# Estimation Activity In Urban Areas Using Passively Collected Data - A Case Study with Smart Street Sensor Project.

Balamurugan Soundararaj
Department of Geography
University College London

I, Balamurugan Soundararaj confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

*Doctor of Philosophy*

2019

# Acknowledgements

- Supervisors - Colleagues - Family and Friends - OpenSource projects (esp. GNU Linux and R)

# Abstract

- data is everywhere but no information - need to understand population in detail - this research connects this two quests - Wi-Fi probes in built environment = detailed footfall information - we conduct experiments collect two sets of data. - we process the data with different methods to get footfall - we establish the application of the data with series of examples

# Impact Statement

- we live in era of explosion of data. - everyone is looking for ways to use data - this research looks in to one such data set - deals with collection, processing and application. - we develop a open-source toolkit for doing all the above - the research resulted in publications - Outputs with C.D.R.C and subsequent use of the data - methods communicated to data partner for industry application - (Gandomi and Haider, 2015)

# List of Outputs

1. Amsterdam Conference
2. Mexico Conference
3. Data natives
4. Retail futures Conference
5. C.D.R.C footfall dashboard
6. C.D.R.C footfall indicator
7. C.D.R.C footfall atlas
8. I.J.G.I.S paper
9. Transfer Entropy Paper
10. Medium data paper
11. GISRUK 2017
12. GISRUK 2019
13. Workshop talk - Tank
14. C.D.R.C book
15. Humans book
16. Footfall package R (to be done)

# Contents

**1 Introduction** **9**

**2 Literature Search** **11**

2.1 Outline . . . . . . . . . . 11

2.2 Major themes . . . . . . . . . . 11

**3 Data Collection** **21**

3.1 Wi-Fi Specification . . . . . . . . . . 21

**4 Data Processing** **22**

**5 Applications** **23**

**6 Conclusions** **24**

**A Software** **25**

**B Code** **26**

**Bibliography** **26**

# List of Figures

# List of Tables

# Chapter 1

# Introduction

We talk about the theory of cities and built environment. We start from how these have been perceived as function of the form and gradually changed to people, activity, economy and information. Built environment is manifestation of information exchange that happens in them. We talk about the change in theory regarding this. We talk about how this information exchange has been becoming more and more open and the opportunity it provides us planners, geographers and researchers to understand these things better.

Along with the information age there is an explosion of open data. The data collection has changed from structured high effort activity to low effort scraping activity. The data generated by scraping is unprecedented and staggering. Most of the Big-data research has gone into this in the past decade. Disadvantages of structured data which the unstructured data fills. This is changing how we view, understand and experience the world. Some of these datasets fall into this unique medium size category as well which are neither big data nor trivial. There is a need for methods and tools to collect, convert and use these data.

Talk about the ubiquity of the mobile technology. Everyone has a device which connects them with world wirelessly. Major ones cellphone and Wi-Fi Wi-Fi is uniquely placed in between Cellphone and Bluetooth. The design of Wi-Fi gives us amazing opportunity. This has been done before for the past decade by loads. The privacy advocacy has become a new thing. The change is from both ends. Collectors are regulated, cellphones are getting sneakier. Need for method to collect data and analyse it without compromising on privacy. The conversion of this unstructured data into something tangible and measurable is not a trivial problem. There are loads of such data and measurements. Examples - banking vs economic activity, oyster card data vs movement.

The potential use of such information is immense. Give examples of smart city paradigm and connected city where real time census is possible. We can not only take snapshot of the state of the city, we can record and understand the built environment as living, breathing organism. The insights we get by combining this information with other similar info is more than sum of their parts.

It can revolutionise understanding , planning, policy etc, urban management and finally industry such as retail, transportation etc. city mapper, sharing economy etc.

