Predictability of Human Mobility from Highly Granular Location Data

Rafaela-Ioana Voiculescu

Kongens Lyngby 2014

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3351
compute@compute.dtu.dk
www.compute.dtu.dk

Summary

TODO - The goals of this thesis is to..

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling at the Technical University of Denmark in fulfilment of the requirements for acquiring an M.Sc. in Informatics.

The thesis deals with ...

The thesis consists of ...

Lyngby, 01-August-2014

Rafaela-Ioana Voiculescu

Acknowledgements

I would like to thank my....

Contents

Chapter 1

Introduction

TODO

Chapter 2

Related work

There is a high interest and a huge amount of work the scientific community dedicates to understanding the patterns of human mobility. The knowledge we can gain from the results of this work has the potential to benefit a wide variety of industries from the modeling and maintenance of the transportation infrastructe, to the medical industry where we can use this knowledge in trying to prevent the spreading of epidemies. [DB08]

Various studies have been conducted in order to gain a better understanding of the human mobility patters. These studies give us results that seem to support each other in the idea that people are less spontaneous than they would like to think themselves and that, indeed, our behaviour shows that we are quite rooted into habbits when it comes to the way we travel.

2.1 Mobility patters uncovered by the disipation on bank notes

Brockmann, Hufnagel and Geisel[DB06] have analyzed the human movement based on the way bank notes were dispersed through the United States (excluding Alaska and Hawaii). Their study shows that a relatively small percentage

of bank notes (23.6%) traveled for more than 800 km, while a fraction of 19.1% did not traveled for more than 50 km even after a year of being observed. The possible explanation the authors have given for these findins are that, in general, people would be less inclined to leave the areas of the large cities or the places they usually conduct their lives.

The problem identified with this approach for tracking individuals is that the bank notes exachange hands and the behaviour which is identified by the way they circulate can't be attributed to a single individual, but rather to different ones that at any moment have had the bank note in their posession. Despite this, the result have a high scientific value as they do identify patterns in human travel behaviours in general.

2.2 Mobility patterns of mobile phone users

A. L. Barabasi, M. C. Gonzalez and C. A. Hidalgo have conducted a study [MCG08] that deals with studying the trajectories of over 100000 mobile phone users with anonymized identities. The study was conducted in order to see if there are any patterns in our mobility habbits. Among the things that have been subjected to testing was the return probability of individuals in the same place as in the past. The study shows there is, in general, a peak in the return probability after 24, 48 or 72 since they have left a particular location. This shows that we humans tend to visit locations periodically. This can be explaineQd by our going to places such as work, school, grocery shops near our home etc.

The authors have also ranked the locations the mobile phone users frequented based on the number of times they have been spotted nearby. The results for this have shown that the probability of finding someone near a location that is ranked for them with a level L can be estimated with $1/L$. Another interesting finding that is mentioned in the paper is that, in general, people seem to be spending the majority of their time in just a few locations, while diving the remaining time just between a limited number of locations that varies for the subjects from as low as 5 to around 50.

There are some note worthy plots that the authors present in the paper. They can be seen in figure and they show that most people travel over short distances, yet there is a small number of people that regularly travel over big distances.

The results of this study are a major indicator that individuals display a high level of regularity and that we have a tendency to spend most of our times in places that are familiar to us, or that require us to visit them regularly (e.g.

home, work).

2.3 Mobility patterns in massive multiplayer online games

R. Sinatra and M. Szell have studied the way in which users of a massive multiplayer online game behave inside the virtuale universe provided by the mentioned game [RS14]. It has been established that the massive multiplayer games provide people with a virtual reality where they can interact with others through their characters and can, in fact, form groups and, as such, display both individual as well as collective behaviour actions that can translate to the non-virtual world [Bal03].

This study gives an interesting insight into the habbits and actions of the characters which are controlled by the players. Among the things the authors have analyzed are the predictability of the characters, the entropy generated by the mobility of the characters in the virtual universe and general strategies or patterns that could be observed.

The game the authors have been using for the study is called Pardus [Par]. This game is quite complexe, as it allows the manifestation of normal real-life activities such as the creation of alliances or friendships, communication between the players, economic related action, or even actions which have a negative conotation such as attack of another user, removal of a friendship link etc. The universe of the game consists in hundrets of nodes which represent cities or sectors in the game. These virtual cities are tied to each other through links which mark the posibility for the users to move their characters from one place to another.

By analyzing the why in which characters have interacted through the years, the authors have observed that the mobility of the characters through the univers is highly predictable, as users in general will seem to be chosing a random location to visit next in just about 10% of the cases.

