Planet-scale Human Mobility Measurement

Pan Hui[†], Richard Mortier[◇], Michal Piórkowski, Tristan Henderson[‡], Jon Crowcroft[⋆]

[†] Deutsche Telekom Laboratories [◇] University of Nottingham
[‡] University of St Andrews [⋆] University of Cambridge

ABSTRACT

Research into, and design and construction of mobile systems and algorithms requires access to large-scale mobility data. Unfortunately, the research community lacks such data. For instance, the largest available human contact traces contain only 100 nodes with very sparse connectivity, limited by experimental logistics. In this paper we pose a *challenge* to the community: how can we collect mobility data from *billions* of human participants? We re-assert the importance of large-scale datasets in communication network design, and claim that this could impact fundamental studies in other academic disciplines. In effect, we argue that planet-scale mobility measurements can help to save the world. For example, through understanding large-scale human mobility, we can track and model and contain the spread of epidemics of various kinds.

1. INTRODUCTION

Human mobility traces are critically important to many disciplines in addition to computer networking, ranging from epidemiology [7] to urban planning [30]. Unfortunately, existing traces of human mobility are flawed: using traditional social science methods to collect data has proven difficult [32] and traces collected using technology methods have suffered from a variety of limitations. These include small size (the largest is 100 nodes [15]), short duration (the longest is 9 months [10]) and high locality (many scenarios limited to campus and conference environments [6]). These datasets may not be enough for large mobile system evaluations, and are definitely insufficient for epidemiology, where planet-wide measurements are needed to track the spread of disease.

As members of the networking community, we have both the tools and methods to conduct large-scale mobility data collection. Furthermore, our contributions will not only benefit the wireless and mobile networking research communities, but will impact fundamental research in other areas allowing more features about human behaviour to be uncovered. We believe that the situation is analogous to that of complex networks research, which has flourished since 1989 when the first large datasets from the Internet (and subsequently the World Wide Web) became available [2]. To achieve similar improvements in mobile networking and other related fields, relevant large-scale datasets must be made available.

In this paper we challenge the community to collect large-scale human mobility traces. We highlight some of the issues in the hope that the community can help find good solutions. In the meantime, we propose some solutions intended to form the basis of initial efforts; the main aim is to raise these issues to gain community support to meet this challenge and make the topic *hot* in the networking community.

2. IMPORTANCE OF LARGE-SCALE MOBILITY DATA

As mentioned above, large-scale datasets are useful for many aspects of research. In this paper we focus only on two of them: mobile system design and validation, and epidemiological studies.

2.1 System design and validation

After its first use in the evaluation of Dynamic Source Routing [19], the random waypoint model (RWP) became the *de facto* standard mobility model in the mobile networking community. For example, of the 10 papers in ACM MobiHoc 2002 which considered node mobility, 9 used RWP [33]. This trend has changed dramatically over recent years after the introduction of real mobility traces for evaluation: of the 10 papers considering node mobility in MobiHoc 2008, 7 used real mobility traces for evaluation.

The community has realized that unrealistic models are harmful for scientific research. Although real traces may suffer from limited numbers of participants, coarse granularity, and short experimental duration, they at least reflect *some* aspects of real life. Thanks to the popularity of Online Social Networks (OSNs), we can now gather large-scale data about the topology and membership information of millions of OSN users and use these to study aspects of the social networks [21, 24]. But where is the large-scale dataset for evaluating, for instance, inter-city ad-hoc communication using mobile computing? Or even a single city-wide mobile communication system? We have very few empirical hints for this. Without the help of real data, we cannot even know whether this kind of system is possible. Even if we extrapolate large-scale mobility traces from small-scale traces, the problem of validating the extrapolation remains.