# Chapter 2

# Literature Search

## 2.1 Outline

The understanding of scale, nature and dynamics of distribution of population in space and time has been the central premise of research in various academic fields such as human geography, sociology and urban planning and is extremely critical for practical decision making in various industries such as real estate, retail and emergency management The major challenge lies in the collection or estimation of detailed and granular data that is as precise and as accurate as possible without being disclosive The proliferation of personal mobile devices has generated considerable interest for research in the past two decades by opening up unprecedented avenues in gathering detailed, granular information on people carrying these devices The general technology landscape that supports this device ecosystem has also been constantly evolving along with the users' acceptance to collection of such data (Kobsa, 2014) In this section we conduct a detailed survey of literature in this topic and we seek the hierarchy of ideas explored and the evolution of techniques and technology used in the previous literature.

## 2.2 Major themes

There have been approximately 250 academic publications concerned with the high-resolution quantification of the spatio-temporal dynamics of urban movements It is clearly a multi-disciplinary topic that has been studied extensively under the fields of geography, urban studies, urban planning and management, emergency planning and management, economics, computer science and engineering, security and transportation planning Fig 1.1. shows the hierarchy themes as a tree map where the size shows the volume of research published We observe that the previous work can be classified into 5 major areas - theory and methods, applications in studying mobility, applications in studying population, privacy and other areas which include de-anonymization, spatial classification, social networks and visualisation.

Traditional and modern geography were usually dominated by the study of centrally collected data acquired through extensive field surveys and remote sensing In the last two decades, a significant paradigm change has been introduced by the availability of unprecedented amount of data generated by unconventional sources such as mobile phones, social media posts etc This move to the postmodern geography (Soja, 1989) has been accompanied by a change in our understanding of the built environment and human geography, from a static point of view to a more dynamic definition based on the bottom up mechanisms which manifest in them, such as economic activity and information exchange (Batty 1990, 1997, 2012, 2013a,b) This transition into the digital age (Graham, 1999; Tranos, 2012; 2013) has changed the politics of space and time (Massey, 1992) and been more pronounced in the study of urban built environment where technology has redefined the concepts of place and space (Graham, 2001; Sassen, 2001; Graham, 2002) With the ability to collect and analyse of data on large complex systems in real-time (Graham, 1997), we are exploring the possibilities of understanding their structure and organisation using concepts of complexity theory (Bettencourt, 2013; Portugali, 2012) with more emphasis on their temporal patterns such as the argument towards finding the pulse of the city (Batty, 2010) With the population getting more and more connected (Castells, 2010), the nature of space/place is being dynamically defined by the population themselves (Giuliano, 1991) and vice versa (Zandvliet, 2006) This movement to the digital era was accompanied not only by optimism in its potential (Thomas, 2001; Nature, 2008) but also by the questions raised on the challenges in handling the diverse, large scale, non standardised data (Miller, 2010; Arribas-Bel, 2014 a) it produces and the usefulness or representativeness of the resulting analysis Nonetheless, availability of such data has impressive uses in urban studies (Bettencourt, 2014) especially with advancement of new technologies (Steenbruggen, 2015) and possibility of distributed, crowdsourced data collection (Lokanathan, 2015).

Visualising the temporal dynamics of data collected on human activities through ecentralised proces es poses significant challenges when approached with traditional cartographic concepts (MacEachren, 2001 Hallisey, 2005). Digital Literature media especially animation has been explored as an option to solve for the temporal dimension (Morrison, 2000; Lobben, 2003) but is bound by the cognitive limits of the viewer (Harrower, 2007). There have been approaches proposed around animations of generated surfaces (Kobayashi, 2011) and network-based visualizations (Ferrara, 2014) leaving gaps in research for new methods in dynamic geo-visualisation (Fabrikant, 2005) and visualising path and flow of phenomena (Thomas, 2005). This provides us with a promising opportunity for research in methods for visualising high frequency, hyper-local pedestrian data within the limits of cognition of the viewer.

As we can see from Figure 2.1, population studies and urban mobility are the most significant themes in this area of research. Population studies include interpolation and estimation of population at small scales by using either traditionally available large-scale data (Sutton, 1997, Yuan, 1997) or newly available mobile device based data (Yuan, 2016) or a combination of both (Rao, 2015). Urban mobility studies include collection and study of trajectories of mobile devices to understand human mobility at small scales and traffic and transportation studies at large scales. Urban mobility research significantly benefited from the decentralised collection of granular data (Castells, 2000) and its augmentation through traditional models of travel behaviour (Janssens, 2013). The large volume of research done under these themes is discussed in detail in the section on techniques and technologies.