2.4 Eigenbehaviours

N. Eagle and A. S. Pentland analyze data of individulas and communities with the purpose of trying to predict and cluster the daily habbits and behaviour of

people [NE09]. The consider that the behaviour of one person throughout a day can be close to a sum of their primary eigenbahaviours throughout that day. The results of the study have shown that when having a weighted sum calculated for the first half of a day, the behaviour of the same person throughout the remaining of the day can actully be approximated with 79% accuracy.

The results have applicability in more fields, as they allow us to consider the possibility of clustering people into various communities based on the similarity of their behaviours. It goes even further, as the findings show that this enables the possibility of calculating similarity for groups as well and thus permiting the a classification that, according to the experiment, can be 96% accurate for determining affiliations in the social network of a particular population.

As a last observation in the paper by N. Eagle and A. S. Pentland it is stated that eigenbehaviours can be used in order to identify the possible friendship ties between people. The observations in this paper have been done based on the Reality Mining dataset that tracked the behavior for 100 individuals at MIT for the duration of one year.

2.5 Human movement recorded through real traces

Studies as the ones with the travel of bank notes or the recorded location of mobile users through telephone is not very exact and does not reflect the real traces for the people. They do provide a very useful estimation, however with the technology that we have access to nowadays, we are able to record mobile phone users' real traces either through GPS or Wifi. The data that can be acquired through these means allows us to conduct studies that can take into consideration a very good approximation of the real location of individuals.

In the paper by M. Kim, D. Kotz and S. Kim [MK06], the authors present us with a method in which the locations of users can be estimated based on the WiFi signals that their devices register. The experiment is conducted cosndiering the data for a duration of 13 months. The user traces that have been used consist of the trace data from the Darmouth College. The mobility traces are defined as the lists of access points that are associated to a user's devices at a given timestamp.

The mobility traces allowed the authors to extract the tracks (locations) of the users. They have explored three methods in which the location can be extracted from the data. The first approach presumed the calculation of the center (intersection of medians) of the traingled defined by the past three access

point associations of the mobile device of the user. This approach has a downside since the devices do not necessarly change the associations in a periodic manner. This lead to the second approach which consisted in considering a time window after which the associations needed to be updated in case new associations have appeared during that time. The thrid and last approach explored the use of Kalman filters [Kal60].

The validation the path extarctors the authors have compared the results with GPS data. This validation has proven that the type of the used device has at the moment a significant importance in how acqurate the results can be as it seems that some devices can be more aggressive in updating the associations with access points while others try to stay associated with the same access points as long as possible before switching to new ones. This leads to problems as different distances between users and access points considered by different devices and as such it affects the estimated paths. The best estimations have been given in this experiment by the approach that used the Kalman filters, however both the other two appraoches have provided fairly good estimations as well.

Another paper which explores the travel patterns from real data is the one written by T. S. Azevedo, R. L. Bezerra, C. A. V. Campos and L. F. M. de Moraes [TSA09]. The authors propose another approach for analyzing the mobility of people. They take into consideration the following movement components: velocity, acceleration, direction angle change and the pause time and they are using the GPS data in order to estimate the locations of individuals. The experiment takes place in a park in Rio de Janeiro and is done based on the data received from around 120 volunteers. The results have shown that people seem to have in general smooth trajectories without abrut changes.

2.6 Entropy and predictability

One step further from understanding the way we travel from place to place is to predict our future locations based on a previous knowledge our our past patterns. There has been an extensive study done in this area of the scientific playground as well and the results which have emerged up until now are remarcable.

In the paper by C. Song, Z. Qu, N. Blumm and A. L. Barabasi [CS10], the authors take up the challenge of studying how predictable people can be. They analyze the mobility patterns of mobile phone users and calculate the entropy of these users. The locations are defined by the telephone towers the users are encountering at hourly intervals and the trajectory of the user is given by the

ordered sequence of these towers. The real entropy of each user i is calculated as $\Sigma_{T_i' \subset T_i} P(T_i') log_2(P(T_i'))$, where $P(T_i')$ represents the probability of encountering a time-ordered subsequence T_i' in the sequence of hourly encountered telephone towers T_i.

The results for this particular study show that, for the considered users, the uncertainty of where they could be at a certain moment, based on the real entropy calculated for them would be very low as they would most probably be in one of two locations.

The authors also take a look into the maximum predictability which can be expected for a user. Their results show that, with the right algorithm, a user's future location can be predicted with between $80 - 93\%$ accuracy. This shows that we are less sponatenous than we might think and that our mobility patterns are, in most cases, rooted into a very well established routine.

