# Towards New Methods for Mobility Data Gathering - Content, Sources, Incentives

Anders Lindgren
Swedish Institute of Computer Science
Box 1263
SE-164 29 Kista, Sweden
[andersl@sics.se]

## ABSTRACT

Over the past decade, huge amounts of work has been done in mobile and opportunistic networking research. Unfortunately, much of this has had little impact as the results have not been applicable to reality, due to incorrect assumptions and models used in the design and evaluation of the systems.

In this paper, we outline some of the problems of the assumptions of early research in the field, and provide a survey of some initial work that has started to take place to alleviate this through more realistic modelling and measurements of real systems. We do note that there is still much work to be done in this area, and then go on to identify some important properties of the network that must be studied further. We identify the types of data that are important to measure, and also give some guidelines on finding existing and potentially new sources for such data and incentivizing the holders of the data to share it.

## 1. INTRODUCTION & MOTIVATION

The past decade has seen much effort being put into mobile network research. Given all this work, it is surprising to see how limited our knowledge still is of some very fundamental aspects that have great effect on such networks.

Since users are no longer stationary and connected to the network at a single point, the characteristics of links and the network depend to a large extent on the mobility and behaviour patterns of the users in the system. Unfortunately, many mobile systems are designed without taking this into account. During early research into mobile ad hoc networks, mobility models such as the random way-point model[13] were developed to be able to evaluate protocols under different levels of mobility Eventually, it was realized that many of these models did not accurately represent the type of mobility that would be seen in real situations where such networks would be deployed, resulting in protocols and systems that cannot perform properly in their intended scenarios. Thus, in order to be able to design protocols and systems that are well adapted for the settings that they will be used in, a fair amount of work has recently been put into designing more realistic mobility models as well as measuring real human mobility, sometimes feeding these measurements into the design of the mobility models[2]. There have been several projects [10, 7, 20, 1, 9], that have collected mobility and contact traces of people in different settings, as well as the CRAWDAD project [26] in which such data is archived for public use. These traces have then been used to analyse the networks they create and see how that affects system performance, including work showing the tradeoffs between using infrastructure and opportunistic forwarding to achieve different types of message delivery in a system[17, 11] and how opportunistic forwarding can significantly reduce system costs[12].

There are however some problems with the data that has been collected up to this point. The measurements have often been done with a very limited amount of participants, and in non-generalizable settings such as conference environments or university campuses. These are situations that are not typical for a majority of the population, and also tend to be in settings that usually are very well-connected. Thus, the measurement results are usually not applicable to more general, larger scale, user populations. Most measurements have also focused on measuring mobility or contact patterns of users by deploying GPS or proximity loggers to be carried by selected participants. This is a good way to collect data, but it is difficult to give incentive to users to carry these loggers, and due to cost issues, the size of the measurement is often limited. One novel approach to gain a massive amount of data on user movement was used in a recent project in which researchers gained access to logs of all trips made on a metro train system in a major city[14].

The other major problem is that most of the data contain only information about the locations of users or the contacts between users. There have been almost no work done on measuring the traffic patterns and network usage of network users. In protocol evaluations, uniformly or exponentially distributed data is often used, and what is worse, sources and destinations are frequently selected at random. In reality, there is however likely to be a correlation between a nodes location, mobility, and interaction with other nodes in the network. This is vital information in order to design network protocols properly and assign the network resources properly to the correct contexts, but it is also a more difficult thing to measure than mobility. The lack of such data is however in part due to the fact that it is hard to measure user behaviour when there is no real network with real applications there for them to use. In one of the few occasions where this have been addressed[17], real-life communication patterns were extracted from the data, and clear differences could be seen in the conclusions that could be drawn using this, as compared to using synthetic communication assumptions. Thus, it is important to gain a greater understanding to these types of user behavior.

Based on the observations above, we conclude that it is necessary to investigate new ways of understanding user behaviour in mobile


*ACM HotPlanet* '09 Krakow, Poland

networks. This should be done both by developing and deploying novel ways to collect new data, as well as studying existing data available in public data repositories. The data that can be collected, together with data collected by other projects should be analysed to provide guidelines and models to be used when designing and evaluating network protocols and applications. This paper does not aim to provide the definite answer to what data to collect and where to get it, but rather to stimulate discussion regarding new types of data and new sources of data that can be used for mobile network analysis.