The rise in personal technology has many opportunities for researchers and industry but at the same time has increased the general concern on privacy (Saponas, 2007; Krumm, 2009). There is immense value in uniquely identifying and profiling information on people for specific purposes such as security (Cutter, 2006) and law enforcement (Dobson, 2003) but also has extreme risks associated when not handled with care (VanWey, 2005). Strictly protecting personal information while ensuring the information is usable for research by maintaining the uniqueness in the data is the major concern which leads us to frameworks for secure practices in confidentially collecting and using the location data (Duckham, 2006; Tang, 2006; Lane, 2014). Some efforts seek to accomplish this through cryptographic hashing algorithms (Pang, 2007) while others aim to thwart identification and tracking at the device level by techniques such as MAC randomisation (Gruteser, 2005; Greenstein, 2008). The consent of users' for the collection and use of such information from their mobile devices is low but with smart devices becoming ubiquitous there is a significantly improved acceptance when the process offers value in return such as discounts and monetary benefits (Kobsa, 2014).

Chronology Figure 1 shows the volume of research done in this topic throughout years categorised based on their major themes discussed above. We can observe that there are distinct trends in the research over time, with the explosion of interest in the last decade. The research was mostly centered around population studies on interpolating larger datasets. The period between 1990-2000 there was interest in potential of the new data generated by the digital age this coincided with mobile phones becoming more popular and ubiquitous with population in urban areas. The next 5 years equal interest is observed in applying the data for population and urban mobility studies and in the development of theory, methods to use the data. Between 2005 and 2010 the 'mobile era' saw significant rise in the volume of research, especially research focused on urban mobility and localisation and identification of the unique devices from the massive cellular network data which was met with an equal interest in privacy and data security.

In 2010 there was a clear increase in the volume of the research concerned with urban mobility, especially on tracking devices for trajectories. This might be due to the emergence and proliferation of 'smartphones' around that time, which made collecting data from these devices directly much easier and less reliant on carrier provided datasets. We also observe a focus on inferring the nature of the spaces these devices occupy and the social interactions between those who own these devices.

With the theoretical limit to predictability in human mobility quantified, the focus on urban mobility has been declining in the past few years. This has been replaced by a renewed interest in population studies at a real-time, hyper-local level. We also see a recent increase in interest in deanonymization in response to the industry adopting anti-tracking mechanisms in their products for example MAC randomisation in iOS devices. Currently the research interest and gaps in the field concern the hyper-local estimation of population and nature along with methods to identify unique devices at this scale.

Techniques and technology When we look at the literature from the technology perspective, we observe that the continuous application of recent technological developments in the pursuit of understanding the distribution of human activity and population spatially and temporally over the past two decades. The distribution of the research in terms of the main technique/ technology used over the years is shown in Figure 1.4 and the total distribution is shown in Figure 1.5. We observe that the earliest attempts started from the exploration of using interpolation and modelling techniques on the available coarse data and as the need for more granular datasets increased there were attempts to devise and utlize bespoke solutions to generate them. When mobile devices became mainstream, the focus shifted to utilize the relevant components of the mobile infrastructure. A significant number of studies were done in utilising data collected from the mobile network, sensors in the mobile devices, especially GPS and WiFi, in addition to the social media content generated from these devices. A detailed account of these studies is given below,

