectives . . . . .	27
<b>4 Experimental Study</b>	<b>28</b>
4.1 IRB Process . . . . .	28
4.2 Mobility Data Collection . . . . .	29
4.2.1 CARMA Description . . . . .	29

4.2.2	Data Collection . . . . .	31
4.3	Content Similarity Data Collection . . . . .	32
<b>5</b>	<b>Analysis and Results</b>	<b>36</b>
5.1	Mobility Related Parameters . . . . .	36
5.1.1	Number of Contacts . . . . .	36
5.1.2	Contact Duration . . . . .	42
5.1.3	Inter-Contact Frequency and Duration . . . . .	46
5.1.4	Hub Location Identification . . . . .	49
5.2	Content Related Parameters . . . . .	52
5.3	Social Networks Relationships . . . . .	55
5.4	Correlation Between Mobility, Content, and Social Networks . . . . .	58
<b>6</b>	<b>Conclusion and Future Work</b>	<b>61</b>
<b>A</b>	<b>Abbreviations</b>	<b>63</b>
<b>B</b>	<b>IRB Informed Consent Form</b>	<b>65</b>
<b>C</b>	<b>Survey Questions</b>	<b>69</b>
<b>D</b>	<b>Survey Answer Statistics</b>	<b>82</b>
	<b>Bibliography</b>	<b>85</b>

# List of Figures

2.1	Device-to-device communication examples . . . . .	3
2.2	Number of connected pairs for the Infocom05 trace: regenerated results (left) and from [13] (right). . . . .	16
2.3	Number of connected pairs for the Rollernet trace: regenerated results (left) and from [13] (right) . . . . .	17
2.4	Number of connected pairs for the KAIST trace: regenerated results (left) and from [13] (right). . . . .	17
2.5	Number of connected pairs for the Sigcomm09 trace: (a) regenerated results (including external devices), (b) regenerated results (excluding external devices), (c) from [14]. . . . .	18
2.6	Number of contacts between devices 1 and 49 from the Sigcomm09 trace over time. . . . .	18
3.1	Decision-making system model for efficient D2D data sharing . . .	24
4.1	CARMA GUI. (a) Dashboard screen. (b) Contacts screen. (c) Neighbors screen. (d) Settings screen. . . . .	30
4.2	Snapshot of the Data table extracted from the database . . . . .	31
4.3	Snapshot of the Neighbors table extracted from the database . . .	31
5.1	Total number of contacts in function of different sets of users with respect to day during (a) spring (b) summer . . . . .	37
5.2	Average number of contacts in function of different sets of users with respect to day during (a) spring (b) summer . . . . .	38
5.3	Total number of contacts in function of different sets of users with respect to time interval during (a) spring (b) summer . . . . .	38
5.4	Average number of contacts in function of different sets of users with respect to time interval during (a) spring (b) summer . . . . .	38
5.5	Empirical aggregated CDF for the normalized number of contacts per day for (a) spring (b) summer . . . . .	39
5.6	Empirical aggregated CDF for the normalized number of contacts per time interval for (a) spring (b) summer . . . . .	40
5.7	Pair-wise number of contacts for (a) spring (b) summer . . . . .	41



5.8	(a) Total and (b) average pair-wise number of contacts with respect to day . . . . .	41
5.9	(a) Total and (b) average pair-wise number of contacts with respect to time interval . . . . .	42
5.10	Pair-wise aggregate CDF of contact duration during (a) spring (b) summer . . . . .	43
5.11	Pair-wise aggregate CDF of contact duration with respect to day for (a) spring (b) summer . . . . .	44
5.12	Pair-wise aggregate CDF of contact duration with respect to time interval for (a) spring (b) summer . . . . .	44
5.13	Aggregated pair-wise inter-contact duration CDF for (a) spring (b) summer . . . . .	46
5.14	Aggregated pair-wise inter-contact duration CDF per day during (a) spring (b) summer . . . . .	47
5.15	Aggregated pair-wise inter-contact duration CDF per time interval during (a) spring (b) summer . . . . .	48
5.16	Aggregate number of (a) contacts [legend: red - green $\rightarrow \leq 50$ ; blue $\rightarrow \geq 125$ ](b) connected pairs [legend: red - green $\rightarrow \leq 3$ ; blue $\rightarrow \geq 7$ ] on AUB campus during spring . . . . .	50
5.17	Aggregate number of (a) contacts [legend: red - green $\rightarrow \leq 50$ ; blue $\rightarrow \geq 125$ ](b) connected pairs [legend: red - green $\rightarrow \leq 10$ ; blue $\rightarrow \geq 17$ ] on AUB campus during summer . . . . .	50
5.18	Aggregate number of (a) contacts [legend: red - green $\rightarrow \leq 50$ ; blue $\rightarrow \geq 125$ ] (b) connected pairs [legend: red - green $\rightarrow \leq 10$ ; blue $\rightarrow \geq 17$ ] in the Engineering Zone during spring . . . . .	51
5.19	Aggregate number of (a) contacts [legend: red - green $\rightarrow \leq 50$ ; blue $\rightarrow \geq 125$ ] (b) connected pairs [legend: red - green $\rightarrow \leq 50$ ; blue $\rightarrow \geq 125$ ] in the Engineering Zone during summer . . . . .	51
5.20	Large aggregate number of (a) contacts [legend: red - green $\rightarrow \leq 110$ ; blue $\rightarrow \geq 150$ ](b) connected pairs [legend: red - green $\rightarrow \leq 14$ ; blue $\rightarrow \geq 22$ ] in the Engineering Zone during spring . . . . .	51
5.21	Large aggregate number of (a) contacts [legend: red - green $\rightarrow \leq 110$ ; blue $\rightarrow \geq 150$ ] (b) connected pairs [legend: red - green $\rightarrow \leq 14$ ; blue $\rightarrow \geq 22$ ] in the Engineering Zone during summer . . . . .	52
5.22	Facebook friendships of spring participants . . . . .	56
5.23	Facebook friendships of summer participants . . . . .	56
5.24	Facebook social relationship strength between CARMA and survey spring participants . . . . .	57
5.25	Facebook social relationship strength between CARMA and survey summer participants . . . . .	58
5.26	Correlation between social relationship strength (top) and content similarity (bottom) during (a) spring (b) summer . . . . .	59

5.27 Correlation between social relationship strength (top), content similarity (middle), and number of contacts (bottom) during (a) spring (b) summer . . . . . 60

# List of Tables

2.1	Characteristics of the most relevant traces used for analysis in the literature . . . . .	9
3.1	Specifications of wireless technologies used in D2D communications [40, 41] . . . . .	20
3.2	System model decision threshold values and constraints . . . . .	26
4.1	CARMA data collection settings during spring and summer 2015 . . . . .	32
5.1	Gamma CDF parameters for the number of contacts . . . . .	40
5.2	Lognormal CDF parameters for the aggregate pair-wise contact duration . . . . .	42
5.3	Aggregate pair-wise contact duration statistics . . . . .	43
5.4	Lognormal CDF parameters for pair-wise contact duration with respect to day and time . . . . .	45
5.5	Pair-wise contact duration per day statistics . . . . .	45
5.6	Pair-wise contact duration per time interval statistics . . . . .	45
5.7	Gamma CDF parameters for pair-wise inter-contact duration with respect to time and day . . . . .	48
5.8	Pair-wise inter-contact frequency and duration statistics per day . . . . .	48
5.9	Pair-wise inter-contact frequency and duration statistics per time interval . . . . .	49
5.10	Aggregate number of contacts and connected pairs statistics . . . . .	49
5.11	User interest survey results . . . . .	53

# Chapter 1

## Introduction

With the widespread use of smartphones, content is available to the user anywhere and at any time. Smartphone interfaces are user friendly, which facilitates the creation and sharing of content. Note that a significant portion of internet traffic is accounted for by content sharing [1]. A direct result of an increase in content generation and sharing is thus an increase in server load and network traffic.

Server and network offloading becomes important as a result [2, 30]. From a networking perspective, a popular solution for the offloading problem is the employment of peer-to-peer networks since they have no single point of failure, and nodes can interact independently without a single entity controlling the network. For a more localized solution, where users are in close proximity to each other, Device-to-Device (D2D) communication is a key attractive technique to data sharing [2]. However, an efficient D2D sharing scheme is still necessary to distribute content efficiently. Such a scheme is not possible to devise without information about user mobility and user content sharing interests. With information about mobility patterns, the prediction of which devices a mobile node is most likely to be in contact with during a specific time interval is possible. Content similarity analysis helps recognize which devices are most likely to have the content a user is seeking. However, privacy and security restrictions make content similarity analysis challenging.

The literature discusses various mobility models, including social-based models and mathematical models. These models were derived by using either large-scale or small-scale data sets. The traces are either self-generated or publicly available. Some works aimed to model mobility data along a certain probability distribution function, whereas others presented future mobility prediction algorithms based on a node's mobility history [6, 11, 13]. Other works model content similarity between users based on the type of files available on their devices, as well as file popularity trends [7, 18, 20].

Our main objective in this work is mobility and content similarity modeling using an experimental approach focused on wireless device-to-device content sharing applications. To that end, we review existing experimental studies related to interaction and contact modeling, and content popularity modeling between mobile users, as well as statistical and mathematical techniques for deriving empirical models. Consequently, a study is conducted on the American University of Beirut (AUB) campus over two different time durations during spring and summer 2015, using a mobile application to collect mobility and content data from a relatively small number of participants (around 25). An additional part to the study is a customized online survey focused on users interests, users' device content, and users' social activity. The data collected will then be analyzed and modeled, with the results used to assess the feasibility of efficient device-to-device data sharing.

This thesis' contribution consists in modeling both mobility and content similarity and correlating them, with the consideration of the impact of social relationships between users as an additional layer. This will be done using our own experimental data collected specifically for the scope of the thesis. A detailed idea will then be obtained about which users, in contact with each other at a certain time, have similar content. Additionally, the strength of the social relationships between users will be a measure of their trustworthiness. These three aspects will then be used to assess whether efficient data sharing via D2D links is feasible.

The thesis is partitioned as follows. Chapter 2 includes a detailed review of the various related modeling techniques available in the literature. Chapter 3 presents the problem definition. Chapter 4 describes the experimental study procedure used for the data collection. Chapter 5 includes analysis and results. Finally, Chapter 6 concludes the thesis.

# Chapter 2

## Background

Device-to-device communication in cellular networks is defined as direct communication between mobile devices without passing through the base station or core network [3]. The same logic applies to D2D communication in Wi-Fi networks, where closely devices can communicate with each other independently of the network infrastructure. There are two types of D2D communication in cellular networks: inband and outband. In the inband case, cellular and D2D communication share the same radio resources, which may result in interference. On the other hand, in the outband case, D2D communication uses the unlicensed spectrum and adopts another wireless technology such as Wi-Fi Direct or Bluetooth, which avoids the interference between cellular and D2D communications [3]. Figure 2.1 shows examples of D2D communication.

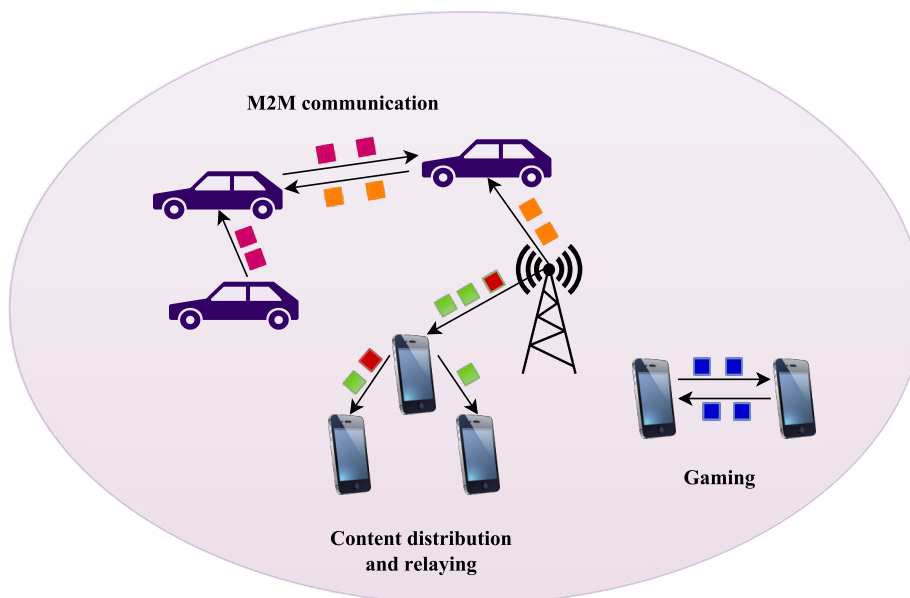


Figure 2.1: Device-to-device communication examples

D2D communication has several advantages such as enhanced spectral efficiency, improved throughput, energy efficiency and delay. In [3], methods were proposed to take advantage of D2D communication's advantages. For instance, clustering cellular users who are in range of Wi-Fi communication is proposed. Following this method, the member with the highest cellular channel quality communicates with the base station, and is responsible to forward the cellular traffic to its clients in the cluster. Another proposition is to cache popular video files in clusters with non-overlapping contents. When a user sends a request to the base station, it checks the availability of the file in the cluster. If the file is not cached, then the user receives the content directly from the base station. Otherwise, the user receives the file from its neighbor in the cluster.

*D2D communication presents opportunities for network offloading when it comes to data sharing, in multimedia cases specifically, especially in the case of non-overlapping content clustering featured in [3]. By modeling mobility and content similarity between mobile devices, one can determine how likely a requested file is available in a neighbor device at a certain time. Thus, in case of file availability, a user can request a file and fetch it from a nearby device, rather than sending the request to the base station and loading the wireless network resources.*

To exploit D2D communication's advantages, an efficient D2D sharing scheme is still necessary to distribute content efficiently. Such a scheme needs information about user mobility and user content sharing interests. Thus, modeling smartphone mobility, contact and inter-contact times between smartphones, and content similarity is of great importance. In other words, it is essential to find mathematical models that explain and represent:

- a mobile user's movements and how its location changes over time: *mobility modeling*
- how many times a pair of devices is in contact and for how long: *contact time modeling*
- how long it takes for the same pair to come into contact again after the current contact ends: *inter-contact time modeling*
- the similarity of content downloaded between users (i.e., type, genre, etc.): *content similarity modeling*

Studies have been conducted using both large-scale and customized data sets. Large-scale data sets contain data collected using a large number of devices for a long period of time (i.e., more than 6 months), in a wide area. Whereas small-scale/customized data sets contain information collected using a relatively small number of devices (i.e. less than a 100) within a small area (i.e., a university

campus or a few neighborhoods of a city), during a short period of time (i.e., days or weeks). These data sets either include mobility information or content information. Since our work will be limited to the university campus and to a time period no longer than a couple of months, we are more interested in small-scale data set analysis. We will also investigate the social-aware modeling techniques available in the literature.

## 2.1 Modeling Using Large Data Sets

### 2.1.1 Mobility Related Modeling

Mobility-related large-scale data was collected in [11] using a client software installed on the smartphone that uploads information via Wi-Fi to a server. Location data was collected from the GPS sensor and Wi-Fi data. This study involved 153 participants over a 17-month time period. The data was used to generate a conditional model for mobility prediction, where a mobile application updates a node's mobility information and predicts the next location in real time. Prediction accuracy is between 0.411 and 0.604, depending on the parameters taken into consideration. An accuracy of 0.411 corresponds to predicting the next location depending on which day of the week it is, whereas an accuracy of 0.604 corresponds to taking into consideration the current location and the current hour of the day. Note that since the model proposed is dynamic, prediction accuracy increases over time. Additionally, the data in [11] served to predict how long a mobile device will stay in the same location. In this case, predictability error was between 0.441 and 0.596, also depending on the parameters taken into consideration.

Mobility can also be modeled according to the social-based approach. Social characteristics of human mobility and social activities are obtained by considering the proximity social networks derived from a real-world mobility trace [16]. Mobile users form social networks by online and proximity communications with social structures and mobility patterns. A trace involving around 100 participants from MIT covering a period of nine months was used. During this experiment, human social interactions and dynamics are studied by analyzing the nearby encounters and interaction behaviors of this trace. The encounter patterns of a community can be used by a user to carry out peer discovery and subsequently data sharing. The strength of the social ties between users is a measure of trustfulness, which alleviates privacy and security concerns: if a user has a history of sharing content with user A more often than with user B, when both users are available for D2D communication, the user will fetch the data from user A, even if user B has more resources in terms of throughput or bit rate. Analysis results show improved performance when using the social-based approach in peer discovery.



*Large scale mobility data sets serve to predict the likelihood of finding and fetching the content requested from a neighbor device. This is done by examining and modeling the history of user mobility patterns. During a certain day of the week, relying on previous user schedules, the most likely devices having the desired content to be found in the vicinity can be predicted. Adding another parameter, the hour of the day for example, increases prediction accuracy. Also, studying the strength of social relationships between users gives a good idea about their individual interests. Thus, knowing which users have similar interests will go a long way towards increasing the efficiency of D2D data sharing: a user searching a neighbor's files for specific content will be more likely to find it and fetch it if the neighbor and user have similar interests. Therefore, studying the social aspect of the traces can be beneficial in the sense that it is more likely to reduce the time needed to probe for a file in neighboring devices.*

## **2.1.2 Content Related Modeling**

In [7], a huge amount of data was collected from user-generated content (UGC) and non-UGC services to obtain the distribution characteristics of requests across videos, the evolution of viewer's focus, and the shifts in content popularity. Data was collected by crawling the websites of four different online video providers and using the public data of one more. It was observed that the more a file is popular, the more it is requested by users. UGC has a power law with a truncated tail for a probability distribution. The tail truncation is due to the fetch-at-most-once user behavior in peer-to-peer environments. The effect of this behavior is amplified when the number of videos is small and/or the average number of requests per user is large. Moreover, it was discovered that when users share videos for a long time period, peer-to-peer sharing always supports 60% of videos with at least 10 current users. Therefore, even when only a small number of users is supported, a peer-to-peer sharing scheme helps decrease server workload considerably.

Chandra and Yu [18] collected a list of shared iTunes songs from 239 users in late April/early May 2006 at University of Missouri School of Journalism, resulting in nearly 2.5TB of data. Assuming that users share their entire music library, analyzing the shared contents gives an indication about the user's music interests. The popularity distribution of song names proved to follow a Zipf-like distribution in a log-log scale. This Zipf distribution of objects means that few objects are extremely popular while many objects are rare. Users are unlikely to find specific objects, but they can find objects belonging to a broader category. Clustering songs by category proved unhelpful since this categorization is more or less subjective, resulting in several categories to actually represent only one. While searching by artist name is promising since only about half of the artists had only one song in the system. User availability information was also collected and the number of unique users seen daily is plotted. Less activity was noticed in the early morning hours than during the rest of the day. User availability is poor

for a large number of users, however about 20% of users exhibited large churn rates. Furthermore, users proved to be predictable in regard to their availability at a certain time. Another conclusion is that for successful serendipitous media sharing, the number of copies of an object should increase.

In [19], Gummadi et al. analyze a 200-day trace of Kazaa peer-to-peer file sharing traffic collected at the University of Washington between May and December 2002, adding up to 20TB of traffic and 1.6 million requests. An important observation of the study is that users request less bytes as they grow older (i.e., with increasing time of using Kazaa). In other words, new clients generate most of the traffic load. Average session lengths are found to be typically small with a median average session length of 2.4 minutes, and most clients have high activity fractions relative to their lifetimes (i.e., the fraction of time a client is transferring content over the client’s lifetime or the entire trace’s duration). Kazaa clients fetch objects at most once since objects are immutable and are not downloaded in a small amount of time. Additionally, object immutability results in a short-lived popularity of Kazaa objects, making the recently born objects the most popular. However, old objects are the most requested.

*Studying content traces over a long period of time gives an idea about their evolution; i.e., the evolution of number of requests, and file popularity over time. While the probability distributions of these parameters would describe the availability of the files, what we hope to do is model the similarities in content between devices, in other words, file names, genre, etc. This will indicate whether a neighbor device is likely to have the requested content or not.*

## 2.2 Modeling Using Small Data Sets

### 2.2.1 Mobility Related Data Collection

In [6], data collection was accomplished by using LifeMap application on mobile devices carried by four graduate students for a duration of 8 weeks in a  $15 \times 20km^2$  area in Seoul, Korea. LifeMap is a mobile application that gathers location-based data. Accelerometer data was also used in [6] to pinpoint accurate locations (e.g. different floors of the same building). Ciobanu et al. [8] gathered traces on a university campus using HYCCUPS tracer [10], a mobile application running in the background that collects information about a device’s encounters with other nodes or access points. Information was gathered via Bluetooth and AllJoyn. Paired devices were scanned via Bluetooth, whereas AllJoyn’s Wi-Fi-based framework allowed information gathering via wireless sessions. The experiment in [8] spanned 64 days and 66 participants. Whereas [6] used GPS and accelerometer data to collect mobility information, [8] employed Bluetooth and Wi-Fi technologies. All four techniques, if utilized together, may give a more

precise idea about a node’s mobility information, which would lead to better modeling. In [9], six different traces were analyzed to study smartphone mobility. Data was gathered via Wi-Fi, GPS, GSM carrier, and Bluetooth. In other words, information about a node’s association with access points and GSM cells, start and end contact time between devices via Bluetooth, and GPS coordinates are logged. However, instead of gathering new traces, one can use traces from online archives, such as CRAWDAD [15], to analyze. Furthermore, in [13], the authors used four datasets available online. Two of them contain location information in a conference setting (via iMotes), one has information gathered during a rollerblade tour in Paris (via iMotes), and the fourth is a set of GPS coordinates. In [14], in addition to two of the traces used in [13], another data set was considered where data was collected during the first day of a conference in Barcelona. The study involved 76 participants using Bluetooth-based smartphones. Two traces are used in [17]: *Intel* and *Infocom06*. The first trace records 128 people’s contacts in one of Intel’s labs for a duration of six days, while the second records 98 people’s contacts in a conference setting.