Instead of using mobility traces directly to run trace-driven simulations, a possible approach is to extract characteristics from the data and build more realistic mobility models. Much work has been done in modeling human mobility for mobile ad hoc network simulation [5]. Researchers have proposed more realistic models by incorporating obstacles [18], social information [25], and clustering features observed in realistic mobility scenarios [27]. Analysis of real traces has demonstrated power-law inter-contact time distributions with cut off [6, 20], levy-flight patterns consisting of lots

ACM HotPlanet '10 San Francisco, USA
Copyright 2010 ACM 978-1-4503-0177-0 ...$5.00.

of small moves followed by long jumps [28], heterogeneous centralities [11] (i.e., popularity) and clustering structure [15]. But again, these results are from small-scale datasets and are limited to specific scenarios with limited time durations. Some researchers have extrapolated from these by assuming, for instance, that the way people move in a city is correlated to the centrality distribution of the city graph [30], but this has yet to be verified empirically. Gonzalez *et al.* [13] extracted levy-walk properties from large-scale mobile phone usage. The limitation is that the conclusions were drawn based on analysis of an extremely coarse-grained dataset, where mobile user location was recorded only up to 12 times per day. Researchers may argue that human behavior should be scale-free in different dimensions, but we need more data for further verification. Moreover, since the data from the [13] study was not released, it is impossible to verify or build on their findings.

We need large-scale human mobility data of high spatio-temporal granularity to verify the properties we mentioned above. Following analogous progress in related fields, it seems likely that we will uncover many more features from such data. We believe that this is crucially important for the mobile computing community.

2.2 Epidemiological Studies

Moving beyond social science, the communication network community has also aided research in many other academic disciplines. For instance, our (computer scientists) methodology and data made the modeling of human dynamics [31] and the development of the field of complex network research [3] possible. Large-scale mobile data can further enable the study of epidemic disease spreading. The current state-of-the-art in epidemic modeling uses data from the International Air Transport Association (IATA) commercial airline traffic database to determine travel between airports and to provide coarse-grained estimates of global spreading patterns [7], as well as data of transportation and commuting patterns in urban areas, which can be used to model a metapopulation mechanism of spreading [8]. Researchers cannot develop more microscopic models of epidemic spreading because of the lack of large-scale fine-grained empirical data.

To take a topical example, consider the current H1N1 outbreak. Scientists have urged governments to map the spread of H1N1 more accurately in order to predict the number of people who may die from it [12]. Current predictions indicate that one in 200 people who get H1N1 badly enough to need medical help could go on to die, but given that vaccines may not be ready until later than hoped, accurate predictions are crucial. Any estimates about H1N1 are subject to a wide margin of error (not everyone who catches it develops symptoms). More accurate mapping of the spread of the virus must be carried out if it is to be effectively managed. Monitoring doctors and hospitals is insufficient since not everyone who is infected with H1N1 will become ill enough to report their case to a doctor.

Figure 1 shows the process of the spreading of epidemics by the mobility of humans from a subpopulation (e.g., a city) to another subpopulation. When a susceptible (S) individual is in contact with a infectious (I) individual (symptomatic or asymptomatic), it will be infected with a certain rate and enter the latent class. When the latent period ends, the individuals become infectious (i.e., able to transmit the infection). After the infectious period, all infectious individuals enter the recovered class. If an infectious individual moves to another city, the subpopulation in the new city will also be infected. Using the IATA data, scientists can roughly model the migration of population across countries. But we need much better granularity of data, instead of assuming a homogeneous mixing in each subpopulation.

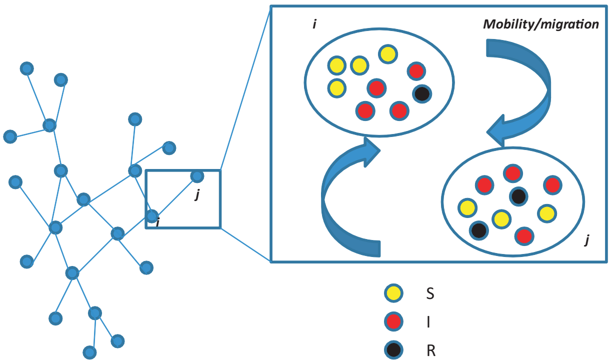

Figure 1: Metapopulation model composed of a network of subpopulation connected by mobility.

