Stumbl: Using Facebook to Collect Rich Datasets for Opportunistic Networking Research

Theus Hossmann, Franck Legendre, George Nomikos
Communication Systems Group
ETH Zurich, Switzerland
lastname@tik.ee.ethz.ch

Thrasyvoulos Spyropoulos
Mobile Communications
EURECOM, France
firstname.lastname@eurecom.fr

***Abstract*—Opportunistic networks use human mobility and consequent wireless contacts between mobile devices to disseminate data in a peer-to-peer manner. Designing appropriate algorithms and protocols for such networks is challenging as it requires understanding patterns of (1) mobility (who meets whom), (2) social relations (who knows whom) and (3), communication (who communicates with whom). To date, apart from few small test setups, there are no operational opportunistic networks where measurements could reveal the complex correlation of these features of human relationships. Hence, opportunistic networking research is largely based on insights from measurements of either contacts, social networks, or communication, but not all three combined.**

In this paper we report an experiment called Stumbl, as a step towards collecting rich datasets comprising social, mobility and communication ties. Stumbl is a Facebook application that provides participating users with a user-friendly interface to report their daily face-to-face meetings with other Facebook friends. It also logs user interactions on Facebook (e.g. comments, wall posts, likes). This way the contact graph, social graph, and activity graphs for the *same set of users* could be compared and analyzed. We report here preliminary results and analyses of a first experiment we have performed.

I. Introduction

The rapid proliferation of small wireless devices creates ample opportunity for novel applications [1], as well as for extending the realm of existing ones [2], [3]. *Opportunistic* or *Delay Tolerant Networking* (DTN) [4] is a novel networking paradigm that is envisioned to complement and extend existing wireless infrastructure such as 3G and WiFi. Nodes take profit of communication opportunities by exchanging data whenever they are within mutual wireless transmission range of each other (*in contact*).

Algorithms and protocols (e.g., routing protocols) for opportunistic networks were originally largely based on random decisions [5], not accounting for heterogeneity in terms of capabilities of devices and behavior of people carrying them. Such random protocols typically require large amount of resources for timely delivery of content (e.g., epidemic spreading of messages). To overcome this, more recent protocols exploit node heterogeneity in order to make educated decisions to provide good performance at limited resource usage. Examples are routing protocols exploiting structure in social ties [6], [7], [8] or structure in mobility ties [9], [10]. Simulations show that efficiency is much better than for random protocols.

Designing and analyzing efficient protocols is challenging, as it requires knowledge about various aspect of human behavior. Relevant questions are: *Which nodes have frequent contacts and hence are good relays? Which nodes are socially related and hence trust each other and are willing to cooperate? Which nodes communicate with each other and need fast routes between them?* In fact, we can assume that these three dimensions of social, communication and contact relations are correlated at least to a certain degree. However, it is largely unknown how strong this correlation is, how it can be exploited for opportunistic networking and how it affects performance of existing protocols.

To date, there are only few small deployments of opportunistic networks [11], [12], [2], [3] from which practical insights of the correlation of social, mobility and communication ties could be gained. Hence, research in this direction is largely based on insights from empirical analysis of datasets typically capturing only one or two of the aspects of relations, but not all three combined.

Example datasets are *mobility traces* (some of which also contain information about social ties between the nodes) from WLAN Access Point associations [13], [14] or Bluetooth contacts [15], [16], [17]. Analysis of such traces has shown that there is some correlation of mobility and social connections [15], [17]. However, these analyses do not consider which nodes would actually actively *communicate* and interact with each other in an opportunistic application (i.e., who is interested in content and whom, who sends messages to whom). To also capture this aspect, we want to collect datasets comprising all *three* dimensions.

