THERE IS NO LOGIC THAT CAN BE SUPERIMPOSED ON THE CITY; PEOPLE MAKE IT, AND IT IS TO THEM, NOT BUILDINGS, THAT WE MUST FIT OUR PLANS.

JANE JACOBS, *THE DEATH AND LIFE OF GREAT AMERICAN CITIES*

INFORMATION IS THE OIL OF THE 21ST CENTURY, AND ANALYTICS IS THE COMBUSTION ENGINE

PETER SONDERGAARD, *SVP, GARTNER*

ERRORS USING INADEQUATE DATA ARE MUCH LESS THAN THOSE USING NO DATA AT ALL

CHARLES BABBAGE *THE EDUCATION INDUSTRY*

BALAMURUGAN SOUNDARARAJ

ESTIMATING FOOTFALL FROM PASSIVE WIFI SIGNALS

CASE STUDY WITH SMART STREET SENSOR PROJECT

DOCTOR OF PHILOSOPHY
UNIVERSITY COLLEGE LONDON - UCL

DOCTOR OF PHILOSOPHY
DEPARTMENT OF GEOGRAPHY, UCL

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.

.....

Submitted on, May 2019

Contents

1 *Introduction* 17

2 *Review of Literature* 19

3 *Collecting Wi-Fi Data* 35

4 *Processing the Data into Footfall* 43

5 *Visualisations and Applications* 73

6 *Discussion and Conclusions* 75

7 *Appendix* 77

Bibliography 81

List of Figures

2.1 Growth of research in the areas of understanding distribution and dynamics of human activity since 1980. 20
2.2 Tree-map showing the volume of research conducted under each major themes and their sub-themes. 21
2.3 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 24
2.4 Growth of research in the topic of '' 25
2.5 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 26

4.1 Comparison of volumes of datasets across various disciplines. 49
4.2 Missing data from five locations at Tottenham Court Road, London on 15 January 2018 demonstrating the veracity of the data. 50
4.3 Number of probe requests collected for every five minute interval at Tottenham Court Road, London on the year 2018 showing the visual complexity of data in the time dimension. 51
4.4 Big data characteristics of the Wi-Fi probe request datasets in their corresponding dimensions 52
4.5 Characteristics of types of Wi-Fi data collection tools at each end of the spectrum compared to an ideal candidate 53
4.6 Exponential increase in the processing time when using traditional methods. 55
4.7 The increase in processing time with the Unix pipeline is linear thus improves the scalability compared to R based processing 57
4.8 The scalability of the processing pipeline could be further improved with parallelising it. 57
4.9 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 59

List of Tables

2.1 Evaluation of advantages of different approaches that can be used for data collection. 32

4.1 Comparison of volume or size of the datasets of Wi-Fi probe requests. 48
4.2 Comparison of velocity or speed of the datasets of Wi-Fi probe requests. 49
4.3 Examples of different types of Wi-Fi based data collection solutions. 53
4.4 Various data storage approaches and their characteristics. 55
4.5 Various types of big data processing tools and corresponding examples. 56
4.6 Tasks in the processing pipeline, corresponding R libraries and equivalent Unix tools 57
4.7 Comparison of clustering algorithms with a sample of 40000 probe requests 66
4.8 Locations where sensors were installes, volume and speed of probe requests collected by the sendor and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format. 68
4.9 Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the filtering process 70

Outputs From The Research

Publications

Book Chapter - Murcio, R., Soundararaj, B., & Lugomer, K. (2018). Movements in Cities: Footfall and its Spatio-Temporal Distribution. In Longley P., Cheshire J., & Singleton A. (Authors), *Consumer Data Research* (pp. 84-95). London: UCL Press.

Journal Paper - Soundararaj, B., Cheshire, J., & Longley, P. (2019) Estimating real-time high-street footfall from Wi-Fi probe requests, *International Journal of Geographical Information Science*.

Conference Paper - Soundararaj, B., Cheshire, J., & Longley, P. (2019). Medium Data Toolkit - A Case study on Smart Street Sensor Project. *In Proceedings of GIS Research UK*, New Castle, United Kingdom.

Conference Paper - Lugomer, K., Soundararaj, B., Murcio, R., Cheshire, J., & Longley, P. (2017). Understanding sources of measurement error in the Wi-Fi sensor data in the Smart City. *In Proceedings of GIS Research UK*, Manchester, United Kingdom.

Conference Paper - Murcio, R., Soundararaj, B., & Lugomer, K. (2018). Trends in urban flows: An information theory approach. *In Proceedings of Conference on Complex Systems*, Cancun, Mexico.

Conference Paper - Soundararaj, B., Murcio, R., & Lugomer, K. (2016). Smart Street Sensor Footfall Project. *In Proceedings of Conference on Complex Systems*, Amsterdam, Netherlands.

Conference Paper - Soundararaj, B., & Lugomer, K. (2016). Smart Street Sensor Footfall Project. *In Proceedings of Oxford Retail Futures Conference*, Oxford, United Kingdom.

Engagement Activities

Event Talk - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Geo+data*, London, United Kingdom.

Event Talk - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Data Natives*, London, United Kingdom.

Conference Workshop - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Smart Urban Policy Futures Workshop*, London, United Kingdom.

Exhibition - Real-time footfall in Bloomsbury. (2017). *Big Data Here - Big Data Week 2016*, London, United Kingdom.

Products

Research Dataset - Smart Street Sensor footfall data, *Consumer Data Research Center*. URL: https://data.cdrc.ac.uk.

Software - Soundararaj, B. (2018). Clicker - an Android app for manually counting pedestrian footfalls with precision and accuracy. *Google Play Store*. URL: https://play.google.com/store/apps?id=com.bala.manualcount.

Awards

Bursary - Yusuf Ali travel bursary (2017) *University College London*.

Dedicated to my parents,
S. Kalavathy and K. Soundararaj.

Prologue

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

Acknowledgements

- Supervisors - Family and Friends - Colleagues - Funding agencies - Open-Source projects (esp. GNU Linux and R)

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.

2
Review of Literature

Understanding the scale, nature and dynamics of distribution of population in space and time has been a central premise of academic research in various fields of study such as human geography, sociology, urban planning and architecture. This granular knowledge of where people are and how they move is critical in practical decision making in various industries such as real estate for valuing places, retail for business planning and emergencies for risk management and evacuation. The first challenge in any of these research is collecting precise and accurate data. This started as research into methods estimating and interpolating highly granular data from existing regional level aggregate datasets and as the technology improved through the later half of twentieth century, the research methodologies adopted the new technologies to utilise these more granular sources of data. Though new technologies provide immense opportunity in collecting large amounts of data which were previously impossible they also introduce their share of uncertainties. Hence it becomes imperative that we understand the evolution of these techniques and methodologies along with the research that used them to build our rationale behind any further research.

Moreover with the proliferation of mobile devices and wireless internet connectivity, even day to day activities are being digitised leading to the creation of large amount of easily accessible data which are generated passively in an unstructured manner. The users' acceptance to the collection and analysis of such data has also been improving until recently [1]. There has also been rising concerns regarding user privacy along with the development of more accurate methods to track them. In this context, the critical task in all these research is to solve the problem of balancing these two by choosing the right technologies and devising the appropriate methods.

In this chapter we conduct a systematic survey of literature in this broad area of 'distribution and dynamics of human activity'. The aim of this survey is to evaluate where we are at the research and understand how we got here. First we conduct a comprehensive survey of over 300 publications which discuss this area of research and we then look at the major themes and trends in the last 70 years. We discuss these themes in detail looking at the aims and achievement of these research while highlighting the opportunities and gaps they leave. We also look at the timeline of their publication and discuss the evolution of the research along with the changes in the technology landscape. We then classify the studies by looking at the technologies employed by them and note the trends in the adoption and phasing out of relevant technologies. We also

[1] Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1_15

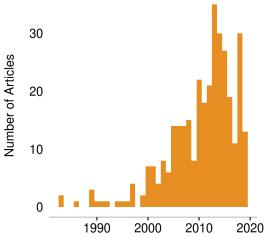

Figure 2.1: Growth of research in the areas of understanding distribution and dynamics of human activity since 1980.

Measured in the number of papers published

try and understand the advantages and disadvantages of these techniques and develop a theoretical framework to understand when and how to use them effectively answer research questions. Finally we summarise the literature survey focussing on the major gaps still left in our understanding and the new, emerging areas where more research is necessary. We also critically evaluate these areas of research in terms of priority and feasibility to develop our research questions and devise a plan for our research.

We begin our literature review with a systematic search for academic publications studying the distribution and dynamics of human activity. We start this search from a set of reviews pertaining to the topic such as Ratti et al. [2006], Jiang et al. [2013], Steenbruggen et al. [2013], Arribas-Bel [2014] and Li et al. [2016]. From these reviews we further expand by navigating the citation networks and finding research which are relevant to our search. Though this does not give us a perfectly comprehensive set of research literature, it does provide us with a representative sample of all the different directions of the research conducted in the area. Through this process, we identified around 325 relevant research publications which deal with the collection, measurement, analysis, visualisation and discussion of population at a granular level. The research started around 1950s looking at estimating day-time urban population at a granular level using existing broader data employing various estimation methods [2,3]. Though this served as a starting point, the pursuit of such granular data and their applications in corresponding fields didn't pick up until the start of the 21st century fuelled by the digital revolution that followed growth of internet. Figure 2.4 shows the yearly volume of research published since 1980. We can observe though there were some research conducted through 80s and 90s the real push forward came around beginning of the millennium when mobile phones adoption skyrocketed. In addition to the early 2000s we also see a substantial increase in interest in the beginning of the next decade fuelled by the smartphone revolution which completely changed the research avenues in-terms of volume and types of data available and methodologies available to tackle them. The area of research is multi-disciplinary encompassing academic interest and commercial applications in Geography, Urban Studies, Urban Planning and Management, Emergency planning and Management, Economics, Computer Science and Engineering etc.

[2] Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732

[3] Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2): 83–85, 1956. ISSN 0002-8991

2.1 *Research Themes*

In this section we look at the major themes and questions tackled by this knowledge base. We start by classifying the research into the major and minor themes explored in them as shown in Figure 2.2. The tree-map shows the volume of research in corresponding themes measured in terms of number of publications. We can observe that the research is conducted in five major areas - population studies focussing on the creating and utilising data on distribution and nature of human activity, mobility and interaction focussing on the changes in these distributions, understanding the nature and function of space from these distribution and change, methods and techniques which can be used to conduct the research and finally issues and solutions related to the privacy of the users while conducting these research. We can also observe

that most of the research apart from developing methods were conducted in the domain of human mobility and social interaction closely followed by the population distribution. In the following sections we discuss these in detail along with their sub themes with the following framework,

1. What are the major lines of questioning?
2. What has been done previously?
3. Where are the opportunities for further research?

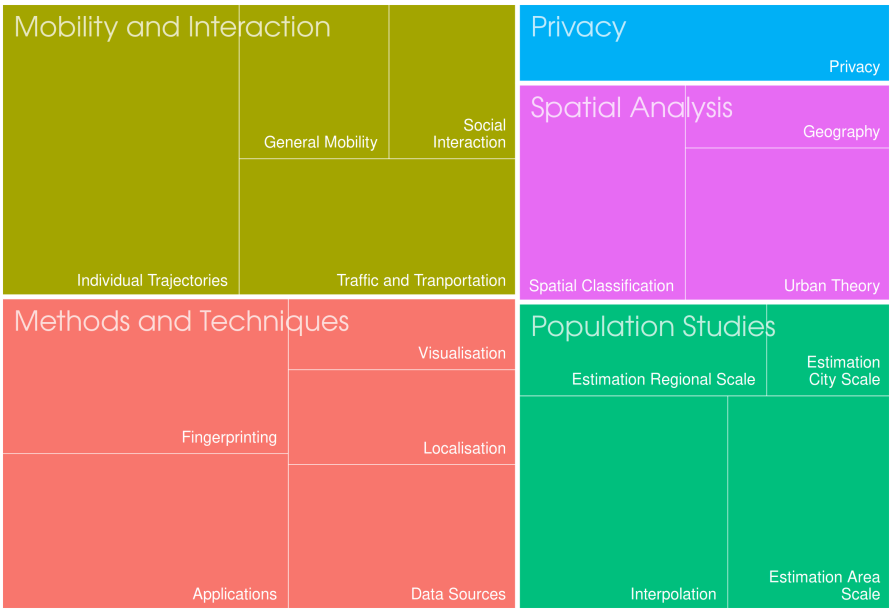

Figure 2.2: Tree-map showing the volume of research conducted under each major themes and their sub-themes.

2.1.1 *Population Studies*

Though Foley [1954] and Schmitt [1956] started this line of research in 1950's with the discussion on estimating daytime population using broader datasets it was not until the 80s significant volume of research kicked off in this area of study. From 80s until mid 2000's numerous studies were conducted on measuring and studying the population at a granular level both spatially and temporally. The focus of the research around this time was primarily on interpolation from the larger datasets created using censuses, regional or national level sample surveys and other centrally collected sources of data. There have been numerous fairly successful attempts with methodologies where a broad dataset such as regional level population summaries and modelling or interpolating more granular data from them by augmenting with other sources of data such as street networks [Reibel and Bufalino, 2005], remote sensing [Sutton, 1997, Yuan et al., 1997, Chen, 2002] etc. Dobson et al. [2000, 2003], Bhaduri et al. [2002, 2007] and [Mennis, 2003, Mennis and Hultgren, 2006] are examples of such research methodology. These studies were almost done on a city scale or above with mostly modelling or interpolation methods since the data sources were few and were centrally collected.

Around 2005, there was a sharp shift in research where the interpolation methods were replaced by highly available granular data collected over cellular network. Studies were conducted on estimating population densities, presence of tourists, general activity pattens using data from cellular networks. Most of these research were conducted at a far larger geographic scale looking at things at an area level [Pulselli et al., 2008, Girardin et al., 2009, Phithakkitnukoon et al., 2010, Yuan and Raubal, 2016]. There were efforts in using device level sensors such as global positioning system(GPS), Wi-Fi and Bluetooth to detect population distribution and socio-geographic routines [Calabrese et al., 2010, Rose and Welsh, 2010, Farrahi and Gatica-Perez, 2010]. There have been studies on looking at people distribution as granular as queue lengths as discussed by [Wang et al., 2013] to city level dynamic population mapping where the limitations of traditional datasets generated through censuses and surveys [4].

[4] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014. ISSN 0027-8424. DOI: 10.1073/pnas.1408439111

Around the 2015, along with the data collected directly from the mobile devices,the data that are generated by the users activity on these devices are became more important. Social media data such as twitter [Lansley and Longley, 2016] and other consumer data such as loyalty cards [Lloyd and Cheshire, 2018], smart cards [Ordonez and Erath, 2012] etc. have also become a significant sources of data for such research. Recently, with increased concerns and legislation on privacy, there have been studies which go back to the effort of interpolating granular data from broader datasets but using more data and processor intensive technologies such as agent based modelling, deep learning, small area estimation [Crols and Malleson, 2019, Shibata and Yamamoto, 2019, Rao and Molina, 2015] etc.. Though there have been a lot of work done in most of the directions in this research area, the clear gap arises due to the absence of a continuous, granular and sufficiently longitudinal data-sets to complement the methodologies that have been developed.

2.1.2 *Human Mobility and Interaction*

This is one of the major areas of research which have significantly benefited from the decentralised collection of data at a granular level [5]. In addition to being useful in their own right, these data were in turn used to augment traditional models of travel behaviour, traffic and transport to provide a better understanding of human movement over time and space [Janssens, 2013]. The major themes of research within this area are, Movement of people in space and time with emphasis on understanding the built environment, social interaction between these people with a sociology perspective and traffic and transportation studies with a infrastructure perspective. There is significant volume of research which dealt with recording and analysing the trajectories of the users to understand their movement patterns enabled by the unprecedented availability of detailed data from mobile devices.