## 2.2.2 Content Related Data Collection

The trace used in [20] consists of collecting 923,000 files from 12,000 clients by crawling the eDonkey network for 3 days. This trace is considered in the small data set category in comparison to the large content traces such as [7, 18, 19], in addition to the short duration of the data collection period. This trace was obtained by crawling the eDonkey network using a crawler running two tasks: discovery of eDonkey clients and scanning their contents. eDonkey client discovery is done by connecting to a maximum number of eDonkey servers and requesting their clients’ lists. The second task is achieved by trying to connect to every eDonkey client discovered. In case of success, the unique client identifier and its list of shared files are obtained.

*Table 2.1 presents a summary of the characteristics of the most relevant traces used in the literature. For each work, it details the number of traces used in the authors’ analysis, the number of participants in the study and its duration, the description of the location where the data was collected, and a brief description of the data collection technique.*

## 2.2.3 Mobility Related Modeling

The approach in [8] consists of analyzing a node’s past encounters and approximating the time series as a Poisson distribution, representing the probability that a node will make  $N$  distinct contacts within a specific time frame. The chi-squared test proves that the Poisson distribution prediction is accepted with a

Table 2.1: Characteristics of the most relevant traces used for analysis in the literature

Ref #	# of traces	# of participants	Duration	Location	Collection Technique
[6]	1	4	8 weeks	$15 \times 20km^2$ area	Accelerometer data and LifeMap app on HTC Hero Cell phone: 908 meaningful places & 1,923 APs discovered
[7]	5	number of videos = 2,193,376	1 trace: 3 years; 4 traces: 1 month	N/A	Crawling sites for meta-information and public information available online
[8,10]	1	66	64 days	University Campus	HYCCUPS tracer, Bluetooth, and AllJoyn
[9]	6	Total number of devices = 726	Between 3 days and 16 months	N/A	Online archives, and privately collected data
[11]	1	153	17 months	European country	Wi-Fi and GPS sensor data collected by a client software (installed on smartphones) that uploads information to a server via Wi-Fi
[12]	4	1000	N/A	N/A	WLAN trace from MIT, USC trace, Dartmouth trace, and Simulated trace
[13]	4	DS1: 41; DS2: 78; DS3: 62; DS4: 131	Between 3.5 hours and 5 days	Conference scenario, Paris, and area of radius 3km	CRAWDAD traces: iMotes (Bluetooth) and GPS coordinates

2.49% risk if computed as an average per hour per day of the week, since academic schedules are repetitive and have an hour as a unit of time. On the other hand, analysis of the traces in [9] shows a power-law distribution of inter-contact time between smartphones up to a characteristic time, then an exponential decay. This distribution holds across various mobility traces. An important observation is that the return time exhibits the same dichotomy as the inter-contact time. Moreover, devices are in contact in a small set of different locations. These two propositions suggest that the dichotomy in the distribution of the return time explains the dichotomy in the distribution of inter-contact time. Furthermore, Hsu et al. [12] proposed a time-variant community-based mobility model based on location preference. Comparing synthetic traces generated from this model and actual traces, the model’s accuracy is below 20%, but is below 10% for most cases. Each of the three works presented here model different characteristics of a user’s mobility pattern. The number of contacts within a specific time frame [8] follows a Poisson distribution with an error risk of 2.49%, whereas a pair’s inter-contact time and a device’s return time to a certain location [9] follow a power-law probability distribution that holds for multiple traces. Finally, a model based on the location parameter [12] has an error below 10% in most cases.

In [14], the number of connected pairs over time was evaluated. However, not only direct contacts were considered, but contacts at different hop counts were taken into consideration as well (e.g. 2-hop contacts, 3-hop contacts, etc.). The minimum distance is evaluated for each pair of nodes according to the number of hops. It was observed that number of nodes in  $k$ -contact is much larger than the number of nodes in direct contact. This presents more data forwarding opportunities than in the direct contact case. The average duration of an interval during which nodes remain at the same hop distance is also evaluated. The conclusion was that this interval decreases as the environment becomes more dynamic.

*Modeling techniques consist of determining the probability distributions of different parameters, such as the intercontact time between smartphones and the number of contacts within a specific time interval. However, modeling should not be limited to 1-hop contacts (direct contacts). By introducing the  $n$ -hop concept as in [14], additional sharing resources will be made available by forwarding the requested content to the reciever via one or multiple intermediate nodes.*

## 2.2.4 Content Related Modeling

Le Fessant et al. [20] measured a file’s popularity by its replication degree. Authors observe that few files are extremely replicated, while most are not replicated at all. Kazaa and eDonkey workloads show similarities with popularity distributions in function of file rank show an intial flat region followed by a linear trend on a log-log scale. Another observation is the correlation between geographical clustering and video files: peers requesting a certain video file are more

liable to download it from peers in their own country, thus reducing the delay and download time. As for audio files, the exploitation of interest-based locality improves peer-to-peer performance. If two peers share interests, the search mechanism is improved significantly if these peers are connected and first send their requests to each other.

## 2.2.5 Mobility Related Prediction and Validation

Talipov et al. [6] used mobility information to devise a routing algorithm for efficient content sharing. Using location information, the algorithm learns the user’s mobility patterns and predicts a node’s future mobility information to estimate whether its movement would lead it to a contact with the destination node. The algorithm’s learning accuracy is above 90%, whereas its prediction accuracy is 75% for a learning period of one week. The data collected was also used to validate the proposed algorithm, as is the case in [13].

Model validation is an important step since it ensures that the model devised is correct and can be used for designing an effective sharing scheme. Approximating the time series of a node’s location information [8] serves to predict the node’s future mobility information according to the corresponding probability distribution. However, if some data points are missing or inaccurate, the result of applying the probability distribution may be wrong [8, 11]. Modeling data also serves to recognize certain dependencies and characteristics of a node’s mobility. For example, we observe mobility time-of-day non-stationarity and dependency [9], and a tendency to periodically visit the same place as well as the preference of few places over others [12]. This model allows the prediction of the expected average time for a node to have contact with the destination node, and of the expected time for two mobile nodes to come into contact with each other. Moreover, in [13], new sharing opportunities are introduced and validated by the  $n$ -ary inter-contact principle, where  $n$  is the distance between two nodes. Thus, even if two nodes are not in direct contact but there exists a path between them, sharing content is possible for these two nodes. Additionally, in [14], given data from previous time windows, predicting the  $k$ -contacts during the next target period is possible. Zhang et al. [17] establish a social-aware peer discovery approach for D2D communication. They use the characteristics of community and centrality to aid ad-hoc discovery. Specifically, they divide devices in the network into several groups according to their centralities to improve the network’s performance. They validated the usefulness of their approach by using two traces and evaluating the performance of the proposed algorithm in regards to peer discovery ratio, data delivery ratio, and delivery delay. The conclusions obtained are: (1) the social-aware approach improved peer discovery. The peer discovery ratio increases with system energy until a certain threshold then remains unchanged. The possibility to improve the discovery ratio by increasing the number of groups also presents itself; (2) the delivery ratio is also increased,

as a natural result of the increased discovery ratio; (3) the data delivery delay is decreased.

In summary, [6] and [8] both use a node’s mobility history to predict its future mobility. However, the prediction is based on an algorithm in [6], but on a time series approximation in [8]. In the latter case, any error in data measurement leads to a wrong prediction, whereas in the former case the algorithm’s prediction accuracy is 75%. Where prediction isn’t applied, certain data dependencies come to light as in [9, 12], or new sharing opportunities are introduced as in [13]. Finally, as is the case in [6, 13], data is also used for validation purposes, which determines the correctness of the model derived. In both [6] and [13], validation proved that the model derived was correct: in [6] correctness is reflected in the learning and prediction accuracies, and in [13] and [14], new sharing opportunities were discovered for more than half of the nodes in inter-contact mode. Furthermore, [17] presents the possibility of using a social-aware peer discovery approach for D2D communication in order to increase the discovery and delivery ratios, while decreasing the data delivery delay.

*Data collected can be used to either predict a user’s future behaviour based on his/her history, or is used to validate an algorithm developed by the authors and prove its correctness. In this section, we summarized what was featured in previous work relating to prediction and validation for the sake of completeness, since our work will focus on the modeling stage.*

## 2.3 Social-Aware Modeling

Since smartphones and mobile devices in general are carried by human beings who form relatively stable social networks, one can argue that D2D file sharing efficiency can be improved by taking advantage of the users’ social behaviors. People who are close in the physical and social domains tend to have more encounters. However, detecting social context is not a simple matter, and information from multiple sources can be utilized to this end, such as application-layer information from existing social networks, but also historical information about past user encounters or communication [26, 30, 31].

Authors in [27] develop a social-aware D2D communication system with centrality-aware peer discovery and community-aware resource allocation. The social relationship between users is assessed by using online social networks and the physically close social networks formed by user contacts. Li et al. [27] visualize the social characteristics of human mobility and social activities by collecting a mobility trace involving 100 users (students and staff) on MIT campus. The data collected consists of user locations, communication behaviors, and device usage behaviors. A social network is formed by taking into consideration user centrality and relationship strength, assessed by analyzing the contacts between users.

Authors found that allocating more spectrum and energy resources to users with strong relationships improves D2D communications by increasing the peer discovery ratio and improving spectral efficiency. Additionally, user trustworthiness can be assessed using the relationship strength, thus addressing privacy and security concerns. Resource allocation in the system proposed by the authors relies heavily on a user’s centrality in the social network. A user with a high degree of centrality plays a key role in data transmission, thus more resources are allocated to this user. Furthermore, a central node tends to have more contacts with nearby devices. Therefore, peer discovery can be improved by sending beacons from the central node to the other nodes in the network, instead of relying on random beaconing.

Cao et al. [28] design a social-aware video multicast (SoCast) leveraging D2D communication. SoCast stimulates efficient cooperation among mobile users by taking advantage of two types of social ties: social trust and social reciprocity. Social trust exists when users are willing to help each other because they are friends. Social reciprocity does not require a social relationship between users to ensure their willingness to help each other. In this system, users form groups to obtain missing packets of a video mutlicast by the base station (BS) from others and restore incomplete video frames, according to the unique video encoding structure, dramatically improving mobile video quality. A component of the SoCast system is a social trust database. This database reflects the social ties between users represented by a graph where the vertices are the users, and the edges connecting them represent the presence of social ties between them (i.e., friendship, kinship, colleague relationship). When missing video frames exist, users broadcast a request that contains the IDs of the missing packets to other users over the channel via a random access manner. Through a matching and feedback process, users can obtain an information table, which contains the information of candidate helpers for video frame restorations, the video frames to be restored, required resources of delivery, and social trust relationship through the local social trust database. Thus, users form either social trust groups or social reciprocity groups. The first type of groups has two members, the helper and the taker; while the latter type has at least two members, which form a reciprocal cycle, in which a user will provide and receive assistance from others.

In [29], Wang et al. aim to generalize social networks and apply graph theory for appropriate resource allocation in wireless networks. They state the situation as a bipartite graph problem where a user aiming to retrieve data from another selects his/her partner depending on whether the partner has the desired content, and whether they are trusted or tightly connected. The authors propose a hierarchical bipartite (HBP) system, which consists of two bipartite layers: the upper layer for partner selection, and the lower layer for resource allocation. In the upper layer, consideration of social ties and interests similarities ensure that users having similar content wish to share their data with higher security. In the lower layer, mutual social relationship and interactions between user pairs



are considered for efficient resource allocation, with efficient interference management between trusted pairs. Users thus need to be clustered according to social trust, interest similarity, caching capacity, and computational capability.

Xu et al. [33] design a socially aware mobile multimedia community-based approach (SMMC) for content sharing, which integrates performance-related factors (PRFs) and analyze how they influence serving capabilities (e.g., resource sharing and delivery) and scalability. SMMC architecture includes community discovery and content delivery. In community discovery, user relationship similarity is measured with regard to user content interests, social interactions, and mobility. For SMMC content delivery, a network context-aware concurrent multipath transmission is proposed, which includes capacity estimation of transmission path, packet loss identification, friendliness, and retransmission control. Community discovery mechanisms were studied by investigating user video playback behaviors, such as the number of watched videos, video switchover, and playback time. The number of watched videos describes the user interest coverage. Video classification uses content similarity to map videos to categories for better interest coverage boundaries. Video switchover denotes a change in user interests in terms of category. Social relationship strength measurement relies on interaction behaviors such as push content to other users, attention which relates to user interest in and acceptance of content pushed by others, and forwarding. The mobility measure is the stay time in a specific area. Thus, Xu et al. [33] construct a community formed by multiple sub-communities depending on whether users share any combination of two or three of the community measures discussed.

*Social network information can be used to enhance D2D data sharing efficiency, whether by determining user trustworthiness, or designing new mechanisms for efficient D2D communications. Note that user centrality is an integral part of social-aware D2D modeling, since the user with the highest degree of centrality has the most number of ties to others in its community. In our work, we will use social network relationships as an added layer insuring user trustworthiness to our system model.*

## 2.4 Results Regeneration of Previous Work

Investigation of modeling techniques used in previous work was done by results regeneration using MATLAB. Results regeneration is done based on [13] and [14]. Three of the four datasets used in [13] are considered, and the number of contacts vs. time is plotted for all traces.

## 2.4.1 Data Sets

The datasets acquired are available on CRAWDAD [15]. Four traces were used: Infocom05, KAIST, and Rollernet [13], and Sigcomm09 [14].

### 2.4.1.1 Infocom05

The data in this file was gathered during a five-day conference using 41 iMotes. The dataset consists of six fields of interest. The first two fields are the ID numbers of the device pair in contact. The first field is the timestamp when the pair comes into contact, and the fourth field is the timestamp when the pair's contact ends. The fifth field is a count of how many times each pair is in contact for the duration of the trace. Finally, the last field determines the inter-contact time between two successive contacts of the same pair. Note that having the two timestamps allows the calculation of the duration of each contact occurrence.

### 2.4.1.2 Rollernet

For the Rollernet dataset, information was gathered during a three-hour rollerblade tour in Paris using 62 iMotes. The format is the same as the Infocom05 dataset.

### 2.4.1.3 KAIST

KAIST data was compiled by GPS logs of 92 nodes on the campus of KAIST University. GPS coordinates are converted into x and y coordinates, and each node is assigned a 10-meter wireless transmission to emulate Bluetooth. The data format consists of three fields: the time in 30 seconds intervals, the x coordinates, and the y coordinates.

### 2.4.1.4 Sigcomm09

The data is gathered during the first day of a conference in Barcelona. The experiment recorded 76 user-relationships using Bluetooth-based smartphones. Each phone logged contacts every 120 seconds. In our study, only three fields of the collected data are used: the timestamp, the device's user ID, and the user ID of the discovered device.

## 2.4.2 Procedure and Results

Results regeneration was conducted using MATLAB and the datasets mentioned in the previous section. For all traces, only direct contact results were regenerated. Direct contact occurs when two devices interact with each other directly, with no intermediate node involved. Direct contact was evaluated in

two ways, depending on the data available in each data set. For *Infocom05* and *Rollernet*, direct contact is determined when the start time and time of a pair’s contact is within the time-step limit considered. For *KAIST*, a device is in direct contact with another if it is within a circle of radius 10m of that device.

*Note that for Figures 2.2, 2.3, 2.4, and 2.5 the comparison is done between the graph on the left and the first surface from the bottom on the rightmost plot. The numbers of the legend correspond to the hop count between sender and receiver, with contact meaning direct contact between the devices.*

### 2.4.2.1 Infocom05

For the Infocom05 trace, as in [13], data for a 12-hour period of the second day of the conference was considered. A 200-seconds time-step was used since the exact time-step by the authors of [13] is unknown. This results in minor variations in the average number and maximum number of contacts. However, Figure 2.2 shows that the contacts number trend obtained by regeneration is the same as the one presented in [13]. The difference in the number of connected pairs (1640 in [13] vs.1530 in our case) may be caused by the fact that the actual 12-hour period of the second day used by the authors is unspecified. A shift in time would affect the number of connected pairs obtained during the 12-hour interval.

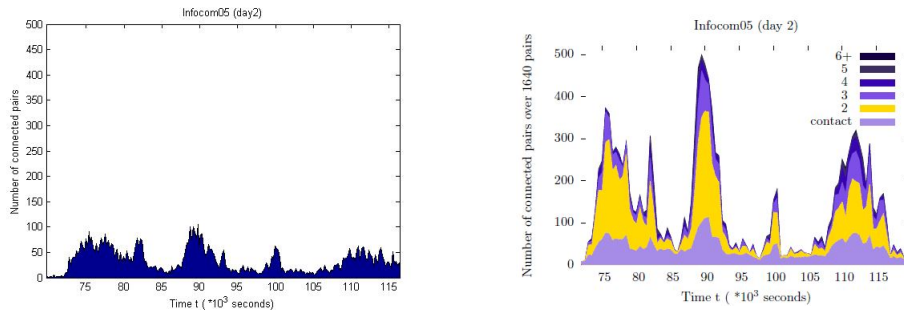


Figure 2.2: Number of connected pairs for the Infocom05 trace: regenerated results (left) and from [13] (right).

### 2.4.2.2 Rollernet

For the Rollernet trace, the entire duration was considered for result reproduction. The total duration of the trace is 1014 seconds. However, the plot in [13] shows the number of contacts up to 5000 seconds. Even considering only the first 5000 seconds of the trace, the number of connected pairs obtained is higher than that produced in [13]. A possible reason for this is that Phe-Nau et al. used only a portion of the trace in their work. Additionally, the time step used by the authors

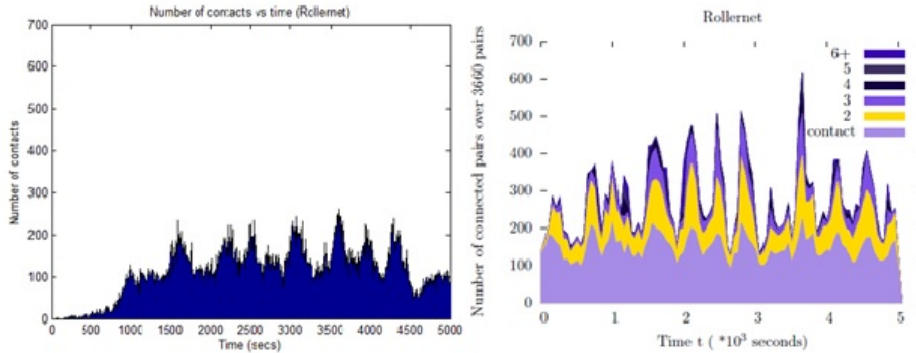


Figure 2.3: Number of connected pairs for the Rollernet trace: regenerated results (left) and from [13] (right)

in unspecified, so a time step of 10 seconds was used in the regeneration. The difference in the average and maximum numbers of contacts we obtained may be due to the time step choice, as in the case of the *Infocom05* trace. Figure 2.3 shows a close resemblance to the plot obtained in [13].

#### 2.4.2.3 KAIST

The KAIST data set was considered in its entirety for result regeneration. We attempted to verify the number of contacts relative to time. Figure 2.4 shows the same trend as in the plot produced in [13]. However, we obtained very different average and maximum numbers of contacts. This is due to the high amount of missing data in the data set obtained. While we chose to ignore the missing data points, it is unclear which data preprocessing technique the authors of the paper.

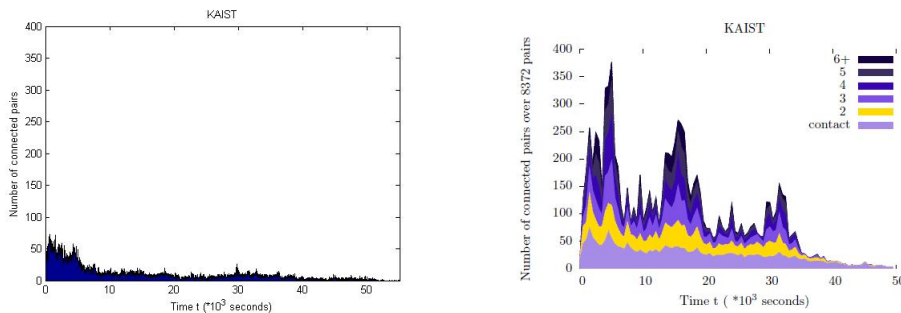


Figure 2.4: Number of connected pairs for the KAIST trace: regenerated results (left) and from [13] (right).

#### 2.4.2.4 Sigcomm09

The Sigcomm09 data set was used to verify the number of connected pairs over time, including or excluding external devices, as well as the number of contacts for each pair of nodes, excluding external devices. Figure 4 below compares the plots obtained for the number of connected pairs over time obtained from the regenerated results to that obtained in [14]. Comparing Figures 2.5(a) and 2.5(c), we notice that the plots are identical for the 1-hop case.

Figure 2.6 shows the number of contacts between device 1 and device 49 over the duration of the trace.

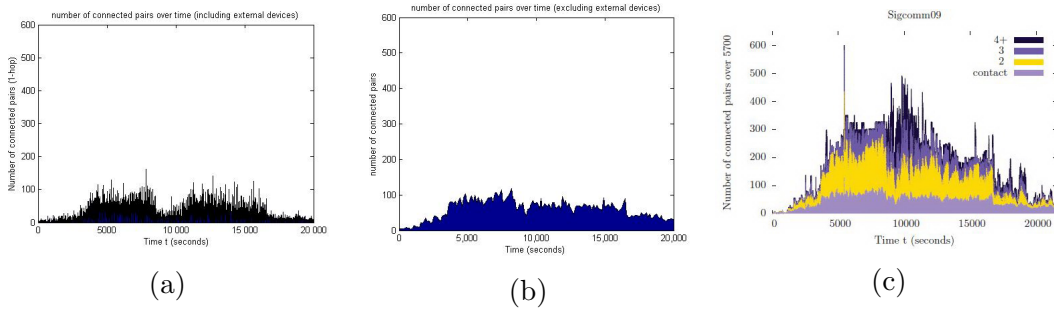


Figure 2.5: Number of connected pairs for the Sigcomm09 trace: (a) regenerated results (including external devices), (b) regenerated results (excluding external devices), (c) from [14].