There have been numerous other methods or experiments conducted in order to analyze or to forcast human mobility patterns. Some of these methods include the Markov chain models [Ros09] [GL96], the neural networks [SCL03] or the Bayesian networks [AS07] as well as some that work with finite automaton [JP04]. Most of the studies support the idea that people's actions and travel behavior is indeed far from being random and thus the science world needs to dedicate further effort and time in order to use this knowledge in order to improve our quality of life and the world we live in.

Chapter 3

Prerequisits and tools

In order to research the way in which people travel we firstly need to have access to a database of information that can be used for this purpose. As it was mentioned in Chapter 2, scientists have been trying in numerous way to identify and work with location information. During our study, we have dedicated our time in working with information about the access points that were visible to the users' mobile phones throught their day. This has allowed use to implement and analyze different ways in which locations can be extarcted from such information.

3.1 SensibleDTU

The data we are using is part of a large-scale study that aims to make observations based on the lives of volunteering students - the Copenhagen Network Study. The data is colected from a variety of sources. Some of them require the volunteers to interact with the system through questionnaires and others track them automatically through their smartphones. The aim of this project is to offer an extensible framework for different studies. The deployments from 2012 and 2013 are based at the Technical University of Denmark and are named SensibleDTU [AS14b].

The students that consented to being volunteers for this ambitious project have received smartphones that are able to track different aspects of their lives and through which they can interact with the system. The big number of volunteers [1] has allowed the gathering of a considerable amount of data regarding the mobile phone users' behaviour.

The data gathered for the SensibleDTU experiment consists in data gathered through questionnaires [2], Facebook data [3], sensor data, qualitative data and Wifi data.

Since the majority of the collected information about the students is sensitive [AS14a], keeping the data secure is and has been a top priority from the beginning of the experiment. The data is annonymzed and stored securely and the students that are part of the experiment have access to tools that allow them to see what data are they sharing, what it is done with this data and that allow them to control how much they want to share.

3.2 Implementation tools

Before starting the work on the present research, we have overviewd possible tools that can be useful in our work.

The scripts that are used for analyzing, transforming and working with the data are developed in Python. The reasons behing using Python instead of any other programming language are numerous. Python is elegant and simple to use, it allows fast development and the code can be easily adapted and reused. Due to its high scalability, it is the perfect choice for both large and small projects, being easily extensible at the same time. Another very important reason for using Python is that there is a large number of libraries that can be used with it and that allow the visualization or handling of big data. [4]

An additional tool that has been used for the present project is Gephi [Gep]. Gephi is a platform that allows the exploration and handeling of various net-

[1] During the second iteration, there have been deployed aproximately 1000 smartphones to students who wanted to take part in the study.

[2] A survey was presented to the participants in 2012 consisting of over 90 questions. In 2013 an addition of over 300 questions were ask per participant. The questioned targeted different aspects from working habbits and various socio-econimic factors to Big Five Inventory measuring personality traits [JS99] and self-esteem.

[3] Participants have the option of allowing the gathering of Facebook data such as friendships and various interactions such as likes, statuses etc.

[4] Examples of libraries and packages used: numpy, matplotlib, pickle, datetime, sympy etc.

works and graphs. Further information on how this tool has proven helpfull can be found in Chapter 5.

Chapter 4

Data analysis and clean up

The present project uses data that has been selected from the database of the SensibleDTU experiment. The data is fully anonymized and the users that have been a part of the study have been chosen randomly from the database.

4.1 Data statistics

We use the data collected from 131 users from the SensibleDTU database. The students that have been selected for the present study had data collected for a period of almost a year. [1]

The application that is installed on the smartphones of the students who are part of the experiment is configured to scan periodically (around every 15 seconds) for Wifi networks, however, it is also set to record the scans which are triggered by any of the other applications that are present on the mobile phone.

[1] The starting time of collection for the 2012 deployment of SernsibleDTU is October 1^{st} 2012 and the end is September 1^{st} 2013.

4.2 Wifi and GPS data

For the present study we are not using all the fields that are accessible from the database of collected information. The study's aim is to analyze the predictability and patterns in the human mobility and as such we need information that can tell us the locations of the users that are part of the study. For this we are accessing information about the Wifi associations for the selected group of users. The results reagrding the users' locations over time are afterwards compared with recorded GPS locations and as such we are accessing this information from the database as well.

The user (first) field gives us information about what user we are currently observing. The real identities of the users are concealed and replaced by an ID which is unique for each of them.

The timestamp (second) field gives us information about the moment of time at which the scan occured and for which the information is gathered. The time format is Unix timestamp. [2] This timestamp can be easily manipulated and converted to any other timestamp format in Python by using the datetime module that can be found in the Python Standard Library [PSL].

The ssid (third) field stands for Service Set Identifier and it represents the unique ID that can be used in order to identify the wireless networks. This identifier is responsible for the correct sending of data when multiple wireless networks overlap.