[5] M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000

2.1.3 *Methodology and Techniques*

[MacEachren and Kraak, 2001] [Hallisey, 2005] [Morrison et al., 2000] [Lobben, 2003] [Harrower, 2007] [Ferrara et al., 2014] [Fabrikant, 2005] [Thomas, 2005]

Visualising the temporal dynamics of data collected on human activities through decentralised processes poses significant challenges when approached

with traditional cartographic concepts (MacEachren, 2001 Hallisey, 2005). Digital 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 geographic 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.

2.1.4 *Spatial Analysis - Theory and Modelling*

Traditional and modern geography was 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 post-modern geography 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 [6]. This definition is based on the bottom-up mechanisms which make human activity such as information exchange and economy to manifest in the physical built environments as argued by [Batty, 1990, 1997, Batty et al., 2012] and [Batty, 2013a,b].

[6] Edward Soja. Postmodern geographies, 1989. URL http://books.google.com/books?id=sNcRAQAAIAAJ

This transition into the digital age [Graham and Healey, 1999, Tranos and Nijkamp, 2012, Tranos, 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 and Marvin, 2001, 2002, Sassen, 2001]. 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 et al., 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 and Small, 1991] and vice versa [Zandvliet and Dijst, 2006]. This flood of hard data [7] was accompanied not only by optimism in its potential [Thomas, 2001] but also by the questions raised on the challenges in handling the diverse, large scale, non standardised data it produces and the usefulness or representativeness of the resulting analysis [Miller, 2010, Arribas-Bel and Sanz-Gracia, 2014].

[7] Nature Editorial. A flood of hard data. *Nature*, 435: 698, 2008. ISSN 0028-0836. DOI: 10.1038/453698a

However, availability of such data has impressive uses in urban studies [Bettencourt, 2014] especially with advancement of new technologies [Steenbruggen et al., 2015] and possibility of distributed, crowdsourced data collection [Lokanathan and Gunaratne, 2015].

2.1.5 *Privacy*

The ubiquity of personal devices and digitisation of day to day activities through these mobile devices [McMeel, 2018] has provided many opportunities for researchers and industry for collecting, analysing and deriving inputs from them. However at the same this also increased the risk of infringement on privacy of the users whose data is being collected [8]. There is immense value in uniquely identifying and profiling information on people for specialised purposes such as security [Cutter et al., 2006] and law enforcement [Dobson and Fisher, 2003] but also has extreme risks associated when not handled with care [VanWey et al., 2005].

[8] T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007; and John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009

Strictly protecting personal information while ensuring the information is usable for research by maintaining the uniqueness in the data is the major concern which was addressed by devising frameworks for secure practices in confidentially collecting and using the location data [Duckham and Kulik, 2006, Tang et al., 2006, Lane et al., 2014]. Some efforts sought to accomplish this task through cryptographic hashing algorithms (Pang, 2007) while others aimed to thwart identification and tracking at the device level by techniques such as MAC randomisation [Gruteser and Grunwald, 2005, Greenstein et al., 2008]. Finally though getting consent of users for the collection and use of such information from their mobile devices is challenging, there is a significantly improved acceptance when the process offers value in return such as discounts and monetary benefits [Kobsa, 2014].

There is opportunity in this area for research in applying the cryptographic solutions along with the privacy preserving frameworks to arrive at methods which can extract useful information out of large personal data while obscuring or anonymising them.

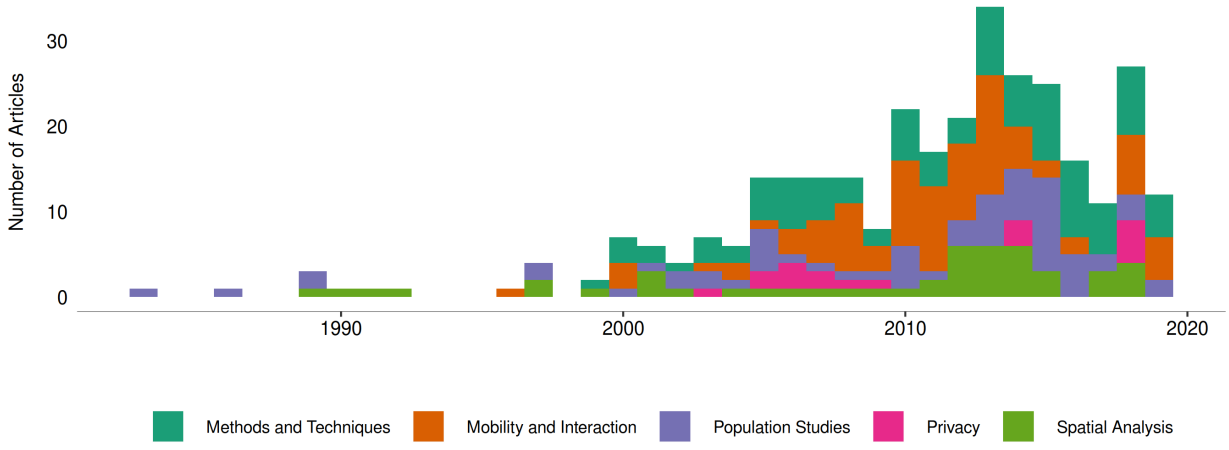

Figure 2.3: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

2.2 *Research Trends*

Figure 2.3 shows the volume of research done in this topic since 1980 categorised based on their major themes discussed earlier. We can observe that there are distinct trends in the research over time, which evolved around the

development of technology in the last two decades. Until 90s the research was mostly centered around population studies on estimating and interpolating granular spatial and temporal information from larger and cross sectional datasets such as census and sample surveys. The period between 2000-2010 there was interest in potential of the new data generated by the digital revolution. We can categorise this as the 'mobile era' where carrying mobile devices become mainstream. This explosion of research coincided with mobile phones becoming more popular and ubiquitous with population in urban areas and was around development of methods and techniques to utilise the data generated from them. There were also extensive studies in using the datasets to understand human mobility along with a rising concern in the privacy of the users who's data which are being used for these studies.

The release of iPhone in 2008 and the increase in the share of 'smartphones' in the next 10 years sparked the 'smartphone' era. The change made sure that all the mobile devices gaining numerous capabilities such as internet connectivity over Wi-Fi and mobile network, location awareness with global positioning system, movement recognition with accelerometers and connectivity other 'wearable' devices through Bluetooth. This also lead to the digitisation of lifestyle where every aspect of the life being done through these devices over internet while generating huge amount of data on these activities. This sparked the large volume of research on the form and function of space by studying this data and on the dynamics of human population in space and time in the next 5 years. These research were particularly centered around tracking the trajectory of people using the mobile devices they carry with them as the smartphones made it easier to collect the necessary data directly from them rather than depending on a centrally collected datasets from mobile carriers. With the theoretical limit to predictability in human mobility quantified [9], the focus on urban mobility has been declining in the past few years which has led to a renewed interest in population studies at a local-local level in real-time. In addition to using the data from the mobile devices, these studies have also been exploring the use of large assemblages of consumer data that are being generated in this connected mobile environment and linking them together to create a fuller picture [10]

Finally, with the increase in use of personal data, there has also been an increase in research regarding the privacy of the users. Along with this, the mobile devices and subsequently the data generated by them are more and more anonymised so that the users cannot be tracked or identified at a personal level. This has given rise to the new trend in research to devise methods to overcome this anonymisation and at the same time research which considers these methods as vulnerabilities and find solutions to make the anonymisation process more robust. There is clear need for methods which anonymise the data sufficiently to protect the identity of the users and at the same time enable us to conduct research in measuring studying population distribution and movement at a granular level.

[9] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327 (5968):1018–1021, 2010. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.1177170

[10] Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018

2.3 *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

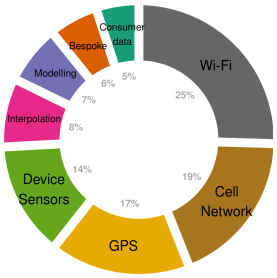

Figure 2.4: Growth of research in the topic of ''
Measured in the number of papers published

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. and the total distribution is shown in Figure 1. . 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,

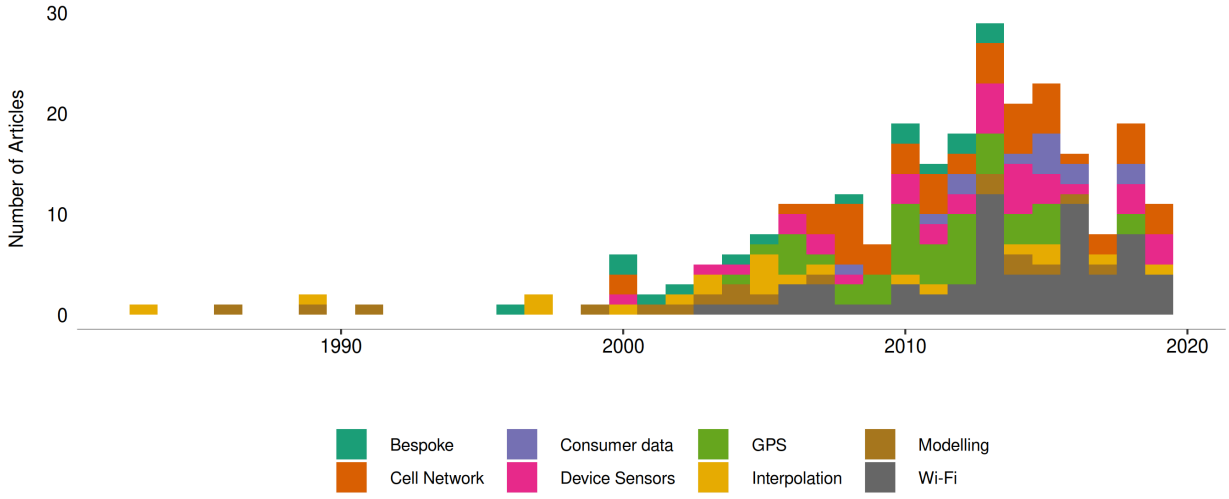

Figure 2.5: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

2.3.1 *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.

2.3.2 *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.

2.3.3 *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.

2.3.4 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.

2.3.5 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.

2.3.6 Wi-Fi

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 and 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).

hough 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 tra-

jectories 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.

2.3.7 *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).

In terms of technology, which can be used for data collection at local level, the summary of the advantages of each is given below,

Table 2.1: Evaluation of advantages of different approaches that can be used for data collection. **independence from secondary data collected by a third party.*

Technology	Interpolation	Bespoke	Cellular	GPS	Wi-Fi
Coverage	Local	City	All	Local	All
Certainty	Very Low	High	Medium	High	Medium
Independence*	Low	Very High	Low	Medium	High
Intrusiveness	Low	Medium	High	High	Medium
Granularity	Very Low	Very High	Medium	High	High
Ease of Collection	Medium	Low	Medium	Low	High
Scalability	Medium	Low	High	Medium	High
Privacy Risk	Low	Medium	High	High	Medium

Table 2. : From Table 2.1 we can see that though it poses some risk on privacy of participants and some uncertainty regarding range, WiFi offers the best possible technology for data collection through mobile devices at smaller scales.

2.4 *Research Gaps and Opportunities*

From the above literature search we can conclude there is a considerable opportunity in the application of mobile phone based data to be used to measure small-scale spatio-temporal dynamics of human activity and potential for research gaps in the following areas: Hyper-local population estimation This research is at the limits of acceptable mobility research in the context of privacy concerns in between the explosion of usage of smartphones and implementation of MAC randomisation. The availability of identifiable device information has led to a considerable effort in urban mobility studies and there are still opportunities in real time small-scale population measurement and estimation. Detailed WiFi data, combined with other data sources such as public transport, weather etc. along with interpolation and modelling

techniques can produce detailed spatio-temporal estimates of population at local level.

2.4.1 *Device fingerprinting*

De-Anonymisation of devices The increase in privacy concerns and the normalisation of device level security measures such as MAC address randomisation has questioned the basic assumptions about the reliability of data collected through mobile devices. There is a definite research need and interest in finding a method to uniquely identify devices with sufficient confidence while not deciphering any personal data from such analysis. Device classification There has been a significant volume of studies in the classification of space based on the temporal patterns in the human activity around them and classification of human activity based on the space it is clustered around. But there is a research opportunity in looking at these devices closely, especially their temporal patterns around the point of interest to make inferences about the nature of the device. This idea is closely related to de-anonymisation discussed about but rather than trying to find unique users or devices from a set of signals, we try to find the groups of similar devices and connecting these groups to meaningful typologies.

2.4.2 *Event Detection*

2.4.3 *Pedestrian Flow and Catchments*

Sensor catchment and flow Having access to a broader sensor data, we can cross-sectionally see the distribution or occurrence of a unique MAC address / device and determine the influence of a sensor. This can be approached in two ways: the first being the pedestrian flow approach where we can model the movement of pedestrian based on these relationships and the second is treating these relationships as network and detecting communities in them. In addition to these, there is a definite opportunity for research in methods for visualising high frequency, hyper-local pedestrian data within the limits of cognition of the viewer.

2.4.4 *Study of Urban Places*

3
Collecting Wi-Fi Data

As we discussed in the previous chapter, out of all the technologies through which we can capture the traces of people moving across the urban environment, Wi-Fi is a well rounded technology and is suitable for our study. In this chapter, we look at Wi-Fi in detail.

3.1 Wi-Fi as a Source of Data

Wi-Fi is ubiquitous. The smart-phone adoption rates are growing. All smartphones try to connect to internet. In addition to mobile networks, Wi-Fi is the second most common way they connect to internet. Most places provide Wi-Fi as the way to connect to internet. Unlike mobile networks Wi-Fi is a general purpose service. There are multiple networks across locations hence the phones are made to be able to move across networks seamlessly. The mobile phone initiates the contact. It sends a special signal called - Probe Requests. This has information about the mobile device. The router replies with a signal called Probe response. This forms a digital handshake between these devices. The devices then carry on with authentication and talking to each other. After authentication the connection is encrypted and private. But the probe request process is unencrypted and open. The probe request is sequential signal which is defined in IEEE standards. The table shows All possible information that can be included in a probe request. The figure shows the structure of a probe request. This is a stream of data broadcast over air from all the phones around a area.

The probe request frame is the signal sent by a WiFi capable device when it needs to obtain information from another WiFi device. For example, a smartphone would send a probe request to determine which WiFi access points are within range and suitable for connection. On receipt of a probe request, an access point sends a probe response frame which contains its capability information, supported data rates, etc. This 'request-response' interaction forms the first step in the connection process between these devices. The structure of a probe request is shown in Figure 3.3. We can observe that the request frame has two parts, a MAC header part which identifies the source device and frame body part that contains the information about the source device. The information that can be included in a probe request shown in Table 3.2. As mentioned earlier, the SmartStreetSensor collects some of the information present in probe request frame relayed by mobile devices, along with the time interval at which the request was collected and the number of

such requests collected during that interval. The actual information present in the data collected by the SmartStreetSensor is shown in the Table 3.3.

3.2 *Initial Experiments*

First setup using the laptop and wireshark in the living room. Second set of experiments in UCL cloisters. To start, we designed a small pilot study to validate the filtering and clustering methodology against the scale and complexity of data collected in an open public area such as a retail high street. We also aimed to find the algorithm which was best suited for the classification of signal strengths as 'low' and 'high' in order to filter out the background noise. The data was collected at Oxford Street, London on 20 December 2017 from 12:30 to 13:00 hrs, Wi-Fi probe requests were collected using the sensor described in Section and pedestrian footfall was manually recorded using the Android app - Clicker bala2018clicker. Being located at one of the busiest retail locations in the United Kingdom, the Wi-Fi sensor captured approximately 60,000 probe requests during the half hour period; 3,722 people were manually recorded walking on the pavement during that time. The surveyor positioned himself at the front of a store while carrying the sensor in a backpack and counted people walking by the store on the pavement (3m wide approximately) using a mobile phone. The sensor was kept as close to the store window as possible, and the manual count was done as a cordon count in front of the store.