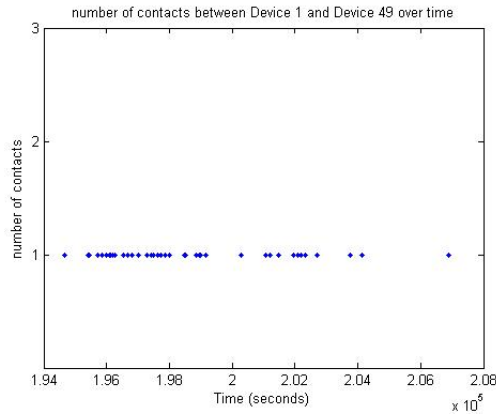


Figure 2.6: Number of contacts between devices 1 and 49 from the Sigcomm09 trace over time.

### 2.4.3 Summary

Modeling contacts between smartphones is the first step towards mobility learning. Knowing the number of nodes a mobile device may be in contact with

during a certain time interval is one parameter in determining whether efficient content sharing is possible. However, a more complete picture of a node's mobility information is needed to devise an efficient sharing scheme. To that purpose, one can take things further and model other parameters characteristic of mobility information. A few examples of the latter would be: contact duration, inter-contact time, movement patterns, most visited locations, and which nodes are most likely to be in contact with a specific node during a certain time interval.

# Chapter 3

## Problem Definition and Thesis Objectives

### 3.1 Problem Definition

For successful D2D data sharing, the users attempting such a connection need to meet two key requirements:

- The users need to be in close proximity, depending on the technologies used in their devices (see Table 3.1).
- The users need to be connected for a sufficient period of time to ensure the complete transfer of media files requested.

These two conditions apply for opportunistic D2D data sharing. However, by themselves, these two requirements are not enough for efficient data sharing, with the goal of network offloading in mind. To increase the efficiency of D2D media file sharing, several more conditions need to be taken into consideration.

Table 3.1: Specifications of wireless technologies used in D2D communications [40, 41]

<b>Technology</b>	<b>Maximum Data Rate</b>	<b>Effective Data Rate</b>	<b>Maximum Outdoor Range</b>	<b>Effective In-door Range</b>
Bluetooth v2.1	3 Mbps	2.1 Mbps	100 m	10 - 12 m
Bluetooth v3.0	24 Mbps	3 Mbps	100 m	10 - 12 m
Bluetooth v4.0	24 Mbps	1 Mbps	50 m	–
Wi-Fi Direct	250 Mbps	40 - 50 Mbps	150 m	20 - 30 m

1. Users' mobility patterns need to be examined. Human mobility is often repetitive due to the mostly fixed schedules people follow in their daily lives. Thus, periodicity in mobility patterns can be established for each user. Furthermore, every user is bound to have a set of locations that s/he visits on a regular basis, and where s/he stays for a significant amount of time; e.g., an office, a classroom, or a lounge area. These locations will then be recognized as meaningful places (MPs) to the user. These MPs, when shared between several users, thus become known as hub locations (HLs). Depending on the users' mobility patterns, when more than one user is present in a hub location at the same time, it is a chance for file sharing. Additionally, the use of intermediate nodes between the sender and the receiver becomes a possibility. As an example, consider the following scenario: three users A, B, and C frequent the same hub location. Users A and C meet for an hour, then users B and C meet half an hour after user A has left the location. In case user B needs a file that user A has, user C can therefore act as an intermediate node between the two, storing the file from A until it can be transferred to B. Hub locations are thus places where D2D media file sharing is more successful than in an opportunistic environment.
2. The number and duration of contacts between a pair of users are important metrics in determining the likelihood of D2D file sharing between the two.
  - (a) If two users' contact frequency ranges from never to only occasionally, then either is not a reliable source of media to the other, since it is not always reachable.
  - (b) If two user devices are in contact for seconds at a time, then neither is not a reliable source of D2D sharing of media files of large size.
3. The number and duration of inter-contacts are also significant measures to consider for efficient data sharing via D2D links. Inter-contact times are defined as time periods where two users are not in proximity to each other, even though they might meet at certain time intervals.
  - (a) If two users are in inter-contact for a long period of time, even if the number of inter-contacts is small, any of the two is not considered as a reliable source of media files in D2D communication.
  - (b) If two users are inter-contact for a relatively long period of time, with a large number of inter-contacts, then these users could share data via D2D links during specific time intervals, inferred from their mobility patterns' history.
  - (c) If two users are in inter-contact for a short period of time, with a small number of inter-contacts, then either one is a reliable source of media files to the other in the context of D2D data sharing.



- (d) If two users are in inter-contact for a short period of time, with a large number of inter-contacts, then the D2D connection between the two is not stable. Thus, these two users are more likely to be able to share a data file of small size, while the transfer of large size files proves problematic.
4. For any two users to be able to share files, they need to have similar content on their devices. Therefore, measuring content similarity between two smartphones is essential in determining whether successful D2D data sharing is possible. The higher the content similarity between two devices, the more likely D2D file sharing is successful. In case where content similarity between two smartphones is low, D2D media file sharing is impossible.
  5. Social relationships between users also have impact on the success of D2D data sharing. The strength of a social relationship between two users is an indicator of their level of trustworthiness to each other. The stronger the social relationship between two users, the more trustworthy one of them is to the other. Note that people tend to request and download data from others they find trustworthy; i.e., friends, colleagues, and family. In a scenario with three users A, B, and C, with A and B, and B and C being friends. Assuming they have the same content similarity measure and the three are in the same location at the same time, B can request files from either A or C. On the other hand, A is more likely to request data from B, instead of C. The same logic applies to user C.

Each of these conditions alone can increase the efficiency of D2D data sharing, albeit not too much. The more conditions satisfied, the more efficient D2D data sharing is. Accordingly, for optimal D2D media file sharing, users need to be in close proximity to each other, meeting in certain hub locations at the same time, have a large number of contacts with long contact duration, as well as a small number of inter-contacts with a small inter-contact duration. They also need to have a high content similarity between their smartphones and a strong social relationship.

Conditions one through three refer to users' mobility information; while condition four refers to content information, and condition 5 to social information. By correlating these three types of information, better efficiency can be achieved in D2D media file sharing. Furthermore, a decision-making system model can be constructed using the measures proposed, illustrated in Figure 3.1. This system model will be followed to determine whether a user will retrieve the file requested from the server using the network core or fetch it from a neighbor via D2D communication. The decision-making process starts when we have two users in contact and follows the steps as ordered below.

1. The content similarity between the two users' devices is checked. If we have

low content similarity, then we decide against D2D file sharing. Otherwise, we continue with the decision chain.

2. The social relationship strength between the two users is assessed.
  - (a) If the social relationship is weak, then we check if the user requesting the file is in contact with another more trusted user. If the answer is yes, then the current decision-making process is stopped and restarted for the new pair of users. Otherwise, we continue with the decision chain.
  - (b) If the social relationship is strong, then we continue with the decision chain.
3. The past inter-contact duration and frequency of these two users are assessed, and four outcomes are possible.
  - (a) If the inter-contact duration is long and the inter-contact frequency is high, then the users are in scheduled D2D file sharing mode. In other words, these users can only share files efficiently via D2D communication during specific time intervals on specific days of the week.
  - (b) If the inter-contact duration is long and the inter-contact frequency is low, we move to assessing the past contact duration between the two users. If it is long, then D2D sharing of any file is possible. Otherwise, only files of small size can be shared via D2D links between these two users.
  - (c) If the inter-contact duration is short and the inter-contact frequency is high, then D2D communication can be used to share only files of small size.
  - (d) If the inter-contact duration is short and the inter-contact frequency is low, then we continue with the decision chain.
4. The past number of contacts between the two users is examined. If the two users usually have a relatively low number of contacts, then opportunistic D2D file sharing is employed. Otherwise, we proceed to the next step.
5. The past contact duration is analyzed. If the two users have a history of short contacts, then only sharing of small size files is reliable via D2D links. Otherwise, D2D communication is employed when sharing all files.

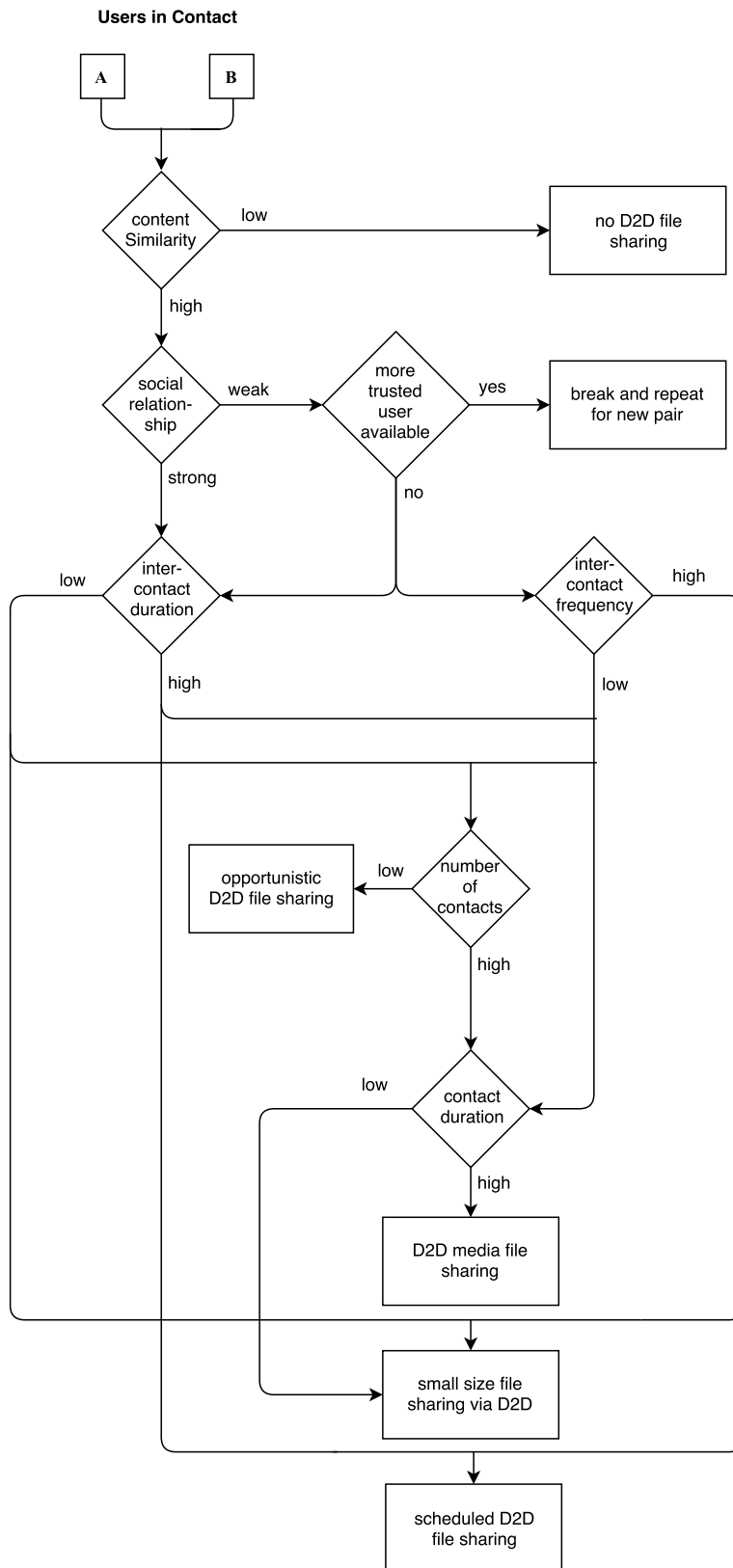


Figure 3.1: Decision-making system model for efficient D2D data sharing

After presenting our system model, we aim to determine the cutoff threshold between high and low for each decision milestone.

1. Content Similarity Milestone: we notice that the content similarity between devices ranges between 0.4 and 0.8 with some repetitions (see Equations 5.10 and 5.11). Thus, we decide that the threshold value should be the most repeated value (MRV) when the content similarity is analyzed for all users in the community.
2. Social Relationship Strength Milestone: the determination of the social relationship strength can be done using several metrics such as the presence of a connection between two users on online social networks, as well as their activities (e.g., the number of posts shared, the number of common pictures, etc.), or the history of the number of physical encounters that users have. In our work, we use Facebook friendships in the assessment of social relationship strength. Therefore, we use the MRV when the social relationship strength is analyzed for all users in the community.
3. Inter-Contact Duration Milestone: our study took place on a university campus, therefore the decision thresholds of our system model should conform to this environment, especially the inter-contact duration decision threshold. University schedules are relatively periodic with a period of one day: MWF schedules are similar, as well as TR schedules, with no classes on the weekends. As a result, we choose the following constraint: the inter-contact duration MRV should be less than 24 hours.
4. Inter-Contact Frequency Milestone: the inter-contact frequency decision should take into consideration the inter-contact duration MRV. If the latter is of the order of days, then the inter-contact frequency MRV should be less than four per week. If the inter-contact duration MRV is of the order of hours, then the inter-contact frequency MRV should be less than 15 per week.
5. Contacts Frequency Milestone: the decision threshold is chosen to be the MRV when the contacts frequency is analyzed for all users in the community.
6. Contact Duration Milestone: the choice of the decision threshold is dependent on the type of the file to be shared:
  - (a) Music files have a size that is usually less than 10 MB. This is a pretty small file size, so this type does not impact much the decision threshold, since 200 MB can be downloaded in one minute using Wi-Fi direct (see Section 5.1.2).

- (b) Document files are also considered small size files, since their size is usually of the order of few MB.
- (c) Image files size estimation: considering a JPEG encoded image file with a resolution of  $1280 \times 720$  pixels, the file size is estimated to be 389.9 KB. For a resolution of  $1920 \times 1080$  pixels, the file size is estimated to be 737.1 KB. Therefore, we consider images to be small size files, and have no impact on the decision threshold.
- (d) Video file size estimation: two types of videos can be shared between smartphones: short clips and movies. Assuming a MPEG2 video encoding with a data rate of 76.8 Mbps, a three-minute 20 fps video clip with a resolution of  $1920 \times 1080$  pixels has a size of 1.44 GB, which necessitates 4.8 minutes to download using Wi-Fi direct. As for an hour and a half 24 fps movie with a resolution of  $1920 \times 1080$  pixels, it has a size of 51.8 GB, which necessitates 2.8 hours to download using Wi-Fi direct.

Since music, document, and image files are considered to be too small to impact the decision threshold, the latter is determined according to the video file size. Two values were obtained for a potential contact duration threshold: 4.8 minutes when sharing short clips, and 2.8 hours when sharing movies. Since most users share short clips rather than movies, the threshold value is chosen to be 4.8 minutes.

However, battery consumption is a concern. According to [42], Wi-Fi direct peer discovery and services consume 10% of the Nexus phones' battery energy every 2.4 hours. Our recommendation would be to refrain from sharing files when the battery energy is less than or equal to 20 %, and sharing movies of large file size when the battery energy is greater than or equal to 60%. These margins are considered since Wi-Fi direct services are not the only ones running on a user's mobile device.

Table 3.2 presents a summary of the threshold decision values and constraints.

Table 3.2: System model decision threshold values and constraints

Milestone	Threshold Value/Constraint	
Content Similarity	Most Repeated Value (MRV)	
Social Relationship Strength	MRV	
Inter-Contact Duration (ICD)	MRV <24 hours	
Number of Inter-Contacts	$ICD \sim days$	MRV <4/week
	$ICD \sim hours$	MRV <15/week
Contact Duration	4.8 minutes	

## 3.2 Thesis Objectives

Following the proposed system model, information related to the three aspects discussed - mobility, content, and social relationships - needs to be analyzed in order to identify the most favorable circumstances for successful D2D data sharing. In other words, the locations, time intervals, and device pairs need to be identified for effective D2D media file sharing. As a consequence, the thesis has three objectives:

- A. **Conducting a study on AUB campus using an Android mobile application:** during this phase, participants will be recruited to participate in a study on AUB campus. This study has two components: data collection using an Android mobile application and an online survey. Participants are asked to install and run the application on their smartphones in order to log their location and neighbor discovery information and upload them to a database. They are also asked to answer an user interest online survey to collect data on their content interests, their smartphones' contents, and their social activity.
- B. **Mobility and content similarity modeling and analysis:** during this phase, pre-processing the data collected is done to avoid error in analysis. The data in the database is cleaned of any outliers and duplicates, and the users anonymized. Correlation between the survey responses and the data collected by the Android application is also done. Then, the data is analyzed to model mobility and content similarity between mobile devices. Modeling would involve identifying hub locations, plotting the probability distributions of the number of contacts between two devices, contact duration, inter-contact duration, as well as determining file popularity and file type popularity. The social relationship between users is also taken into consideration, adding an additional insight.
- C. **Assessing the potential for successful D2D data sharing:** with analysis done, it is now possible to identify the optimal set of conditions for efficient device-to-device data sharing according to the results obtained. Results and number figures obtained in the modeling and analysis stage of this work will provide insight regarding the efficiency of D2D data sharing in light of the conditions met. They will also enable us to determine the number values of the decision thresholds in our system model when applied to a university environment.

# Chapter 4

## Experimental Study

In order to collect the data necessary for mobility and content similarity analysis inherent to our work, we conduct our own study on AUB campus. This study consists of two phases: a mobility data collection phase and a content similarity data collection phase. Each phase was conducted different means; the first being an Android mobile application installed on participants' smartphones to collect mobility information such as GPS coordinates and neighbor discovery logs, and the second being an online survey customized to our needs focused on content similarity between smartphones and user interests.

### 4.1 IRB Process

Due to the nature of the data collected, especially when it comes to the mobile application, privacy and confidentiality concerns have to be addressed. Thus, an application was submitted to the Institutional Review Board (IRB) for approval before any data could be collected. The first form submitted is entitled *Application for Exemption from IRB Review* focused on the phase related to the Android application and was approved in mid December 2014. A consent form, found in Appendix B, was submitted along the IRB application, to be signed by each participant prior to starting the study.

To ensure the participants' privacy, it was ensured that the smartphone's content are not accessible to the Android application. Data collected about files stored on the device is limited to the file's name, extension, path, and last modification date. Therefore, the file's contents are not accessible. URL names of websites will be stored for analysis; however, the user's actions when visiting a certain website will not be tracked. Moreover, the user has the option to disable indexing files and browser history from the application settings. Furthermore, participants in this study will be anonymous and unidentifiable. Any identifying information collected about smartphone identifiers will be one-to-one mapped to arbitrary identifiers to remove any link to the participants' devices. Thus, no

personal data will be stored in the database that relates to the identity of the user.

As for confidentiality, users' records for this study will be kept confidential. Only the investigators will have access to personal information that relate to the participants; this information will not be stored, shared, or utilized in the study. Participants will remain anonymous.

The first IRB application was later amended in October 2015 to include the addition of an online survey to the study in order to address the content similarity data collection of our experiment. Note that the content similarity data collection was to be done using the Android application via the file indexing option. However, most participants chose to disable it on their smartphones. Therefore, no reliable content data was collected which necessitated the use of a survey. This amendment was approved in November 2015, provided the conditions set in the consent form (see Appendix C) are met. Specifically, the participants were asked to provide their names when answering the survey in order to correlate their responses with the data collected by the Android application on their smartphones. However, enforcing participant identification was not approved by the IRB, thus making it optional for the participant. In case, s/he mentioned his/her name, only the investigators will have access to this information. Participants' involvement will remain anonymous in any results and publications.

In what follows, a detailed discription of our experimental approach will be provided.

## 4.2 Mobility Data Collection

Mobility data was collected using an Android mobile application called *Context Aware Resource Management App*, or CARMA for short. It collects mobility and content data from smartphones and stores them in a database on a server. Subsection 4.2.1 presents a description of the CARMA application, while Subsection 4.2.2 details the experiment parameters and its approach.

### 4.2.1 CARMA Description

CARMA has a simple graphical user interface (GUI) consisting of four screens: Dashboard, Contacts, Neighbors, and Settings (see Figure 4.1). The Dashboard screen serves to activate/deactivate the CARMA background service and view the service's statistics. The Contacts screen enables the user to connect to his different social network accounts, thus permitting the storage of his contacts' information in the database. The Neighbors screen shows a list of the neighboring devices detected by the application. Finally, the last screen is the Settings screen where a multitude of parameters can be adjusted following user preferences.

The data collected by this application consists of:



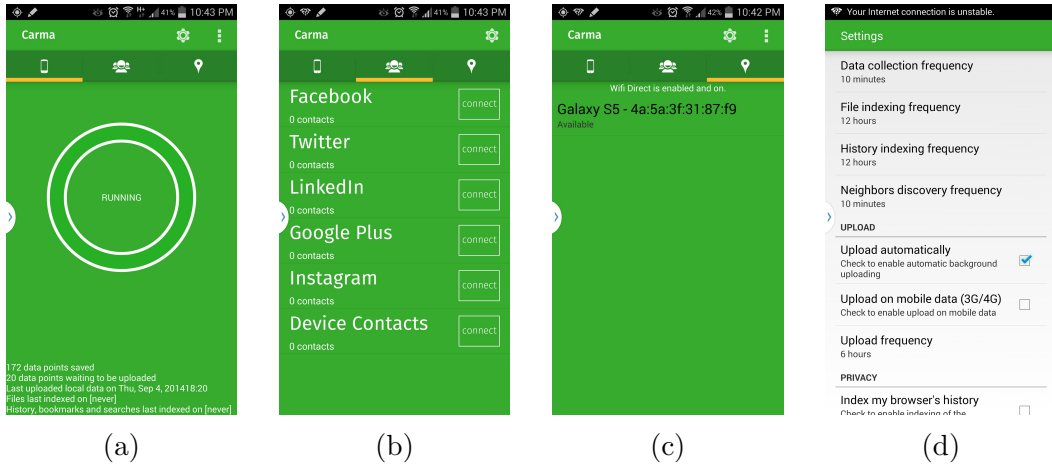


Figure 4.1: CARMA GUI. (a) Dashboard screen. (b) Contacts screen. (c) Neighbors screen. (d) Settings screen.