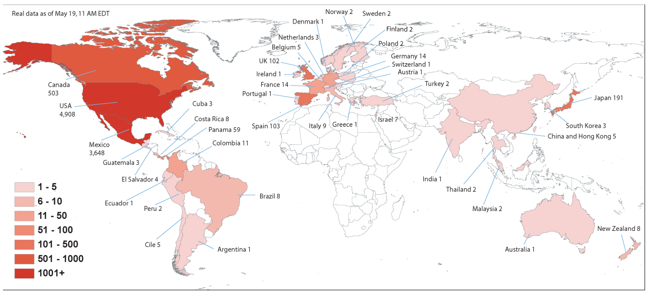

Figure 2: Confirmed number of H1N1 cases worldwide on 19 May 2009 (from GLEaMviz.org).

Consider the H1N1 epidemic in early 2009 (cf. Figure 2). It first started in Mexico and then spread to other countries by human mobility. Figure 2 shows that the worst countries besides Mexico are its neighbors (USA and Canada). Spain was the worst in Europe as it has a lot of connections with Mexico. However there is a clear need for more accurate data to build fine-grained models, capable of predicting the dynamics of disease spreading.

Mobile computing can help to fight epidemics in at least two ways:

Case 1: If we can track real-time or nearly real-time human health status, we can provide advice and precautions for each users, accurately estimate the number of asymptomatic infectious individuals, predict the spreading process, identify the hotspots of the pandemic, and effectively isolate the infectious victims. This may be possible by using a personalised epidemic software. Users can self identify their health status (e.g., cough, cold) and embed this status in a Bluetooth service. Users periodically run Bluetooth service discovery and log the devices discovered, the health status of each encountered user, and if possible also their geographical locations. Users can upload their log files to the server, which analyses results and provide effective feedback.

Case 2: If we do not have the health status of each users but only the contact log and the geographical location of certain encounters, we can understand the mixing properties of each subpopulation, model contact and mobility processes, and identify the social hotspots. With this understanding, we can accurately predict and emulate the spreading of diseases.

3. CHALLENGES IN COLLECTING DATA

3.1 High experimental cost

In general the cost of conducting large-scale mobility experiments is high. It includes equipment, software, human resources and generating incentives for people to participate. For example, for the iMote experiments, carried out by the Haggle Project [15], the development, the hardware cost, the participation incentives and the human resources spent on assembling and distributing devices and monitoring the experiments add up to $12,000 for a small-scale experiment (only 50 participants). This is clearly not scalable to experiments involving billions of people.

3.2 Privacy and government regulations

The law in many jurisdictions strictly regulates privacy and thus making large-scale data collection even more challenging [14]. Before data collection can begin, an explicit consent of participants is required, substantially increasing the administrative burden. Further, telephone operators are restricted in what customer data they can store, for how long, and for what purpose, and the dissemination of such data is even more tightly controlled. This dramatically increases the difficulty of obtaining data from operators, which otherwise is a good way to reduce collection cost and increase dataset size.

3.3 Lack of motivating applications

It is clear that giving out hardware for large-scale experiments does not scale. Instead we must rely on useful or interesting applications to motivate participation of users that already own their own hardware. For example, there are many applications developed for iPhones but no key application exists that enables large-scale data collection. An application able to scale up to millions of users while collecting data would be incredibly valuable to the research community (as well as economically!). Equal value might be obtained through many applications with smaller (but still large) user communities: it is not a strict requirement that such a large dataset consist of a single community, and indeed, it might be valuable in avoiding bias if the overall billion-sized dataset were composed of numerous smaller (multi-million sized) components.

3.4 Lack of business models

To motivate a large amount of participation, we need good business models. They can motivate operators to share their data, and users to participate in experiments. If all parties (operators, users and researchers) can benefit from participating in a system, it is more likely to succeed.

3.5 Lack of organisation

CAIDA (caida.org) exists to aid Internet traffic data collection, but there is no such organisation or group for data collection in mobile or wireless networks. The closest is CRAWDAD (crawdad.org), but that was established only to *archive* wireless data and, though it has performed this role well, it does not currently coordinate or lead data collection. An organisation for initiating, motivating, and coordinating mobile data collection would be extremely valuable. If such an organisation cannot be founded then, given the distributed and large-scale nature of the problem, crowd-sourcing might be utilised to achieve the same goal.

4. WHAT CAN WE DO?

Here we propose the following guidelines that should help researchers and practitioners to collect large-scale mobility data.