3.3 *Pilot Study*

The methodology set out above was implemented in five different Central London locations at different times. % Sensors were installed and data collected for extended periods of time. ensors were installed and data collected for extended periods of time. go We then applied the methodologies discussed earlier to arrive at estimated pedestrian footfall and compared them with the corresponding manual counts. We finally evaluated the effectiveness of the processes with the Mean Absolute Percentage Error (MAPE) at the locations and report our findings below.

Locations where sensors were installes, volume and speed of probe requests collected by the sendor and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format. The locations at which the data were collected are shown in Table . The locations were chosen for their diverse site conditions and unique sources of noise around the potential location of the sensors. The position of the sensor at these locations with respect to the context is shown the Figure We can see that Location 5 is the 'cleanest' with one clear stationary source of noise (phone shop) while location 2 is the most complex due to the proximity of seating areas to the sensor. The sensors were operational through out February and March, while manual counts were conducted in these locations in half hour sessions on at least two different days. For the purposes of comparing with ground truth, we considered the data from sensors which correspond to the 12 sets of available manual counts. The schedule of data collection is shown in Figure .

3.4 *Smart Street Sensor Project*

The SmartStreetSensor project is one of the most comprehensive study carried out on consumer volume and characteristics in retail areas across UK. The project has been organised as a collaboration between Local Data Company (LDC) and Consumer Data Research Centre, University College London (CDRC, UCL). The data for the study is generated independently within the project through sensors installed at around 1000 locations across UK. When completed, the project will serve as the first and unique comprehensive research into the patterns of retail activity in UK high streets.

As a first step, various locations for the study were identified by CDRC to include a wide geographical spread, different demographic characteristics and range of retail centre profiles. A custom footfall counting technology using WiFi based sensors was also designed, developed by LDC and the sensors were installed the identified locations. The sensor monitors and records signals sent by WiFi enabled mobile devices present in its range. In addition, the number of people walking by the sensor was counted manually for short time periods during the installation. The project aims to combine these two sets of data to use as a proxy for estimating footfall at these locations. The potentially identifiable information collected on the mobile devices is hashed at sensor level and the data is sent to central server via encrypted channel for storage. This data is then retrieved securely for the preparation of the commercial dashboards by LDC and for research purposes by CDRC users. The project began on July 2015 with the first sensor installation and has grown to an average of 450 daily active sensors as of January 2017.

The primary aim of the project is to improve our understanding of the dynamics of the high street retail in UK. As we saw in our literature search, unlike online retail, this involves quantification and measurement of human activity at small scales, such as high streets which already the subject of active research. The key challenge in this area is the collection of data at smallest scales possible with minimal resources while not infringing on people's privacy. This challenge when solved can provide immense value to occupiers, landlords, local authorities, investors and consumers within the retail industry. The project aims to facilitate decision making by stakeholders in addition to the tremendous opportunities for academic research.

The data is collected through set of SmartStreetSensors (shown in Figure 3.1), a WiFi based sensor which when installed acts as a WiFi access point and collects specific type of packets (probe requests) relayed by mobile devices which are which are within the device's signal range and are searching for available access points. The sensor is usually installed on partnering retailer's shop windows so that its range covers the pavement in front of the shops. The installation and calibration of device with respect to the shop window and the pavement is illustrated in Figure 3.2. There is also a small percentage (3%) of the devices which are installed within large shops to monitor internal footfall. Each device collects data independently and uploads the collected data to a central container at regular interval 5 minutes through a dedicated 3G mobile data connection. The sensor hardware has been improved over the course of the project and currently has built in failure prevention mechanisms such as, backup battery for power failures, automatic reboot capabilities and in-

device memory for holding data when internet is not available. The hardware versions and the corresponding features are detailed in Table 3.1.

3.5 *Uncertainties in Data*

Having set up the data collection process, organised the data for quick and easy retrieval and satisfied with the consistency of the data collection infrastructure, the next step is the identification of further uncertainties in the data and formation of informed assumptions to move forward with the analysis. The major source of uncertainties we encounter and assumptions we undertake are as follows:

Range of the sensor: Since the strength of the signal from a mobile device to the WiFi access point depends on various factors such as distance between them, the nature and size of obstructions between them, interference from other electromagnetic devices etc., the exact delineation of the range of the sensors is almost to impossible. Moving forward in the research we assume that the range of the sensor is equal in all directions and is linearly indicated by the RSSI (received signal strength indicator) reported by the mobile devices in range.

Probe request frequency: The frequency of probe requests generated by device varies widely based on the manufacturer, operating system, state of the device and the number of access points already known to the device as illustrated in Figure 3.9 and 3.10 (Freudiger, 2015). These requests are also generated in short bursts rather than at regular intervals. Moreover android devices send probe requests even when the WiFi is turned off. With the large number of different devices available, it is impossible to predict and create a general model for this probing behaviour. For simplicity, we assume that for a probe request received which has a MAC address with a known OUI, there is a corresponding device present within the range of the sensor at that time interval, irrespective of the number of such requests received in the mentioned interval. Essentially we are just looking for unique MAC addresses within a time period rather than the total number requests made by them.

MAC address collisions: From the initial analysis we have observed that there are few instances of MAC address collisions reported where a device known to be in some place has been reported somewhere else. This might be occurring due to rogue MAC randomisation by certain devices and the hashing procedure done at two different places. Due to the negligible volume of such collisions (2%), for the purpose of this report, we ignore these collisions and treat all distinct hashed MAC addresses with know OUID to be the same device.

4
Processing the Data into Footfall

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

4.1 Data Toolkit

Big data and its analytics promises huge benefits in terms of value realisation, cost reduction, insights but it also introduces a numerous pitfalls [1]. With developments in information technology, mobile communications and internet of things, large assemblages of data are readily available leading to immense possibilities in research. But when we analyse these data at such scale, we also encounter a large amount of added complexity and cost. Hence it is important to be careful in choosing the methods and tools in dealing with big data where we should look to devise right methods and tools for the right problems. Moreover in several disciplines, such as statistics and geography etc., the existing methods and tools are already developed for dealing with large scale data. These methods along with improvements in hardware has made the processing big data in these disciplines possible without a major changes in workflow. In the current environment of constant change and growth of sources of data, we cannot afford to lose the opportunity to extract information from them while trying to create a perfect, future proof approach in dealing with them. We need to move fast with a pragmatic approach where we look at other disciplines and adopt best practices and solutions in them and develop consistent approach for our needs rather than reinventing the wheel.

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

In the previous chapters we looked at various methods we devised to collect and process data from Wi-Fi probe requests emitted by phones. Though we discussed the methods conceptually, we left out the rationale behind choosing the toolkit employed to implement those methods. In this section we elaborate the thought process and rationale behind these decisions. We start by discussing the concept of 'Big Data' in general and look at previous literature to understand its definition, nature and the challenges they pose. Then we look at the data-sets we collected through the pilot studies and the 'Smart Street Sensor' project and evaluate them in terms of the dimensions of the big data. We also discuss the challenges faced in dealing with our dataset in detail and try to understand the requirements for devising a toolkit for it. Finally we put together a toolkit to suit our datasets built from simple small UNIX tools. [2]

[2] "Write programs that do one thing and do it well. Write programs to work together. Write programs to handle text streams, because that is a universal interface.", Doug McIlroy on UNIX philosophy.

4.1.1 What is 'Big Data'?

With the proliferation of internet enabled personal devices, we have quickly moved from data sparse environment to a data rich one. We can even confidently say that we are in an age of data deluge where the amount of data which are collected and stored are increasing exponentially in a very short period of time [3]. As we saw in the previous chapters collecting large amount of data is quick and easy. Technological advancements have enabled us to be able to think about utilising such large assemblages of data which would have been impossible even in the recent past. By providing unprecedented coverage, these large assemblages of data - 'Big data', provide us with insights which were not possible before. They often change our approach and methods employed in entire disciplines. For example, In computer science, fuelled by the explosion of collected user data, there is a paradigm shift in

[3] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1): 2053951714528481, 2014

Artificial Intelligence with the use of data mining, machine learning and deep learning. It is only time before this approach pervades social sciences research as well. In addition to the above advantages, Big data because of their nature also introduce numerous challenges in their collection, storage, analysis and visualisation. This is not including the enormous additional overhead and complexity introduced when we try to employ big data methods and tools. If we are not careful, using big data tools and methods for solving 'normal' problem can be counter productive where the advantages realised don't justified the overheads introduced. Hence it is important to understand the 'Big data' nature of the datasets we are dealing with at a granular level and choose the tools and methods without any presumptions.

The first and foremost challenge we face while discussing big data is its definition. It is hard to clearly and objectively define 'Big data' as it can vary widely based on the discipline and perspective. What may be 'big' in one discipline may not be in another. The nature of data can also be evaluated in various dimensions and can exhibit different properties in those dimensions. 'Big data' is generally defined within the context of disciplines, as data which cannot be managed with traditional methods and tools in those disciplines and requires substantial change in the approach of the practitioners. This definition is too subjective and falls short of giving us more understanding of 'Big data'. One of the most subscribed definition is to define the scale of the data in the dimension of volume - size of the data, velocity - speed of the data and variety - the complexity of the data [4]. This has also been extended to include more dimensions such as, veracity - the reliability or truthfulness of the data, visualisation - the complexity in visual interpretation and presentation of the data, and others such as visibility validity, variability, volatility and value. There have also been other alternative dimensions proposed such as Cardinality, continuity and complexity [5]. However we can consider the core dimensions of data - volume, velocity, variety, veracity and visualisation for evaluating our datasets. Since not all data is 'Big' in all these dimensions, we need to evaluate the 'bigness' of the data in each dimension and consider the associated challenges and solutions.

[4] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001

[5] Shan Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):70–73, 2014

The second set of challenges arise while we process the big data, its acquisition, storage, extraction, cleaning, annotation, integration, aggregation, modelling, analysis, visualisation and interpretation. Challenges in each one of these processing activity arises due to the data being big in one or more dimensions. The data being big in volume, velocity and variety poses challenges in data acquisition, aggregation, cleaning and analysis [6]. These challenges make traditional methods impractical and introduce the need for distributed, crowdsourced collection of data, heavily parallelised computing and application of functional programming concepts. The unstructured nature of the big data also introduces notable biases which mandate careful consideration, proper calibration and weighting during analysis so that we can understand and remove any uncertainties arising from them. The data being big in veracity dimension poses significant challenges in its analysis and modelling. Since simple methods such as linear regression fails in such scenarios, we require complex methods such as support vector machines, neural networks and hidden Markov models which compensate the lack of structure with the volume of data. With such computationally intensive meth-

[6] Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016

ods, heavily parallelised high performance computing techniques such as GPU processing become indispensable. We also face significant challenge in visualising such complex features and methods which not only supports critical decision making but also is indispensable in exploratory analysis. The volume and velocity of big data makes them hard to visually simplify and digest. They are especially complex to interpret in the time dimension unless presented in small parts. Geographic information systems do a good job in visualising complex geographic data but struggle to maintain legibility and meaning when dealing with the temporal dimension. The visualisations of big data need to be highly processed, simplified and interactive to present meaning to the viewer. They have to balance between functionality, aesthetics and performance. Finally, because of the variety, big data creates need for consistent, well engineered standards so that multiple approaches and tools can be employed in tandem.

Apart from these processing challenges, we also have management challenges associated with big data such as privacy and security, data governance and ownership, data and information sharing, and cost[7]. Since these big datasets are usually comprehensive, securing them and protecting the privacy of the users becomes a central consideration in any project dealing with them. In many cases, though the data collected itself may not contain personal information but at these scales, in conjunction with other datasets, it can be used to infer them. The overall approach, methods, tools should comply with relevant legislation such as GDPR as well as the research ethics of all the stakeholders. This is especially challenging since these large unstructured datasets exhibit ambiguity of their ownership as well which calls for a clear, transparent and secure way to share them with other stakeholders along with publications of results in a timely, accessible manner. The associated project management and tracking tools need to be capable of handling these data ownership and sharing concerns as well.

[7] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014

Finally, the biggest challenge we face with big data is the cost in terms of money, resources and time. Though most of the big data tools are developed openly and distributed freely there can be lot of incidental, non-direct costs associated with collecting, processing and managing data with them. For example, there are the operational costs collecting data at such scale, network costs moving them, server costs storing and processing them, cost of procuring and supporting specialised tools and the human resource cost in hiring and training people who are capable for dealing with them. Though there are economies of scale at larger scales, the overall resources required to manage big data effectively can be several folds of what is needed for a traditional dataset. This makes it important to look at the data in our hands closely and carefully so that we can make informed decisions on how 'big' it is and choose the methods which are the most suited for such dataset.

4.1.2 How big are the Wi-Fi probe request datasets?

In this section we take a detailed look at the three sets of Wi-Fi probe requests collected as described in chapter on data collection using the 5Vs big data framework. Our aim is to understand the nature of the data in each dimension and thus evaluate the challenges we face in that specific dimension leading to

a bespoke solution. We look at each set of data in each dimension and try to answer the following questions,

1. How can this dimension be measure objectively?
2. How big is the data in terms of the defined measurement?
3. How does it data compare with datasets in other disciplines?
4. How can we describe the size of the data?

We then combine these isolated evaluations to form a combined description of the datasets. This is then used as the basis for developing a list of requirements for designing the data processing and management toolkit.

Volume

Probe requests data, being dynamic and continuous, cannot be quantified as an absolute static number in terms of volume. Hence we use a long term measurement - yearly rate, for each location instead. On shorter datasets such as the pilot study, we estimate the yearly volume linearly from the available data. We standardise this measure as the amount of disk space needed to store the collected data when encoded in text form. It is important to note that this can be reduced many folds by using compression or binary formats but we chose text since it the de-facto standard for exchanging data.

Table 4.1: Comparison of volume or size of the datasets of Wi-Fi probe requests.

Study	Maximum*	Minimum*	Average*	Total**
Pilot Study	134	3	54	48.3
Main Study	6.1	2.4	4.42	4.1
Smart Street Sensor	5.4	0.001	0.8	0.8

** Measured/ Estimated for each location in gigabytes per year. ** Measured/ Estimated for 920 locations in terabytes per year*

We can see that there is a lot of variability in the volume of probe requests generated at a given location. This mostly depends on how many mobile devices are present around the location. We observe that when we collect most of the information present in the probe requests in a busy area such as Oxford street in the Pilot studies, we generate around 50 terabytes of data in a year. But in a more real world setting such as the Smart Street Sensor project where the sensors fail at times and the amount of data collected is optimised, the volume is around a 1 gigabyte. The total volume of data we deal with in the case of a national scale project with around 920 sensors running for around 4 years is around 2 terabytes. A comparison of this to datasets from other disciplines is shown in Figure 4.1. It is key to note that the y-axis is scaled exponentially.

We can see that the probe requests data is not truly 'Big data' as experienced in other fields. It is only when we reach a complete coverage, i.e, putting a sensor at each retail establishment in UK, our estimated data volume reaches around 250 petabytes which is comparable to scales experienced in other fields such as particle physics and world wide social networks. At the same time, the scale of probe request data is not small either. The volume of 2 terabytes is more than the memory available in any desktop systems and is more than any of them can process in a timely manner. Summarising from the above, we can confidently say that the probe request datasets are 'Medium Data' - especially the dataset collected by the smart street sensor project. Though it

has potential to scale into a truly big dataset, for the purposes of this research we can classify it as 'Medium data' in the volume dimension.