- Data consumption: number of bytes transmitted and received, and over which network (Wi-Fi or mobile networks).
- Battery status: percentage and charging status.
- What interface the device is connected to: Wi-Fi, mobile network or none.
- Device location: latitude, longitude, location provider, location accuracy.
- Neighboring devices: device identifier number, MAC address.
- Files stored in memory: name, absolute path, size, date last modified, type extension (this is optional and can be disabled by user). Note that the file indexing occurs only in one specific directory called "CARMA", where the user saves files that s/he is willing to share.

For CARMA to function properly, users need to turn on the GPS service and be connected to a Wi-Fi network at all times. The application works in the background, discovers neighboring devices via Wi-Fi direct, and uploads all collected data to the server using a Wi-Fi connection.

After data collection, the information will be uploaded to the server, either at regular intervals (specified by the settings) or manually (done periodically by the user). The information is then arranged in the database in tables for easy access and analysis. The data stored in two tables are essential to the analysis required for this thesis: the Data table and the Neighbors table. Snapshots of these two tables are provided in Figures 4.2 and 4.3 respectively.

id	localDbId	timeMs	wifiBytesR	wifiBytesT	mobileBytesR	mobileBytesT	batteryPercent	isCharging	connectedTo	latitude	longitude	locProvider	locAccuracy
1	1	1407154374900	0	0	11845	28992	100	1	Mobile	33.8588912	35.5814183	network	39
2	2	1407154434903	0	0	692	804	100	1	Mobile	33.8588912	35.5814183	network	39
3	3	1407155197185	0	0	614	3634	100	1	Mobile	33.8588826	35.5814203	network	34.5
4	4	1407155257186	1094	1094	1261	1242	100	1	Mobile	33.8588128	35.5814974	network	42
5	5	1407155317241	0	0	594	3619	100	1	Mobile	33.8580197	35.5817664	network	139.5
6	6	1407155377186	0	0	1999	11732	100	1	Mobile	33.85831	35.5816566	network	130.5
7	7	1407155437186	0	0	420	4445	100	1	Mobile	33.8588189	35.5814367	network	40.5
8	8	1407155497186	0	0	0	52	100	1	Mobile	33.8588189	35.5814367	network	40.5
9	9	1407155557186	0	0	947	6099	100	1	Mobile	33.85831	35.5816566	network	130.5
10	10	1407155617184	0	0	144	513	100	1	Mobile	33.8581423	35.5817209	network	114
11	11	1407155677185	880	880	15744	13131	100	1	Mobile	33.8589006	35.5814193	network	36

Figure 4.2: Snapshot of the Data table extracted from the database

id	localDbId	TimeMs	deviceName	deviceAddress	Latitude	Longitude	Accuracy	LocationProvider	model	uploadedTime	ipAddress
496	48	1409501325953	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
497	49	1409501871611	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
498	50	1409501909871	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
499	51	1409501911608	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
500	52	1409502469826	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
501	53	1409502472033	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
502	54	1409502511978	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
503	55	1409503069835	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
504	56	1409503074117	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
505	57	1409503113184	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...
506	58	1409503671526	Galaxy S5	4a:5a:3f:31:87:f9	33.7956421619	35.4839382444	16	gps	Samsung SM-N900	2014-09-01 08:...	109.110.10...

Figure 4.3: Snapshot of the Neighbors table extracted from the database

## 4.2.2 Data Collection

The mobility data collection using CARMA on Android devices was done in two stages: the first during the 2015 spring semester, and the second during the 2015 summer semester. During both stages, participants were only required to run the application while on AUB campus. Two data collection stages were undertaken to infer any seasonal changes in the trends of the parameters to be analyzed, assuming that the fall and spring semesters in a university environment have similar trends based on similar schedules followed by the students. Participants were recruited independently for each stage, although a fraction of them participated in both stages.

The first stage started on March 9, 2015 and ended on May 20, 2015, for a duration of 11 weeks. 37 participants were recruited, all part of the Electrical and Computer Engineering department. 23 of these participants were undergraduate students, 8 graduate students, and 5 PhD candidates. Data collection frequency was set to 10 minutes, resulting in 21,289 neighbor discovery data points and 354,528 Data table points collected and stored in our database.

The second stage started on June 11, 2015 and ended on July 6, 2015, for a duration of approximately 4 weeks. 13 participants were recruited for this stage, distributed among first and second year undergraduate ECE students,

as well as graduate students, PhD candidates, and post-doc. Neighbor discovery interval was set to 1 minute, while location information is logged every 2 minutes. Consequently, 34,920 neighbor discovery points and 326,539 Data table points are collected.

For both stages, data collected is uploaded automatically to the server every hour. Additionally, eight participants participated in both stages. Table 4.1 summarizes the experiment setup during the two stages.

Table 4.1: CARMA data collection settings during spring and summer 2015

<b>Settings</b>	<b>Spring</b>	<b>Summer</b>
<b>Area</b>	AUB campus	AUB campus
<b>Start date</b>	March 9	June 11
<b>End date</b>	May 20	July 6
<b>Trace duration (days)</b>	77	25
<b>Location data collection frequency</b>	10 minutes	2 minutes
<b>Neighbor discovery frequency</b>	10 minutes	1 minute
<b>Data points upload frequency</b>	1 hour	1 hour
<b># Participants recruited</b>	37	13
<b># Undergraduate participants</b>	23	10
<b># Graduate participants</b>	8	1
<b># PhD participants</b>	5	1
<b># Post-doc participants</b>	0	1
<b># Common participants</b>	8	
<b># Points in data table</b>	21,289	34,920
<b># Points in neighbors table</b>	354,528	326,539
<b># Participants correlated with survey</b>	8	4

### 4.3 Content Similarity Data Collection

Since most of the participants in the first stage of the experiment chose not to index their files, the content similarity data collection was done via an online survey that we designed. The survey is titled *Device-to-Device Media File Sharing - User Interest Survey*, and has 28 questions that focus on user interests and content stored on their smartphones, downloaded or shared, as well as on social network activity. Specifically, the questions were divided into five groups, each centering around a certain aspect of content similarity.

The first group of questions focus on the type of files stored on users' devices, as well as on user interests in the five categories of files mentioned: documents, images, videos, and movies/TV series. Group 1 questions are below:

1. Name
2. On average, how many media files per day do you store on your hand-held device (e.g., smartphone, tablet, etc.)?
3. What type of files do you store in your hand-held device?
4. For the document category, which holds the most interest to you, provided that you would store it on your hand-held device?
5. For the image category, which holds the most interest to you, provided that you would store it on your hand-held device?
6. For the video category, which holds the most interest to you, provided that you would store it on your hand-held device?
7. What are your favorite genres of music?
8. What are your favorite genres of movies/TV series?

The second group inquires about users' sources of media files. Group 2 questions are:

9. What source of music do you use the most on your hand-held device?
10. What source of videos do you use the most on your hand-held device?

Group 3 concentrates about the number of files stored and streamed on users' hand-held devices such as smartphones and tablets. Group 3 questions are:

11. How many music files do you have stored on your hand-held device?
12. How many document files (.pdf, .epub, .docx, .lit, .mobi) do you have stored on your hand-held device?
13. How many image files do you have stored on your hand-held device?
14. How many video files do you have stored on your hand-held device?
15. On average, how many music files do you stream per week on your hand-held device?
16. On average, how many video files do you stream per week on your hand-held device?

Group 4 serves to assess users' willingness to download and share their files via D2D communication. Group 4 questions are:

17. Given the opportunity to acquire files from nearby devices via device-to-device sharing, what kind of files are you willing to download and share using this technique?
18. Are you willing to share files with others via device-to-device technology without any incentives?
19. Are you willing to share files with others via device-to-device technology without any incentives?
20. In case your answer to the previous question is YES, what incentives would most appeal to you?

The final group focuses on users' social activity regarding four social networks: Facebook, Instagram, Twitter, and LinkedIn. Group 5 questions are:

21. Do you have a Facebook account?
22. If your answer to the previous question is YES, how many times do you access your Facebook account?
23. Do you have a Twitter account?
24. If your answer to the previous question is YES, how many times do you access your Facebook account?
25. Do you have an Instagram account?
26. If your answer to the previous question is YES, how many times do you access your Instagram account?
27. Do you have a LinkedIn account?
28. If your answer to the previous question is YES, how many times do you access your LinkedIn account?

The survey detailed questions can be found in Appendix C.

Based on the IRB's request, the request to fill the survey was issued to a big pool of students and not only our participants, since the survey was not part of the original study. Responders, whether they participated in the study or not, were asked to fill this survey only once. They were asked to mention their names - although not all of them complied with this request - in order to correlate the responses of those who participated in the first phase of the experiment with the data collected using the CARMA application. We received a total of

31 responses, with only 8 participants during the spring and 4 during summer mentioning their names for correlation purposes. The 31 total responses are not enough to get a thorough insight about user interests and device content where the pool of participants is more than a hundred students per class/year. However, compared to the subsets of people who participated in the CARMA study, this many responses are enough to draw our conclusions. As for the identifiable responses, those obtained let us draw some conclusions when correlating mobility and content similarity data, but are not enough for generalization.

# Chapter 5

## Analysis and Results

In this chapter, the data collected using the two experiment phases described in Chapter 4 is analyzed in order to derive conclusions about the efficiency of D2D file sharing considering the system model proposed. The mobility, content similarity, and social networks dimensions are studied individually, before being correlated to study their combined effect on D2D file sharing efficiency.

The mobility related parameters studied focus on the number of contacts, contact duration, inter-contact frequency and duration, and the identification of hub locations. As for the content similarity analysis, the survey responses will be analyzed to determine how this similarity will be measured. Finally, social network relationship will be studied using Facebook friend lists.

Note that all deductions and conclusions presented in the section apply to the university environment only.

### 5.1 Mobility Related Parameters

For the analysis of mobility related parameters, the data collected by the CARMA mobile application will be used. Specifically, the data in the Neighbors table is of interest.

#### 5.1.1 Number of Contacts

The number of contacts between two devices is a good measure to determine whether D2D file sharing is likely between the two. If two devices contact frequency ranges from never to only occasionally, then either is not a reliable source of media to the other, since it is not always reachable.

First, the normalized total number of contacts, aggregated over all users, is analyzed with respect to day and time. During the spring semester the trace has 11 weeks (77 days); while the data collection during the summer semester lasted a little over three weeks (25 days). With 37 and 13 participants recruited during

the spring and summer semester respectively, the effect of the number of users on the total number of contacts is investigated. Thus, we consider different sets of users of size ranging from 5 to 25 for the spring semester, and from 5 to 13 for the summer semester. We start with a set of 5 users, then incrementally increase this size by 5 additional users up to 25 or 13 corresponding with the appropriate semester. The total and average number of contacts with respect to day and time interval are plotted in Figures 5.1 through 5.4. As expected, the larger the set of users, the higher the number of contacts obtained. However, the average number of contacts per user is dependent on the selected set.

Comparing the results obtained between spring and summer, we notice similar trends. Contacts mostly occur on weekdays, with rare occurrences during weekends. Additionally, the trends following the days of the week are highly similar, with the highest number of contacts occurring on Tuesdays during both semesters. During the spring semester, the average number of contacts is relatively high from 6:00 to 18:00 hours. On the other hand, the average number of contacts during the summer is relatively high from 6:00 to 14:00 hours. This discrepancy can be justified by the difference in work day hours between the two semesters.

Moreover, on average, a set of 20 users during spring and a set of 13 users during summer present the highest number of contacts with respect to both day and time. Coupled with location coordinates, the set size can be used to determine hub locations, since they are defined as places where, for most of the time, the number of users found is greater than or equal to the set size value with the highest average number of contacts. Thus, looking at the results above, a location is considered a hub location if, for most of the time, the number of users found in this location is equal to 20 or 13 depending on the semester.

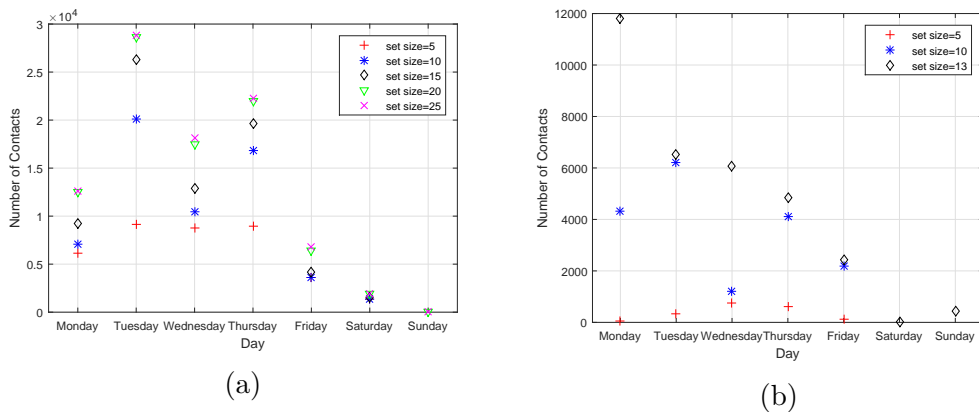
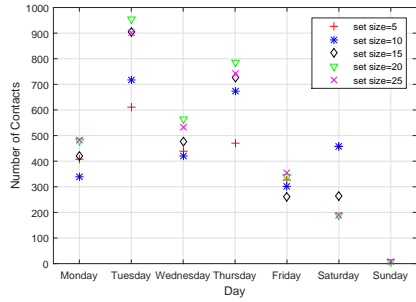
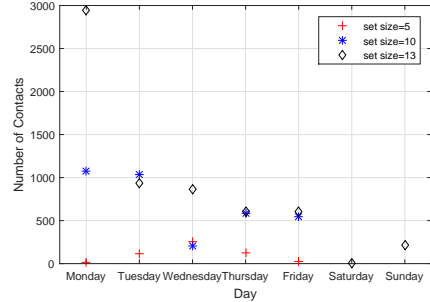


Figure 5.1: Total number of contacts in function of different sets of users with respect to day during (a) spring (b) summer



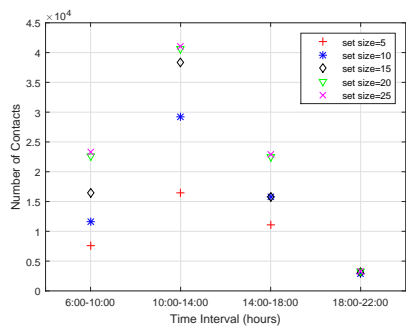


(a)

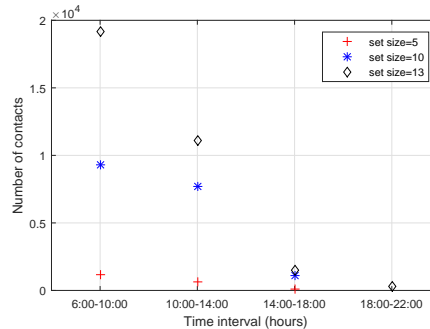


(b)

Figure 5.2: Average number of contacts in function of different sets of users with respect to day during (a) spring (b) summer

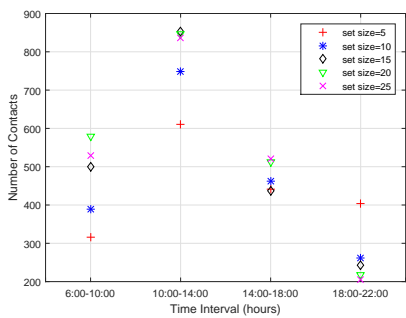


(a)

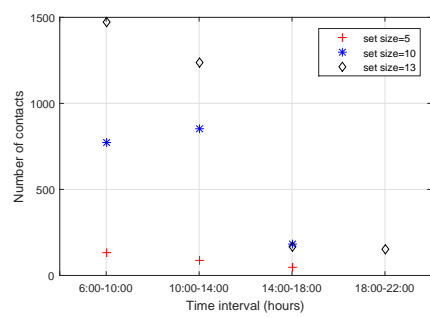


(b)

Figure 5.3: Total number of contacts in function of different sets of users with respect to time interval during (a) spring (b) summer



(a)



(b)

Figure 5.4: Average number of contacts in function of different sets of users with respect to time interval during (a) spring (b) summer

Next, the cumulative distribution function (CDF) of the normalized number of contacts per day, and that of the normalized number of contacts per time interval are derived empirically (see Figures 5.5 and 5.6). For weekends, the number of data points collected is not enough to determine the corresponding CDF, which is to be expected since no classes are held on AUB campus during weekends. Therefore, we consider the CDF for weekdays only. The empirical aggregated CDF for the normalized number of contacts per day, as well as per time interval, follows the Gamma distribution [43]

$$F(x) = \frac{\Gamma(\beta, x/\alpha)}{\Gamma(\beta)} \quad \text{for } x > 0 \quad (5.1)$$

where  $\alpha$  is the shape parameter,  $\beta$  the scale parameter, and  $\Gamma(\beta) = \int_0^\infty t^{\beta-1} e^{-t} dt$  the incomplete gamma function. The Gamma distribution parameters for both semester are presented in Table 5.1.

Referring to Figure 5.5, 80% of the number of contacts recorded are less than 61 contacts per day during the spring, and less than 2,118 contacts per day during the summer. Referring to Figure 5.6, 80% of of the number of contacts recorded are less than 15 contacts every four hours during the spring, and less than 125 contacts every four hours during the summer, provided that these four-hour time intervals fall during the workday hours. We notice that the distribution of the number of contacts is steeper in the summer in comparison with the spring. This means that, during the spring, the number of contacts is well distributed within the ranges  $[1, 61]$  or  $[1, 15]$  depending on the season. On the other hand, during the summer, the number of contacts is distributed unevenly, with values having mostly small or large values, with a few values being in the middle of the ranges  $[1, 2118]$  and  $[1, 125]$  depending on the semester.

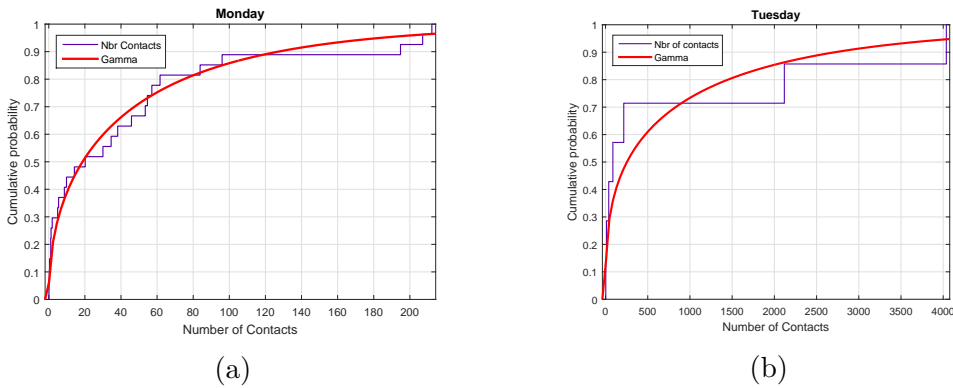


Figure 5.5: Empirical aggregated CDF for the normalized number of contacts per day for (a) spring (b) summer

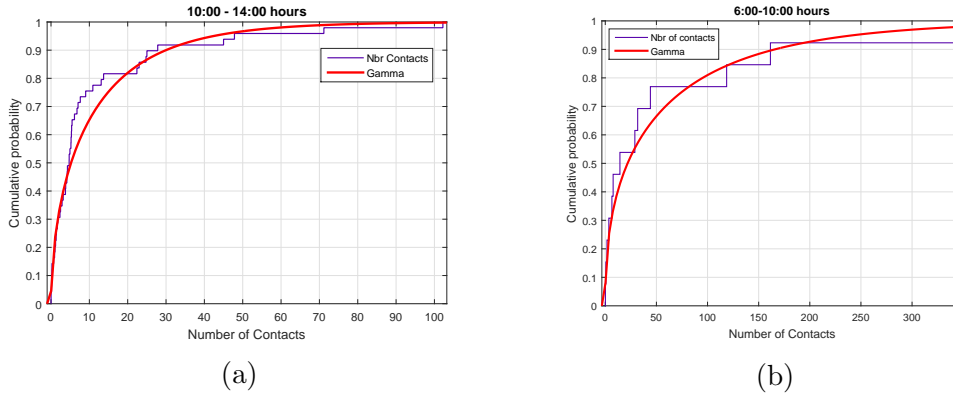


Figure 5.6: Empirical aggregated CDF for the normalized number of contacts per time interval for (a) spring (b) summer

Table 5.1: Gamma CDF paramters for the number of contacts

	Spring		Summer	
	per day	per time interval	per day	per time interval
<b>Shape (<math>\alpha</math>)</b>	0.445	0.517	0.322	0.407
<b>Scale (<math>\beta</math>)</b>	103.114	21.572	2889.52	144.675
<b>% fit</b>	90%	90%	76%	88%

Finally, we calculate the pair-wise number of contacts for both semesters (see Figure 5.7). We obtain an average of 155 contacts per pair, with a minimum of 1 contact per pair, and a maximum of 1867 contacts per pair during the spring. During this semester, five pairs show an exceptionally high number of contacts: (5;29), (12;17), (12;19), (17;19), and (21;23). As for the summer, the maximum pair-wise number of contacts is 1965, the minimum is 2, and the average is 265. Three pairs show an exceptionally high number of contacts: (3;34), (38;39), and (39;40). Additionally, user 3, who participated in both data collection phases, has a high number of contacts with its neighbors. Thus, user 3 is considered a reliable source in D2D media file sharing. Note that pairs who have a relatively low number of contacts can only take advantage of opportunistic D2D file sharing when dealing with each other.