# 2. DATA OF INTEREST

At present, most data that have been gathered have been mobility and contact data for urban and campus settings. While this is useful, a wide range of other types of data should also be recorded in order to be able to do a more complete analysis of the network. In this section, we outline some of the different types of data that must be measured and analyzed to be able to understand the network.

## 2.1 Mobility Data

Mobility data usually comes in either the form of traces of physical locations (with varying resolution – this can be either GPS traces[22], or contacts with infrastructure nodes with known locations[20]) or as traces of contacts between devices[4]. This is where the main focus of mobile network monitoring has been and a plethora of such information has already been gathered. A problem with the data that has been collected so far is that most data sets have been very limited in size, both in terms of the number of nodes studied, and the duration of the measurement. The data has also mainly been collected at university campuses, scientific conferences, and some urban settings. These tend to be well-connected scenarios with properties that often are quite different from the scenarios common to, for example, delay tolerant and opportunistic networking protocols, which the collected data sets have often been used to evaluate.

## 2.2 Application Usage and Data Traffic Models

So far, most protocol analysis for opportunistic networks have been made using very simplistic data traffic models. The source and destination of messages are usually selected at random, and the data arrival rate is uniform, or at best modelled by a Poisson process. As the performance of routing protocols and other network mechanisms can be highly dependent on the characteristics of the network traffic, this is of course not good. This type of data is however quite difficult to collect, in large part due to the fact that there are very few large real deployments of such opportunistic networks in which users are running real applications. Network usage is likely to be quite different in opportunistic networks than traditional ones (e.g., data generation can often be expected to be correlated with user mobility as a user is more likely to generate data when stationary than mobile). Some such information could be garnered from information about cellular network usage, but as that communication model still is more like traditional Internetworking with continuous end-to-end connectivity, it will still be different from what can be expected in an opportunistic network.

### *2.2.1 Social Network Data*

Another aspect that will influence the usage of the network is the social network between users of mobile nodes. It is plausible to believe that communication between users with an existing social relationship will be more frequent than between complete strangers (at least for end-to-end addressed data – for content centric approaches, this will be less apparent). In many cases, there will also be correlations between the social network and the mobility and contact patterns observed as users with a social relationship tend to be in physical contact more often as well. This types of correlations should be studied further to find out if data from social networking communities and similar data sets can be used to improve mobility and communication models.

# 3. FINDING THE DATA AND GAINING ACCESS TO IT

Once we have determined which types of data are of interest to us in order to better understand mobile networks, it is time to consider where such data can be found. In addition to this, it is important to consider how to collect such data, or how to provide incentives to key players in order to allow us to gain access to that data. As much of the data can be considered highly sensitive, both to operators and users, many of the issues encountered will not be technical, but will be policy-based and deal with ethics, anonymization, and trust.

## 3.1 From the Network Operators

Mobile phone operators can be a great source of information about the behavior of mobile users. As their networks are large operational business entities that turn around large amount of money each year, operators have already studied their networks to optimize their performance. They would also be able to provide network researchers with different types of data. First of all, they have information about how users move and which cells they associate with. While such mobility data is on a cell level (thus neglecting small-scale mobility), it would capture the most important aspects of mobility, and also allow for very large-scale mobility models with a large number of users, both in urban and non-urban scenarios.

In addition to mobility data, operators would also be able to provide information about social relationships between users through traffic patterns in phone call and text messages. This would be very useful both to infer how social relationships affect the mobility interactions of users, but it would be even more useful for creating more realistic traffic models for protocol evaluations.

### *3.1.1 Challenges and Incentives*

As outlined above, mobile phone operators would be able to provide a great source of data for researchers to analyze. There is however a problem. Operators are business that want to make money and run efficient services. They cannot risk making their customers upset, or reveal business secrets on their network operation to competitors.

Thus, it is not very easy to gain access to such data. First of all, there are major privacy and personal integrity concerns, as their customers might not want others to be able to track them. This can to some extent be circumvented by the anonymization of traces, but for an eager investigator, even an anonymized trace can be used to figure out information about individuals (it will be possible to figure out who some id X really represents by looking at combinations of locations and time periods - someone who spends nights in my house and days in my office is highly likely to be me; even if absolute location coordinates are removed, relative locations of popular gathering points in conjunction with knowledge of the general area where the data was gathered can still be used to infer the exact locations with high probability)[8].