While mobility and social connections can be measured, the question of who communicates with whom using opportunistic applications is difficult to answer, as there are only few – and mostly small – deployments of opportunistic applications [3], [11], [12]. However, we assume that opportunistic applications are of social nature and we speculate that they would create similar communication patterns like today's online social network and Web 2.0 platforms, such as Facebook or Twitter. In fact, current online social networks could be run over opportunistic networks [2], [3]. Facebook is a typical, and to date the most widely used, representative of an online social networking service, fostering communication and distribution of (user generated) content among friends. It

provides an API for application development, allowing us to create an application – called Stumbl – to measure all three dimensions of interest. Using the Facebook API, the Stumbl application records communication and social ties of its users. Additionally, it asks participants to report their meeting data on a regular basis, to also cover the mobility dimension of users' relations (i.e., how often, how long and in what context users meet their Facebook friends).

Our contributions can be summarized as follows. (1) We discuss Stumbl as a new methodology of collecting combined social, communication and mobility datasets, relying on self-reported as well as automatically measured data. (2) We analyze the dataset from a preliminary Stumbl experiment with special focus on how mobility, communication and social ties relate to each other. In particular, we find that we can expect communication ties to be one order of magnitude stronger for friends who see each other face-to-face. (3) We discuss implications of these results for opportunistic routing and traffic modeling.

The rest of this paper is structured as follows. In Sec. II we describe the Stumbl application and characterize the data collected in the experiment. Sec. III presents an empirical analysis of how social tie type, meetings and communication patterns relate to each other, along with discussions about implications of the findings for routing and traffic modeling. Finally, we conclude and discuss future work in Sec. IV.

II. Stumbl Application and Dataset

To measure contacts, social ties and communication, we have implemented Stumbl as a Facebook application. In this section we briefly discuss the Stumbl application (II-A) as well as the Stumbl experiment and resulting dataset (II-B). Finally, we also discuss limitations of the methodology and collected data (II-C). A more detailed description of the application and experiment can be found in [18].

A. The Stumbl Application

Facebook provides an API for authorized (by the user) applications to access user data. Our Stumbl application[1] uses this API to retrieve the user's social connections and Facebook communication events. Additionally, we ask the users to regularly report whom of their friends they meet face-to-face, by filling in a survey form in the Stumbl Facebook application. One big benefit of integrating Stumbl as an application in the Facebook.com website is that it is a convenient way for many people to regularly report their meeting data: Since visiting the Facebook website is part of the daily routine for many people, the barrier to fill in the survey is small.

When a user joins the Stumbl experiment, there are two main phases of participation.

Initialization Phase: In a one time initialization step, the user is asked to select a subset of her Facebook friends which she meets face-to-face regularly (at least once a month). We will refer to this subset of Facebook friends as the *Stumbl friends*. The reason for selecting a subset of the friends for the survey is two-fold. First, most users have large number of Facebook friends, many of which living far away. These pairs typically have only very rare meetings (weak ties). In order to keep the effort for reporting data as small as possible, we wanted to exclud them from the input interface. Second, we are mainly interested in the meeting patterns of people who see each other frequently (strong ties), as such meetings are more predictable than the random occasional meetings[2].

Fig. 1: Stumbl screen shot. For each Stumbl friend, context, number and total duration of meetings can be reported (for the previous day). Options are chosen to capture a range of different meeting behaviors.

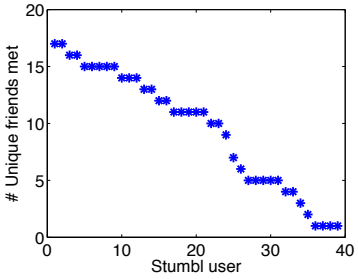
In our experiment we have limited the number of Stumbl friends to 20. Typically, a user regularly sees less than 20 of her Facebook contacts, as we will report later. The selection of 20 friends hence does not narrow the data we gather. Note that the users have the option to change their selection of Stumbl friends during the experiment.

To complete the initialization step, Stumbl asks the user to classify the relationship type to each of the Stumbl friends as one or more of *family, friend, colleague* or *acquaintance*. As "friendship" on Facebook is a very broad term characterizing a wide range of actual social relationships, we use this classification for a more fine tuned analysis of the social dimension of relations.