Interpolation and Modelling Attempts in using the existing data collected through traditional methods such as census and large scale sample surveys to create spatially and temporally granular and detailed estimates were carried out by applying various interpolation methods such as pycnophylactic, dasymetric interpolation (Tobler, 1979; Mennis, 2003; Mennis, 2006; Hawley, 2005; Tapp, 2010) along with spatial (Lam, 1983; Martin, 1989) and temporal interpolation techniques (Glickman, 1986). These methods along with supplementary data such as remote sensing imagery (Sutton, 2001; Chen, 2002) and street networks (Reibel, 2005) were shown to be useful in producing detailed granular population maps at various scales with varying degree of success (Dobson, 2000; Bhaduri, 2002; Dobson, 2003; Bhaduri, 2005; Bhaduri, 2007). These approaches have been employed in various applications such as econometric studies (McDonald, 1989), studies on public health (Hay, 2005), emergency management (Kwan, 2005) and flood risk estimations (Smith, 2016). In addition to these interpolation techniques classic modelling techniques can also be used to estimate daytime populations and demographic structure at hyper local scales (Jochem, 2013; Jia, 2014), urban scales (Alahmadi, 2013; Abowd, 2004) and regional scales (Foley, 1954; Schmitt, 1956, Singleton, 2015). The granular data created with such modelling techniques are shown to be useful in urban planning and management (Parrott, 1999), emergency management (Alexander, 2002; Cutter, 2006) and in modelling traffic and transportation (Lefebvre, 2013). These interpolation and modelling techniques along with granular data produced are also used in classifying spatial areas and hence understanding the structure of cities in general (McMillen, 2001; McMillen, 2004; Lee, 2007; Arribas-Bel, 2014 b). Though being useful, these techniques are still shown to have limitations and uncertainties (Nagle, 2014), which mostly arise from the nature of the input data employed. This leads us to the need for more detailed and frequent collection of data.

Bespoke technologies Following this need, there has been efforts to use bespoke/specialised technologies such as cameras (Cai, 1996 Heikkilä, 2004 Kröckel, 2012), Lasers (Zhao, 2005 Arras, 2008) and radio frequency receivers  (Bahl, 2000 Yang, 2013 Chothia, 2010 Bulusu, 2000 Dil, 2011) to measure human activity. But the major problem with such solutions is the cost and effort involved in implementing them at large scales. Moreover, being specialised and centralised they tend to be challenging to maintain and update. In the addition, the rise of mobile phones as ubiquitous personal devices for the broader population has made them a viable alternative for collecting such data with greater granularity at large scales.

Mobile infrastructure consists of both the 'network part', built and managed by the service providers, and the 'user part', which is the phones owned by the users' themselves. The network part, in addition to providing connectivity to the users, also collects information on these devices actively (calls, messages) and passively (tower to tower handover). The mobile devices themselves have a variety of sensors (accelerometer, compass, barometer etc) and capabilities (cellular, WiFi, NFC etc) that can be sources of data themselves. With the growth of mobile devices and the infrastructure surrounding it, there has been significant effort in utilising data generated by every component of this complex infrastructure.

Cellular/ Mobile network The use of cellular network data is relevant for urban studies (see Jiang, 2013; Steenbruggen, 2015; Lokanathan, 2015; Calabrese, 2015; Reades, 2007) even though it is acknowledged to have inherent biases such as ownership bias across particular demographic groups (Wesolowski, 2013). Visual exploration of use of such data using interactive interfaces to evaluate quality of service and scenario testing has been tested for the optimisation of public transport (Sbodio, 2014). Such network data with the active and passive information collected from them can be used to create trajectories of people (Schlaich, 2010), detect their daily routine (Sevtsuk, 2010) and classify those routes (Becker, 2011). It was also demonstrated to be useful in understanding overall mobility and flow of people and information (Candia, 2008; Krings, 2009; Simini, 2012; Zhong, 2016). It can be used to identify asymmetry in flow of people spatially (Phithakkitnukoon, 2011), estimate volume and pattern of road usage (Bolla, 2000; Wang, 2012) and by augmenting the topology to optimise operations (Puzis, 2013). Such datasets have been extensively used in traffic and transportation research to derive origin-destination matrices (Caceres, 2007; Mellegard, 2011; Iqbal, 2014), travel time estimation (Janecek, 2012) and traffic status estimation (Demissie, 2013; Grauwin, 2015)