Further, operators also highly treasure their mobility data for business reasons. They do not want their competitors to know how

their network is structured and how their customer base behave. Thus it is important to be able to analyze the data that the operators have, without revealing personal details about customers, or business secrets.

#### 3.1.2 An Anonymity Protecting Data Analysis Framework

In order to enable the use of the vast amounts of data that operators possess, but are not willing or able to share, we propose to define an anonymity protecting data analysis framework. This framework would specify a predefined set of operations that can be performed on the data, and researchers will then create their analysis scripts and give them to the operators who inspect them to ensure that no unauthorized data is extracted, and then the analysis tools are run by the operators, with only the final results being given out to the researchers. Depending on the level of sensitivity of the data, different levels of detail can be allowed to be extracted from it.

In future work, we will in collaboration with existing operators define the framework in more detail, with security levels that are acceptable for the operators.

### 3.2 From the Users

Given the policy problems that abound when trying to get data from operators, it can often be easier to go directly to the users to get different types of data. This is the approach that has been taken by most mobility measurement projects up to this date[10, 15, 7]. Researchers have either given users special tracking devices, or installed special software on their mobile devices, which then record data about their location (either in terms of absolute coordinates, or which base station they are associated with), and/or contacts with other mobile devices. The main problem with this approach lies in creating large deployments. If special-purpose hardware devices are to be deployed, all experiment participants must be physically met and given the device, and it also incurs a cost that grows with the size of the deployment. Since this hardware does not give the user any benefits, it is also likely that users will forget to carry it on their person at all times, creating inaccuracies in the data collected[15].

If only software is required, it is easier to reach a potentially large user-base as the software could be downloaded by users all over the world (free software made available for platform such as the iPhone or Android have been known to be downloaded and installed on hundreds of thousands of devices in a short time span, so there is great potential here). Here, the problem is to motivate why they should do that. If it only measures their mobility without any added benefit, there are multiple problems. First of all, they might not want to be tracked. Secondly, it will consume their resources. Thus, the software should provide some value adding service.

#### 3.2.1 Providing Incentives to the Users

When using software installed on the regular devices of users, it is important that this software provides some added benefit to the user for two main reasons. First of all, it will be easier to motivate users to install and run the program if they actually feel that it is useful to them instead of just consuming battery power and memory. It is important that the data stored is sufficiently secured so that users feel comfortable with this being shared. Data that is stored at the device must also be sufficiently encrypted and protected to avoid people with physical access to the device to gain unauthorized access to the data (jealous spouses etc). Authorized access can however be of interest to the user in different types of social applications (e.g., allowing a certain group of friends to know when you are in the vicinity). Depending on how the stored data will be used, different methods for data retrieval will also be needed. When mobility and usage data is only stored for future research and analysis, it is sufficient to store it on the device and upload it to a central server when the device is in contact with a local network (if such capabilities are available) or when synchronizing with a computer. On the other hand, if data is needed in real-time, considerations on how much data to send and how frequently to update it must be taken in order to reduce potential cost of cellular data communication. The types of applications that could be used are many, but obvious choices would be instant messaging services [18], social networking services [24], and mobile sensing applications [3].

Many of the applications that are proposed for delay tolerant and opportunistic networking are applications for developing countries and regions[21, 6, 19, 5]. However, all the efforts on collecting data on mobility and user behavior have been performed in developed countries, and usually in urban areas. It is to be expected that the mobility patterns and behavior of people in a poor rural area are dramatically different from those of the typical users in a rich metropolitan area. Similar problems of regulatory issues and personal integrity as in other regions must also be addressed here, but in these regions it might be somewhat easier to collect the data. If the communication system is provided for free or at a very reduced cost, and is the only way for the community to gain this type of network access, it is easier to convince the end users to assist in providing this type of data in return, especially as it is to be used to improve the performance of the systems that they use.

Secondly, this may also enable researchers to log more data that is even more useful. Not only can it be used to log for instance device proximity or cell tower associations, but also traffic patterns and application usage, which we concluded earlier in the paper is one of the most important types of data to gather.

There are however other challenges with running such tracking software on normal mobile phones. On many mobile phones, such as the iPhone, it is difficult, or impossible, to run user-installed software in the background when using the device to perform other tasks. Users cannot be expected to restart the software manually all the time and keep it running if it interferes with normal usage of the device. This is a problem that needs to be addressed by manufacturers.