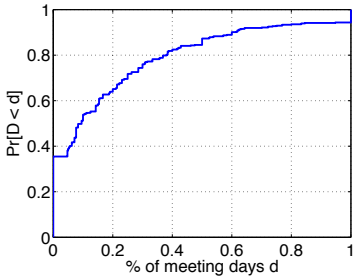
Reporting Phase: After the initialization step follows the recurring report of face-to-face meetings. As automated measuring of face-to-face meetings typically requires special equipment (iMotes [16] or phones equipped with special software [17]) and is costly and complex, we rely on self-

[1] http://apps.facebook.com/_stumbl/

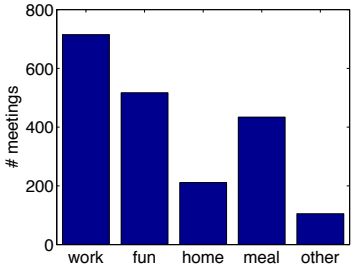
[2] Note that the occasional random meetings of weak ties can be very beneficial, for instance for opportunistic routing protocols as "short cuts". However, they are typically not predictable and protocols can not rely on them. Decisions have to be made depending on strong and predictable mobility ties.



(a) Ranked numbers of unique friends met by Stumbl users.



(b) CDF of percentage of days the pairs meet during experiment.



(c) Number of meetings reported per context.

Fig. 2: Overview of Stumbl meeting statistics.

reported data to assess the mobility dimension of relations. Correlating self-reported and measured (via Bluetooth) proximity has shown that the quality of self-reported proximity data drops when reporting events more than seven days back in time [17]. To ensure a good level of accuracy for the reported information, we choose a reporting interval of one day: The Stumbl users are asked every day (reminded by E-Mail) to visit the Stumbl application and fill in the questionnaire about whom of their Stumbl friends they met the previous day[3]. Thus, the collected data has a temporal resolution of meetings of one day.

For each friend a user reports a meeting, additional information has to be provided about (i) how often she saw the friend, (ii) for how long in total these meetings lasted, and (iii) the contexts of the meetings (given the options *work*, *fun*, *home*, *meal*, *other* for selection). These additional features allow us to make a more fine grained analysis of the contact data.

Fig. 1 shows the input interface as participants see it. We designed the interface such that we can collect a maximal amount of data with as small an effort as possible by the user. From experience and user reports, we know that the input requires less than 5 Minutes per day, a target we had set to motivate daily participation.

In order to capture communication between a user and her Stumbl friends, the application uses the Facebook API to query for interaction events, every time meeting data is submitted. We collect the following three types of interaction to which the API provides access[4]. *Wall posts:* Users post content (messages, photos, videos, links, etc.) on each others wall. *Comments:* Wall posts can be commented on. *Likes:* As a brief sign of approval, any item on the wall can be "liked".

These communication events are time stamped. They are all directed (e.g., a user writes on an other user's wall), and we collect both, incoming and outgoing events.

[3]Note that with the check-in service *Places*, Facebook also provides a platform for recording user location and meetings (tagging people at the same location). However, this would require users to check-in and tag people at every meeting and is too cumbersome to ask. Also, since check-ins and tags show up in the user profile, this methodology of recording meetings would have serious privacy issues.

[4]A fourth communication mechanism, private messages, is not accessible by the API for obvious privacy reasons.

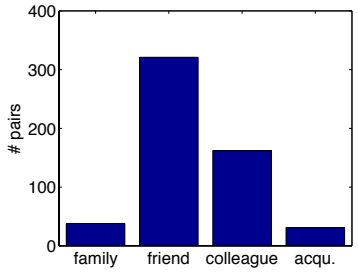


Fig. 3: Number of pairs per social tie type.