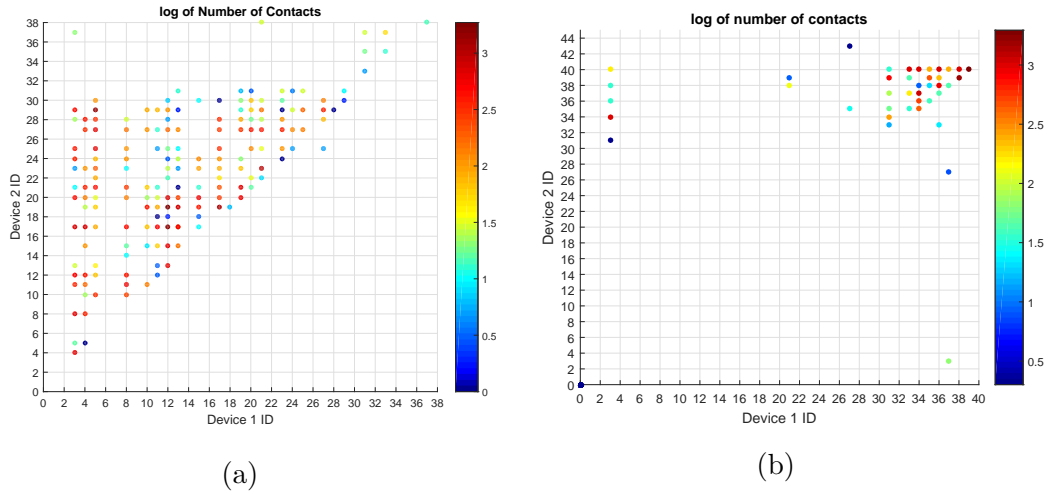


Figure 5.7: Pair-wise number of contacts for (a) spring (b) summer

We also consider the pair-wise number of contacts in function of day and time (see Figures 5.8 and 5.9). The trends coincide with those of the number of contacts. In this case, we have similar trends for both semesters. Looking at Figure 5.9b, we note that for the spring the semester, we have the highest average pair-wise number of contact between 10:00 and 14:00 hours, while this number is highest between 6:00 and 10:00 hours during the summer. This can be easily justified by the fact that AUB classes tend to start and end earlier in the summer, with students leaving campus mostly before noon.

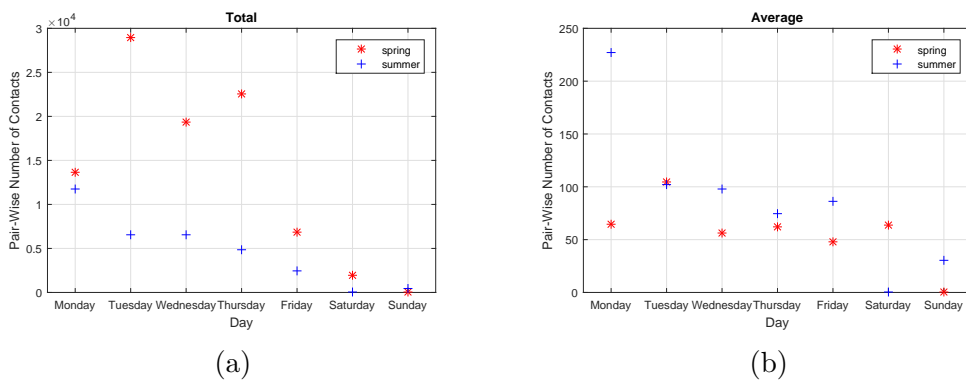


Figure 5.8: (a) Total and (b) average pair-wise number of contacts with respect to day

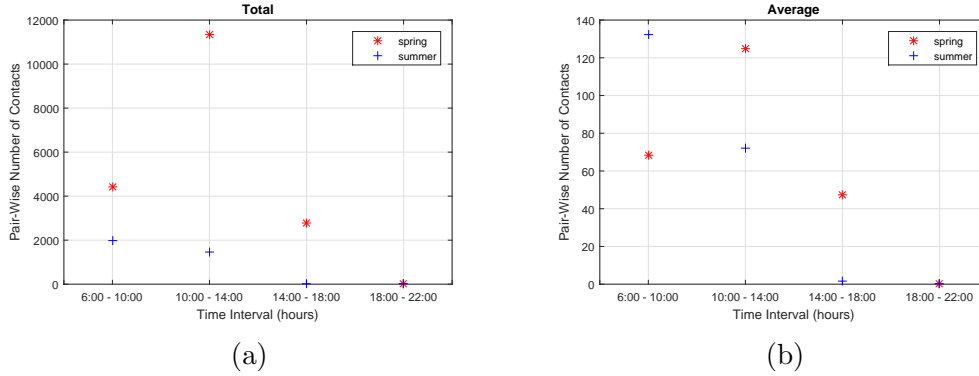


Figure 5.9: (a) Total and (b) average pair-wise number of contacts with respect to time interval

### 5.1.2 Contact Duration

The contact duration is also an important factor in determining the efficiency of D2D file sharing. If two devices are in contact for seconds at a time, then neither is not a reliable source of D2D sharing of media files of large size. Thus, the pair-wise contact duration is studied using the collected experimental data. For each pair of devices we calculate the contact duration for each contact established, and then we plot the aggregate CDF of the contact duration, along with some relevant statistics presented in Table 5.3. As for the aggregate CDF (see Figure 5.10), it follows a lognormal distribution [44]

$$F(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{e^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}}{t} dt \quad (5.2)$$

with  $\mu$  being the log location, and  $\sigma$  the log scale. The parameter values of the lognormal distribution for both spring and summer semesters are presented in Table 5.2. 80% of the pair-wise contact duration recorded is less than 130 minutes during the spring, and less than 15 minutes during the summer.

Table 5.2: Lognormal CDF parameters for the aggregate pair-wise contact duration

	Spring	Summer
$\mu$	3.94	1.54291
$\sigma$	0.95	1.25157
% fit	95%	97%

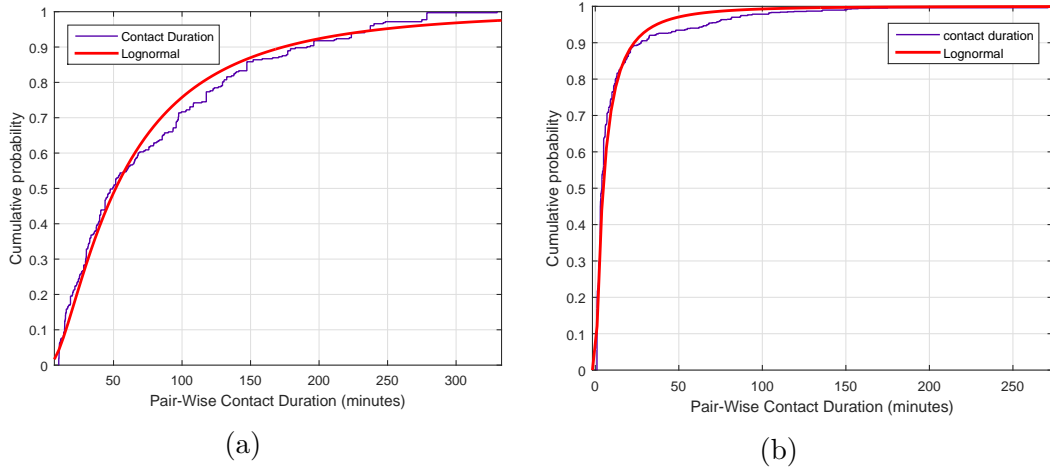


Figure 5.10: Pair-wise aggregate CDF of contact duration during (a) spring (b) summer

Table 5.3: Aggregate pair-wise contact duration statistics

	<b>Spring</b>	<b>Summer</b>
<b>Maximum</b>	5.5 hours	4.5 hours
<b>Minimum</b>	10.03 minutes	1 minute
<b>Average</b>	78.1 minutes	12.6 minutes
<b>Most repeated value (MRV)</b>	43.8 minutes	3.9 mins

On average, the contact duration for both semesters exceeds 10 minutes (see Table 5.3). Assuming an effective Wi-Fi direct data rate of 40 Mbps (see Table 3.1) and a contact duration of 10 minutes, the largest file that can be shared is 3 GB. For a minimum of 1 minute of contact, a user can download a file of maximum 300 MB, which is still a relatively considerable file size. Thus, D2D media file sharing should be possible if the other conditions are met.

The pair-wise aggregate contact duration CDF is further studied with respect to day and time (see Figures 5.11 and 5.12). In both cases, the CDF follows a lognormal distribution (5.2), which parameters are presented in Table 5.4. Statistics about pair-wise contact duration per day and time interval can be found in Tables 5.5 and 5.6.

Using Figure 5.11, it can be concluded that 80% of the pair-wise contact duration registered is less than 125 minutes per day during the spring, and less than 20 minutes during the summer. Figure 5.12 gives insight on the contact duration per time interval. 80% of the pair-wise contact duration registered is less than eight minutes every four hours during the spring, and less than 12 minutes

every four hours during the summer, provided that these four-hour time intervals occur during the workday hours. Due to the steeper curves observed during the summer, the contact duration is mostly very short or very long relatively, with a few occurrences presenting a contact duration somewhere in the middle of the range; whereas the contact duration is more evenly distributed during the spring semester.

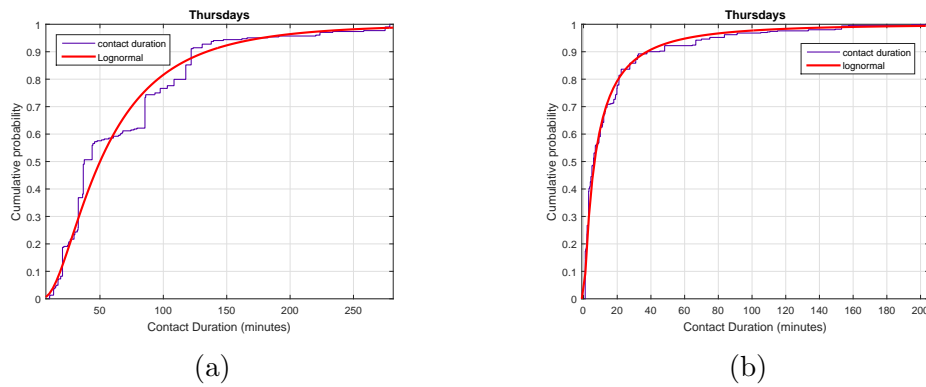


Figure 5.11: Pair-wise aggregate CDF of contact duration with respect to day for (a) spring (b) summer

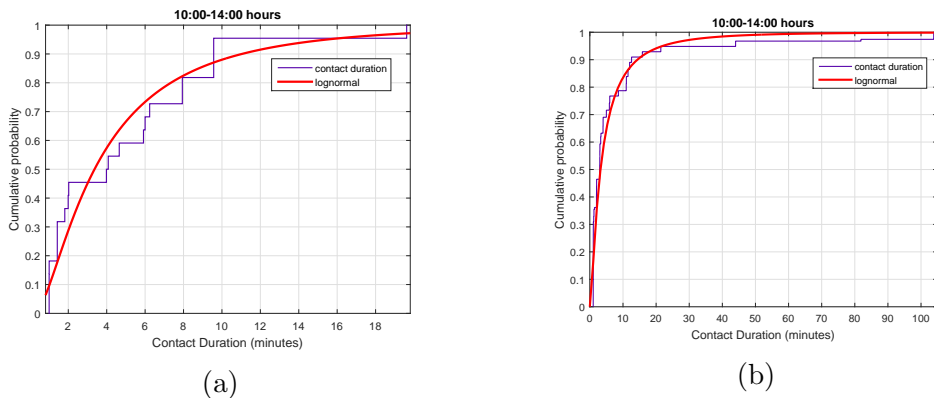


Figure 5.12: Pair-wise aggregate CDF of contact duration with respect to time interval for (a) spring (b) summer

Table 5.4: Lognormal CDF parameters for pair-wise contact duration with respect to day and time

	Spring		Summer	
	per day	per time interval	per day	per time interval
$\mu$	3.9132	1.87737	1.21328	1.16356
$\sigma$	0.769	1.36512	0.92659	1.17026
% fit	95%	95%	80%	90%

Table 5.5: Pair-wise contact duration per day statistics

		M	T	W	R	F	Sat	Sun
Max	Spr	4.11 h	4.43 h	7 h	4.64 h	5.5 h	1.7 h	0
	Sum	75.08 m	4.5 h	6.55 h	3.4 h	81 m	0	37.3 m
Min	Spr	10.5 m	10.03 m	10.23 m	10.03 m	10.25 m	10.24 m	0
	Sum	1	0	0	1	1	0	1.04 m
Mean	Spr	60.71 m	71.63 m	72.07 m	67.38 m	1.94 h	31.62 m	0
	Sum	11.74 m	16.7 m	12.15 m	17 m	9.34 m	0	13.63 m
MRV	Spr	97.26 m	31.36 m	18.95 m	32.91 m	15.68 m	10.23 m	0
	Sum	1 m	1.02 m	3 m	12 m	1 m	0	37.34m

Table 5.6: Pair-wise contact duration per time interval statistics

		6:00-10:00	10:00-14:00	14:00-18:00
Max	Spring	19.05 m	19.62 m	9.74 m
	Summer	75.95 m	1.73 h m	68 m
Min	Spring	2.18 m	1 m	1.03 m
	Summer	1 m	1 m	68 m
Mean	Spring	8.81 m	4.96 m	4.11 m
	Summer	11.39 m	8.13 m	68 m
MRV	Spring	2.18 m	1 m	1.36 m
	Summer	1.02 m	1 m	68 m



### 5.1.3 Inter-Contact Frequency and Duration

Inter-contact frequency and duration are good measures to determine whether D2D file sharing is possible. Depending on the combination of these two measures, can will be able to either share small size files via D2D links, opportunistically share files, use scheduled D2D file sharing (i.e., share files during specific time intervals on certain days), or exploit D2D file sharing fully.

The aggregated pair-wise inter-contact duration is plotted for both spring and summer in Figure 5.13. The CDF follows a Gamma distribution with a 98% fit in both cases. The CDF follows a distribution  $\Gamma \sim (0.245219, 6517.22)$  during spring, and  $\Gamma \sim (0.236987, 195.965)$  during summer. 80% of the pair-wise inter-contact duration recorded is less than 450 minutes, approximately 7.5 hours, during the spring, and less than five minutes during the summer.

During the spring, 479 inter-contacts occurred during the whole trace, with a minimum duration of 10 minutes, a maximum duration of approximately 58 days, and an average duration of approximately 26 hrs. During the summer, 1111 inter-contacts occurred during the whole trace, with a minimum duration of 1 minute, a maximum duration of approximately 5 days, and an average duration of 46.441 minutes.

The average case during the spring has a long inter-contact duration coupled with a small number of inter-contacts. On the whole, we conclude that during the spring D2D file sharing can be employed efficiently. It remains to be seen whether the users can exploit these links fully or only be able to download small size files, requiring an investigation of the contact duration. On the other hand, the average case during the summer has a short inter-contact duration coupled with a large number of inter-contacts. Thus, users can only share small size files via D2D links during this semester.

For a more involved outlook, the inter-contact frequency and duration per day and time interval are considered. For both semesters, no inter-contacts were

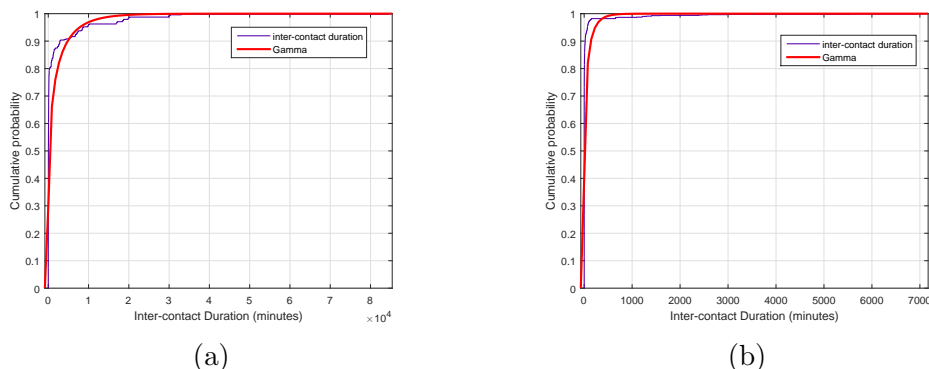


Figure 5.13: Aggregated pair-wise inter-contact duration CDF for (a) spring (b) summer

recorded on Wednesdays and on weekends, as well as after 18:00 hours. Statistics on inter-contact frequency and duration can be found in Tables 5.8 and 5.9.

Studying the inter-frequency frequency and duration per day (see table 5.8), for the average case, the spring semester has short pair-wise inter-contact duration coupled with a small number of inter-contacts. In this case, the question of D2D file sharing efficiency relies on the contact number and duration. As for the summer, the inter-contact duration is short and the number of inter-contacts is large. Therefore, users can only share small size files during the summer. The same trends and logic apply for the pair-wise inter-contact frequency and duration per time interval (see Table 5.9), especially during workday hours.

The aggregated pair-wise contact duration CDF per day and time interval, for both semesters, is also plotted (see Figures 5.14 and 5.15). The CDF parameter values can be found in Table 5.7.

From Figure 5.14 we infer that 80% the pair-wise inter-contact duration registered is less than three minutes per day during the spring, and less than one minute during the summer. Referring to Figure 5.15, 80% of the pair-wise inter-contact duration recorded is less than 4.6 minutes every four hours during the spring, and less than 2.7 minutes every four hours during the summer, provided these four-hour time intervals occur during the workday hours.

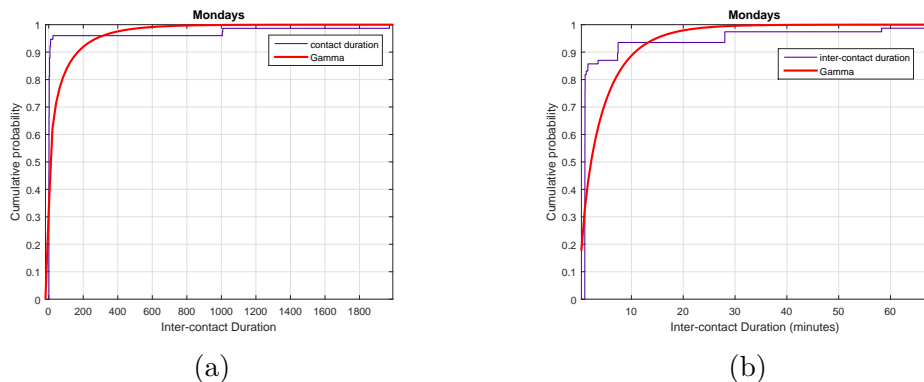


Figure 5.14: Aggregated pair-wise inter-contact duration CDF per day during (a) spring (b) summer

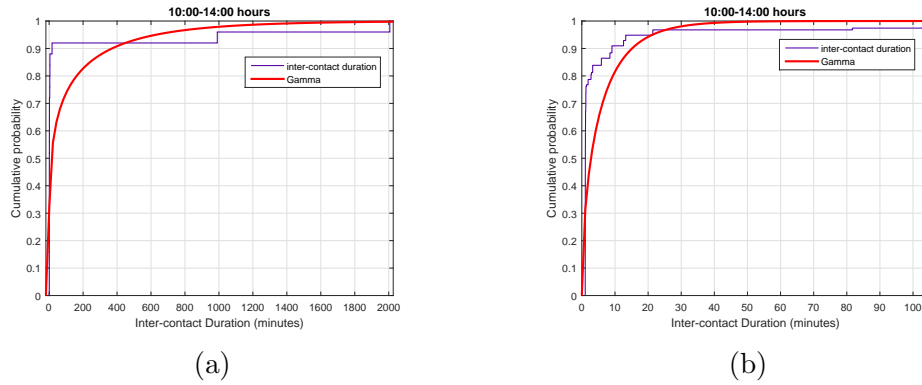


Figure 5.15: Aggregated pair-wise inter-contact duration CDF per time interval during (a) spring (b) summer

Table 5.7: Gamma CDF parameters for pair-wise inter-contact duration with respect to time and day

	Spring		Summer	
	per day	per time interval	per day	per time interval
$\alpha$	0.229093	0.20165	0.613853	0.53444
$\beta$	243.005	606.347	= 6.73697	10.6159
% fit	97%	95%	91%	95%

Table 5.8: Pair-wise inter-contact frequency and duration statistics per day

		<b>M</b>	<b>T</b>	<b>R</b>	<b>F</b>
<b>Max</b>	<b>Spring</b>	33 h	6 days	1.5 days	2.1 days
	<b>Summer</b>	1.11 h	6.8 days	7 days	1.35 h
<b>Min</b>	<b>Spring</b>	1 m	1 m	1 m	1.02 m
	<b>Summer</b>	1 m	1 m	1 m	1 m
<b>Mean</b>	<b>Spring</b>	55.67 m	2.75 h	1.45 h	4 h
	<b>Summer</b>	4.14 m	49.04 m	69.07 m	7 m
<b># inter-contacts</b>	<b>Spring</b>	75	142	331	122
	<b>Summer</b>	583	476	512	46

Table 5.9: Pair-wise inter-contact frequency and duration statistics per time interval

		<b>6:00-10:00</b>	<b>10:00-14:00</b>	<b>14:00-18:00</b>
<b>Max</b>	<b>Spring</b>	16.7 h	1.4 days	16.75 h
	<b>Summer</b>	75.3464 m	1.73 h	68 min
<b>Min</b>	<b>Spring</b>	1.1813 m	1.0026 m	1.0055 m
	<b>Summer</b>	1 m	1.0009 m	68 m
<b>Mean</b>	<b>Spring</b>	4.2 h	2.04 h	69.26 m
	<b>Summer</b>	7.58 m	5.67 m	68 m
<b># inter-contacts</b>	<b>Spring</b>	2	25	15
	<b>Summer</b>	56	155	1

#### 5.1.4 Hub Location Identification

Hub locations (HLs) are places where a large number of users spends a relatively long time. These locations are called hub since they provide great contact opportunities, thus chances to share all types of media files. The determination of HLs is dependent on two factors: the number of contacts occurring in a particular location, and the number of connected pairs at that location. Note that in Section 5.1.1 we derived the optimal number of users that in a specific location for it to be considered a hub location: 20 for spring and 13 for summer.

The aggregate numbers of contacts and connected pairs are evaluated for both semester, with the results featured in Table 5.10.