4.1 Build common research platform and novel applications

Currently there are several research groups involved in human mobility measurements [4, 6, 20, 28, 29]. We observe that more and more researchers are moving into the mobility data collection area, e.g. the field of geosocial networking research has recently become very popular. In order to motivate researchers to create a crowdsourcing effect, we propose the development of an open platform for social network and mobility experiment. Researchers can create their own OSNs for their projects by defining the fields of users profiles according to the experiment needs, e.g. name, email addresses, and Bluetooth ID. Separate projects can have different users, but the platform itself will merge the database from all projects. When a new project starts the central server informs all users about this project and invites them to participate. The user interface and format for each project are similar, and projects can be merged on the platform. The difference is that each project has a database, and manages its own data independently. This will save a lot of effort and administrative hassle when collecting and interpreting data, and conducting experiments.

In order to convince users to join those new OSNs we need novel, inciting applications. Consider an application that should help to solve a common sociological problem in metropolitan cities, i.e., the isolation. Thanks to embedded short-range radios (Bluetooth) mobile devices are able to detect other devices in proximity. In fact they can sense people we meet everyday within the radio range and also detect the duration of the proximity - this helps to notice the *familiar strangers* around us. We suggest a platform including both mobile phone software and a web-based application, allowing the users to build an OSN based on the proximity information detected. Mobile users can create a profile page on the web server by registering their Bluetooth ID. The profile page can be similar to a Facebook page, but having additional features, allowing the user to preview statistics about the people he met, and propose related strategies for subsequent encounters. The user can request addition of a particular owner of a Bluetooth ID to his friend list as on Facebook. This could open a completely new way of socializing, e.g. a user could use his mobile phone to detect someone whom he sees on the subway everyday, but to whom he is too scared to talk. This could enable him to initiate contact, while leaving the other party in control of any communication. This application scenario may seem socially unlikely in the Western world but it is a common pattern in Asia. But note that a single Asian population, however large, is also unrepresentative: many suitable applications, encouraging participation from different continents, countries and cultures, may be necessary. Recently, researchers have run experiment and collected data in the wild using Apple's App Store [23]. This is an encouraging evidence of this approach to us.

4.2 Collaborate with local government and media

Local governments are powerful entities for assisting with data collection. They can help to push applications into reality. Some governments seek to develop infrastructure and facilities to improve people's life in cities. By collaborating with these governments, we can quickly access the resources and deploy the facilities. The local media can be also a good way to gather mobility information as they are often interested in new technologies, wanting to use them in future campaign activities. For example, to market the movie *Artificial Intelligence*, an augmented reality game based on the movie, called *Beasts*, was created. The game was conceived as an elaborate murder mystery played out across hundreds of websites, email messages, faxes, fake advertisements, and voicemail messages, and involved over three million active participants.

Collaborating in such activities can gain us datasets of millions of people. The UK government for the H1N1 case can also be a good collaborator for the data collection.

4.3 Request more data from various operators

We have two ways to request data from the operators: either access to anonymised data e.g., via collaborative research projects; or full access to data as a commercial partner, e.g., by providing commercial value to the operator through data analysis. An example of the former is the access of the Google metropolitan Wi-Fi dataset [1]. This might be possible if the data can help to improve their services or provide them better revenues, e.g., understanding human mobility may help in Wi-Fi hotspot deployment and placement. For the latter approach, good examples are applications like Qiro (www.qiro.net) or SenseNetworks (www.sensenetworks.com), both of which use collaboration with operators to access location information to provide additional services to the users. Qiro uses information from T-Mobile, E-Plus, Vodafone and O2 to help users to locate nearby friends, and facilities such as bicycle rental. Recently, several research groups were able to obtain some useful data from the operators [13] [17], which add successful stories to this approach.

4.4 Leverage on existing location-based service providers

Novel and useful communication and networking applications can be one efficient way to motivate participation. For example SenseNetworks provides mobile application for real-time nightlife discovery and social navigation, answering the question: "*Where is everybody going right now?*" So far it has attracted around 100'000 users in North America. Unfortunately, as with other companies, the data are not available to the public but it seems that developing useful applications might be a viable way to collect large-scale datasets for research purpose. Another example are applications designed to encourage users to share their mobile phones [22] or calling minutes and text messages [16]. Such applications provide incentives for usage and could be used to motivate participation in experiments.