Summarizing, Stumbl records social ties (friend, family, colleague acquaintance), Facebook communication (wall posts, likes, comments, tags) and meeting data (number, duration and context of meetings) which allows us to get insight in three dimensions of the relationships of a Stumbl user.

B. The Stumbl Experiment

In this paper, we report results from a preliminary experiment using the Stumbl application, which we used to gain experience with application and user behavior – and which also led to a first interesting (but limited in size) dataset. The experiment ran for three weeks between August 16th 2010 and September 6th 2010. At the beginning of the experiment, we recruited participants mainly by personal invitations, which led to a total of 39 users providing useful information. In order to provide incentives for these users to persistently report their meeting data during the experiment, the users participated in a raffle. To provide the right incentives, the chances of winning were dependent on two factors: the number of days the application was visited, and the number of their friends who registered as Stumbl users. While these raffles should be incentives to provide data regularly, they should not provide incentives to provide false data. In the following we provide an overview of the dataset we collected during this experiment.

During the 21 days of the experiment, on average 22 of

the 39 participants reported meeting data. This means that users were quite persistent in participating and shows that the incentives for participating regularly worked well. In the following, we report some general statistics about the collected data to provide a general impression of the dataset.

On average, users selected 14 *Stumbl friends* in the initialization step. 11 users selected the maximum allowed 20 users. The number of Stumbl friends the user actually reported meetings with throughout the experiment is lower than the number of Stumbl friends, as shown in Fig. 2a. On average a user reported meeting 9.5 unique Stumbl friends during the experiment. The maximum is at 17 unique friends and hence lower than the 20 allowed. We conclude that the selection of 20 friends does not narrow the number of pairs for which we receive meeting reports.

In total, we have 498 pairs of Facebook users[5] in our Stumbl dataset. Fig. 2b shows the cumulative distribution function of how often these pairs met. As users did not report their meetings every day, we divide the number of days a pair meets by the number of days we have self-reported meeting data for the given pair (including days where they report no meeting). Thus, the figure shows the percentage of days the pairs met. Roughly 65% of the pairs met at least once (35% had zero meetings) and almost 5% of pairs report meeting every day.

Further, we want to analyze the contexts (*work*, *home*, *fun*, *meal*, *other*) in which the meetings happen. Fig. 2c shows how the meetings are split among the different contexts. We observe that most meetings happened at work, but also for the other contexts we have quite large numbers of meetings reported.

Next, we want to provide an overview of the social tie types our measurements cover. Fig. 3 shows how the 498 Stumbl pairs are divided into *family*, *friends*, *colleagues* and *acquaintances*. We observe that most of the pairs are classified as friend or colleague. Relatively few pairs are of types family or acquaintance. Note that the user can specify more than one type of social tie per Stumbl friend. Hence, the number of pairs per type sum to more than 498.

In terms of communication events, Tab. I summarizes the number of events we recorded during the experiment. With a total number of 643 communication events, we have a large enough sample to provide statistics about communication. In total, we saw communication between 91 or 18% of the 498 pairs.

	Posts	Comments	Likes	Total
Nr. of Events	199	341	103	643

TABLE I: Total number of registered Facebook communication events between Stumbl friends per event type.

These statistics give a separate overview about each of the three dimensions of relationship we measure. In Sec. III we will analyze how the different aspects of relationship correlate with each other. We will look at questions like: *How does the type of social tie affect meeting probabilities and communication probabilities? How do meetings relate to probabilities of communicating?*

C. Limitations and Validation of Dataset

We now want to adress potential bias and limitations of our dataset and the general methodology of collecting data (especially self-reported meeting data) using a Facebook application.

The 39 users have an average of 252 friends in their Facebook social graph. This is considerably more the average friend count of 130 reported by Facebook [19]. We assume that the large number of Facebook friends does not mean that the average Stumbl user is more sociable than an average person. Rather, it means that the Stumbl users are more active Facebook users. While this may cause a bias in the measurement, we believe that the Stumbl users may actually be more representative users of opportunistic networks, as we expect them to be well-versed users of new technologies.