It has been shown that mobile network data can be used to uncover nature of the population such as tourists in specific areas (Girardin, 2008) and the interaction between the people in the study area. The structure (Onnela, 2007 a, b) geography (Lambiotte, 2008) and dynamics (Hidalgo, 2008) of such networks have been studied and demonstrated to be useful in predicting their change (Wang, 2011). The network data and its spatio-temporal structure can also be used for classification of land use (Pei, 2014), assessment of spatial patterns (Reades, 2009; Steenbruggen, 2013) and understanding the spatial structure of cities (Louail, 2014; Arribas-Bel, 2015). The data collected from the cellular network measured at the smallest scales such as web chatting, mobile calls and so on can be used to create estimations of micro site level population density (Pulselli, 2008), characteristics (Girardin, 2009) and the nature of the activity (Phithakkitnukoon, 2010). Aggregated human activity measured from the data can be used to measure and model population dynamics and land use density and mix at large scales (Jacobs-Crisioni, 2014; Tranos, 2015). The spatial patterns understood can then be applied to urban planning (Becker, 2011) whilst the temporal patterns have particular utility for the likes of epidemiology where population influxes measured from changes in mobile network usage can be used to model spread of diseases (Buckee, 2015).

Though the mobile network provides much more granular and accurate data than interpolation techniques, it is not without its limitations. The network distribution usually follows the purpose of service coverage and commercial decisions which introduces systematic biases in the data passively collected through them, while the data actively collected through them has bias based on the volume of usage of services by the customers which can vary widely based on location and demography. This makes collection of data directly from the devices using the sensors available a more robust option.

Mobile Sensors The major sensors and capabilities present in mobile devices that can be used for distributed urban sensing are cellular radio, Bluetooth, WiFi, GPS, accelerometer and a compass. Since cellular radio is managed by the cellular network and covered in mobile network data, we explore the research done with other sensors. In contrast to planned actively collected data, data passively collected via a distributed network of general purpose devices tends to be larger and more temporally dynamic. For example, an organised survey conducted every month to understand interpersonal communications between people in a team of 50 will result in a 2500 records a month. The same task is done through collecting data on email communication sent by them will result in a same volume records in a day. The challenges and solutions on collecting and analysing such large-scale longitudinal data are discussed by (Laurila, 2012; Antonic, 2013). The real time nature of such data also gives us the opportunity to monitor and understand the city in much smaller temporal scales (Townsend, 2000; O'Neill, 2006) and the representativeness of such datasets have also been explored (Shin 2013; Kobus 2013). Data generated from communication networks can be used to understand the structure of urban systems which are becoming increasingly borderless (Bertolini, 2003). Similar to the network based data, it can help in understanding human mobility

(Asgari, 2013; Amini, 2014; Zhang, 2014) through mining trajectory patterns (Giannotti, 2007) and socio geographic routines (Farrahi, 2010). It is also useful in various traffic and transportation applications for monitoring roads (Mohan, 2008) and estimating traffic (Cheng, 2006), uncovering regional characteristics (Chi, 2014) and extracting land use patterns (Shimosaka, 2014). Apart from GPS and WiFi, there have been efforts in exploring other possibilities such as Bluetooth for location (Bandara, 2004) and aggregate detected Bluetooth activity to monitor freeway status (Haghani, 2010). There have also been successful implementations of frameworks to predict movement of people by combining WiFi and Bluetooth (Vu, 2011). But owing to shorter range and requirement of active engagement from the user (device pairing) Bluetooth is much less preferable for large-scale data collection than GPS/ WiFi. The research on GPS and WiFi based studies are discussed in more detail below.

Global Positioning System In addition to providing a user's location to applications such as Google Maps, the GPS capability in mobile devices working with the WiFi can maintain a continuous list of locations visited by the device over long periods of time. It works mostly in the background and requires almost no active input from the user to operate. Though very convenient for collecting data, due to the privacy risks associated with it GPS is often one of the resources in a device that requires explicit user permission to be accessed. The concepts and methodologies for collecting such data were set out by Asakura (2004) and there have been attempts to collect this rich data from volunteers at a large scale along with ancillary data (Kiukkonen, 2010) and provide a location based service application for the collection of data (Ratti, 2006; Jiang, 2006; Ahas, 2005).