Table 5.10: Aggregate number of contacts and connected pairs statistics

	<b>Spring</b>		<b>Summer</b>	
	<b># contacts</b>	<b>#connected pairs</b>	<b># contacts</b>	<b>#connected pairs</b>
<b>Max</b>	4119	75	2922	30
<b>Min</b>	1	1	1	1
<b>Average</b>	40	25	50	6
<b>MRV</b>	6	2	6	2

Considering a range of 20 meters for Wi-Fi direct (refer to Table 3.1), we plot the aggregate number of contacts and the number of connected pairs on an AUB map for the spring and summer semesters (see Figures 5.16 and 5.17). The highest concentration of locations with the highest number of contacts and connected

pairs centers on the Engineering Zone, as defined by three buildings: Bechtel Building, Irany Oxy Engineering Complex, and Raymond Ghosn Building. Thus, in Figures 5.18 and 5.19, we plot the number of contacts and connected pairs during spring and summer in the Engineering Zone only. While these two figures offer a better insight about the most visited locations in the Engineering Zone with the most number of contacts and connected pairs, the range of values remains large enough that we further need to minimize the number of points plotted. Therefore, in Figures 5.20 and 5.21, we only mark the locations on the map with a high aggregated number of contacts (i.e., greater than or equal to 75) and a relatively high aggregated number of connected pairs (i.e. greater than or equal to 7).

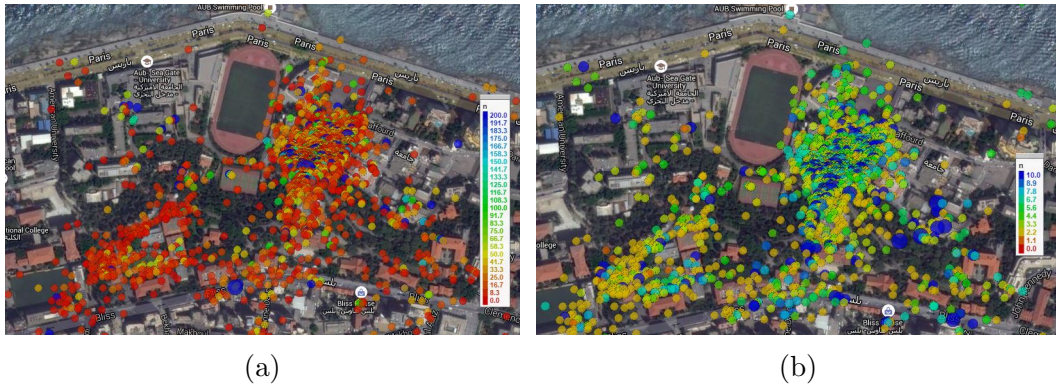
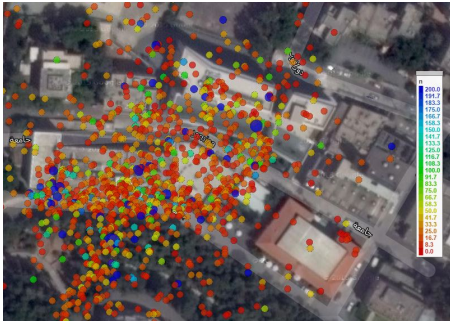


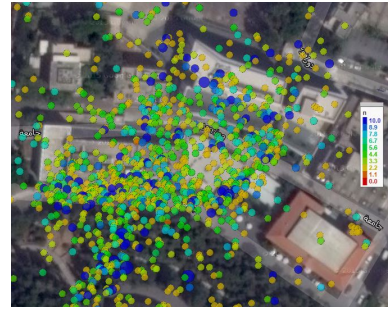
Figure 5.16: Aggregate number of (a) contacts [legend: red - green  $\rightarrow$   $\leq 50$ ; blue  $\rightarrow$   $\geq 125$ ](b) connected pairs [legend: red - green  $\rightarrow$   $\leq 3$ ; blue  $\rightarrow$   $\geq 7$ ] on AUB campus during spring



Figure 5.17: Aggregate number of (a) contacts [legend: red - green  $\rightarrow$   $\leq 50$ ; blue  $\rightarrow$   $\geq 125$ ](b) connected pairs [legend: red - green  $\rightarrow$   $\leq 10$ ; blue  $\rightarrow$   $\geq 17$ ] on AUB campus during summer

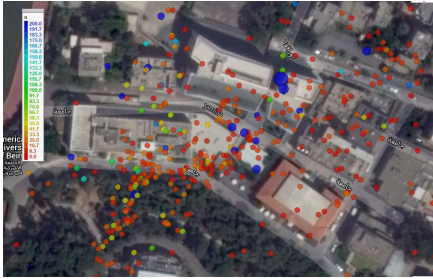


(a)



(b)

Figure 5.18: Aggregate number of (a) contacts [legend: red - green  $\rightarrow \leq 50$ ; blue  $\rightarrow \geq 125$ ] (b) connected pairs [legend: red - green  $\rightarrow \leq 10$ ; blue  $\rightarrow \geq 17$ ] in the Engineering Zone during spring

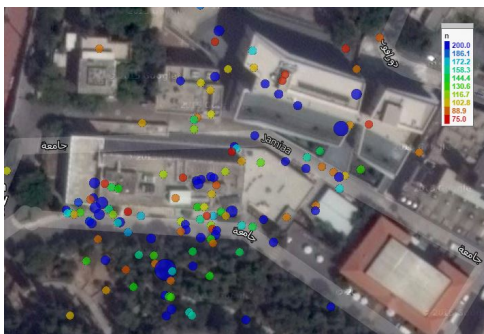


(a)

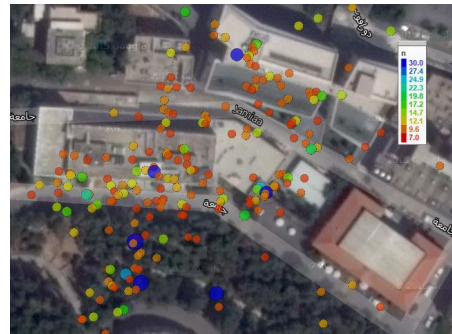


(b)

Figure 5.19: Aggregate number of (a) contacts [legend: red - green  $\rightarrow \leq 50$ ; blue  $\rightarrow \geq 125$ ] (b) connected pairs [legend: red - green  $\rightarrow \leq 50$ ; blue  $\rightarrow \geq 125$ ] in the Engineering Zone during summer



(a)



(b)

Figure 5.20: Large aggregate number of (a) contacts [legend: red - green  $\rightarrow \leq 110$ ; blue  $\rightarrow \geq 150$ ] (b) connected pairs [legend: red - green  $\rightarrow \leq 14$ ; blue  $\rightarrow \geq 22$ ] in the Engineering Zone during spring

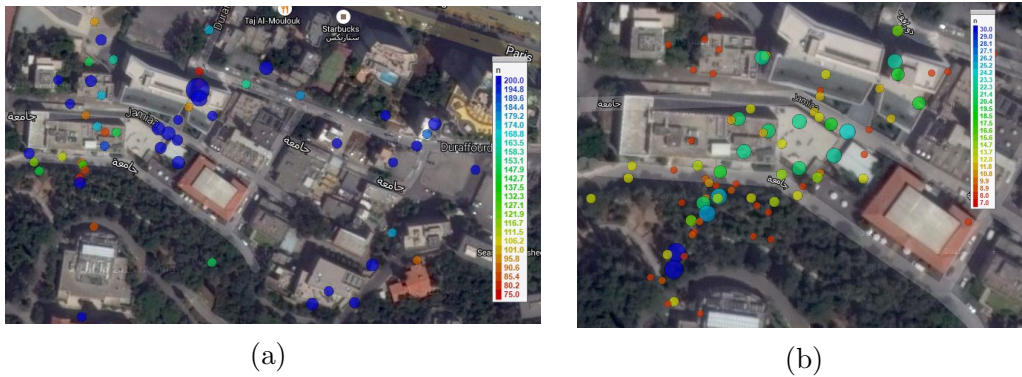


Figure 5.21: Large aggregate number of (a) contacts [legend: red - green  $\rightarrow \leq 110$ ; blue  $\rightarrow \geq 150$ ] (b) connected pairs [legend: red - green  $\rightarrow \leq 14$ ; blue  $\rightarrow \geq 22$ ] in the Engineering Zone during summer

During spring, several locations spanning the Bechtel Building and Irany Oxy engineering complex are considered hub locations. As for the summer, HLs are more concentrated to the East sides of these two buildings. To narrow these locations to specific rooms in the buildings, further smartphone sensing capabilities are needed.

## 5.2 Content Related Parameters

The content similarity analysis is solely based on the survey responses. Response statistics show that the majority of the responders store about four to six media files per day on their hand-held devices. The files they store are distributed to include all seven categories considered (personal documents, public documents, personal images, public images, personal videos, public videos, and music) with high percentages. Even though personal files occupy the majority of files stored by the participants, public media files also occupy a high percentage of user storage with the most popular file type being images, followed by music. Table 5.11 presents a summary of survey results relating to user interest.

When asked about their source of music, responders mostly use online sharing (download) websites such as Rapidgator, rather than streaming websites. However, the majority of responders prefer streaming videos rather than downloading and storing them on their devices. In general, most our responders stream less than 10 music and/or video files per week. As for the total number of music files stored on their devices, the majority of our responders have less than 50 music files stored on their device. The same applies to document and video files. As for image files, the majority of the responders (83.87%) have more than 300 files stored on their device.

Furthermore, when asked about D2D media file sharing, responders were willing to share and/or download all file categories, i.e. documents, images, videos,

and music; with music (90.32%) and videos (87.1%) being the most popular choices. As for their willingness to share files using D2D 61.29% were favorable when no incentives were offered. In the case where incentives are offered, this percentage increased to 96.77% favorable to D2D sharing of their files. The most popular incentives were extra internet (3G/4G) quota (90.32%), followed by free call minutes for each file exchanged (58.06%), and credits for online shopping (54.84%).

Additional and more detailed survey results statistics can be found in Appendix D.

Table 5.11: User interest survey results

Category	Popular Genres	Percentage
Documents	Lectures and course material	58.06%
	Articles of interest	51.61%
	Technical documents	45.16%
Images	Photos from the internet (e.g. 9gag, reddit)	67.74%
Videos	Short humor clips/ stand-up comedy	58.06%
	Music Videos	25.81%
Music	Blues and Jazz/Classical	48.39%
	Rock/R&B/Rap/Hip-Hop	45.16%
	Arabic music	41.94%
Movies and TV	Comedy	80.65%
	Action	54.84%
	Science Fiction/Fantasy/Paranormal	48.39%

For content similarity analysis, two parts of the survey are relevant here: the one related to user interest and the one related to the device content. A weight is calculated for each part by taking the number of common answers between any two participants divided by the corresponding total number of questions. The final weight is then calculated by adding the two obtained weights and dividing the outcome by 2. These weights are only calculated for the responders who participated in the CARMA data collection phase during both semesters. Responses are available for only 8 CARMA spring participants and 4 CARMA summer participants, since the rest didn't include their names in their responses.

The spring participants' IDs in question are: {4, 10, 12, 17, 19, 27, 29, 30}.

The summer participants' IDs in question are: {27, 34, 35, 36}.

First, the similarity matrices pertaining to device contents are derived, and designated as  $CONT_{spr}$  (5.4) and  $CONT_{sum}$  (5.5) for spring and summer respec-



tively. Each weight  $CONT(i, j)$  is derived as in Equation (5.3)

$$CONT(i, j) = \frac{\# \text{ common answers between users } i \text{ and } j}{\text{total } \# \text{ device content related questions}} \quad (5.3)$$

where  $i, j \in \{4, 10, 12, 17, 19, 27, 29, 30\}$  for  $CONT_{spr}$ ; and  $i, j \in \{27, 34, 35, 36\}$  for  $CONT_{sum}$ .

$$CONT_{spr} = \begin{bmatrix} 1 & 0.67 & 0.67 & 0.58 & 0.75 & 0.67 & 0.25 & 0.58 \\ 0.67 & 1 & 0.67 & 0.5 & 0.67 & 0.75 & 0.42 & 0.58 \\ 0.67 & 0.67 & 1 & 0.42 & 0.67 & 0.75 & 0.42 & 0.58 \\ 0.58 & 0.5 & 0.42 & 1 & 0.42 & 0.5 & 0.33 & 0.42 \\ 0.75 & 0.67 & 0.67 & 0.42 & 1 & 0.67 & 0.42 & 0.58 \\ 0.67 & 0.75 & 0.75 & 0.5 & 0.67 & 1 & 0.5 & 0.75 \\ 0.25 & 0.42 & 0.42 & 0.33 & 0.42 & 0.5 & 1 & 0.33 \\ 0.58 & 0.58 & 0.58 & 0.42 & 0.58 & 0.75 & 0.33 & 1 \end{bmatrix} \quad (5.4)$$

$$CONT_{sum} = \begin{bmatrix} 1 & 0.67 & 0.75 & 0.75 \\ 0.67 & 1 & 0.67 & 0.67 \\ 0.75 & 0.67 & 1 & 0.58 \\ 0.75 & 0.67 & 0.58 & 1 \end{bmatrix} \quad (5.5)$$

Second, the similarity matrices pertaining to user interests are derived, and designated as  $INT_{spr}$  (5.7) and  $INT_{sum}$  (5.8) for spring and summer respectively. Each weight  $INT(i, j)$  is calculated as in Equation (5.6)

$$CONT(i, j) = \frac{\# \text{ common answers between users } i \text{ and } j}{\text{total } \# \text{ user interest related questions}} \quad (5.6)$$

where  $i, j \in \{4, 10, 12, 17, 19, 27, 29, 30\}$  for  $INT_{spr}$ ; and  $i, j \in \{27, 34, 35, 36\}$  for  $INT_{sum}$ .

$$INT_{spr} = \begin{bmatrix} 1 & 0.45 & 0.7 & 0.62 & 0.66 & 0.55 & 0.62 & 0.48 \\ 0.45 & 1 & 0.62 & 0.55 & 0.66 & 0.62 & 0.55 & 0.7 \\ 0.7 & 0.62 & 1 & 0.72 & 0.76 & 0.79 & 0.72 & 0.59 \\ 0.62 & 0.55 & 0.72 & 1 & 0.55 & 0.59 & 0.72 & 0.59 \\ 0.66 & 0.66 & 0.76 & 0.55 & 1 & 0.76 & 0.55 & 0.55 \\ 0.55 & 0.62 & 0.79 & 0.59 & 0.79 & 1 & 0.52 & 0.66 \\ 0.62 & 0.55 & 0.72 & 0.72 & 0.55 & 0.52 & 1 & 0.52 \\ 0.48 & 0.7 & 0.59 & 0.59 & 0.55 & 0.66 & 0.52 & 1 \end{bmatrix} \quad (5.7)$$

$$INT_{sum} = \begin{bmatrix} 1 & 0.76 & 0.62 & 0.62 \\ 0.76 & 1 & 0.59 & 0.66 \\ 0.62 & 0.59 & 1 & 0.66 \\ 0.62 & 0.66 & 0.66 & 1 \end{bmatrix} \quad (5.8)$$

Finally, the similarity weights  $SIM(i, j)$  are calculated by adding the device content and user interests weights, and dividing the outcome by 2 (see Equation (5.9)). The similarity weights are found in the matrices designated  $SIM_{spr}$  (5.10) and  $SIM_{sum}$  (5.11) for spring and summer respectively.

$$SIM(i, j) = \frac{CONT(i, j) + INT(i, j)}{2} \quad (5.9)$$

$$SIM_{spr} = \begin{bmatrix} 1 & 0.558 & 0.678 & 0.602 & 0.703 & 0.609 & 0.435 & 0.533 \\ 0.558 & 1 & 0.644 & 0.526 & 0.661 & 0.686 & 0.484 & 0.637 \\ 0.678 & 0.644 & 1 & 0.57 & 0.713 & 0.772 & 0.57 & 0.585 \\ 0.602 & 0.526 & 0.57 & 1 & 0.484 & 0.543 & 0.529 & 0.501 \\ 0.703 & 0.661 & 0.713 & 0.484 & 1 & 0.713 & 0.484 & 0.568 \\ 0.609 & 0.686 & 0.772 & 0.543 & 0.713 & 1 & 0.509 & 0.703 \\ 0.435 & 0.484 & 0.57 & 0.529 & 0.484 & 0.509 & 1 & 0.425 \\ 0.533 & 0.637 & 0.585 & 0.501 & 0.568 & 0.703 & 0.425 & 1 \end{bmatrix} \quad (5.10)$$

$$SIM_{sum} = \begin{bmatrix} 1 & 0.713 & 0.685 & 0.685 \\ 0.713 & 1 & 0.626 & 0.661 \\ 0.685 & 0.626 & 1 & 0.619 \\ 0.685 & 0.661 & 0.619 & 1 \end{bmatrix} \quad (5.11)$$

The most repeated value in  $SIM_{spr}$  is 0.484 and that for  $SIM_{sum}$  is 0.685. These values thus constitute the decision threshold values for the content similarity milestone in our system model. However, these values apply to our scenario only and cannot be considered as a general measure since the set of identifiable responses is too small to be representative of a university setting. Additionally, the similarity weights are higher for the summer participants compared to the spring similarity weights. This can be due to the fact that the set of participants in the summer was too small to have a mix of students of different classes and/or who similar interests.

### 5.3 Social Networks Relationships

The social relationship strength is analyzed using our participants' Facebook friend lists. In Figures 5.22 and 5.23, we graph the Facebook friendship relationship for the users who participated in the CARMA data collection during spring and summer respectively. The vertices represent the participants IDs, and the presence of an edge means that the two participants are Facebook friends. Participants with no Facebook friendships are not visualized in these graphs for simplicity. The vertices are arranged according to their degree of centrality, with the node with the highest degree of centrality being in the center and that with the lowest degree of centrality being on the periphery. The degree of centrality quantifies how many ties a node has in a network. 18 out of 37 participants

during the spring and 7 out of 13 participants during the summer have Facebook friendships.

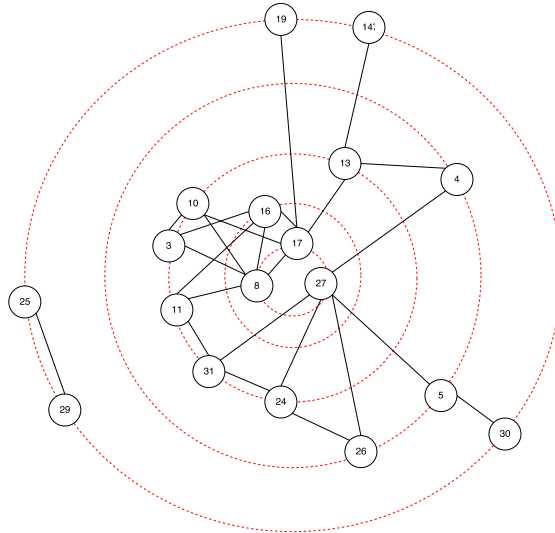


Figure 5.22: Facebook friendships of spring participants

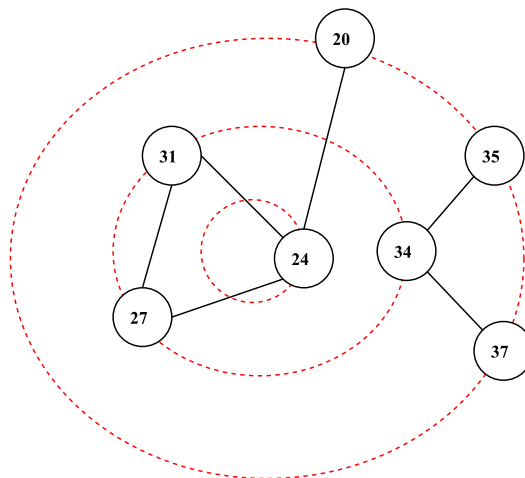


Figure 5.23: Facebook friendships of summer participants

As mentioned in Section 5.2, only a portion of the CARMA participants mentioned their names in their survey responses. The spring participants' IDs in question are  $\{4, 10, 12, 17, 19, 27, 29, 30\}$ ; and the summer participants' IDs in question are  $\{27, 34, 35, 36\}$ . Looking at the graphs above, only pairs  $(4;27)$ ,  $(10;17)$  and  $(17;19)$  (spring), and participants 34 and 35 (summer) are Facebook friends. In order to correlate mobility and content similarity results with participants' social relationships, the number of mutual friends is considered to assess

the social relationship strength between any two participants. The number of mutual Facebook friends between participants is presented in the matrices  $MUT_{spr}$  (5.12) and  $MUT_{sum}$  (5.13) for spring and summer participants respectively. The number of mutual friends compared to oneself is considered to be zero for graph construction simplicity.

$$MUT_{spr} = \begin{bmatrix} 0 & 0 & 0 & 11 & 11 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 & 2 & 14 & 7 & 9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 11 & 9 & 0 & 0 & 16 & 1 & 0 & 0 \\ 11 & 2 & 0 & 16 & 0 & 1 & 0 & 0 \\ 0 & 14 & 0 & 1 & 1 & 0 & 7 & 19 \\ 0 & 7 & 0 & 0 & 0 & 7 & 0 & 7 \\ 0 & 9 & 0 & 0 & 0 & 19 & 7 & 0 \end{bmatrix} \quad (5.12)$$

$$MUT_{sum} = \begin{bmatrix} 0 & 1 & 3 & 14 \\ 1 & 0 & 9 & 2 \\ 3 & 9 & 0 & 3 \\ 14 & 2 & 3 & 0 \end{bmatrix} \quad (5.13)$$

The graphs illustrating the adjacency matrices in (5.12) and (5.13) are presented in Figures 5.24 and 5.25. The edges have different thicknesses illustrating the strength of the social relationship between any two participants according to the number of their mutual friends. The vertices are arranged by degree of centrality as before. In this instance the edges are weighted according to (5.12) and (5.13). Therefore, the degree of centrality is calculated as the sum of weights of outbound edges from a certain node to all adjacent nodes.