As Stumbl users were recruited based on personal invitations by the authors of this study and by word-of-mouth recommendation, the Stumbl users present a rather local group of people (most are researchers or students living in few cities). In the future, we plan to extend Stumbl and use it for experiments with broader audience.

Another concern is that the self-reported meeting data may be erroneous because the user does not recall meetings correctly or decides to provide wrong information. In order to estimate the severity of these effects, we validate the data where possible. We do so by looking at the 47 pairs of users for which we have mutual meeting data. We find that in 86% of the cases the reports whether or not there was a meeting between a pair matches (i.e., both Stumbl users report that there was a meeting or both report there was no meeting). This seems a quite good correlation. For the cases where both report that there was a meeting, we further check whether their reported meeting counts[6], meeting duration[7] and meeting contexts[8] match. We find that this is the case in 57% of meeting counts, 66% of durations and 87% of contexts. While not perfect correlation, we conclude that the reports are accurate enough, especially those of meetings or not on a given day and the context in which the meetings happen.

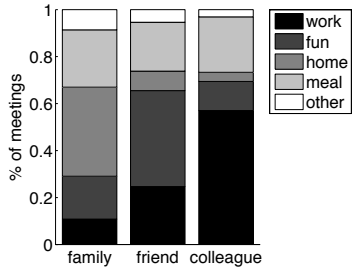
A limitation inherent to the methodology of self-reported mobility data is that Stumbl can only capture meetings between friends. Random encounters of strangers or meetings between familiar strangers cannot be recorded. Thus, on one hand we are limited to the analysis of properties of *strong* mobility ties. On the other hand, Stumbl provides very faceted information for these strong ties, allowing us to make very detailed analyses of the strong *backbone* of opportunistic

[5]For 47 of these pairs, we have mutual meeting reports data, i.e, both nodes participate in the Stumbl experiments. For the rest only one node reported data.

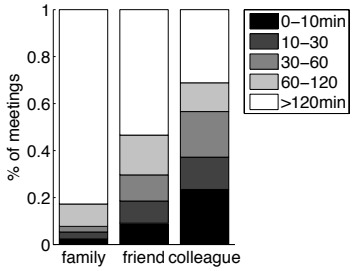
[6]For meeting count the selects between 1 or $2-3$ or $4-5$ or >5.

[7]For duration, the options are $0-10$min or $10-30$min or $30-60$min or $60-120$min or >120min

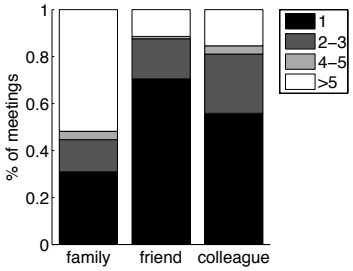
[8]For context the options are work, fun, home, meal, other



(a) Tie types vs. context of meetings.



(b) Tie types vs. duration of meetings.



(c) Tie types vs. frequency of meetings.

Fig. 4: Dependence of meeting patterns on social tie type.

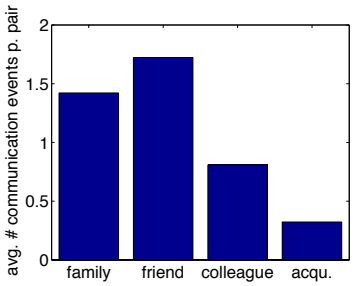


Fig. 5: Tie types vs. communication events.

networks. Note that for analyzing contacts this limitation can also be an advantage: Typically, in automatically recorded contact traces, it is hard to distinguish strong and weak ties and it is not a priori clear whether a contact is a random encounter or part of a more "meaningful" mobility tie.

III. Social Ties vs. Meetings vs. Communication

In Sec. II we have seen an overview of the Stumbl dataset. We now present an empirical analysis of how the aspects social tie type, meetings and communication, relate to each other. We also discuss the hints our findings provide for opportunistic routing and traffic modeling.