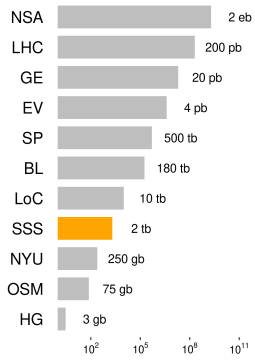

Figure 4.1: Comparison of volumes of datasets across various disciplines.

NSA - National Security Agency, LHC - Large Hadron Collider, GE - Google Earth, EV - Event Horizon project, SP - Spotify music, BL - British Library data store, LoC - Library of Congress, SSS - Smart Street Sensor, NYU - New york city Uber trips 2009-15, OSM - Open Street Map and HG - Human Genome Project

Velocity

Velocity is the rate at which the data is collected over time. It is significant when evaluating big data since some data which may not scale in terms of absolute volume but the speed at which they are received makes them challenging to deal with. A perfect example is the comparison between data generated by the Large Hadron Collider project by European Council for Nuclear Research and a world wide social network such as Facebook. Though their total volumes are comparable at 200 petabytes, the data from LHC is generated in concentrated experiments at a rate of 3 petabytes in 5 seconds while Facebook generates the same about in about a day or two. Since the size of an individual Wi-Fi probe request doesn't vary widely, we define the velocity of this dataset as the number of requests received at a given location at a given location within a given time interval. Though the precision of the time measured during data collection is in microseconds, the practical data transfer resolution in all the datasets is around 5 minutes. Hence we measure velocity of out datasets in terms of number of requests every 5 minutes. Table 4.2 compares the datasets we collected on Wi-Fi probe requests in terms of their volume.

Study	Maximum*	Minimum*	Average*	Total**
Pilot Study	8577	188	3469	3.20
Main Study	1362	534	782	0.72
Smart Street Sensor	5024	6	408	0.27

Table 4.2: Comparison of velocity or speed of the datasets of Wi-Fi probe requests.

** Measured/ Estimated for each location in number of requests per 5 minutes. ** Measured/ Estimated for 920 locations in Millions of requests per 5 minutes*

We observe that locations can receive up to 8500 requests in 5 minutes or can get no request at all depending on the time and how busy it is. But we can see that on average a national scale project with around 900 locations generates around a million requests every 15 mins. Compared to the LHC's 180 billion records or Google's 190 million searches per 5 minutes this seems to be not high speed data. However, this is much faster compared to traditional data sources such as census or geographical surveys which are updated anywhere between 6 months to 10 years.

To summarise, in terms of velocity, the Wi-Fi probes data can be described as 'Medium' at best. The methods dealing with the data should be time sensitive and be able to deal with a continuous stream of data but at the same time need not be real time or need sub-second latency. Since the Wi-Fi probe requests don't have actual location information the mobile devices, it does not have the similar value in real-time analytics as shown in comparable location or movement based datasets.

Variety

Variety is defined by the amount of variance in the type and characteristics of the data. Since variety is hard to quantify and compare across disciplines we evaluate the dataset subjectively for the variety present in it. The data transmitted in a Wi-Fi probe request is defined by the 802.11 Wi-Fi specification

[8] and every probe request has to have a set of mandatory fields for Wi-Fi to work. This set of fields is the same everywhere across the world and the specification, especially the probe request part, has remained stable over years. Though there is some variability allowed within the specification, being part of a global standard, the data collected is heavily structured in general.

The first set of variety present in the Wi-Fi probes data set arises from the 'information elements' part of the probe request. The structure of a probe request is discussed in detail in the data collection chapter and is summarised in Figure ?. Essentially the information about the capabilities and type of the mobile device is encoded in the information elements part of the probe request and this information is optional and is implemented at the discretion of the manufacturers. As this information elements are demonstrated to be useful in successfully fingerprinting the mobile devices [9], mobile devices increasingly don't include any information in them. Emergence of manufacturers with large market share and narrow set of device models such as Apple and Samsung also reduce further variability in them. The second set of variety in the dataset arises from the rate at which these probe requests are generated by the mobile devices. Unlike devices which generate data on events or at regular intervals, mobile phones generate probe requests at a rate based on various factors. Though this leads to some challenges in counting footfall from these probe requests the variability exhibited here is neither so large nor so complex that traditional methods could not deal with them.

Comparing with some of the big data encountered in unstructured data collected over web such as social networks or other sensor based methods, the variability here can be considered trivial. Further when we convert these probe requests in to footfall counts, the variety in the dataset drops almost to zero as it becomes just an ordinal data point varying in geography and time. Summarising the above, we can confidently say that the Wi-Fi probe request data does not exhibit any 'big data' properties in the variety dimension.

Veracity

Veracity is defined as the amount of abnormality present in the data in the form of inaccuracies, biases and noise. Similar to variety, veracity is hard to quantify hence required a subjective evaluation. Being sensor collected data, veracity is the dimension where the data exhibits most 'big data' properties.

First set of veracity in the dataset arise from the fact that it is collected through sensors located in multiple locations which communicate to the central server using 3G mobile data connectivity. We know from experience that the sensors are unreliable and fail to send back data regularly due to various reasons. More over the sensors are installed and uninstalled regularly as partners join and leave the project. This results in a data stream which is often erratic and incomplete with large gaps in them. In addition to this the sensors need to be rebooted regularly due to issues or updates leading to small gaps as well. Since the sensors are part of retail establishments they can be switched on and off regularly in some of them as well. Figure 4.2 demonstrates the veracity of the data in terms of missing data for a sample of locations in London. All the above pose immense challenges when we attempt to aggregate the data where we have to estimate and fill these gaps.

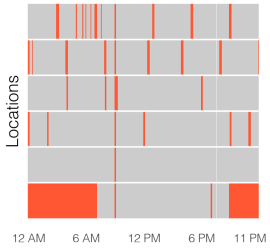

Figure 4.2: Missing data from five locations at Tottenham Court Road, London on 15 January 2018 demonstrating the veracity of the data.

[8] IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534, Dec 2016

[9] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: 10.1145/2897845.2897883

There is also a lot of variability in the physical location of the sensors and the area of measurement. The sensors may report higher or lower count due to their configuration and the context of their location as discussed in chapters pertaining to data cleaning. This leads to a situation where the accuracy of the data collection varying quite widely across location and times [10]. It is often not clear if the change in the data is due to actual changes at the location or just the change in the configuration of the device. For example, Opening of a mobile shop next door to the sensor can increase the estimated footfall without any change in actual footfall at the location.

[10] Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017

Finally we also have to work within the changing mobile landscape. Though the Wi-Fi probe requests are standardised by IEEE, the mobile manufacturers have started adopting obfuscation techniques to protect the privacy of the users. This started with randomisation of MAC addresses, removal of information elements and generally getting more sophisticated with new versions of operating system. There is also bias in terms of operating system adoption and change in market share between manufacturers. There is no inherent structure or information on what is changed and how often these changes occur which leads to questions on the continuity of the data over long periods of time.

Summarising from the above, we can confidently conclude that Wi-Fi probe requests dataset shows 'Big data' characteristics in terms of its veracity and requires appropriate tools and methods when aggregating, analysing and modelling it.

Visualisation

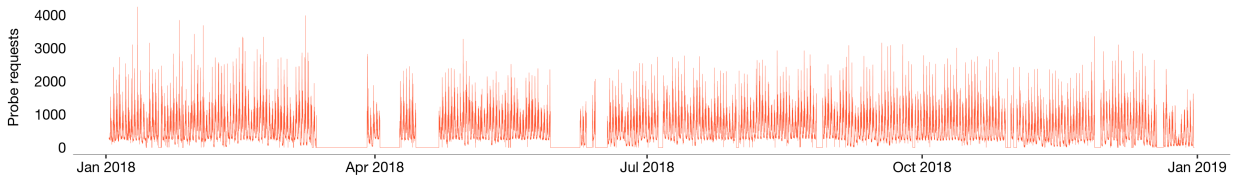

Figure 4.3: Number of probe requests collected for every five minute interval at Tottenham Court Road, London on the year 2018 showing the visual complexity of data in the time dimension.

Visualisation is closely related to volume, velocity and variety of the data. The Wi-Fi data due to its non-trivial volume and velocity, exhibits similar characteristics and challenges in terms of visualisation. Since there is not much variety in the dataset, when we process the raw data into footfall counts we are left with just the time, location and footfall count for each data point. Out of these, location and footfall counts are easy to visualise but time exhibits big data properties. This is primarily due to its granularity at 5 minute intervals and longitudinal nature of the data collection. The major challenge with Wi-Fi data is to simplify and visualise them in a legible way while showing change in term of time. The veracity of the data presents challenges in simplifying them and the volume poses challenges in maintaining legibility. We also have to take the 'near real time' aspect of the data into consideration while visualising them. There is a clear need for always on, interactive, real time dashboards with geographic capabilities in addition to the capabilities of traditional desktop GIS. There is also need for multiple linked dynamic visualisation platform

for separating the scope of the visualisation into manageable units. Figure 4.3 demonstrates the illegibility of simple visualisations of the data due to granularity, variability and veracity. We can safely say that the Wi-Fi probe requests dataset is at best 'Medium' in the visualisation dimension.

Summarising the above discussion, we can conclude that the datasets collected from Wi-Fi probe requests are at best of 'medium'. They show the most big data characteristics in terms of their veracity. In rest of the dimensions the datasets are not truly big data and we need to look at tools and methods appropriate to their size. The toolkit we devise need to be able to deal with their mid-size volume, velocity and visualisation dimensions and at the same time need to able to deal with the large amount of veracity of in them. Figure 4.4 illustrates the summary our discussion. This leads us to devise a 'medium data toolkit' which can be used without incurring the extra cost and complexity introduced by big data tools while be able to handle the data at hand.

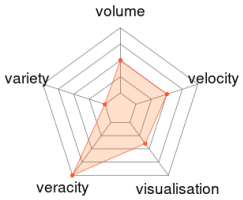

Figure 4.4: Big data characteristics of the Wi-Fi probe request datasets in their corresponding dimensions

4.1.3 *A Survey of Methods and Tools*

Having classified the Wi-Fi probes dataset as a 'Medium' sized data, in this section, we survey the tools and methods available at various stages of the data processing and management process - data collection, storage and retrieval, processing and analysis, visualisation. We first survey the tools available in each stage and specifically look at their suitability for Wi-Fi probe request datasets in terms of the following characteristics,

- *Performance* - How much data can be processed in a given time?
- *Flexibility* - How easy it is to change the scale and scope?
- *Complexity* - How many components or parts are involved?
- *Cost* - How much money or infrastructure do they require?

We then discuss the principles of UNIX philosophy and how it helps in solving similar sized problems in computer science. Finally we pick and connect the tools to devise our toolkit which is best suited for out Wi-Fi probe request dataset.

Collection

We discussed various technologies used in collecting passive data on ambient population and pedestrian movement in the literature search. In this section we look at tools and methods used to collect Wi-Fi based data passively. The primary considerations for evaluating data collection strategy are the scale of the infrastructure, expertise and effort required to implement it and cost involved.

There have been numerous sensors, tools and associated software platforms made available for data collection under the umbrella of 'internet of things'. We start by looking at different approaches in the Wi-Fi data collection tools and try to reason the most appropriate solution for our research. On one end there are low level and low cost bespoke solutions which require lot of effort to implement and maintain. On the other end there are turn key solutions which doesn't require lesser effort but costs considerably more. The key is finding a

balance between both while satisfying the requirements of the project. Since the Wi-Fi data is medium sized in terms of volume and velocity, we can deal with solutions with less than optimal scalability but since the data is 'big' in terms of veracity the toolkit has to give us most flexibility. Essentially, we are looking for a data collection methodology which prioritises flexibility and cost while performing moderately in terms of scalability and complexity as illustrated in Figure 4.5.

Type of solution	Examples
Bespoke	Micro-controllers with Wi-Fi modules e.g. Audrino + ESP8266
Turn-key	End to end commercial services e.g. Blix, Euclid, Pygmalios etc.
Ideal	**General purpose hardware e.g. Raspberry Pi, Repurposed mobile devices - Tablets, Phones etc.**

Table 4.3: Examples of different types of Wi-Fi based data collection solutions.

In terms of hardware, an example of a highly customised solution would be a micro-controller, such as Audrino, coupled with dedicated Wi-Fi module and programmed with custom software to collect the exact data needed. Designing and implementing of such system is time consuming, cumbersome and usually involves significant cost but it can also be highly flexible, efficient and cheap to deploy. On the other end of this spectrum, we have end-to-end solutions such as Blix, Walkbase, Ecuclid, Retail next, pygmalios etc. where the data is collected through multiple sensors and sources and syndicated into a clean footfall information by a third party service provider. These platforms for footfall data collection and analysis have the advantage of being quick and easy to develop and deploy while they can also be highly inflexible for changes and turn out to be costly when scaled up. A middle ground here is to use a general purpose hardware such as single board computers or repurposed mobile devices, augment them with additional hardware modules and use general purpose scripting languages to write software for them. This way we avoid low level hardware or software design and implementation while maintaining good amount of flexibility. Table 4.3 shows some examples of such systems while highlighting an ideal system.

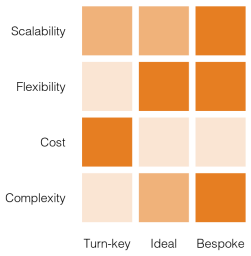

Figure 4.5: Characteristics of types of Wi-Fi data collection tools at each end of the spectrum compared to an ideal candidate
(Darker colors show higher score)

The Smart Street Sensor project uses its own proprietary sensor system designed and instrumented by the data partner. The design and implementation decisions were made with the commercial application in mind and is not entirely relevant to our discussion in the context of our research. For the research conducted with the data, it is necessary to understand the data collection process and make sure it aligns and integrates with the rest of the toolkit. As discussed in the data collection chapter, the methodology used in the smart street sensor project satisfies our requirements. The toolkit we designed to collect other datasets are in-part inspired by this methodology or a modified version to include more flexibility. The toolkit consists of Raspberry Pi, Linux, tcpdump or tshark [11] and nodejs. Raspberry Pi and the Linux OS provides a general purpose base system and hence the flexibility. On top of this we built our data collection system by assembling open source and free network analysis tools such as tcpdump and tshark along with other tools providing functions such as scheduling, personal data obfuscation and data

[11] Gerald Combs and Contributors. Wireshark - network protocol analyzer. https://www.wireshark.org/about.html, 2018

transmission with scripting languages like nodejs and bash.

Storage

Data storage technology is one of the most diverse landscape in terms of both methods and tools available. It has been constantly in research and development since the beginning of computing and is one of the fastest changing landscapes with the advent of big data paradigm. A comprehensive review of storage solution warrants a chapter in itself so we restrict our survey to an outline of most significant approaches and corresponding systems and tools.

At one end of the spectrum is one of the most underappreciated for of data storage - File systems. Though they seem like a low level interface for storing data, file systems have their advantages as well. When the data is not complex or inter-related, flat text files in file systems could be the fastest way to store, search and retrieve data. Since operating systems are usually optimised to manage storage media through file systems, they involve no additional overhead and are extremely reliable. The hierarchical file systems use in most of the operating systems act as an index with hierarchical data. The major disadvantage of file systems is that they are not useful for managing data with any kind of complexity. This is the primary reason why database management systems are developed on top of file systems.