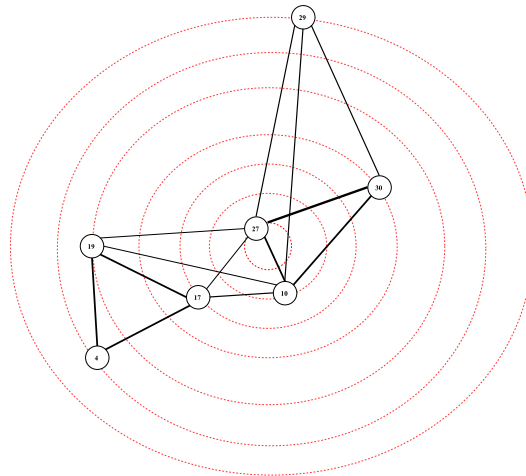


Figure 5.24: Facebook social relationship strength between CARMA and survey spring participants

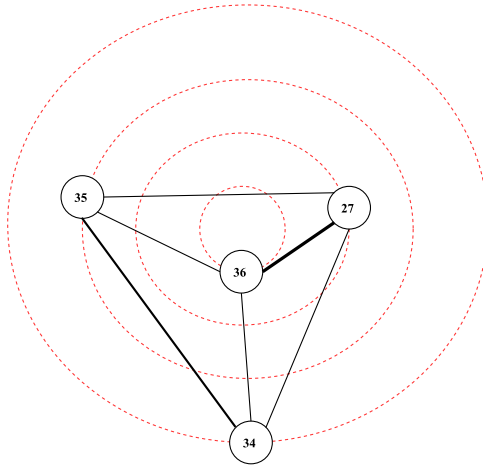


Figure 5.25: Facebook social relationship strength between CARMA and survey summer participants

## 5.4 Correlation Between Mobility, Content, and Social Networks

According to the system model proposed, three factors need to be satisfied apart from close user proximity: a high content similarity, a relatively strong social relationship, and favorable mobility patterns as discussed in Section 5.1. In this section, the social relationship strength, the content similarity, and the number of pair-wise contacts are correlated to identify favorable quantifiable conditions for successful D2D file sharing.

First, the pair-wise number of contacts is correlated with the Facebook network friendships. In other words, Figures 5.7a and 5.22, and 5.7b and 5.23 are correlated. Referring to Section 5.1.1, the average pair-wise number of contacts is 155 in the spring and 265 in the summer. When a Facebook friendship exists between two users, 66% of the time the pair-wise number of contacts exceeds 155 in the spring and 265 in the summer. When no Facebook friendship exists between two users, 29% of the time during the spring the pair-wise number of contacts exceeds 155, and 15.5% of the time during the summer the pair-wise number of contacts exceeds 265. Therefore, when people are connected on online social networks, they tend to be in contact with each other much more often than unconnected people.

The pair-wise number of contacts is then correlated with the number of mutual friends between a user pair. In other words, Figures 5.7a and 5.24, and 5.7b and 5.25 are correlated. When a user pair have mutual Facebook friends, 50% of the time the pair-wise number of contacts exceeds 155 in the spring and 265 in the summer. When no mutual friends exist between two users, 32.34% of the time during the spring the pair-wise number of contacts exceeds 155, and 15.71% of

the time during the summer the pair-wise number of contacts exceeds 265. Thus, even when people aren't connected on online social networks, they tend to be in contact with each other much more often when they have mutual connections on online social networks.

Second, the content similarity and the social relationship strength are correlated. Specifically, matrices (5.10) and (5.12), and (5.11) and (5.13) are correlated. To visualize this correlation, graphs composed of two tiers are plotted in Figures 5.26a and 5.26b. The tier on the top represents the social relationship strength between users as stated in the *MUT* matrices, where the edges' thickness reflects the strength of the relationship. The tier on the bottom represents the content similarity between two users' devices as stated in the *SIM* matrices. The thicker the edge connecting two nodes, the higher the content similarity between the two. Even though it doesn't apply to all nodes, there exists a certain correlation between these two measures. In general, when two people have a strong social relationship, they tend to have a higher percentage of similar content stored on their devices. This is somewhat justified, since relationships on online social networks mostly fall into two categories: professional acquaintances and connections, and friends and family. The correlation between social relationship strength and content similarity is then dependent on the type of connection and on the social network nature. For example, Facebook targets friends and family more than professional acquaintances. Thus, a stronger correlation between these two measures is expected between users who are family and friends rather than professional acquaintances. The inverse is true for a social network such as LinkedIn, which targets professional acquaintances and connections.

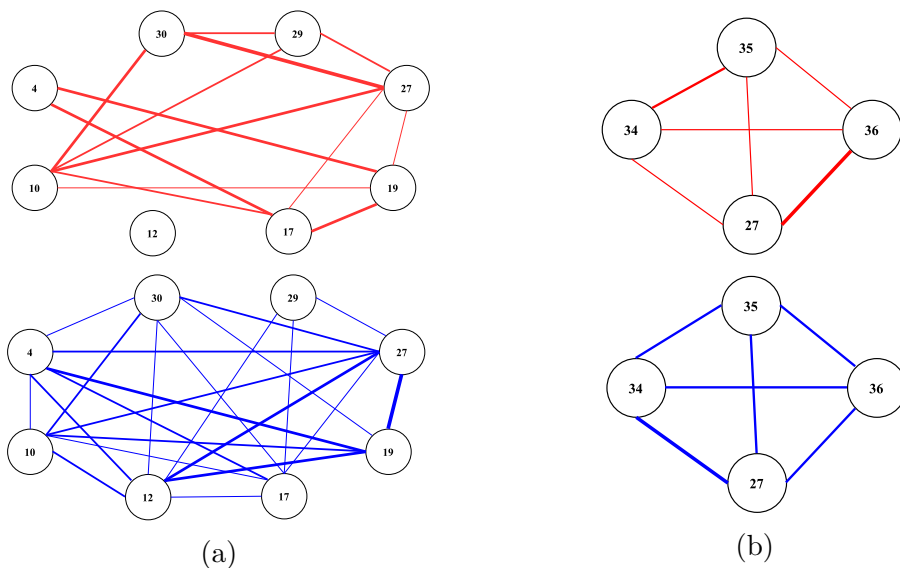


Figure 5.26: Correlation between social relationship strength (top) and content similarity (bottom) during (a) spring (b) summer

Finally, the number of contacts, the content similarity, and the social relationship strength are correlated in Figure 5.27. Looking at this figure, for each user pair one can roughly determine whether they have the potential to successfully share files via D2D links. When the edges connecting a pair of users are thickest on the three levels simultaneously, a D2D link has the highest chance of being successful while keeping in mind the constraints pertaining to the contact duration, the inter-contact duration, and the inter-contact frequency. Assuming the users are in a favorable situation concerning these three constraints, pairs (4;17), (4;19), and (19;27) in the spring, and pair (34;35) in the summer have the highest chance of success in D2D data sharing.

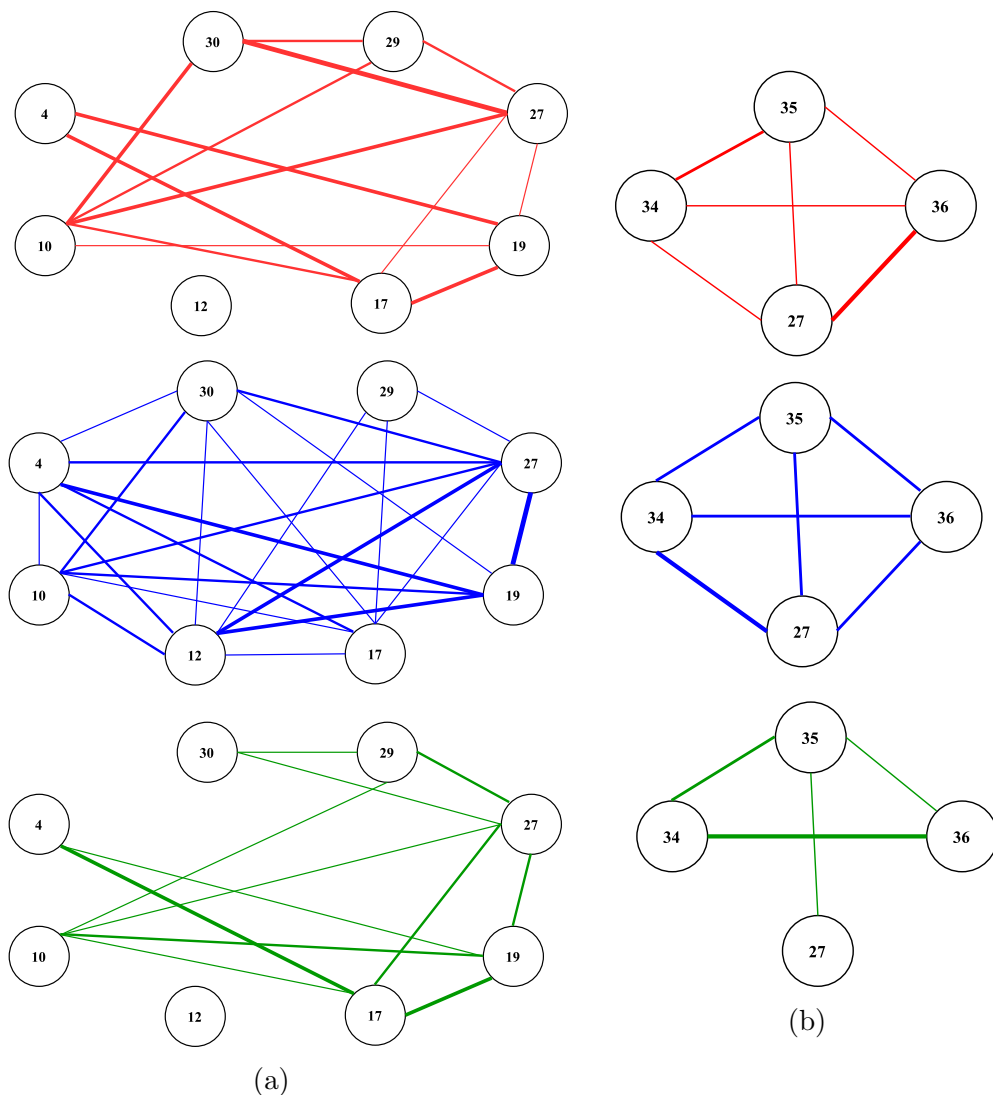


Figure 5.27: Correlation between social relationship strength (top), content similarity (middle), and number of contacts (bottom) during (a) spring (b) summer

# Chapter 6

## Conclusion and Future Work

With the high traffic demands, network offloading becomes necessary with device-to-device communications as an attractive solution, especially with the widespread use of smartphones providing mobility, connectivity, and user friendly interfaces to users. In this work, we model mobility and content similarity between smartphones using an experimental study focused on wireless D2D content sharing applications. In this study, engineering students at AUB were recruited to participate in a mobility data collection phase using the Android application CARMA, and a content data collection phase using a customized users interests survey. The study was done during two different semesters, spring and summer 2015, to account for seasonal trend changes.

Before the trace data was analyzed, the conditions necessary for successful D2D file sharing were determined and discussed. A system model is then derived, reflecting the efficiency of D2D file sharing when different conditions are met.

First, mobility parameters were analyzed, including the number of contacts between devices, the contact duration, the inter-contact frequency and duration, as well as the identification of hub locations. The number of contacts and the inter-contact duration proved to have a cumulative distribution function following the Gamma distribution; while the contact duration has a CDF following the lognormal distribution. The results reflected specific trends with respect to days of the week and specific time intervals.

Second, statistics were derived using the users interests survey responses. For users who included their names with their survey responses, similarity scores were derived using two types of questions: questions relating to their devices' contents and questions relating to their interests in music, images, videos, documents, and movies. For each question type the score was calculated, then the two scores are added and the outcome divided by two to obtain to total similarity score.

Finally, social relationship ties were investigated and mapped to graphs using Facebook's friendships and mutual friendships. Afterwards, the social relationships were correlated with the number of contacts, then with the contact similarity, before correlating the three measures simultaneously. A strong corre-



lation between the number of contacts and the social relationship strength was discovered. As for the correlation between the content similarity and the social relationship strength, a strong correlation exists between some users while the opposite is applicable to others. This can be rationalized by the social tie's nature coupled with the online social network target audience type. Furthermore, by correlating the number of contacts, the content similarity, and the social relationship strength, one can determine user pairs with potential to successful D2D file sharing.

As future work, further smartphone sensing capabilities can be used to better determine hub locations, especially in indoor situations. GPS coordinates alone are not enough to isolate specific rooms in buildings in order to determine hub locations accurately, as we have seen in the course of our research.

Also, our study was confined to a university environment. The conclusions derived in the course of our work were specific to a university environment. Moreover, our study involved a small number of participants, especially when considering a university environment. Therefore, conclusions were also dependent on the set of participants chosen. As a result, a more large scale study in terms of participants and duration is needed for a more precise generalization concerning the university environment. Additionally, studies in different environments, such as company offices and public transport environments, need to be conducted in order to ultimately generalize our proposed decision making system model with the right decision threshold values and constraints.

Future research can also focus on a more in depth study of the D2D file sharing effect on battery life. Energy efficiency is a big concern, especially for mobile phones and hand-held devices in general. A study of how D2D file sharing affects battery drainage is necessary, not just using Wi-Fi direct, but other wireless technologies as well such as Bluetooth and NFC. Additionally, D2D peer discovery is also an energy consuming process based on beaconing. Therefore, a mechanism needs to be designed to ensure energy efficient peer discovery.

# Appendix A

## Abbreviations

#	Number
AUB	American University of Beirut
BS	Base Station
CARMA	Context Aware Resource Management App
CDF	Cumulative Distribution Function
D2D	Device-to-Device
F	Friday
fps	Frames Per Second
GB	Gigabyte
GPS	Global Positioning System
GUI	Graphical User Interface
HBP	Hierarchical BiPartite
HL	Hub Location
IRB	Institutional Review Board
M	Monday
Max	Maximum
MB	Megabyte
Min	Minimum
MIT	Massachusetts Institute of Technology
MP	Meaningful Place
MRV	Most Repeated Value
MWF	Monday, Tuesday, & Wednesday
Nbr	Number
PRF	Performance-Related Factor
R	Thursday
Sat	Saturday
SMMC	Socially aware Mobile Multimedia Community-based approach
SoCast	Social-aware video multiCast
Spr	Spring
Sum	Summer

Sun	Sunday
T	Tuesday
TR	Tuesday & Thursday
UGC	User-Generated Content
W	Wednesday

# Appendix B

## IRB Informed Consent Form

American University of Beirut  
Faculty of Engineering and Architecture  
Electrical and Computer Engineering Department  
Informed Consent for Non-Medical Research  
**Smartphone Sensing Study for Mobile to Mobile Content Sharing  
and Cooperation**

You are invited to participate in a research study conducted by Prof. Zaher Dawy and his graduate student Miss Lynn Aoude at AUB, because your daily smartphone usage and social interaction with your friends make you a perfect fit for the study. Your participation is voluntary. You should read the information below, and ask questions about anything you do not understand, before deciding whether to participate. Please take as much time as you need to read the consent form. You may also decide to discuss participation with your family or friends. If you decide to participate, you will be asked to sign this form. You will be given a copy of this form.

### **PURPOSE OF THE STUDY**

In this study we seek to model the different interactions that occur among smartphones. In particular, our objective is to model mobility information and content similarity between smartphones. The developed Context Aware Resource Management App (CARMA) Android application will be used during this experiment for data collection

### **STUDY PROCEDURES**

If you volunteer to participate in this study, you can either use your own smartphone or borrow a smartphone from the investigators with a preloaded application that will be used to collect data from the device for analysis and modeling purposes. The application is called Context Aware Resource Management App

(CARMA). The collected data includes the lists of neighboring devices, location information, data consumption, battery status, names of stored files (optional), and browser history (optional). After data collection, the information will be uploaded to a server and arranged in a database for easy access and analysis. The application only requires a Wi-Fi connection and enabling the GPS settings. There are no costs incurred to participate in this study irrespective if you use your own device or borrow a device from the investigators. The experiment will have duration from six to eight weeks.

### **POTENTIAL RISKS AND DISCOMFORTS**

The research study entails no risk or discomfort for you. The data collection process will take place in the background and will not interfere with your daily phone usage and activities.

### **POTENTIAL BENEFITS TO PARTICIPANTS AND/OR TO SOCIETY**

This project does not have any direct personal benefit to the participant. As this is a research study, the benefits are contingent upon the obtained results. Anticipated long-term benefits include a better management of the network resources that will take into account the smartphone usage behavior. This will also lead to better quality of experience for the smartphone users in general.

### **PAYMENT/COMPENSATION FOR PARTICIPATION**

You will not be paid for participating in this research study.

### **PRIVACY**

The investigators will not be able to access the contents of your smartphone. Data collected about files stored on the device is limited to the files name, extension, path, and last modification date. Therefore, the files contents are not accessible to investigators. URL names of websites will be stored for analysis; however, your actions when visiting a certain website will not be tracked. Moreover, you have the option to disable indexing files and browser history from the application settings.

Participants will be anonymous and unidentifiable in this study. The identifying information collected will be the devices MAC address, its IP address, and the unique ID that the app generates for the device; these addresses will be one-to-one mapped to arbitrary identifiers to remove any link to the participants devices. So, no personal data will be stored in the database that relates to the identity of the user.

### **CONFIDENTIALITY**

We will keep your records for this study confidential. Only the investigators will

have access to personal information that relate to the participants; this information will not be stored, shared, or utilized in the study. Participants will remain anonymous; only device ID, and IP and MAC addresses will be collected and one-to-one mapped to arbitrary identifiers.

After data collection, the information will be uploaded to a server and arranged in a database for easy access and analysis. The data will be kept for the duration of this research study at minimum as it can be utilized for other future research studies.

### **PARTICIPATION AND WITHDRAWAL**

Your participation is voluntary. Your refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled. You may withdraw your consent at any time and discontinue participation without penalty. You are not waiving any legal claims, rights or remedies because of your participation in this research study. Your participation can also be terminated by the investigator without regard to your consent if the test devices are used for any activity banned by the American University of Beirut. The device usage should abide by the university regulations at all times.

### **INVESTIGATORS CONTACT INFORMATION**

If you have any questions or concerns about the research, please feel free to contact

Principal Investigator: Prof. Zaher Dawy [zd03@aub.edu.lb]

Co-Investigator: Miss Lynn Aoude [lwa05@aub.edu.lb]

### **RIGHTS OF RESEARCH PARTICIPANT IRB CONTACT INFORMATION**

If you have questions, concerns, or complaints about your rights as a research participant or the research in general and are unable to contact the research team, or if you want to talk to someone independent of the research team, please contact the Institutional Review Board. Tel: 01350000-5445 or irb@aub.edu.lb

### **SIGNATURE OF RESEARCH PARTICIPANT**

I have read the information provided above. I have been given a chance to ask questions. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

Name of Participant:

Signature of Participant:

Date:

**SIGNATURE OF INVESTIGATOR**

I have explained the research to the participant and answered all of his/her questions. I believe that he/she understands the information described in this document and freely consents to participate.

Name of Person Obtaining Consent:

Signature of Person Obtaining Consent:

Date:

# Appendix C

## Survey Questions

On the following page start the survey questions, preceded by an informed consent form specific to the survey, of which the participant can keep a copy.



## Device-to-Device Media File Sharing – User Interest Survey

You are asked to participate a research study in light of your participation in the previous CARMA experiment. You should read the information below, and ask questions about anything you do not understand, before deciding whether to participate. You may also decide to discuss participation with your family or friends. You can keep a copy of this form. Your participation is voluntary. Your refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled. You may withdraw your consent at any time and discontinue participation without penalty. You are not waiving any legal claims, rights or remedies because of your participation in this research study.

This survey aims to study user interests regarding downloading and sharing content using hand-held mobile devices (smartphones/tablets). Results concluded from the survey's responses will be used in the course of the research, and correlated to the results of the previous CARMA experiment. The purpose of this study is to determine how much user content interests and similarity is beneficial in increasing the efficiency of device-to-device data sharing.

Identification is necessary to connect your responses to the data collected during the CARMA experiment. However providing your name is not mandatory, you are free to choose to fill your name or not. Your involvement and participation will remain anonymous in any results and publications. Furthermore, records will be monitored and may be audited by the IRB without violating confidentiality.

You need to fill this survey only once. This survey contains 28 questions pertaining to user interests and social network interactions. The expected duration to complete the survey is no longer than 15 minutes. Note that multiple answers will be possible for some of the questions. In case you choose the "Other" option, please provide your input via keywords or brief descriptions.

If you have any questions or concerns about the research, please feel free to contact:

Principal Investigator: Prof. Zaher Dawy [[zd03@aub.edu.lb](mailto:zd03@aub.edu.lb)]

Co-Investigator: Miss Lynn Aoude [[lwa05@aub.edu.lb](mailto:lwa05@aub.edu.lb)]

If you have questions, concerns, or complaints about your rights as a research participant or the research in general and are unable to contact the research team, or if you want to talk to someone independent of the research team, please contact the Institutional Review Board. Tel: 01350000-5445 or [irb@aub.edu.lb](mailto:irb@aub.edu.lb)

By submitting your answers to this survey, you understand and agree to the terms set above. Thank you for your participation; it will be highly important for the research that we are conducting in this area.