4.5 Exploit the generosity of on-line, geo-aware masses

Together with the proliferation of GPS-enabled devices and location-based services we observe that more and more people are tagging their shared content with geographical information. Geotagged data is very often public, therefore the problem related to privacy does not exist as users usually opt-in for such services. Also, the geo-scope of this data is becoming worldwide and as such it is a good source for experimenting with planet-scale mobility. There are different types of services that could be exploited for careful mobility analysis (cf. Table 1).

So far only few research groups have identified the opportunity in exploiting such data [9, 26]. What we need is a common framework for combining the data from different on-line sources, such that we will be able to create detailed mobility profiles for cities, groups (cf. Figure 3).

5. CONCLUSION

In this paper we challenge the networking community to collect planet-scale human mobility datasets. We explained why such datasets are important for networking research and how they could impact fundamental research in other academic disciplines. We identified the challenges and difficulties, and further proposed potential methods to achieve this goal.

Table 1: Novel mobility data on-line sources

Name	**Type**	**Scale** (March 2010)	**API**
Flickr (www.flickr.com)	photo sharing	>90'000'000 geotagged photos	Yes
Foursquare (www.foursquare.com)	geosocial networking	>200'000 users	Yes
Geocaching (www.geocaching.com)	GPS cache hunt	>990'000 geocaches	No
Nokia Sports Tracker (sportstracker.nokia.com)	outdoor activities	>2'500'000 users	No
Twitter (twitter.com)	micro-blogging	>5'500'000 users	Yes

We in no way claim that we have the ideal strategies for collecting and managing such data: we would go so far as to say that this is an impossible mission for a single research group. Our intent is to draw the attention of the community to this problem, enabling the collective intelligence of the whole community to be brought to bear on these crucial problems.

With these kind of datasets, we believe that we will completely change the understanding of human dynamics, potentially opening many new fields of academic study, as the availability of Internet and web data allowed the study of complex networks and systems to flourish, further impacting the understanding of biological structures.We urge the community to address these challenges to make this possible, and in doing so perhaps we can help to save the people from worldwide epidemics.

6. REFERENCES

[1] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren. Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan. In *Proc. of IMC '08*, pages 85–98, 2008.

[2] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[3] R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the world wide web. *Nature*, 401(6749):130–131, 1999.

[4] G. Bigwood, D. Rehunathan, M. Bateman, T. Henderson, and S. Bhatti. Exploiting self-reported social networks for routing in ubiquitous computing environments. In *Proc. of SAUCE 2008*, pages 484–489, 2008.

[5] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing*, 2(5):483–502, 2002.

[6] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

[7] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*, 5:34, 2007.

[8] V. Colizza and A. Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling

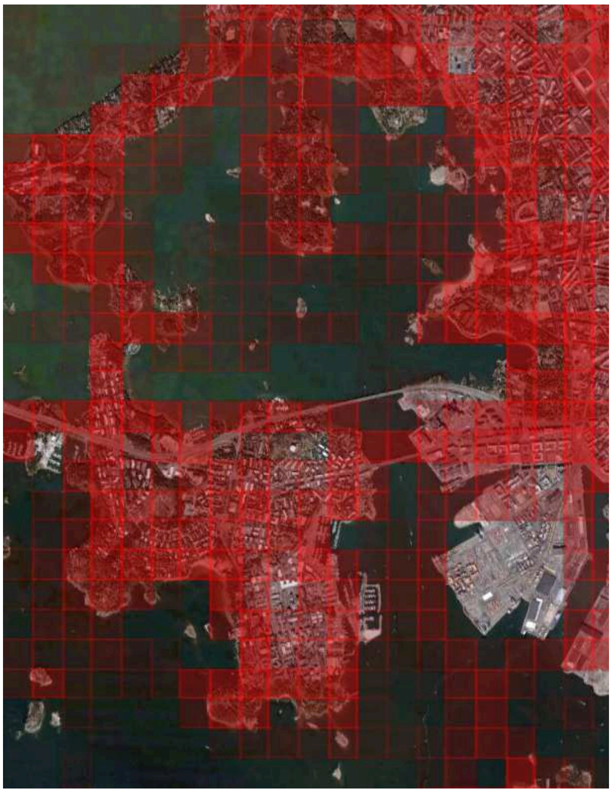

Figure 3: Place popularity among Nokia Sports Tracker users in Helsinki (zoomed view of SW side of Helsinki). Opacity level of each red-coloured cell corresponds to the number of location updates generated by athletes between July 2007 and September 2008.