Database systems can be broadly divided into relational and document based. The relational databases are optimised to deal with relational data and usually enforce strict structure for the data In general they can handle large number of rows and are designed to scale vertically. Most relational database systems try to guarantee ACID [12] compliance and hence used in critical systems such as financial operations, sales etc [13]. The document based databases are optimised to deal with unstructured data and can doesn't need a strictly defined scheme. In general they can handle large number of columns and are designed to be distributed and scaled horizontally. Being distributed, most document based databases try to pick a focus and compromise on others as specified in CAP theorem [14]. There a numerous databases systems which prioritise different things and the right solution depends on the properties of the data and the requirements of the project.

Since the publication of the paper on 'Google file system' by google [15]. There have been significant effort in designing and building 'big data' file storage systems which can large data in the range of petabytes. These systems are designed to be distributed and optimised for high throughput for queries on them. Hadoop Distributed File System (HDFS) is one such file system which is also the most widely adopted. There are numerous cloud based, third-party solutions built with these file systems making them easy to use. There are also numerous tools, libraries and frameworks which emulate the features of database systems on these distributed file systems making them easier to use further. The primary advantage of these systems is the sheer scalability they provide when it comes to data volume. The primary disadvantage is the associated overheads in terms of cost and time incurred in learning, designing and implementing them. Unless the project is sufficiently large, the advantages gained usually do not justify the overheads introduced.

[12] Atomicity, Consistency, Isolation and Durability are properties which make sure that the data in the database is valid even during failures.

[13] Theo Haerder and Andreas Reuter. Principles of transaction-oriented database recovery. *ACM Comput. Surv.*, 15(4):287–317, December 1983. ISSN 0360-0300. DOI: 10.1145/289.291. URL http://doi.acm.org/10.1145/289.291

[14] Brewer's theorem or CAP theorem states that it is impossible to simultaneously guarantee consistency, availability and partition tolerance in a distributed data store.

[15] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY, 2003

Table 4.4 summarises the above discussion along with relevant examples.

Approach	Data size	Examples	Comments
File system	< 10 TB	ext, ntfs, zfs, btrfs	Simple and efficient. Best for hierarchical data. Cannot handle complex connected data.
Relational DB	< 5 TB	MySQL, PostgreSQL	Handles structured and relational data. Optimised for large amount of rows and tries to guarantee validity.
Document DB	<10 TB	MongoDB, Cassandra	Handles unstructured data. Optimised for large number of fields and distribution to multiple clusters. Tries to focus on any two guarantees of the Brewer's theorem.
Distributed FS	> 10 TB	HDFS, Ceph, GFS	Optimised for really large datasets which need to be distributed over multiple nodes.
Cloud Storage	> 10 TB	AWS, Azure, SWIFT	Implements distributed file systems on the cloud. Has more reliability and scalability than local implementations.
Data Warehouse	> 10 TB	Hive, Hbase, Impala, Presto	Interfaces built on top of distributed file systems to emulate capabilities of relational databases on them.

Table 4.4: Various data storage approaches and their characteristics.

We saw that the Wi-Fi probe request datasets are 'Medium' sized hence we can safely eliminate distributed file systems for storing them. Though the smart street Sensor project uses Azure Blob Storage, when the data is downloaded to the local servers at the university, we can just store them in the file system because of their size (2TB approx.) and the hierarchical nature. The folder structure of - year/month/day/location/interval/ with individual text file, enable us to query the data for any given location at any interval nearly instantaneously without any further database operations. When this raw data is processed into 5 minute counts, we require more relational queries. For this purpose a relational database is sufficient as volume is quite small (25GB approx.). We chose PostgreSQL because of the PostGIS extension which gives us flexibility in handling geographic data.

Processing

We saw that the data is medium in terms of volume and velocity and shows big data properties in terms of veracity. Hence we require tools which are capable of dealing with the veracity of the data while being able to manage the volume and velocity. The traditional approach to deal with such dataset is to load it into a general purpose analysis tool such as R or a GIS packages and process it. The size of the dataset and the lack of meaningful complexity of geography element in the data eliminates the use of GIS packages. Scripting languages such as R and Python can deal with the dataset and its requirements but the time taken to do so increases exponentially with the size of the data as the size of objects in memory increases. Figure 4.6 illustrates the increase in processing time with respect to number of location for a simple exercise where a day's worth of raw data is parsed and aggregated into number of probe

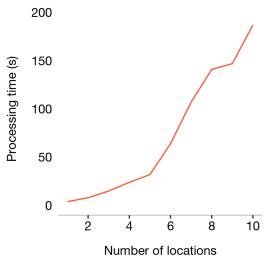

Figure 4.6: Exponential increase in the processing time when using traditional methods.

The processing involves parsing JSON data received for a single day at each location and aggregating them as number of probes requests received in every five minute intervals.

requests per 5 minute intervals (The code used to produce these benchmarks are detailed at Section 7.1.1). This becomes prohibitively expensive as the number of locations and complexity of the processing increases. Though this can be improved with more efficient coding practices, the margin of improvement is quite limited hence creating the need for better techniques. It is important to note that data processing is done in two stages - the first stage where the raw Wi-Fi probe requests are filtered, cleaned and aggregated into footfall counts and second stage where the footfall counts are in turn analysed to produce reports and visualisations. The traditional methods are sufficient for the second stage of the processing and the first stage is the one which requires a better solution.

On the other end we have big data analysis tools which are built for dealing with extremely large amount of data. Since the publication of the paper on MapReduce, there have been immense developments in the Big data analysis landscape. There numerous distributed programming tools to use the data stored within a distributed storage system each focussing on specific type of data and analysis. A concise, non-comprehensive list of types of data or specialities and corresponding big data tools is shown in Table 4.5.

Table 4.5: Various types of big data processing tools and corresponding examples.

Tools	Speciality
General purpose	MapReduce, Spark
Real-time streams	Flink, Pulsar
Events or messages data	Storm, Kafka, Flume
Networked or graph data	Tinkerpop, Corona
Scheduling	Oozie, Falcon, Azkaban
Turn-key platforms	SpringXD, Cask Data

We can rule out the necessity of the above big data tools since our dataset is neither big enough nor fast enough. The dataset does not have any specialised structure such as graph or network but just a very minimal component of geography to it. Using any of specialised big data tools is just going to introduce immense overheads without any added benefits. We need something in-between the above two approaches where we is sufficiently fast and flexible for our datasets.

This is where we come across the possibility of using standard Unix tools along with connecting them to create a processing pipeline. In some cases, a data processing pipeline made using command line Unix tools have been demonstrated to be 230 times faster than using big data toolkits [16]. The command line tools were developed as parts of Unix operating system for processing text. They are developed in line with the Unix philosophy which focuses on modular and minimal software development. The core tenants of the Unix philosophy has been summarised by Doug McIlroy as below,[17],

1. Make each program do one thing well. To do a new job, build afresh rather than complicate old programs by adding new "features".
2. Expect the output of every program to become the input to another, as yet unknown, program. Don't clutter output with extraneous information. Avoid stringently columnar or binary input formats. Don't insist on interactive input.
3. Design and build software, even operating systems, to be tried

[16] Command-line tools can be 235x faster than your hadoop cluster, Jan 2014. URL https://bit.ly/2s2XZYI

[17] Malcolm D McIlroy, Elliot N Pinson, and Berkley A Tague. Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904, 1978

early, ideally within weeks. Don't hesitate to throw away the clumsy parts and rebuild them.

4. Use tools in preference to unskilled help to lighten a programming task, even if you have to detour to build the tools and expect to throw some of them out after you've finished using them.

These principles along with the 'pipe' operator gives us necessary tools to build more complex tools. We can replace most of the libraries we used in the R implementation of our processing with a corresponding command line tools and connect them together with a text interface to achieve similar pipeline. The first advantage of such design is that it is much more efficient than a monolith design. These tools being actively developed for since their invention are compiled as native binaries and are usually extremely optimised resulting in a much faster pipeline. Because of the design of the pipe operator, the individual parts of the pipeline are executed in parallel as chunks of data are passed through them thus avoiding the need to load entire datasets into memory which results in an exponential increase processing time with the size of the data. Being modular, we can even introduce process level parallelism to parts of the pipeline without any major change in the overall design. Finally the modular structure also gives us the advantage of using the best tool for any part of the pipeline.

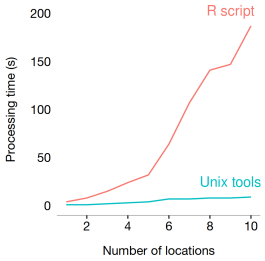

Figure 4.7: The increase in processing time with the Unix pipeline is linear thus improves the scalability compared to R based processing

Tools	R Library	Unix tool(s)
Move data to and from Azure blog storage, SQL server and Postgres	AzureR, odbc , RPostgreSQL	azcopy, mssql, psql
Convert data from JSON format to CSV	jsonlite	jq
Encrypt raw data for secure storage	Rcrypt	gnupg
Anonymise personal data into cryptographic hash	digest	openssl
Transform and manipulate tabular data	dplyr	find, cat, cut, grep, sed, awk, sort, uniq, column, paste, join
Impute missing value using time series analysis	imputeTS	Rscript
Visualise the results into maps and charts	ggplot2	Rscript
Create and manipulate geographic data	sf, rgdal	postgis, gdal

Table 4.6: Tasks in the processing pipeline, corresponding R libraries and equivalent Unix tools

All of the gives us an extremely minimal and efficient toolkit to process the raw Wi-Fi probes data into counts in a scalable way. Figure 4.7 compares the processing times of such Unix toolkit with the traditional R based toolkit as the data size increases. We can see that Unix toolkit performs extremely well and the performance gains are significant as the size of the data increases. For example, to process data for 25 locations, R based toolkit takes around 20 minutes while the Unix toolkit gets it done in 20 seconds (The code used to produce these benchmarks are detailed at Section 7.1.1). Table 4.6 shows the activities in our pipeline and corresponding libraries in the traditional R workflow along with the equivalent Unix tools. It is important to note

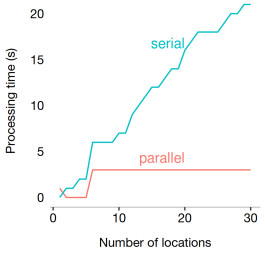

Figure 4.8: The scalability of the processing pipeline could be further improved with parallelising it.

that tools for doing specialised actions such as statistical analysis, machine learning and time-series analysis are built on top of scripting languages such as R and Python.

These can be embedded into our Unix pipeline as scripts running in corresponding front-ends such as Rscript or python. This toolkit can be further accelerated by parallelising parts of the pipeline using gnu-parallel [18]. For example, the previous example pipeline can be parallelised by spawning a pipeline for each location this reduces the processing time for a set of 25 locations from 18 seconds to 3 seconds. This done by utilising every processor cores available in the CPU. Figure 4.8 compares the processing times of the Unix toolkit with a parallelised implementation (The code used to produce these benchmarks are detailed at Section 7.1.1). Finally all the Unix tools discussed in this toolkit are open source and free software which has almost no cost in terms of resources. Since these tools are part of the POSIX specification [19] for operating systems, the expertise in their design and use are transferable to and from other disciplines thus reducing researcher time learning and using these tools.

[18] Ole Tange. *GNU Parallel 2018*. Ole Tange, March 2018. ISBN 9781387509881. DOI: 10.5281/zenodo.1146014. URL https://doi.org/10.5281/zenodo.1146014

[19] Stephen R. Walli. The posix family of standards. *StandardView*, 3(1):11–17, March 1995. ISSN 1067-9936. DOI: 10.1145/210308.210315. URL http://doi.acm.org/10.1145/210308.210315

Visualisation

In the last section we saw that the visualisation dimension of the data shows some level of complexity. The primary source of this complexity arises from the longitudinal nature of the data and the noise due to granularity of the data. For the processed dataset, traditional visualisation and mapping libraries with R is sufficient while the visualisation of raw data across long time periods for either for exploratory analysis or for communication needs some form of interactivity or simplification to be able to legible. Data driven documents (D3) [20] and Dimensional charting (DC) provides us with both of these requirements. Both of these tools can accept text based input and can fit with other Unix tools discussed earlier. In case of binary file output such as images or documents, they could be directed to the file system and then read into other programs.

[20] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL http://vis.stanford.edu/papers/d3

4.1.4 *The Bespoke 'Medium data toolkit'*

In this chapter we saw how the advent of internet and internet enabled devices has lead to significant increase in the amount of data generated and collected across disciplines. This data deluge and improvements in the capabilities of computing hardware has fuelled an explosion of research and development in tools and methods to deal with these 'Big data'. Though these big data tools promise huge improvements in processing capabilities, when used under wrong circumstances they can lead to unwanted overheads and costs. Thus we need a framework to examine and understand the scale of the data that is being used so that we use the right tools for the right purposes and the 5Vs of 'Big data' - Volume, Velocity, Veracity, Variety and Visualisation provides us with such frame work. Every dimension of big data poses unique set of challenges and we need make right decisions in choosing specialised tools and methods to overcome them.

We then closely examined the Wi-Fi probes data we collected with this framework and found that the data, though posed significant challenges with

traditional data processing techniques, do not exhibit 'big data' properties in all its dimensions. Only veracity of the data was found to have any meaningful big data properties, while volume and velocity was found to be 'medium' at best. The datasets lacked any variety and posed minimal challenge in the visualisation dimension because of it high temporal granularity. Thus we arrived at the requirements for a bespoke 'medium data toolkit' which is able to deal with these challenges.

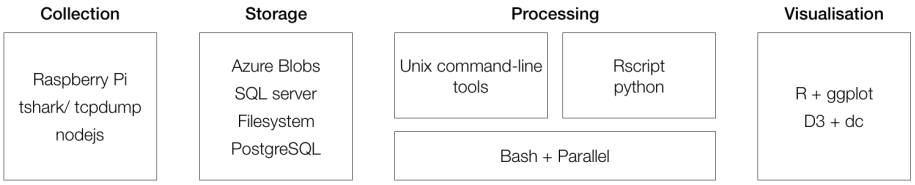

Figure 4.9: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

We undertook a brief survey of tools available for collecting, storing, processing and visualising the Wi-Fi probe request data and with the understanding of the data from the previous analysis chose the ones which are perfect for the datasets. For collecting Wi-Fi probes data in a scalable way, we chose a general purpose single board computers such as Raspberry-Pi along with open source tools such as tcpdump and tshark in a Linux environment. For data storage we narrowed in on using just the file system for the raw data and relational database management system for the processed counts. To process the raw data we chose to devise a processing pipeline using an assortment of standard Unix command line tools linked together using a shell scripting language and parallelised at the process level with gnu-parallel. We also demonstrated that this processing pipeline can be 400 times faster than the using a monolithic pipeline even with a small sample of locations. For visualisation we chose D3 and DC as the solution for communicating time information legibly. Finally we arrive a 'medium data toolkit', illustrated in Figure 4.9, which is best suited for the Wi-Fi probes dataset which we employ to process and examine the data further.

4.2 *Device fingerprinting*

In the past decade, Wi-Fi has emerged as one of the most commonly used technologies in providing high speed internet access to mobile devices such as smartphones, tablets and laptops in public and private spaces [**?**]. This has resulted in multiple Wi-Fi networks being available at almost every location in dense urban environments. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi network interfaces regularly broadcast a special type of signal known as 'Probe Requests' in order to discover the Wi-Fi networks available to them. This helps these devices to connect and switch between the Wi-Fi networks seamlessly.