There are 28 questions in this survey

### Group 1

<p><b>Name:</b></p> <p>Please write your answer here:</p> <input type="text"/>
--

**On average, how many media files per day do you store on your hand-held device (e.g., smartphone, tablet, etc.)?**

Please choose **only one** of the following:

- Zero
- 1 - 3
- 4 - 6
- 7 - 10
- 11 - 15
- More than 15

**What type of files do you store in your hand-held device?**

Please choose **all** that apply:

- Personal documents
- Public documents
- Personal images
- Public images
- Personal videos
- Public videos
- Music
- Other:

**For the document category, which holds the most interest to you, provided that you would store it on your hand-held device?**

Please choose **all** that apply:

- Novels, books, and the like
- Lectures and course material
- Technical documents
- Articles of interest
- Not applicable
- Other:

**For the image category, which holds the most interest to you, provided that you would store it on your hand-held device?**

Please choose **all** that apply:

- Personal photos
- Photos downloaded from the net (e.g. from websites such as 9gag, reddit, etc.)
- Landscapes
- Not applicable
- Other:

**For the video category, which holds the most interest to you, provided that you would store it on your hand-held device?**

Please choose **all** that apply:

- Music videos
- Scientific videos
- How-to videos
- TV series/Movies
- Short humor clips/stand-up comedy
- Personal videos
- Not applicable
- Other:

**What are your favorite genres of music?**

Please choose **all** that apply:

- Rock/ R&B/ Rap/Hip-Hop
- Metal
- Alternative
- Electronic/Dubstep
- Blues and Jazz/Classical
- Arabic music
- Not applicable
- Other:

**What are your favorite genres of movies/TV series?**

Please choose **all** that apply:

- Action
- Comedy
- Drama
- Science fiction (e.g. futuristic environment)/ Fantasy/Paranormal
- Horror
- Animation
- Historical/historical fiction
- Musicals
- Not applicable
- Other:

## Group 2

### What source of music do you use the most on your hand-held device?

Please choose **only one** of the following:

- Online streaming (e.g. Spotify, Pandora, Amazon...)
- Online sharing websites (e.g. Rapidgator, 4shared, torrents...)
- Online paying websites (e.g. iTunes)
- Not applicable
- Other

### What source of videos do you use the most on your hand-held device?

Please choose **only one** of the following:

- Online streaming (e.g. Youtube, vimeo, metacafe, hulu...)
- Online sharing websites (e.g. Rapidgator, 4shared, torrents...)
- Online paying websites
- Not applicable
- Other

### Group 3

#### How many music files do you have stored on your hand-held device?

Please choose **only one** of the following:

- Zero
- Less than 20
- Between 20 and 50
- Between 51 and 150
- Between 151 and 300
- More than 300

#### How many document files (.pdf, .epub, .docx, .lit, .mobi) do you have stored on your hand-held device?

Please choose **only one** of the following:

- Zero
- Less than 50
- Between 51 and 150
- Between 151 and 300
- More than 300

#### How many image files do you have stored on your hand-held device?

Please choose **only one** of the following:

- Zero
- Less than 50
- Between 51 and 150
- Between 151 and 300
- More than 300

**How many video files do you have stored on your hand-held device?**

Please choose **only one** of the following:

- Zero
- Less than 50
- Between 51 and 150
- Between 151 and 300
- More than 300

**On average, how many music files do you stream per week on your hand-held device?**

Please choose **only one** of the following:

- Zero
- Less than 10
- Between 10 and 30
- Between 31 and 50
- More than 50

**On average, how many video files do you stream per week on your hand-held device?**

Please choose **only one** of the following:

- Zero
- Less than 10
- Between 10 and 30
- Between 31 and 50
- More than 50

## Group 4

**Given the opportunity to acquire files from nearby devices via device-to-device sharing, what kind of files are you willing to download and share using this technique?**

Please choose **all** that apply:

- Documents
- Images
- Videos
- Music
- Not applicable
- Other:

**Are you willing to share files with others via device-to-device technology without any incentives?**

Please choose **only one** of the following:

- Yes
- No

**Are you willing to share files with others via device-to-device technology if incentives are offered?**

Please choose **only one** of the following:

- Yes
- No



**In case your answer to the previous question is YES, what incentives would most appeal to you?**

Please choose **all** that apply:

- Credits for online shopping
- Extra credit for each file shared (e.g. more account credits for calls and SMS)
- Free call minutes for each file shared
- Free SMS messages for each file shared
- Extra internet (3G/ 4G) quota
- Not applicable
- Other:

## Group 5

### Do you have a Facebook account?

Please choose **only one** of the following:

- Yes
- No

### If your answer to the previous question is YES, how many times do you access your Facebook account?

Please choose **only one** of the following:

- Never
- Once a week
- 2 to 3 times a week
- Daily
- Multiple times per day
- Not applicable

### Do you have a Twitter account?

Please choose **only one** of the following:

- Yes
- No

**If your answer to the previous question is YES, how many times do you access your Twitter account?**

Please choose **only one** of the following:

- Never
- Once a week
- 2 to 3 times a week
- Daily
- Multiple times per day
- Not applicable

**Do you have an Instagram account?**

Please choose **only one** of the following:

- Yes
- No

**If your answer to the previous question is YES, how many times do you access your Instagram account?**

Please choose **only one** of the following:

- Never
- Once a week
- 2 to 3 times a week
- Daily
- Multiple times per day
- Not applicable

**Do you have a LinkedIn account?**

Please choose **only one** of the following:

- Yes
- No

**If your answer to the previous question is YES, how many times do you access your LinkedIn account?**

Please choose **only one** of the following:

- Never
- Once a week
- 2 to 3 times a week
- Daily
- Multiple times per day
- Not applicable

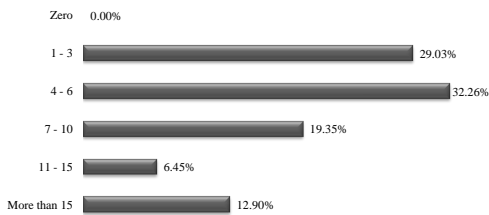
# Appendix D

## Survey Answer Statistics

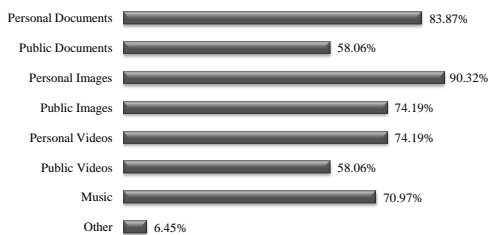
### Q1. Name



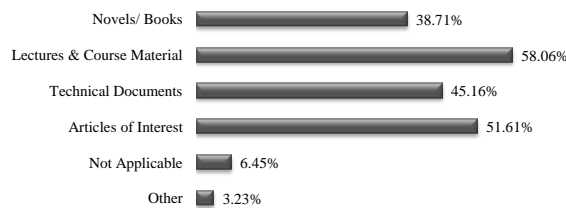
### Q2. Number of media files per day stored on hand-held device



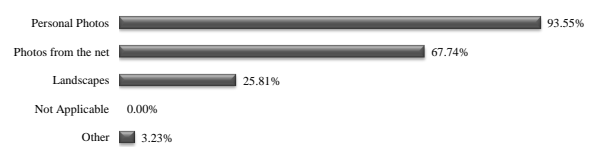
### Q3. Type of files stored on hand-held device



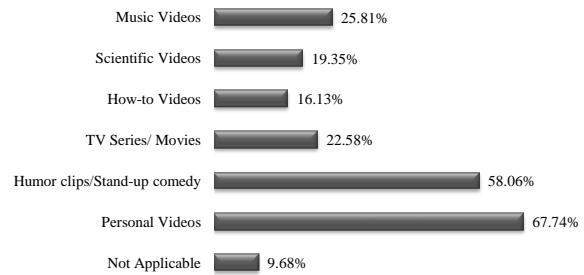
### Q4. Most popular document type stored



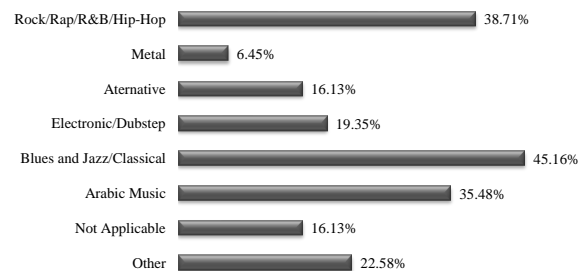
### Q5. Most popular image type stored



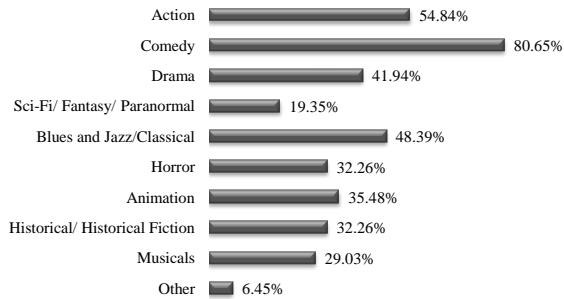
### Q6. Most popular video type stored



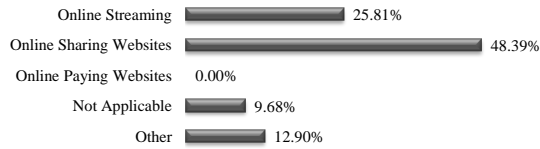
### Q7. Favorite music genres



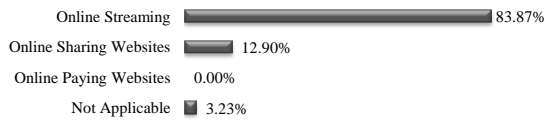
**Q8. Favorite Movie/ TV series genres**



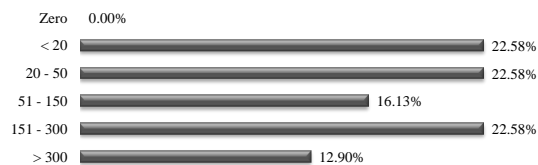
**Q9. Most used music source**



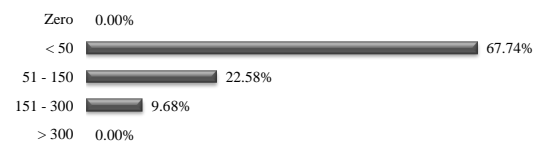
**Q10. Most used video source**



**Q11. Number of music files stored**



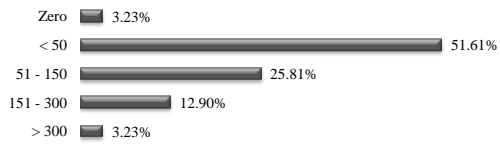
**Q12. Number of documents stored**



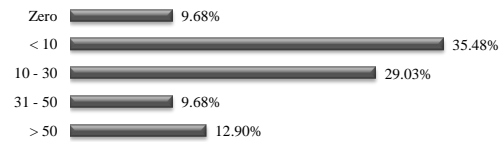
**Q13. Number of images stored**



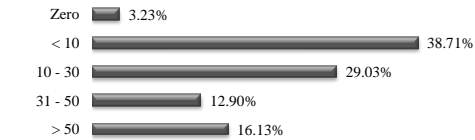
**Q14. Number of videos stored**



**Q15. Number of music files streamed per week**



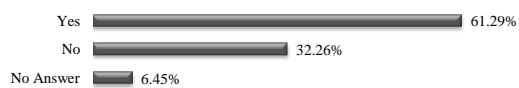
**Q16. Number of video files streamed per week**



**Q17. Type of files willing to share via D2D communication**



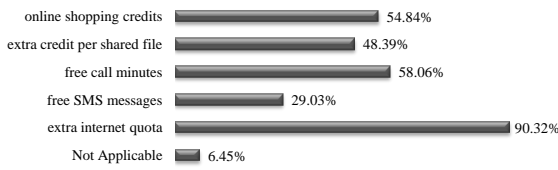
**Q18. Willingness to share files via D2D communication *without* incentives**



**Q19. Willingness to share files via D2D communications *with* incentives**



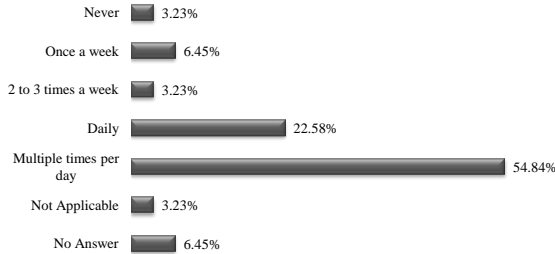
**Q20. Appealing incentives for D2D communication**



**Q21. Facebook account**



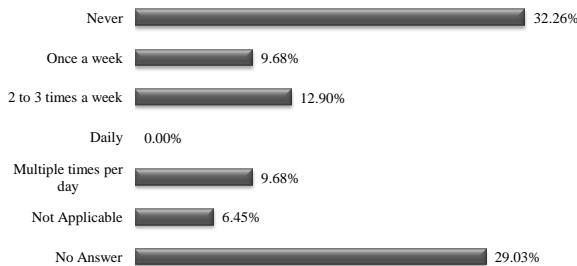
**Q22. Facebook access frequency**



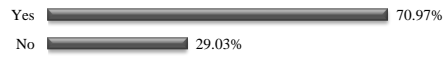
**Q23. Twitter account**



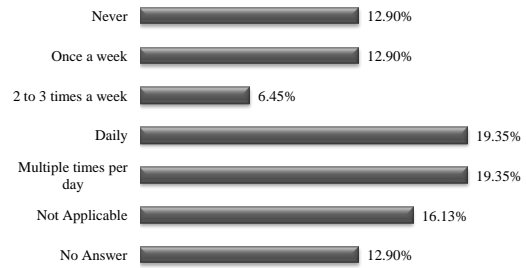
**Q24. Twitter access frequency**



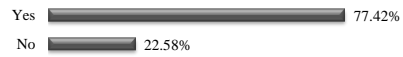
**Q25. Instagram account**



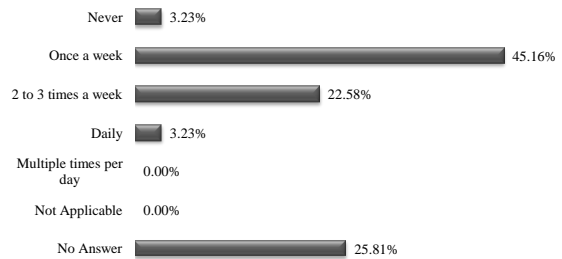
**Q26. Instagram access frequency**



**Q27. LinkedIn account**



**Q28. LinkedIn access frequency**



# Bibliography

- [1] “White paper: Cisco visual networking index: Forecast and methodology, 20152020,” CISCO, Tech. Rep., June 2016. [Online]. Available: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html>
- [2] V. Mancuso and O. Gurewitz, “Special issue on d2d-based offloading techniques,” *Physical Communication*, vol. 19, pp. 133 – 134, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1874490716300040>
- [3] A. Asadi, Q. Wang, and V. Mancuso, “A survey on device-to-device communication in cellular networks,” *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.
- [4] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, “Social network aware device-to-device communication in wireless networks,” *Wireless Communications, IEEE Transactions on*, vol. 14, no. 1, pp. 177–190, 2015.
- [5] Y. Zhang, L. Song, W. Saad, Z. Dawy, and Z. Han, “Exploring social ties for enhanced device-to-device communications in wireless networks,” in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 2013, pp. 4597–4602.
- [6] E. Talipov, Y. Chon, and H. Cha, “Content sharing over smartphone-based delay-tolerant networks,” *IEEE Transactions on Mobile Computing*, vol. 12, no. 3, pp. 581–595, 2013.
- [7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, “I tube, you tube, everybody tubes: Analyzing the world’s largest user generated content video system,” in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC ’07. New York, NY, USA: ACM, 2007, pp. 1–14.
- [8] R.-I. Ciobanu, R.-C. Marin, and C. Dobre, “Interaction predictability of opportunistic networks in academic environments,” *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 8, pp. 852–864, 2014.



- [9] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnovic, “Power law and exponential decay of intercontact times between mobile devices,” *IEEE Transactions on Mobile Computing*, vol. 9, no. 10, pp. 1377–1390, 2010.
- [10] R.-C. Marin, C. Dobre, and F. Xhafa, “A methodology for assessing the predictable behaviour of mobile users in wireless networks,” *Concurrency and Computation: Practice and Experience*, vol. 26, no. 5, pp. 1215–1230, 2014.
- [11] T. M. T. Do and D. Gatica-Perez, “Contextual conditional models for smartphone-based human mobility prediction,” in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp ’12. New York, NY, USA: ACM, 2012, pp. 163–172.
- [12] W. jen Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, “Modeling time-variant user mobility in wireless mobile networks,” in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, May 2007, pp. 758–766.
- [13] T. Phe-Neau, M. Dias de Amorim, and V. Conan, “Fine-grained intercontact vharacterization in disruption-tolerant networks,” in *Computers and Communications (ISCC), 2011 IEEE Symposium on*, June 2011, pp. 271–276.
- [14] A. Tatar, T. Phe-Neau, M. Dias de Amorim, V. Conan, and S. Fdida, “Beyond contact predictions in mobile opportunistic networks,” in *WONS 2014. 11th Annual Conference on Wireless On-Demand Network Systems and Services. IEEE*, Obergurgl, April 2014, pp. 65–72.
- [15] “Crawdad: Community resource for archiving wireless data at dartmouth,” <http://crawdad.org>.
- [16] Y. Li, T. Wu, P. Hui, D. Jin, and S. Chen, “Social-aware d2d communications: qualitative insights and quantitative analysis,” *Communications Magazine, IEEE*, vol. 52, no. 6, pp. 150–158, June 2014.
- [17] B. Zhang, Y. Li, D. Jin, P. Hui, and Z. Han, “Social-aware peer discovery for d2d communications underlying cellular networks,” *Wireless Communications, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.
- [18] S. Chandra and X. Yu, “An empirical analysis of serendipitous media sharing among campus-wide wireless users,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7, no. 1, pp. 6:1–6:23, 2011.
- [19] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, “Measurement, modeling, and analysis of a peer-to-peer file-sharing workload,” in *Proceedings of the Nineteenth ACM Symposium on*

*Operating Systems Principles*, ser. SOSP '03. New York, NY, USA: ACM, 2003, pp. 314–329.

- [20] F. Le Fessant, S. Handurukande, A.-M. Kermarrec, and L. Massouli, “Clustering in peer-to-peer file sharing workloads,” in *Peer-to-Peer Systems III*, ser. Lecture Notes in Computer Science, G. Voelker and S. Shenker, Eds. Springer Berlin Heidelberg, 2005, vol. 3279, pp. 217–226.
- [21] K. Sripanidkulchai, B. Maggs, and H. Zhang, “Efficient content location using interest-based locality in peer-to-peer systems,” in *Twenty-Second Annual Joint Conference of the IEEE Computer and Communications INFOCOM 2003*, vol. 3, March 2003, pp. 2166–2176.
- [22] C. Wang, Y. Li, and D. Jin, “Mobility-assisted opportunistic computation offloading,” *IEEE Communications Letters*, vol. 18, no. 10, pp. 1779–1782, October 2014.
- [23] Y. Li, D. Jin, P. Hui, and L. Zeng, “Modeling the communication contacts in roadside unit aided vehicles opportunistic networks,” in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 2376–2380.
- [24] Y. Li, Z. Wang, D. Jin, and S. Chen, “Optimal mobile content downloading in device-to-device communication underlying cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3596–3608, July 2014.
- [25] F. Hao, M. Jiao, G. Min, and L. T. Yang, “A trajectory-based recruitment strategy of social sensors for participatory sensing,” *IEEE Communications Magazine*, vol. 52, no. 12, pp. 41–47, December 2014.
- [26] P. A. Frangoudis and G. C. Polyzos, “Security and performance challenges for user-centric wireless networking,” *IEEE Communications Magazine*, vol. 52, no. 12, pp. 48–55, December 2014.
- [27] Y. Li, T. Wu, P. Hui, D. Jin, and S. Chen, “Social-aware d2d communications: Qualitative insights and quantitative analysis,” *IEEE Communications Magazine*, vol. 52, no. 6, pp. 150–158, June 2014.
- [28] Y. Cao, T. Jiang, X. Chen, and J. Zhang, “Social-aware video multicast based on device-to-device communications,” *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1528–1539, June 2016.
- [29] L. Wang, H. Wu, W. Wang, and K. C. Chen, “Socially enabled wireless networks: resource allocation via bipartite graph matching,” *IEEE Communications Magazine*, vol. 53, no. 10, pp. 128–135, October 2015.

- [30] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, “Social network aware device-to-device communication in wireless networks,” *Wireless Communications, IEEE Transactions on*, vol. 14, no. 1, pp. 177–190, 2015.
- [31] Y. Zhang, L. Song, W. Saad, Z. Dawy, and Z. Han, “Exploring social ties for enhanced device-to-device communications in wireless networks,” in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 2013, pp. 4597–4602.
- [32] —, “Social network enhanced device-to-device communication underlaying cellular networks,” *arXiv preprint arXiv:1510.04684*, 2015.
- [33] C. Xu, S. Jia, L. Zhong, and G. M. Muntean, “Socially aware mobile peer-to-peer communications for community multimedia streaming services,” *IEEE Communications Magazine*, vol. 53, no. 10, pp. 150–156, October 2015.
- [34] A. Roy, P. De, and N. Saxena, “Location-based social video sharing over next generation cellular networks,” *IEEE Communications Magazine*, vol. 53, no. 10, pp. 136–143, October 2015.
- [35] T. C. Tsai and H. H. Chan, “Nccu trace: social-network-aware mobility trace,” *IEEE Communications Magazine*, vol. 53, no. 10, pp. 144–149, October 2015.
- [36] P. Y. Chen, S. M. Cheng, P. S. Ting, C. W. Lien, and F. J. Chu, “When crowdsourcing meets mobile sensing: a social network perspective,” *IEEE Communications Magazine*, vol. 53, no. 10, pp. 157–163, October 2015.
- [37] W. Moreira and P. Mendes, “Pervasive data sharing as an enabler for mobile citizen sensing systems,” *IEEE Communications Magazine*, vol. 53, no. 10, pp. 164–170, October 2015.
- [38] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor, “Context-aware small cell networks: How social metrics improve wireless resource allocation,” *Wireless Communications, IEEE Transactions on*, vol. 14, no. 11, pp. 5927–5940, 2015.
- [39] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford large network dataset collection,” Jun. 2014. [Online]. Available: <http://snap.stanford.edu/data>
- [40] I. Paul, “Wi-fi direct vs. bluetooth 4.0: A battle for supremacy,” October 2010. [Online]. Available: <http://www.pcworld.com/article/208778/Wi-Fi-Direct-vs-Bluetooth-4.0-A-Battle-for-Supremacy.html>
- [41] [Online]. Available: <http://www.differen.com/difference/Bluetooth-vs-Wifi>

- [42] Jukka, “Random thoughts on mobile programming.” [Online]. Available: <http://www.drjukka.com/blog/wordpress/?p=95>
- [43] [Online]. Available: <http://www.math.wm.edu/~leemis/chart/UDR/PDFs/Gamma.pdf>
- [44] [Online]. Available: <http://www.mathworks.com/help/stats/logncdf.html>