GPS is one of the most used technologies for mobility studies. It has been used to analyse and understand individual mobility patterns (Gonzalez, 2008; Neuhaus, 2009), which have been shown to have a high order of regularity in spite of the complexity (Brockmann, 2006; Song, 2010 b). There have been efforts to use this regularity to predict the future location of people (Monreale, 2009; Calabrese, 2010). The limitations of predictions have also been quantified (Song, 2010a). There have been successful efforts in extracting behaviours and patterns from such trajectory data (Liu, 2010; Cho, 2011; Hoteit, 2013; Pappalardo, 2013) along to understand individual patterns from large assemblages (Giannotti, 2011; Calabrese, 2013) and vice versa (Wirz, 2012). In traffic and transportation, GPS trajectory from mobile devices is used to estimate (Calabrese, 2011) and expand (Jing, 2011) OD matrices, detect the mode of travel (Gong, 2012; Rossi, 2015) and calibrate existing spatial interaction models (Yue, 2012).

Since the data is collected at the device level and depends on the activity of the individual, it can be de-anonymised to reveal the nature of the owner of the devices. The possibilities of detecting the activity of the individual from trajectory information is demonstrated by (Liao, 2006; Krumm, 2007). Patterns (Jiang, 2012) and structures in routines (Eagle, 2009) can be extracted from these trajectories and can be used for socio geographic analysis of the population (Licoppe, 2008). It can also utilised in classification of the population at a particular location at a given time (Pappalardo, 2015). Being inherently spatial and activity driven, GPS trajectories have been shown to be useful to identify (Bao, 2012), characterise (Wan, 2013) and automatically label (Do, 2014) significant places of interest. It can also be used for land use detection (Toole, 2012), classification (Jiang, 2015) and the study of urban morphology (Kang, 2012). These GPS trajectories have been shown to be useful in estimating population dynamics at local level and within short durations during social events (Calabrese, 2010; Kim, 2014; Deville, 2014). When combined with other data sources can be useful to understand relationship between spatial areas (Long, 2015).

From the literature we see that GPS is one of the most precise and accurate user side methods of collecting location of mobile devices. In addition, the data collected is well understood and collection methodologies can be scaled up with minimum resources. That said, it is well known that urban sensing methods using GPS of mobile devices also has problems of enhanced risk of breach of privacy when done passively and need for user engagement when done actively.

WiFi WiFi is a wireless network connection protocol standardised by IEEE, 2013. It is a distributed server-client based system where the client connects to access points (AP). Every device in the network has a unique hardware specific MAC address, which is transmitted between the device and AP before the connection is made. The key feature of WiFi infrastructure is that the network is distributed i.e. the APs can be set up and operated by anyone locally unlike mobile networks. Since they are primarily used for Internet service provision, the protocol has priority for continuity of connectivity so the devices constantly scan for new and better connections. This is done through a probe request, which is detailed in later sections. With this background we can see that WiFi provides a fair middle ground between an entirely network driven approach such as cellular network to an entirely user driven approach such as GPS. Since the network infrastructure is distributed and deployed for Internet it offers near complete coverage, is very resilient, and can encapsulate and reinforce civic space in cities (Torrens, 2008).

Though WiFi is a location less technology, there are reliable methods to triangulate the location of the device by the signal strength and the known locations of APs (He, 2003; Moore, 2004; LaMarca, 2005). This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments (Zalampas, 2006; Kawaguchi, 2009; Xi, 2010). Utilising this, we can easily and quickly estimate trajectories of the mobile devices just using the WiFi communication the device has with multiple known APs (Sorensen, 2006). This can be used similar to the GPS trajectories to understand individual travel patterns (Kim, 2006; Rekimoto, 2007; Sapiezynska, 2015), crowd behaviour (Abedi, 2013; Mowafi, 2013), vehicular (Lu, 2010) and pedestrian movement (Xu, 2013; Fukuzaki, 2014; Wang, 2016). It can also be used in transportation planning and management to estimate travel time (Musa, 2011) and real time traffic monitoring (Abbott-Jard, 2013).