### 3.3 Finding New Sources of Data

Above we have described some traditional methods for collecting data about mobile users. While these provide meaningful data, we also want to investigate alternative methods for data collection that might provide data on a larger scale or from a different perspective. In this section we study some such possibilities.

#### 3.3.1 Public Transportation

Many large cities in the world have advanced public transportation systems that transport hundreds of thousands, or even millions, of people every day. In many of these places, electronic smartcards are now replacing traditional paper tickets and this also means that it is possible to log where users enter and exit the transport system. Thus, it is also possible to estimate the path that was taken within the system, and create an idea of how the mobility in the city happens on a very large scale. This is information that will already be logged by most transportation companies as they are interested in using it to optimize the planning of train schedules. In a novel approach to gain a massive amount of data on user movement, researchers at University College London gained access to logs of all trips made on the metro train system of a major city[14]. By combining the information about the time and location of users entering

and exiting the system with train schedules and knowledge about the train network, it was possible to convert this data into a massive set of journeys that could later be used for protocol analysis.

Gathering this type of data can be quite useful, as it requires little technical effort (as the data is often already stored) and can yield mobility data that is at a much larger scale than all other methods that have been used. One must however remember that this only provides a high-level picture of the mobility in the system with a very low resolution. While it can allow for a reasonable idea on when users are at a certain station, it will not be able to provide detailed interactions between mobile users. It will also only provide at most a few data points per day and user (most users will mainly use the public transport system to get to and from work).

When building a model, it would be interesting to combine large-scale measurements such as this to create the overall flow of nodes, but also include measurements from a more local setting to capture the individual node interactions.

### *3.3.2 City Planning Authorities*

Using a similar way of thinking as when exploiting data from public transport systems, it can also be very useful to be in contact with city planners in the local government. In most newly built urban areas, much work has been put into designing the environment such that it can cope with the flow of people at different times. Thus, much studies have been made on how people are likely to move. Vukadinovic *et al* [25] made an initial attempt at using such data by using a pedestrian simulator used to dimension sites such as the Sydney Olympics to create a mobility model and investigate its properties. Other similar tools and sources of information, such as the vehicular traffic simulator used by Leontiadis and Mascolo [16], should also be investigated.

### *3.3.3 User Schedules*

Different types of calenders and schedules may often contain location information regarding where a user will be for a certain meeting. If such information could be extracted for a large number of users, it could be used to create a high-level mobility model. One such attempt was made by Srinivasan *et al* [23], in which the class schedules for all students on a university campus was extracted over a period of time, and a simplistic contact analysis could be done.

## 4. CONCLUSIONS

In this paper, we have outlined some of the problems of the assumptions of early research in the field, and some of the initial work that has started to take place to alleviate this through more realistic modelling and measurements of real systems. There is however still a large amount of work to be done in this area, as much of the data that has been collected so far has been too limited in scope, or collected from inappropriate settings. In this paper, we defined some main types of data that must be collected to gain a better understanding of mobile and opportunistic networks. Further, we noted that it is vital to find new data sources and identified some such exciting new avenues for getting useful data. Finally, we briefly discuss the problems of gaining access to the data that is out there, and how incentives can be provided to allow researchers to collect and use the necessary data, including a proposal of a anonymous data analysis framework.

## Acknowledgements

This work has been performed within the SICS Center for Networked Systems funded by VINNOVA, SSF, KKS, ABB, Ericsson, Saab Systems, TeliaSonera and T2Data.

This work has been carried out in the IST 7th Frame- work Programme Integrated Project 4WARD, funded in part by the Commission of the European Union.

## 5. REFERENCES

[1] M. Balazinska and P. Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proceedings of the First International Conference on Mobile Systems, Applications, and Services (MobiSys 2003)*, 2003.

[2] R. Calegari, M. Musolesi, F. Raimondi, and C. Mascolo. Ctg: A connectivity trace generator for testing the performance of opportunistic mobile systems. In *Proceedings of the European Software Engineering Conference and the International ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE07)*, Dubrovnik, Croatia, September 2007. ACM Press.

[3] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn. The rise of people-centric sensing. *IEEE Internet Computing Special Issue on Mesh Networks*, June/July 2008.