A. Social Ties vs. Meetings

First, we look at the dependence of meeting behavior on the type of social tie. From experience and intuition about human mobility, we expect that meeting patterns of colleagues, family, friends and acquaintances have quite different characteristics in terms of context, frequency and duration.

In a first step, and as a sanity check, we look at social tie type and meeting contexts. Naturally, we expect the tie type to influence the context of meetings: We meet colleagues at work, family at home, friends for fun, etc. Fig. 4a confirms this by showing the percentage of meetings happening in a given context, split by social tie type[9].

[9]We do not show acquaintance relationships here since we observe too few meetings between acquaintances in the dataset to make reliable statements.

Fig. 4b and 4c show how long and how often pairs meet per day (given that they meet at least once that day). We observe that meetings between family are generally long and frequent. Between friends, meetings are still quite long but typically only once per day. For colleagues, meetings are generally shorter. Such short meetings of colleagues may be just crossing each other, talking briefly or drinking a short coffee during breaks.

Summarizing, we find that the social tie type has very strong impact on meeting characteristics in terms of context, duration and frequency of meetings. These results are not surprising. Yet, they have implications for example for DTN routing protocols where routing decisions are based on social networks [6], [7], [8]. If the *type* of social link is known to such protocols, this might be useful information, without necessary having to sample actual contact times. Different conclusions and strategies may be applicable to different tie types: Typically, a tie with frequent meetings is a good carrier in terms of short delivery delay. However, if the frequent meetings are short, the capacity of the contacts may be too small to deliver a large amount of data. For large data transfers, long meetings may be more desirable.

B. Social Ties vs. Communication

In a next step, we want to investigate how the social tie type is related to communication patterns. Fig. 5 reports the average number of communication event per pair during the experiment, split by social tie type. We note that friends and family are the most communicative. Colleagues communicate much less and for acquaintances we find an average of merely 0.3 communication events per pair, not even one fifth of the communication events an average friend pair shows.

Not all nodes with social ties communicate with the same frequency. Instead, communication, or traffic, between pairs of nodes depend on their type of social tie. This is something to consider when simulating opportunistic network traffic. Realistic traffic models should incorporate heterogeneity of social ties and how this reflects in different communication patterns.

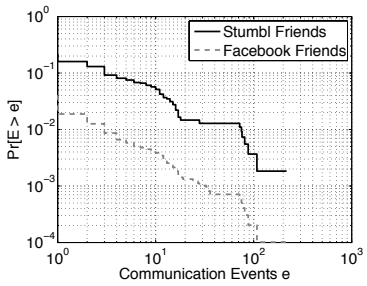


Fig. 6: CCDF of number of communication events for Facebook friends and Stumbl friends (log-log scale).

C. Meetings vs. Communication

The last question we want to answer is how the meeting and communication patterns correlate. *Are we more likely or less likely to communicate with friends to whom we have strong mobility ties? In other words, do we communicate with friends we see face-to-face (e.g., to discuss common experiences) or with remote friends (e.g., to stay in touch)?* To answer this, we compare the number of communication events of Stumbl friends (as representatives of friends to whom we have strong mobility ties) to the number of communication events with general Facebook friends (including strong and weak mobility ties). Fig. 6 shows the complementary cumulative distribution functions of the pairwise number of communication events, for Stumbl friends, compared to Facebook friends. The plot shows that the number of communication events between Stumbl friends is indeed much higher than between "normal" Facebook friends. In fact, on average a user communicates about 10 times more often with a Stumbl friend.

These are only preliminary results and the matter requires further research. However, already with the present data we can point out some implications. First, the finding that communication is more "local" than social connections is a strong argument in favor of opportunistic networks. In the future, more detailed analysis could provide answers to where opportunistic network are useful and in which cases infrastructure is required (i.e., for combined opportunistic and infrastructure networks). Second, in order to model data traffic in opportunistic networks, we should consider that pairs with strong mobility ties are more likely to communicate. Thus, realistic traffic models should be combined with realistic mobility models.