pattern: theory and simulations. *Journal of Theoretical Biology*, 251:450, 2008.

[9] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[10] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, V10(4):255–268, 2006.

[11] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[12] T. Garske, J. Legrand, C. A. Donnelly, H. Ward, S. Cauchemez, C. Fraser, N. M. Ferguson, and A. C. Ghani. Assessing the severity of the novel influenza A/H1N1 pandemic. *BMJ*, 339(b2840), 2009.

[13] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[14] T. Henderson and F. ben Abdesslem. Scaling measurement experiments to planet-scale: Ethical, regulatory and cultural considerations. In *Proc. of ACM HotPlanet '09*, 2009.

[15] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: Social-based forwarding in delay tolerant networks. In *Proc. of MobiHoc '08*, 2008.

[16] P. Hui, R. Mortier, K. Xu, J. Crowcroft, and V. O. Li. Sharing airtime with shair avoids wasting time and money. In *Proc. of HotMobile*, 2009.

[17] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *HotMobile '10: Proceedings of the Eleventh Workshop on Mobile Computing Systems &#38; Applications*, pages 19–24, New York, NY, USA, 2010. ACM.

[18] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri. Towards realistic mobility models for mobile ad hoc networks. In *Proc. of MobiCom 2003*, pages 217–229, 2003.

[19] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. In *Mobile Computing*, pages 153–181. Kluwer Academic Publishers, 1996.

[20] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnović. Power law and exponential decay of inter contact times between mobile devices. In *Proc. of MobiCom 2007*, pages 183–194, 2007.

[21] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30(4):330–342, 2008.

[22] Y. Liu, A. Rahmati, Y. Huang, H. Jang, L. Zhong, and Y. Zhang. xshare: enabling impromptu sharing of mobile phones. In *Proc. of MobiSys 2009*, 2009.

[23] D. McMillan, A. Morrison, O. Brown, M. Hall, and M. Chalmers. Further into the wild: Running worldwide trials of mobile systems. In P. Floréen, A. Krüger, and M. Spasojevic, editors, *Proceedings of the 8th International Conference on Pervasive Computing*, pages 210–227, Berlin, Heidelberg, May 2010. Springer Berlin Heidelberg.

[24] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of IMC'07*, 2007.

[25] M. Musolesi and C. Mascolo. Designing mobility models based on social network theory. *Mobile Computing and Communications Review*, 11(3):59–70, 2007.

[26] M. Piorkowski. Sampling urban mobility through on-line repositories of GPS tracks. In *Proc. of ACM HotPlanet '09*, 2009.

[27] M. Piorkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser. A Parsimonious Model of Mobile Partitioned Networks with Clustering. In *Proc. of COMSNETS*, 2009.

[28] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the levy-walk nature of human mobility. In *Proc. of INFOCOM*, Phoenix, USA, 2008.

[29] V. Srinivasan, M. Motani, and W. T. Ooi. Analysis and implications of student contact patterns derived from campus schedules. In *Proc. of MobiCom 2006*, pages 86–97, 2006.

[30] E. Strano, A. Cardillo, V. Iacoviello, V. Latora, R. Messora, S. Porta, and S. Scellato. Street centrality vs. commerce and service locations in cities: a Kernel Density Correlation case study in Bologna, Italy, 2007. arXiv:physics/0701111v1.

[31] A. Vazquez, J. G. Oliveira, Z. Dezso, K. I. Goh, I. Kondor, and A. L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73:036127, 2006.

[32] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[33] J. Yoon, M. Liu, and B. Noble. Random waypoint considered harmful. In *Proc. of INFOCOM*, pages 1312–1321, 2003.