Being a general network protocol designed to be used by mobile devices, WiFi devices relay a range of public signals known as probe request frames on regular intervals throughout its operation, for the purpose of connecting and maintaining a reliable and secure connection for the mobile device (Freudiger, 2015). These signals can be captured using inexpensive customised hardware, non-intrusively and in turn to be used for numerous applications. In addition to a uniquely identifiable MAC address, these signals include a range of other information which when combined with the temporal signatures of the signals received can help us understand the nature and identify the devices which are generating these signals. These device/user fingerprinting techniques are demonstrated by Franklin (2006) and Pang (2007) and the unique MAC addresses and associated information can successfully track people across access points (Cunche, 2014a), their trajectories (Musa, 2012), the relationship between them (Cheng, 2012 Barbera, 2013 Cunche, 2014b) and predict which of them will be most likely to meet again (Cunche, 2012). Using the semantic information present in these probe requests it is possible to understand the nature of these users at a large scale (Di Luzio, 2016). Using the received signal strengths from pre placed devices we can monitor the presence and movement of entities that are not even carrying a WiFi enabled device (Elgohary, 2013).

Because of the security and privacy risks posed by the WiFi protocol's use of hardware based MAC address, various methods to strengthen the security have been proposed (Pang, 2007; Greenstein, 2008). The randomisation of MAC addresses has become more mainstream in mobile devices with the introduction of it as a default operating system behaviour in iOS 8 by Apple Inc. Since MAC randomisation is not a perfect solution (Cunche, 2016) there have been numerous attempts to fingerprint unique devices from the randomised anonymous information present in the probe request frames for the purposes of trajectory tracking and access point security. The methods used are decomposition of OUIs where detailed device model information is estimated by analysing an already known dataset of OUIs (Martin, 2016); Scrambler attack where a small part of the physical layer specification for WiFi is used (Bloessl, 2015); and finally, the timing attack where the packet sequence information present in the probe request frame is used (Matte, 2016; Cheng, 2016). A combination of these methodologies has been proven to produce de-anonymised unique device information from randomised MAC addresses (Vanhoef, 2016). In addition to tracking, WiFi probe requests can be aggregated to uncover the urban wireless landscape (Rose, 2010) and used to reveal human activity at large scales (Qin, 2013), pedestrian numbers in crowds (Schauer, 2014; Fukuzaki, 2015) and also counting people in hyper local scales such as queues (Wang, 2013). With enough infrastructure we can aim to generate a real-time census of the city (Kontokosta, 2016) and also predict the amount of time a device will spend around the sensor as well (Manweiler, 2013). Similar to GPS data this can be used as an additional control layer for interpolation techniques such as map merging (Erinc, 2013). A comparison of various approaches was done by Pinelli (2015) where through experiments on a telecom operator dataset, it was showed that using network-driven mobile phone location data is more advantageous compared to the widely used event-driven ones.

Social Media In addition to the direct data from the sensors themselves the content generated from the mobile devices in social media can provide a viable proxy for estimating the level and nature of human activity. The use of geo located tweets on the study of small area dynamic population estimation (Ordonez, 2012; Marchetti, 2015; McKenzie, 2015), geotemporal demographics (Bawa-Cavia, 2011; Longley, 2015; Lansley, 2016) and global mobility (Hawelka, 2014) has been thoroughly explored. These data sources are shown to be useful in social sciences (Crane, 2008), abnormal event detection (Chae, 2012) and analysing urban environments (Sagl, 2012). It can also be used as a control layer for interpolation techniques we discussed earlier (Lin, 2015).

Chapter 3

# Data Collection

## 3.1 Wi-Fi Specification

## Chapter 4

# Data Processing

## Chapter 5

# Applications

## Chapter 6

# Conclusions

# Appendix A

# Software

# Appendix B

# Code

# Bibliography

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.