Probe requests are low level signals standardised by IEEE 802. 1 specification [IEEE, 2016] for service discovery, and are implemented in any Wi-Fi capable device irrespective of the manufacturer or the model. This ubiquity and standardisation makes them an excellent source of open, passive, continuous, and wireless data generated by Wi-Fi capable devices present at any given time and location. Considering the unprecendented levels of mobile device ownership in recent years, we can, in turn use this data to understand the population distribution in highly dynamic urban environments with high spatial and temporal granularity [Freudiger, 2015, Kontokosta and Johnson, 2017]. While a Wi-Fi based method to collect data offers us various advantages such as, easy scalability and efficiency in terms of cost and time, it also introduces few systematic biases and uncertainities in the collected data along with the serious risk of infringing on the privacy of the mobile users. In this section, using the set of probe requests and manual counts collected at various high street locations across London, we demonstrate that pedestrian footfall at these locations can be estimated with considerable precision and accuracy while protecting the privacy of the pedestrians.

Unlike GPS, the location of the Wi-Fi enabled mobile device cannot be directly inferred from Wi-Fi, however there are reliable methods to triangulate the location of mobile devices from the locations of known access points (AP) and the signal strength reported by them [He et al., 2003, Moore et al., 2004, LaMarca et al., 2005]. This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments [Zarimpas et al., 2006, Kawaguchi, 2009, Xi et al., 2010]. Utilising this, we can easily and quickly estimate trajectories of the mobile devices [Musa and Eriksson, 2012] which can be used similary to the GPS trajectories to understand individual travel patterns [Rekimoto et al., 2007, Sapiezynski et al., 2015], crowd behaviour [Abedi et al., 2013, Mowafi et al., 2013], vehicular [Lu et al., 2010] and pedestrian movement [Xu et al., 2013, Fukuzaki et al., 2014, Wang et al., 2016]. Such data can also be used in transportation planning and management to estimate travel time [Musa and Eriksson, 2011] and real time traffic monitoring [Abbott-Jard et al., 2013]. Using techniques demonstrated by [Franklin et al., 2006] and [Pang et al., 2007], along with information present in the probe requests, one can even model interactions between the users [Cheng et al., 2012, Barbera et al., 2013, Cunche et al., 2014] such as predicting which of them are most likely to meet again [Cunche et al., 2012]. Using the semantic information present in these probe requests it even is possible to understand the nature of population at a large scale [Di Luzio et al., 2016] .

Although extensive research has been carried out on this subject with feasible and favorable results, in recent years, one of the major challenges faced in such attempts has been the increasing attempt by mobile phone manufacturers to protect their users' privacy by anonymising the globally identifiable portion of the probe requests [Greenstein et al., 2008]. Various methods have been devised to overcome this anonymisation process such as estimating the device model information from a known dataset of manufacturers and device behaviours [Martin et al., 2016]; Scrambler attack using a small part of the physical layer specification for Wi-Fi [Vo-Huu et al., 2016, Bloessl et al., 2015]; and timing attack where the packet sequence information along with information elements present in the probe request frame is used [Matte et al., 2016, Cheng and Wang, 2016]. A combination of these methodologies has been proven to produce de-anonymised globally unique device information [Vanhoef et al., 2016, Martin et al., 2017]. These approaches usually result in serious risk of breach of privacy of the users of the mobile devices by revealing their identifiable personal information.

There is a clear gap in the research for exploring methodologies for estimating the number of unique mobile devices from a set of anonymised probe requests, without the need to reveal their original device information. Such a technique has various applications such as uncovering the urban wireless landscape [Rose and Welsh, 2010], revealing human activity at large scales [Qin et al., 2013], estimating pedestrian numbers in crowds [Schauer et al., 2014, Fukuzaki et al., 2015], and even counting people in hyper local scales such as queues [Wang et al., 2013]. With enough infrastructure to collect such information we can even aim to generate a real-time census of the city [Kontokosta and Johnson, 2017]. With this background, we set out to devise and implement a methodology to reliably estimate human activity such as pedestrian footfall from Wi-Fi probe requests without risking a breach of privacy of the users involved.

Methodology

The primary aim of this research was to enable us to collect a series of probe requests and process them into a usable pedestrian footfall count. We did this by using a Wi-Fi receiver to collect probe requests broadcast by mobile devices, filtering out the background noise, and aggregating them based on the device that generated them. In this section, we examine the characteristics of probe requests in detail, devise a methodology to collect these probe requests in public areas, examine the systemic biases and uncertainties in the data collection method, and devise data processing methods to overcome these challenges. Finally, we compare the processed footfall counts to the ground truth recorded by primary surveys.

Probe requests are a special type of management packet broadcast by Wi-Fi enabled devices as part of their various functions such as scanning for available APs and quick geolocation by triangulation based known APs, etc. These are broadcast by all Wi-Fi enabled devices regardless of the manufacturer, type or model of the devices, although there is some variation in the frequency and the content of the information transmitted through them. In some cases, such as Android devices, these are broadcast even when the Wi-Fi functionality

has been turned off by the user so that the device can immediately connect to networks when the functionality is switched back on. Since some devices even use the probe requests as a less accurate form of localisation, they continuously send probe requests when Wi-Fi has been switched off. Thus, these signals can be used to reliably identify the presence of Wi-Fi enabled mobile devices. Being a first step of connection initiated by the mobile device, these packets have information regarding the characteristics of the mobile device itself. Some of the key information we can infer from these requests are,

1. **Media Access Control (MAC) address** which is a name identifying the wireless hardware of the mobile device,
2. **Sequence number** of the request for the mobile device to keep track of the responses,
3. **Time stamp** at which the request was received by the AP,
4. Total **length** of the request in number of bits, and
5. The **strength of the signal** received by the mobile device.

The MAC address is the primary identifier for the mobile device and has two parts. The first part is the Organisationally Unique Identifier (OUI) which provides information about the manufacturer of the device and the second part is the identifier for the device. In modern devices, to protect users' privacy, the second part of the MAC address can also be randomised and hence may not be unique to devices. When the MAC address is randomised, it is marked as such by setting a specific bit in the probe request packet as 1. Although sequence number of the packet is strictly unique to a mobile device, we hypothesize that we can use them to estimate the number of unique devices as demonstrated by [Vanhoef et al., 2016]; where optional information present in the probe requests - Information Elements (IE) along with the sequence numbers, have been used to fingerprint the devices. This approach has become increasingly difficult as mobile phone manufacturers have severely limited the IEs present in the probe requests thus leading us to explore methods which use only the sequence numbers. This also affects the established commercial solutions using Wi-Fi probe requests such as Blix, Walkbase, Euclid Analytics, RetailNext etc. There has been another solution proposed by [Hong et al., 2018] where the authors tried to solve the similar problem using a hidden markov models based trajectory inference algorithm but the scope of this research was limited to enclosed, exit controlled public spaces such as shopping malls, railway stations, etc.

Data collection was done with the help of custom sensors built from modifying the hardware used in Smart Street Sensors [CDRC, 2016] and updating them with custom software. The sensor is essentially a Raspberry-Pi device with Wi-Fi and 3G modules. It keeps the Wi-Fi module in 'Monitor mode' and uses the open source software - Wireshark [Combs and Contributors, 2018] to passively collect all packets sent to 'broadcast', marked with type as 'management', and subtype 'probe requests'. The MAC address in these probe requests is obfuscated at the device level using a cryptographic hashing algorithm and transmitted through 3G connection to a central database via web-sockets protocol, where it is stored in a PostgreSQL database for further

analysis. The random salt used in the hashing algorithm was rotated regularly to further mitigate the risk of de-anonymisation of the hash. Though hashing cannot completely ensure anonymisation as discussed in [Demir et al., 2014], it can sufficiently obfuscate the data; which along with a secure process of data handling, gives us reasonable security. An overall schematic of the data collection and storage process is shown in Figure **??**. The ground truth on the number of pedestrian footfall was recorded using a custom Android application - Clicker [Soundararaj, 2018]. This app logs accurate timestamps each time the surveryor records a pedestrian crossing the designated cordon line at the location. In addition to counting the pedestrians manually, this procedure results in the device broadcasting probe requests regularly, which in turn, gives us a 'known device' to calibrate our methodology against.

After collecting data, we began estimating the footfall or pedestrian activity from them by identifying the following potential uncertainties arising from our data collection method:

1. **Background noise** - since the extent to which Wi-Fi signals travel differs subject to various factors such as interference and humidity, it is close to impossible to restrict our data collection to a finite area of interest. This can lead to a significant background noise at certain locations. For example, a phone shop or a bus stop located next to the study area can artificially increase the number of probe requests received by the sensor. It is important to note that this method may not work effectively on study locations with complex configurations such as the source of noise and the area of study being located at the same distance from the sensor. This aspect is explored in detail in the broader case study in the following sections.

2. **MAC randomisation** - mobile devices in recent years have been using randomised 'local' MAC addresses for probe requests to protect the users from being tracked. This makes it impossible to tell if the probe requests are being sent by the same mobile device. This along with the previous problem can further increase the magnitude of error by several fold.

3. **Mobile ownership** - since the rate of mobile ownership can vary widely across geography and demography, we cannot assume that every mobile device translates to one pedestrian footfall. In addition to this, there is a long term overall increase in mobile ownership which may affect the number of probe requests collected overtime.

We propose the following internal and external validation methods to tackle each of these uncertainties.

Filtering with Signal Strength

One of the clues that we can use to estimate the distance between the mobile device and the sensor is the strength of the signal received by the sensor. The obvious approach was to first try and establish a relationship between the signal strength and distance and to use this to filter out the unwanted probe requests. However this approach was found to not be feasible, since the decay of signal strength with respect to distance is not always constant. For instance, signal strength varies with atmospheric conditions, the presence of

obstructions between the source and the target, the nature of these obstructions, and the strength (power level) of the source. This severely limits our ability to establish a simple conversion between reported signal strength and distance. As such, there was a need for a method which takes in to account all of these variables across the various locations.

We assumed that, in configurations where a specific source of background noise was at a constant distance, there must be a distinct pattern in the number of probe requests reporting signal strength corresponding to that distance. For example, if there was a phone shop next to our sensor where hundreds of phones regularly sent probe requests, there should be a sharp rise in the of number of probe requests with reported signal strength corresponding to the distance between the sensor and the phone shop, irrespective of the local conditions as shown in Figure **??**. We could identify these breaks in the data using traditional one dimensional clustering algorithms such as 'jenks natural breaks', 'k-means', 'quantile' and 'hierarchical clustering', etc. Since we were only looking for the break in the data and not for absolute values, the methodology should apply for all the variations due to micro site conditions reducing the overall noise in the collected data.

Clustering with sequence numbers

Since our primary unique identifier - MAC addresses are being anonymised by new devices, we needed to find other information present in the probe requests for use as a unique identifier. The obvious approach was to establish a factor of randomisation, and adjust the counts for the probe requests based on this factor. We found this approach to not be feasible since the proportion of devices which randomise the MAC addresses increased over time. There was also a wide variation in the frequency at which the devices randomised the MAC addresses and the method used for the process. This led us to look for a more generalisable approach which was independent of the device model. From our initial look at the data, we found that OUI and the sequence number of the packet was the most promising information to achieve this. First we divided our dataset into sets of probe requests with randomised and non-randomised MAC addresses by looking at the second character of the vendor part of the MAC address; if it was E, A, 2 or 6, then those addresses were identified to be randomised. We kept the MAC address as the unique identifier for the non-randomised requests and further divided the randomised ones in to sub-categories based on their OUI. We then identified unique mobile devices from within those sets, and assigned a unique identifier to each device.

The proposed algorithm created a graph where the probe requests represented the nodes; links were created between them based on the following rules:

- A link could go only forward in time.
- A link could go from low to high sequence numbers.
- A link could exist between nodes with a maximum time difference of α - time threshold.

- A link could exist between nodes with a maximum sequence number difference of β - sequence threshold.
- A node could have only one incoming link and one outgoing link, which is the shortest of all such possible links in terms of both time and sequence number.

The nodes were then assigned a unique ID based on the unique connected component they belonged to as shown in Figure **??**. This unique identifier was used in the place of MAC addresses for aggregation of the anonymised probe requests. Although the recycling of sequence numbers after 4096 led to multiple unique IDs being reported from a single device, a sample consisting of all randomised probe requests sent by "Google" devices showed that only 0. % of the sample had their sequence number reset. This led assume this to be inconsequential.

Calibrating with Ground Truth

Since proportions of mobile device ownership was an external uncertainty to our study and could arise from variety of spatio - temporal and demographic factors, we aimed to solve this by using a manual sample count at each location. We then calculated an adjustment factor, or an 'offset' for each location by comparing the sensor-based counts and ground truth. In turn it was then used to adjust the data reliably to reflect the ground truth in absolute numbers. This calibration can be carried out periodically at these locations to improve the quality of the estimation.

4.2.1 pilot study

Table 4.7: Comparison of clustering algorithms with a sample of 40000 probe requests

Algorithm	Time (s)	MAPE (%)
Quantile	0.002	27 %
K-Means	0.007	-23 %
Hierarchical Clustering	172.520	-9 %
Bagged Clustering	0.135	-30 %
Fisher	3.034	-30 %
Jenks Natural Break	556.279	-30 %

To start, we designed a small pilot study to validate the filtering and clustering methodology against the scale and complexity of data collected in an open public area such as a retail high street. We also aimed to find the algorithm which was best suited for the classification of signal strengths as 'low' and 'high' in order to filter out the background noise. The data was collected at Oxford Street, London on 20 December 2017 from 12:30 to 13:00 hrs, Wi-Fi probe requests were collected using the sensor described in Section **??** and pedestrian footfall was manually recorded using the Android app - Clicker [Soundararaj, 2018]. Being located at one of the busiest retail locations in the United Kingdom, the Wi-Fi sensor captured approximately 60,000 probe requests during the half hour period; 3,722 people were manually recorded walking on the pavement during that time. The surveyor positioned himself at the front of a store while carrying the sensor in a backpack and counted

people walking by the store on the pavement (3m wide approximately) using a mobile phone. The sensor was kept as close to the store window as possible, and the manual count was done as a cordon count in front of the store.

As a first step we aggregated the probe requests by their MAC addresses for every minute to generate a minute by minute count of the number of people near the sensor. We assumed that each MAC address corresponded to a mobile device and hence a pedestrian. We then compared this preliminary 'footfall' count to the actual number of pedestrians recorded manually to check for it's robustness. We used Mean Absolute Percentage Error (MAPE) as a measure of robustness of the count, since it provided a simple and quick measurement and the street conditions ensured that there are no intervals without any footfall. We found that the MAPE in the raw counts compared to the ground truth was around 425%. This suggests the presence of large amount of noise in the data which may have been generated by the sources of uncertainties discussed in Section **??** thus demonstrating the need for filtering the data.

We then classified the probe requests as 'high signal strength' and 'low signal strength' using various one dimensional clustering algorithms such as k-means, quantile, hierarchical clustering, bagged clustering, fisher and jenks natural breaks with the number of clusters set as 2. The results are shown in Table 4.7. We found that while hierarchical clustering and jenks gave us fairly low errors, they were too resource intensive for practical use with a larger dataset. We also found that k-means gave the quickest results with the lowest MAPE, closely followed by quantile algorithm. The cut-off point or threshold for the collected data with which we could classify as high and low was -71 dBm. We then removed all the probe requests which reported 'low signal strength' and repeated the same aggregation process as before to produce footfall count. This process resulted in a footfall count with a net MAPE of 30%. Although the results are encouraging we are still not completely confident that our filtering process is removing noise or has any correlation the configuration of sensor or position of the mobile devices. These concerns need to be addressed with a larger survey with multiple locations of varying orientations.