IV. Conclusion

We have presented Stumbl, a Facebook application to collect contact, social and interaction graphs for the same set of users. Stumbl automatically collects interaction events using the Facebook API and relies on user reports about the type of their social relationships and the face-to-face meetings.

The analysis of the dataset from a preliminary experiment has revealed that all three dimensions of tie strength depend on each other. (1) The type of social tie (friend, family, colleague or acquaintance) has strong impact on context, duration and frequency of meetings. Consequently, we argue that having this information is valuable information for instance for opportunistic routing protocols. (2) The number of Facebook communication events differs for different relationship ties, a fact which should be considered when modeling traffic in opportunistic network. (3) People use communicate preferentially with friends they also have face-to-face meetings. Thus, communication ties are more local than social ties.

In the future, we plan to run bigger Stumbl experiments with more participants. The challenge is to provide incentives to the users to regularly report true data about their face-to-face meetings. Using game mechanisms, if designed carefully, could be a promising approach to spread the application.

Our goals are to better understand (1) how traffic patterns in opportunistic networks relate to mobility patterns and social ties, and (2), the extent to (or conditions under) which the social graph correlates with or dictates physical mobility. We believe that measurements frameworks/experiments like Stumbl could offer valuable insight into these questions.

References

[1] The Aka Aki Network. `http://www.aka-aki.com/`. [Online]. Available: http://www.aka-aki.com/

[2] B. Distl, G. Csucs, S. Trifunovic, F. Legendre, and C. Anastasiades, "Extending the reach of online social networks to opportunistic networks with PodNet," in *MobiOpp*, 2010.

[3] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, "Mobiclique: Middleware for mobile social networking," in *WOSN*, 2009.

[4] "Delay tolerant networking research group," *http://www.dtnrg.org*.

[5] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *WDTN*, 2005.

[6] A. Mtibaa, M. May, C. Diot, and M. Ammar, "PeopleRank: Social opportunistic forwarding," in *IEEE INFOCOM*, 2010.

[7] C. Boldrini, M. Conti, and A. Passarella, "Contentplace: social-aware data dissemination in opportunistic networks," in *ACM MSWiM*, 2008.

[8] P. Hui and J. Crowcroft, "How small labels create big improvements," in *ACM CoNEXT*, 2006.

[9] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble Rap: Social-based forwarding in delay tolerant networks," in *ACM MobiHoc*, 2008.

[10] T. Hossmann, T. Spyropoulos, and F. Legendre, "Know thy neighbor: Towards optimal mapping of contacts to social graphs for DTN routing," in *IEEE Infocom 2010*, March 2010.

[11] X. Zhang, J. Kurose, B. N. Levine, D. Towsley, and H. Zhang, "Study of a Bus-Based Disruption Tolerant Network: Mobility Modeling and Impact on Routing," in *Mobicom*, 2007.

[12] A. Lindgren, A. Doria, J. Lindblom, and M. Ek, "Networking in the land of northern lights: two years of experiences from dtn system deployments," in *WiNS-DR*, 2008.

[13] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," in *ACM MOBICOM*, 2004.

[14] C. Tuduce and T. Gross, "A mobility model based on WLAN traces and its validation," in *IEEE INFOCOM*, 2005.

[15] A. Mtibaa, A. Chaintreau, J. LeBrun, E. Oliver, A.-K. Pietilainen, and C. Diot, "Are you moved by your social network application?" in *WOSN*, 2008.

[16] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Pocket switched networks and human mobility in conference environments," in *WDTN*, 2005.

[17] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *PNAS*, 2009.

[18] G. Nomikos, "Studying social-driven mobility: Comparing face-to-face meetings to online social network activity," Master's thesis, ETH Zurich, 2010.

[19] Facebook Statistics. [Online]. Available: http://www.facebook.com/press/info.php?statistics