The bssid (forth) field stands for Basic Service Set identifier and it represents the MAC address of a wireless access point.

The rssi (fifth) field stands for Received Signal Strength Indication and it represents the strength for a signal picked up by the mobile phone from an access point. The rssi values in our case are registered as the real signal strength recorded in dBm and are therefore negative values. As such, the signal is stronger when the value recorded for it is closer to 0.

The context (sixth) field is based in the ssid and it translates to the possibilities presented in

[2] The Unix time stamp represents a way in which time can be tracked as the total number of seconds starting from January 1^{st}, 1970 at UTC and a particular date and time.

4.3 Noise elimination

1. why working with wifi data;
2. data structure and fields;
3. how much data was analized
4. how was the data prepared (noise elimination)

Chapter 5

Locations

1. previous studies on what does a location is considering wifi fingerprints
2. analazying what can determine a location fingerprint
    - Plot access points' presence over time considering their signal strength
    - Plot number of samples of each access point over time
    - Plot the average signal strength for various time windows for each access point identified for a user
    - Plot the running average signal strength for 2,5 and 10 min time windows for each access point identified for a user
    - Plot of access point presence in time bin (5 mins used for time bin) without considering its signal strength
3. identifying locations
    - Networks
    - Hidden Markov Models
    - Further improvements
4. matching locations locations
    - Percentage similarity

- Keeping track of previous locations
- Creating fingerprints

5.

Chapter 6

Entropy and predictability

1. Calculating entropy for users
2. Calculating predictability for users
3. Observations

Chapter 7

Comparing results with GPS data

Chapter 8

Results and observations

Chapter 9

Future work

Chapter 10

Conclusions

Appendix A

Appendix

Appendix ...

Bibliography

[AS07] Sherif Akoush and Ahmed Sameh. Mobile user movement prediction using bayesian learning for neural networks. In *Proceedings of the 2007 international conference on Wireless communications and mobile computing*, pages 191–196. ACM, 2007.

[AS14a] Alex Pentland David Lazer Sune Lehmann Arkadiusz Stopczynski, Riccardo Pietri. Privacy in sensor-driven human data collection: A guide for practitioners. *CoRR*, abs/1403.5299, 2014.

[AS14b] Piotr Sapiezynski Andrea Cuttone Mette My Madsen Jakob Eg Larsen Sune Lehmann Arkadiusz Stopczynski, Vedran Sekara. Measuring large-scale social networks with high resolution. *CoRR*, abs/1401.7233, 2014.

[Bal03] P. Ball. The physical modelling of human social systems, 2003.

[CS10] 1 2 3 Nicholas Blumm 1 2 Albert-László Barabási 1 2 * Chaoming Song, Zehui Qu. Limits of predictability in human mobility. *Science*, 327, 2010.

[DB06] T. Geisel D. Brockmann, L. Hufnagel. The scaling laws of human travel. *Nature*, 439, 2006.

[DB08] F. Theis D. Brockmann. Money circulation, trackable items, and the emergence of universal human mobility patterns. *Pervasive Computing, IEEE*, 7, 2008.

[Gep] Gephi - the open graph viz platform.

[GL96] G. Maguire Jr G. Liu. A class of mobile motion prediction algorithms for wireless mobile computing and communication. *Mobile Networks and Applications*, 1, 1996.

[JP04] W. Trumler T. Ungerer L. Vintan J. Petzold, F. Bagci. Global state context prediction techniques applied to a smart office building, 2004.

[JS99] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems, 1960.

[MCG08] A. L. Barabasi M. C. Gonzalez, C. A. Hidalgo. Understanding individual human mobility patterns. *Nature*, 453, 2008.

[MK06] S. Kim M. Kim, D. Kotz. Extracting a mobility model from real user traces. *The IEEE INFOCOM Proceedings*, 2006.

[NE09] A. S. Pentland N. Eagle. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63, 2009.

[Par] Pardus game.

[PSL] The python standard library.

[Ros09] S. M. Ross. *Introduction to probability models.* Academic Press, 2009.

[RS14] M. Szell R. Sinatra. Entropy and the predictability of online life, 2014.

[SCL03] H. C. Lu S. C. Liou. Applied neural network for location prediction and resources reservation scheme in wireless networks. *International Conference on Communication Technology Proceedings*, 2, 2003.

[TSA09] C. A. V. Campos L. F. M. de Moraes T. S. Azevedo, R. L. Bezerra. An analysis of human mobility using real traces, 2009.

[XL13] N. Bharti A. J. Tatem L. Bengtsson X. Lu, E. Wetter. Approaching the limit of predictability in human mobility, 2013.