[4] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the performance of opportunistic forwarding algorithms. In *Proceedings of INFOCOM 2006*, 2006.

[5] M. Chetty, W. Tucker, and E. Blake. Developing locally relevant applications for rural areas: A south african example. In *SAICSIT*, 2004.

[6] A. Doria, M. Udén, and D. P. Pandey. Providing connectivity to the saami nomadic community. In *Proceedings of the 2nd International Conference on Open Collaborative Design for Sustainable Innovation (dyd 02), Bangalore, India*, Dec 2002.

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

[8] P. Golle and K. Partridge. On the anonymity of home/work location pairs. In *Pervasive 2009*, pages 390 – 397, 2009.

[9] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *Proceedings of the Tenth Annual International Conference on Mobile Computing and Networking (MobiCom 2004)*, 2004.

[10] P. Hui, A. Chaintreau, R. Gass, J. Scott, J. Crowcroft, and C. Diot. Pocket switched networking: Challenges, feasibility, and implementation issues. In *Proceedings of WAC 2005*, Oct 3-5 2005.

[11] P. Hui and A. Lindgren. Phase transitions of opportunistic networking. In *Proceedings of the ACM SIGMOBILE workshop on Challenged Networks (CHANTS 2008)*, September 2008.

[12] P. Hui, A. Lindgren, and J. Crowcroft. Empirical evaluation of hybrid opportunistic networks. In *Proceedings of the First International Conference on COMmunication Systems and NETworkS (COMSNETS 2009)*, January 2009.

[13] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. In Imielinski and Korth, editors, *Mobile Computing*, volume 353, chapter 5, pages 153–181. Kluwer Academic Publishers, 1996.

[14] L. McNamara, C. Mascolo and L. Capra. Media Sharing based on Colocation Prediction in Urban Transport. In *Proc. of Mobicom 2008*, 2008.

[15] J. Leguay, A. Lindgren, T. Friedman, J. W. Scott, and J. Crowcroft. Opportunistic content distribution in an urban

setting. In *Proceedings of the ACM SIGCOMM workshop on Challenged Networks (CHANTS 2006)*, September 2006.

[16] I. Leontiadis and C. Mascolo. Opportunistic Spatio-Temporal Dissemination System for Vehicular Networks. In *In Proceedings of the First International Workshop on Mobile Opportunistic Networking (ACM/SIGMOBILE MobiOpp 2007). Colocated with Mobisys07*, Puerto Rico, USA, June 2007.

[17] A. Lindgren, C. Diot, and J. W. Scott. Impact of communication infrastructure on forwarding in pocket switched networks. In *Proceedings of the ACM SIGCOMM workshop on Challenged Networks (CHANTS 2006)*, September 2006.

[18] A. Lindgren and A. Doria. Experiences from deploying a real-life dtn system. In *Proceedings of the 4th Annual IEEE CONSUMER COMMUNICATIONS and NETWORKING CONFERENCE (CCNC 2007)*, January 2007.

[19] 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 *Proc. of ACM WiNS-DR*, September 2008.

[20] M. McNett and G. M. Voelker. Access and mobility of wireless pda users. *Mobile Computing Communications Review*, 9(2):40–55, April 2005.

[21] A. Pentland, R. Fletcher, and A. A. Hasson. A road to universal broadband connectivity. In *Proceedings of the 2nd International Conference on Open Collaborative Design for Sustainable Innovation (dyd 02), Bangalore, India*, Dec 2002.

[22] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the levy-walk nature of human mobility. In *INFOCOM*, pages 924–932. IEEE, 2008.

[23] V. Srinivasan, M. Motani, , and W. T. Ooi. Analysis and implications of student contact patterns derived from campus schedules. In *Proceedings of MobiCom 2006*, 2006.

[24] N. Vallina-Rodriguez, P. Hui, and J. Crowcroft. Goose: Social network services for developing worlds. To appear in Proceedings of the First Extreme Workshop on Communication (ExtremeCom 2009), August 2009.

[25] V. Vukadinovic, O. Helgasson, and G. Karlsson. A mobility model for pedestrian content distribution. In *Proceedings of Simutools 2009*, Mar. 2009.

[26] J. Yeo, D. Kotz, and T. Henderson. CRAWDAD: A community resource for archiving wireless data at dartmouth. *ACM SIGCOMM Computer Communication Review*, 36(2):21–22, April 2006.