The next challenge was to identify the probe requests which are generated by the same device irrespective of the MAC randomisation process. We used the algorithm defined in Section **??** and assigned a unique identifier or signature to each probe request, independent of their the MAC addresses. Since we didn't know the nature or frequency of the MAC address randomisation process, we used the surveyor's mobile device as a reference. As the surveyor's device was being actively used to count pedestrians and it's Wi-Fi module was kept active without establishing connection to any network, the device was known to be continuously probing for new networks. We also knew that the OUI of the device was 'Google' and the device was regularly randomising it's MAC address, thus providing us an excellent reference with which we could optimise the parameters for our clustering algorithm. We then, by increasing the thresholds in steps of 1, found the minimum possible threshold for time and sequence numbers at which the algorithm clusters the reference device properly without over clustering the other probe requests. This process is shown in Figure **??**. We observed that the threshold for time

α and the threshold for sequence numbers, β are 16 seconds and 60 respectively This was undertaken on top the filtering done based on signal strength, and only for the probe requests with randomised MAC addresses. Figure **??** shows the results of this clustering process on a small set of randomised probe requests. The probe requests with different randomised MAC address are shown by the coloured points and the lines joining them show that those probe requests were most likely be generated by the same device. We finally aggregated the probe requests as we did before but with the device signature rather than MAC addresses. This results in a footfall count with a MAPE of -18% compared to the manual count. A comparison of minute by minute counts resulting from different filtering processes along with the ground truth is shown in Figure **??** illustrating the promising effectiveness of the methods.

To conclude, from the pilot study we found that both the filtering and the clustering methods we devised worked on complex real world data and resulted in final pedestrian counts within a MAPE of 20%. We also found that 'k-means' and 'quantile' are best algorithms for clustering signal strengths. Finally, we observed that the best thresholds for time and sequence numbers in the clustering algorithm is around 16 and 60 respectively.

4.2.2 *main study*

The methodology set out above was implemented in five different Central London locations at different times. Sensors were installed and data collected for extended periods of time. We also carried out manual counting at these locations across different times of the day. We then applied the methodologies discussed earlier to arrive at estimated pedestrian footfall and compared them with the corresponding manual counts. We finally evaluated the effectiveness of the processes with the Mean Absolute Percentage Error (MAPE) at the locations and report our findings below.

Table 4.8: Locations where sensors were installes, volume and speed of probe requests collected by the sendor and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format.

ID	Location	Type	Installation notes	Probe Requests	Footfall
1	Camden High Street	Phone Shop	Bus stop in front	9.9 (297)	3683 (33)
2	Central St.Giles	Restaurant	Seating area on both sides	3.9 (169)	0346 (05)
3	Holborn Station	Info. Kiosk	Overlooks station entrance	4.3 (303)	2956 (46)
4	Brunswick Center	Fast Food	Has seating area on one side	3.4 (210)	0960 (12)
5	The Strand	Tea Shop	Has phone shop next door	8.4 (382)	1969 (21)

The locations at which the data were collected are shown in Table 4.8. The locations were chosen for their diverse site conditions and unique sources of

noise around the potential location of the sensors. The position of the sensor at these locations with respect to the context is shown the Figure **??**. We can see that Location 5 is the 'cleanest' with one clear stationary source of noise (phone shop) while location 2 is the most complex due to the proximity of seating areas to the sensor. The sensors were operational through out February and March, while manual counts were conducted in these locations in half hour sessions on at least two different days. For the purposes of comparing with ground truth, we considered the data from sensors which correspond to the 12 sets of available manual counts. The schedule of data collection is shown in Figure **??**.

We begin by looking at the distribution of the signal strength reported by the probe requests across the locations. From the density plot shown in Figure **??**, we can observe that there is clear relation between the distribution of the signal strength and the distance and complexity of the source of noise. We can see that while location 5 shows clean difference between low and high signal strengths, location 2 is almost normally distributed. Intuitively we expected that location 2 and 4 would be harder to classify than locations 1, 3 and 5. We ran the k-means clustering algorithm and filtered out the probe requests which were randomised and had signal strengths less than the second break (threshold). It is important to note that we were dealing with relative thresholds of signal strengths which can vary with location and time of the analysis. We then aggregated the probe requests by counting the number of Unique MAC addresses present in every minute. We also removed devices that dwelled around the sensor by removing the MAC addresses which reappeared within the previous hour.

The results of the first stage of the filtering process along with the thresholds are shown in Table 4.9. Confirming our intuition, we see that the location 2 has the most MAPE followed by location 4, while the rest of them have highly reduced MAPE. It is significant that this method alone reduces our margin of error by 50 - 100% from the raw counts without any cleaning. This makes the signal strength filtering a quick and ideal method for practical applications, one which doesn't require absolute numbers such as creating large aggregated indexes to show long-term trends. We also found that the success of the signal strength filtering can be improved significantly by installing sensors so that the pedestrians and source of noise are at different distances from the sensor. This ensures that the distribution of signal strengths within the field of measurement is distinct from that of the surroundings.

We then ran the sequence numbers based clustering process on the rest of the probe requests to reduce the MAPE by almost 50 - 100% on all the sensors except for location 3. Location 3 is an outlier among all the other sensors since it is the only one with large amount of pedestrians very close to the sensor. This may be the reason behind the over filtering observed in the previous process. We finally ran the calibration process where we calculated the adjustment factors from the ratio between the manual counts to the sensor based counts for the sample period as shown in Table 4.9. We used them to adjust the counts to achieve a MAPE ranging from 10 - 50%. We observed that the sensors with people moving right next to them tend to under-count with our methodology, while sensors with seating next to them tend to over-count significantly. However, using the filtering process, we can reduce the error to

almost 10% closer to that of the ground truth.

Table 4.9: Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the filtering process

Sensor	Signal strength threshold	Adjustment factor	MAPE without any	MAPE after filtering signal	MAPE after filtering sequence	MAPE of final adjusted
1	-70	1.25	259	22	-13	9
2	-74	0.51	928	396	206	55
3	-72	1.60	87	-19	-31	10
4	-70	0.88	498	142	52	33
5	-72	0.80	473	84	38	11

4.2.3 *Conclusions*

Sentient technologies make measurement of the human activities that are the life blood of the smart city possible. Yet the data that they harvest are frequently relevant only to the sub-groups within society that avail themselves of particular goods and services – such as social media applications, transport modes or retail offers. In each of these cases, it is necessary to remember that the resulting data are by-products of consumer transactions, and will as a consequence, only pertain to users of the relevant goods or services. If the smart city is to be socially inclusive, it therefore follows that sentient data must represent entire populations, whether by design or by triangulation with external, population wide, sources. This is a non-trivial task, since the ebbs and flows of smart device-enabled citizens rarely pertain to any clearly defined population in either administrative or functional terms [Massam, 1975].

Our objective here has been to collect, rather than re-use, data on smart city functioning, by recording Wi-Fi probes and ultimately reconciling them with manual counts in order to infer ambient populations. The internal validation methodology set out in the technical sections of this paper, allied to external validation from pedestrian counts, renders the method inclusive and robust when recording activity levels in retail centres in real time. We have described the collection and processing of a novel consumer Big Dataset that enables valid measures of levels of footfall activity which has been scaled across a wide network of sensors [Longley et al., 2018]. In both conceptual and technical terms, it illustrates the ways in which passively collected consumer data can be 'hardened' to render them robust and reliable by using related procedures of internal and external validation.

Internal validation addresses the issues of screening out device probes that do not indicate footfall, and the further screening of device probes to 'fingerprint' the effects of MAC randomization. It is important to note that the filtering process work based solely on the information present in the probe requests and their temporal distribution. This ensures that although the mobile devices were uniquely identified, there was no further personal data generated by linking the probe requests to the users of the mobile devices. This method essentially gave us a way to estimate the footfall in real-time without identifying or tracking the mobile devices themselves. External validation then entailed reconciling adjusted counts with the footfall observed at sample locations. This procedure makes it possible to generalise from locations at which manual footfall surveys are conducted to all others in the system, and

to develop a classification of device locations that are more or less susceptible to noise generation.

This Wi-Fi based footfall counting methodology offers a large number of applications and benefits for real time spatial analysis. Since Wi-Fi based sensors are inexpensive and the data model is scalable, it is possible to use this methodology for a large network of sensors to gather granular data on pedestrian footfall. A snapshot showing a week's worth of precise footfall in area around Charring cross, London is shown in Figure **??** in order to demonstrate the potential for such a dataset. Projects such as SmartStreetSensors [Longley et al., 2018], may utilise this methodology to overcome the challenges introduced by the implementation of MAC address randomisation.

The vicissitudes of MAC randomisation, and the provisions of privacy legislation such as EU General Data Protection Regulations mitigate against tracking individuals across the smart city using this approach. This can be modelled using agent-based methods [Heppenstall et al., 2011], however. In our own research we have also begun to link store time-lagged till receipts to footfall, and have used such data to better understand the dwell times that characterise such different retail uses as stores with window displays and fast food restaurants. Such analysis not only provides a more nuanced picture of movement through retail areas, but also enables valorisation of micro sites within retail centres. In the UK, for example, this is of immediate practical importance in evaluating business rates on properties, and has still wider implications for the setting of retail unit rental values. There are obvious extensions to understanding the ebbs and flows of activities in the 24-hour smart city such as understanding urban mobility [Gariazzo et al., 2019] and conceptualising them with a people dimension [Nam and Pardo, 2011].

More broadly still, extensions to this strand of smart city research are likely to seek to differentiate the quality of different elements within footfall according to mission e.. travel to adjacent workplace zones, leisure, etc. and personal characteristics such as spending power. In this respect, future research may not only simulate linkage of harmonised footfall counts between sensor locations, but also link these in turn to disaggregate origin-destination matrices for bikeshare and other public transport modes. Our own investigations will consider these and other challenges to understanding the functioning of the sentient city.

5
Visualisations and Applications

5.1 Footfall Indices

5.2 Events Detection

5.3 Pedestrian Flows

6
Discussion and Conclusions

7
Appendix

7.1 Code used in the research

This section documents all the code that has been used in the research. The programming languages used are including but not limited to R, Bash, JavaScript and SQL.

7.1.1 Benchmarking the Data Toolkit

Implementation in R

This R script lists all files in a given folder, parses them as JSON data serially, aggregates the records for each time interval and finally writes it to disk as a CSV file.

```
#! /usr/bin/Rscript
suppressMessages(library(tidyverse))
suppressMessages(library(RJSONIO))

day <- "~/unorganised-files/ff_sample/2018/01/01"
sensors <- paste(day, dir(day), sep = "/")[1:25]
probes <- NULL

for(sensor in sensors) {
  files <- paste(sensor, dir(sensor), sep = "/");
  for( file in files ) {
    records <- fromJSON(file);
    location <- vector(); signal <- vector();
    mac <- vector(); packets <- vector();
    vendor <- vector(); type <- vector(); time <- vector();
    for(record in records) {
      t <- strsplit(strsplit(file, '\\.')[[1]][1], '/')[[1]][8]
      l <- strsplit(strsplit(file, '\\.')[[1]][1], '/')[[1]][7]
      signal <- append(signal, record$Signal);
      mac <- append(mac, record$MacAddress);
      packets <- append(packets, record$PacketCount);
      type <- append(type, record$PacketType);
      vendor <- append(vendor, record$VendorMacPart);
```

```
      time <- append(time, t);
      location <- append(location, l);
    }
    recordsdf <- data.frame(location, time, signal,
                            mac, packets, type, vendor);
    if(is.null(probes)) { probes <- recordsdf; }
      else { probes <- rbind(probes, recordsdf); }
  }
}

probes %>%
  group_by(location, time) %>%
  summarise(count = length(unique(paste0(vendor, mac)))) %>%
  write.csv("output-old.csv", row.names = FALSE);
```

Serial implementation in bash

This bash script lists all the files in a given folder, parses them into JSON data *serially*, aggregates the resulting records for each time interval and finally writes it to disk as a CSV file.

```
#! /bin/bash
awkc="awk -vFPAT='[^,]*|\"[^\"]*\"' -v OFS=','"
FOLDER="/home/ucfnbso/unorganised-files/ff_sample/2018/01/01/"
SENSORS=`ls $FOLDER | head -n $1`

for SENSOR in $SENSORS;
do
  jq_string=".[] | \
    [\"$SENSOR\",\
    ( (input_filename/\"/\" | .[ .|length-1 ] )/\".\" | .[0]),\
    .VendorMacPart+.MacAddress] \
    | @csv";
  cmd="jq -r '$jq_string' $FOLDER$SENSOR/*.pd \
    | sort | uniq \
    | $awkc '{print \$1,\$2}' \
    | sort | uniq -c";
  echo "$(eval $cmd)" > output-new.csv;
done
```

Parallel implementation in bash

This bash script lists all the files in a given folder, parses them into JSON data *in parallel*, aggregates the resulting records for each time interval and finally writes it to disk as a CSV file.

```
#! /bin/bash
awkc="awk -vFPAT='[^,]*|\"[^\"]*\"' -v OFS=','"
folder="/home/ucfnbso/unorganised-files/ff_sample/2018/01/01/"
```

```
sensors=`ls $folder | head -n $1`

jq_string=".[] | \
  [\"{}\",\
  ( (input_filename/\"/\" | .[ .|length-1 ] )/\".\" | .[0]),\
  .VendorMacPart+.MacAddress] \
  | @csv";
cmd="jq -r '$jq_string' $folder{}/*.pd \
  | sort | uniq \
  | $awkc '{print \$1,\$2}' \
  | sort | uniq -c";

echo "$sensors" \
  | parallel "$cmd" \
  > output-new-parallel.csv
```

7.1.2 *Wi-Fi Sensor*

7.1.3 *Data Processing Pipeline*

Bibliography

Command-line tools can be 235x faster than your hadoop cluster, Jan 2014. URL https://bit.ly/2s2XZYI.

Michael Abbott-Jard, Harpal Shah, and Ashish Bhaskar. Empirical evaluation of bluetooth and Wi-Fi scanning for road transport. In *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia*, page 14, 2013.

Naeim Abedi, Ashish Bhaskar, and Edward Chung. Bluetooth and Wi-Fi mac address based crowd data collection and monitoring: benefits, challenges and enhancement. 2013.

Daniel Arribas-Bel. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49:45–53, 2014. ISSN 01436228. DOI: 10.1016/j.apgeog.2013.09.012.

Daniel Arribas-Bel and Fernando Sanz-Gracia. The validity of the monocentric city model in a polycentric age: Us metropolitan areas in 1990, 2000 and 2010. *Urban Geography*, 35(7):980–997, 2014. ISSN 0272-3638.

Marco V Barbera, Alessandro Epasto, Alessandro Mei, Vasile C Perta, and Julinda Stefa. Signals from the crowd: uncovering social relationships through smartphone probes. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 265–276. ACM, 2013. ISBN 145031953X.

Michael Batty. Invisible cities. *Environment and Planning B: Planning and Design*, 17(2):127–130, 1990.

Michael Batty. Virtual geography. *Futures*, 29(4-5):337–352, 1997. ISSN 00163287. DOI: 10.1016/S0016-3287(97)00018-9. URL http://www.sciencedirect.com/science/article/pii/S0016328797000189.

Michael Batty. The pulse of the city. *Environment and Planning B: Planning and Design*, 37(4):575–577, 2010. ISSN 0265-8135.

Michael Batty. *The new science of cities*. Mit Press, 2013a. ISBN 0262019523.

Michael Batty. The future cities agenda. *Environ. Plann. B Plann. Des*, 40: 191–194, 2013b.

Michael Batty, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012. ISSN 1951-6355.

Luís M A Bettencourt. The origins of scaling in cities. *science*, 340(6139): 1438–1441, 2013. ISSN 0036-8075.

Luis M A Luís M.A. Bettencourt. The uses of big data in cities. *Big Data*, 2(1):12–22, 2014. ISSN 2167-6461. DOI: 10.1089/big.2013.0042. URL http://online.liebertpub.com/doi/abs/10.1089/big.2013.0042.

Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Jerome Dobson. Landscan. *Geoinformatics*, 5(2):34–37, 2002.

Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L. Urban. Landscan usa: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. In *GeoJournal*, volume 69, pages 103–117. Springer, 2007. ISBN 0165-0009. DOI: 10.1007/s10708-007-9105-9.

Bastian Bloessl, Christoph Sommer, Falko Dressier, and David Eckhoff. The scrambler attack: A robust physical layer attack on location privacy in vehicular networks. In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 395–400. IEEE, 2015. ISBN 1479969591.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL http://vis.stanford.edu/papers/d3.

Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*, pages 22–37. Springer, 2010.

M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000.

M Castells. *The Rise of the Network Society*, volume I. 2010. ISBN 9781405196864. DOI: 10.2307/1252090. URL http://www.lavoisier.fr/livre/notice.asp?depuis=e.lavoisier.fr{&}id=9781405196864.

CDRC. Smart street sensor project. goo.gl/E4tR8o, 2016. [Online; accessed 31-January-2018].

K Chen. An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*, 23(1):37–48, 2002. ISSN 0143-1161.

Linsong Cheng and Jiliang Wang. How can i guard my ap?: non-intrusive user identification for mobile devices using wifi signals. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 91–100. ACM, 2016.

Ningning Cheng, Prasant Mohapatra, Mathieu Cunche, Mohamed Ali Kaafar, Roksana Boreli, and Srikanth Krishnamurthy. Inferring user relationship from hidden information in wlans. In *Military Communications Conference, 2012-MILCOM 2012*, pages 1–6. IEEE, 2012. ISBN 1467317314.

Gerald Combs and Contributors. Wireshark - network protocol analyzer. https://www.wireshark.org/about.html, 2018.

Tomas Crols and Nick Malleson. Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility. *Geoinformatica*, pages 1–20, 2019.

Mathieu Cunche, Mohamed Ali Kaafar, and Roksana Boreli. I know who you will meet this evening! linking wireless devices using wi-fi probe requests. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*, pages 1–9. IEEE, 2012. ISBN 1467312398.

Mathieu Cunche, Mohamed-Ali Kaafar, and Roksana Boreli. Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, 11:56–69, 2014. ISSN 1574-1192.

S Cutter, Douglas B Richardson, and Thomas J Wilbanks. The changing landscape of fear. *The geographical dimensions of terrorism*, pages 1–5, 2006.

Levent Demir, Mathieu Cunche, and Cédric Lauradoux. Analysing the privacy policies of Wi-Fi trackers. In *Proceedings of the 2014 workshop on physical analytics*, pages 39–44. ACM, 2014.

Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014. ISSN 0027-8424. DOI: 10.1073/pnas.1408439111.

Adriano Di Luzio, Alessandro Mei, and Julinda Stefa. Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests. In *Proceedings - IEEE INFOCOM*, volume 2016-July, 2016. ISBN 9781467399531. DOI: 10.1109/INFOCOM.2016.7524459.

Jerome E Dobson and Peter F Fisher. Geoslavery. *IEEE Technology and Society Magazine*, 22(1):47–52, 2003. ISSN 0278-0097.

Jerome E Dobson, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley. Landscan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, 66(7): 849–857, 2000. ISSN 0099-1112.

Jerome E Dobson, Eddie A Bright, Phil R Coleman, and Budhendra L Bhaduri. Landscan2000: A new global population geography. Technical report, 2003.

Matt Duckham and Lars Kulik. Location privacy and location-aware computing. *Dynamic & mobile GIS: investigating change in space and time*, 3:35–51, 2006.

Nature Editorial. A flood of hard data. *Nature*, 435:698, 2008. ISSN 0028-0836. DOI: 10.1038/453698a.

Sara Irina Fabrikant. Towards an understanding of geovisualization with dynamic displays: Issues and prospects. In *AAAI Spring Symposium: Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance*, pages 6–11, 2005.

Katayoun Farrahi and Daniel Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing*, 4(4):746–755, 2010.

Emilio Ferrara, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. Visualizing criminal networks reconstructed from mobile phone records. In *CEUR Workshop Proceedings*, volume 1210, 2014.

Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732.

Jason Franklin, Damon McCoy, Parisa Tabriz, Vicentiu Neagoe, Jamie V Randwyk, and Douglas Sicker. Passive data link layer 802.11 wireless device driver fingerprinting. In *USENIX Security Symposium*, volume 3, pages 16–89, 2006.

Julien Freudiger. How talkative is your mobile device?: An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, pages 8:1–8:6, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3623-9. DOI: 10.1145/2766498.2766517. URL http://doi.acm.org/10.1145/2766498.2766517.

Yuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. A pedestrian flow analysis system using wi-fi packet sensors to a real environment. *Proceedings of UbiComp-Adjunct*, pages 721–730, 2014. DOI: 10.1145/2638728.2641312. URL http://ubicomp.org/ubicomp2014/proceedings/ubicomp{_}adjunct/workshops/HASCA/p721-fukuzaki.pdf.

Yuuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. Statistical analysis of actual number of pedestrians for wi-fi packet-based pedestrian flow sensing. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 1519–1526, 2015. ISBN 978-1-4503-3575-1. DOI: 10.1145/2800835.2801623. URL http://doi.acm.org/10.1145/2800835.2801623.

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

Claudio Gariazzo, Armando Pelliccioni, and Maria Paola Bogliolo. Spatiotemporal analysis of urban mobility using aggregate mobile phone derived presence and demographic data: A case study in the city of rome, italy. *Data*, 4(1):8, 2019.

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY, 2003.

Fabien Girardin, Andrea Vaccari, Alexandre Gerber, Assaf Biderman, and Carlo Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

Genevieve Giuliano and Kenneth A. Small. Subcenters in the los angeles region. *Regional Science and Urban Economics*, 21(2):163–182, 1991. ISSN 01660462. DOI: 10.1016/0166-0462(91)90032-I.

Stephen Graham. Cities in the real-time age: The paradigm challenge of telecommunications to the conception and planning of urban space. *Environment and Planning A*, 29(1):105–127, 1997. ISSN 0308518X. DOI: 10.1068/a290105.

Stephen Graham and Patsy Healey. Relational concepts of space and place : Issues for planning theory and practice. *European Planning Studies*, 7(5): 623–646, 1999. ISSN 0965-4313. DOI: 10.1080/09654319908720542. URL http://www.tandfonline.com/doi/abs/10.1080/09654319908720542.

Stephen Graham and Simon Marvin. *Splintering Urbanism*. 2001. ISBN 0203452208. DOI: 10.4324/9780203452202. URL http://books.google.com/books?id=6IdAAY9xqlgC{&}pgis=1.

Steve Graham and Simon Marvin. *Telecommunications and the city: Electronic spaces, urban places*. Routledge, 2002. ISBN 1134813937.

Ben Greenstein, Damon McCoy, Jeffrey Pang, Tadayoshi Kohno, Srinivasan Seshan, and David Wetherall. Improving wireless privacy with an identifier-free link layer protocol. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*, pages 40–53. ACM, 2008. ISBN 1605581399.

Marco Gruteser and Dirk Grunwald. Enhancing location privacy in wireless lan through disposable interface identifiers: a quantitative analysis. *Mobile Networks and Applications*, 10(3):315–325, 2005. ISSN 1383-469X.

Theo Haerder and Andreas Reuter. Principles of transaction-oriented database recovery. *ACM Comput. Surv.*, 15(4):287–317, December 1983. ISSN 0360-0300. DOI: 10.1145/289.291. URL http://doi.acm.org/10.1145/289.291.

Elaine J Hallisey. Cartographic visualization: an assessment and epistemological review*. *The Professional Geographer*, 57(3):350–364, 2005. ISSN 0033-0124.

Mark Harrower. The cognitive limits of animated maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(4): 349–357, 2007. ISSN 0317-7173.

Tian He, Chengdu Huang, Brian M Blum, John A Stankovic, and Tarek Abdelzaher. Range-free localization schemes for large scale sensor networks. In *Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 81–95. ACM, 2003. ISBN 1581137532.

Alison J Heppenstall, Andrew T Crooks, Linda M See, and Michael Batty. *Agent-based models of geographical systems*. Springer Science & Business Media, 2011.

Hande Hong, Girisha Durrel De Silva, and Mun Choon Chan. Crowdprobe: Non-invasive crowd monitoring with Wi-Fi probe. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):115, 2018.

IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534, Dec 2016.

HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

Davy Janssens. *Data Science and Simulation in Transportation Research*. IGI Global, 2013. ISBN 1466649216.

Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, page 2. ACM, 2013. ISBN 1450323316.

Nobuo Kawaguchi. Wifi location information system for both indoors and outdoors. In *International Work-Conference on Artificial Neural Networks*, pages 638–645. Springer, 2009.

Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481, 2014.

Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1_15.

Constantine E Kontokosta and Nicholas Johnson. Urban phenology: Toward a real-time census of the city using Wi-Fi data. *Computers, Environment and Urban Systems*, 64:144–153, 2017.

John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, and Fred Potter. Place lab: Device positioning using radio beacons in the wild. In *International Conference on Pervasive Computing*, pages 116–133. Springer, 2005. ISBN 3540260080.

Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press, 2014. ISBN 1316094456.

Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

Guy Lansley and Paul A Longley. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58:85–96, 2016. ISSN 0198-9715.

Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016.

Alyson Lloyd and James Cheshire. Detecting address uncertainty in loyalty card data. *Applied Spatial Analysis and Policy*, Jan 2018. ISSN 1874-4621. DOI: 10.1007/s12061-018-9250-1. URL https://doi.org/10.1007/s12061-018-9250-1.

Amy Lobben. Classification and application of cartographic animation. *The Professional Geographer*, 55(3):318–328, 2003. ISSN 0033-0124.

Sriganesh Lokanathan and Roshanthi Lucas Gunaratne. Mobile network big data for development: Demystifying the uses and challenges. *Digiworld Economic Journal*, (97):75–94, 2015. ISSN 11578637.

Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018.

Henghui Lu, Sheng Zhang, Xingchuan Liu, and Xiaokang Lin. Vehicle tracking using particle filter in wi-fi network. In *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, pages 1–5. IEEE, 2010. ISBN 1424435730.

Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017.

Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001. ISSN 1523-0406.

Jeremy Martin, Erik Rye, and Robert Beverly. Decomposition of mac address structure for granular device inference. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 78–88. ACM, 2016.

Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C Rye, and Dane Brown. A study of mac address randomization in mobile devices and when it fails. *arXiv preprint arXiv:1703.02874*, 2017.

Bryan H Massam. *Location and space in social administration*. E. Arnold, 1975.

Doreen Massey. Politics and space/time. *New Left Review*, I(196):249–272, 1992. ISSN 0028-6060. DOI: 10.1049/el:19990302. URL https://newleftreview.org/I/196/doreen-massey-politics-and-space-time.

Célestin Matte, Mathieu Cunche, Mathy Vanhoef, Franck Rousseau, and Mathy Vanhoef. Defeating mac address randomization through timing attacks. In *ACM WiSec 2016*, pages 15–20. ACM, 2016. ISBN 1450342701.

Malcolm D McIlroy, Elliot N Pinson, and Berkley A Tague. Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904, 1978.

Dermott John James McMeel. The dark web of urban data: Fitness data ecosystems, urban design and privacy in the modern city. *International Journal of Art, Culture and Design Technologies (IJACDT)*, 7(2):12–25, 2018.

Jeremy Mennis. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55(1):31–42, 2003. ISSN 0033-0124.

Jeremy Mennis and Torrin Hultgren. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, 33(3):179–194, 2006. ISSN 1523-0406.

Harvey J. Miller. The data avalanche is here. shouldn't we be digging? *Journal of Regional Science*, 50(1):181–201, 2010. ISSN 00224146. DOI: 10.1111/j.1467-9787.2009.00641.x.

David Moore, John Leonard, Daniela Rus, and Seth Teller. Robust distributed network localization with noisy range measurements. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 50–61. ACM, 2004. ISBN 1581138792.

Julie Bauer Morrison, Barbara Tversky, and Mireille Betrancourt. Animation: Does it facilitate learning. In *AAAI spring symposium on smart graphics*, pages 53–59, 2000.

Yaser Mowafi, Ahmad Zmily, Dhiah el Diehn Abou-Tair, and Dirar Abu-Saymeh. Tracking human mobility at mass gathering events using wisp. *Future Generation Communication Technology (FGCT), 2013 Second International Conference on*, pages 157–162, 2013. DOI: 10.1109/FGCT.2013.6767212. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6767212.

A. B. M. Musa and Jakob Eriksson. Tracking unmodified smartphones using wi-fi monitors. *Proceedings of the 10th ACM Conference on Embedded Networked Sensor Systems (SenSys '12)*, pages 281–294, 2012. DOI: 10.1145/2426656.2426685. URL http://dl.acm.org/citation.cfm?doid=2426656.2426685.

ABM Musa and Jakob Eriksson. Wiflow: real time travel time estimation using Wi-Fi monitors. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 429–430. ACM, 2011.

Taewoo Nam and Theresa A Pardo. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*, pages 282–291. ACM, 2011.

S. Ordonez and A. Erath. Estimating dynamic workplace capacities using public transport smart card data and a household travel survey. *Proceedings*

of the 17th International Conference of Hong Kong Society for Transportation Studies, HKSTS 2012: Transportation and Logistics Management, pages 505–512, 2012. ISSN 0361-1981. DOI: 10.3141/2344-03. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84896879480{&}partnerID=tZOtx3y1.

Jeffrey Pang, Ben Greenstein, Srinivasan Seshan, and David Wetherall. Tryst: The case for confidential service discovery. In *HotNets*, volume 2, page 1, 2007.

Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *International Workshop on Human Behavior Understanding*, pages 14–25. Springer, 2010.

Juval Portugali, Han Meyer, Egbert Stolk, and Ekim Tan. *Complexity theories of cities have come of age: an overview with implications to urban planning and design*. Springer Science & Business Media, 2012. ISBN 3642245447.

R Pulselli, P Ramono, Carlo Ratti, and E Tiezzi. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *Int. J. of Design and Nature and Ecodynamics*, 3(2):121–134, 2008.

Weijun Qin, Jiadi Zhang, Bo Li, and Limin Sun. Discovering human presence activities with smartphones using nonintrusive wi-fi sniffer sensors: The big data prospective. *International Journal of Distributed Sensor Networks*, 2013, 2013. ISSN 15501477. DOI: 10.1155/2013/927940.

John N K Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015. ISBN 1118735722.

Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006. ISSN 02658135. DOI: 10.1068/b32047.

Michael Reibel and Michael E Bufalino. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37(1):127–139, 2005. ISSN 0308-518X.

Jun Rekimoto, Takashi Miyaki, and Takaaki Ishizawa. Lifetag: Wifi-based continuous location logging for life pattern analysis. In *LoCA*, volume 2007, pages 35–49, 2007.

Ian Rose and Matt Welsh. Mapping the urban wireless landscape with argos. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 323–336. ACM, 2010. ISBN 1450303447.

Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PloS one*, 10(7):e0130824, 2015. ISSN 1932-6203. DOI: 10.1371/journal.pone.0130824. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130824.

T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007.

Saskia Sassen. *The global city: New York, London, Tokyo*, volume 15. 2001. ISBN 9780691070636. DOI: 10.2307/2152688. URL http://books.google.com/books?id=PTAiHWK2BYIC.

Lorenz Schauer, Martin Werner, and Philipp Marcus. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 171–177, 2014. DOI: 10.4108/icst.mobiquitous.2014.257870. URL http://eudl.eu/doi/10.4108/icst.mobiquitous.2014.257870.

Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2):83–85, 1956. ISSN 0002-8991.

Kyosuke Shibata and Hiroshi Yamamoto. People crowd density estimation system using deep learning for radio wave sensing of cellular communication. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 143–148. IEEE, 2019.

Edward Soja. Postmodern geographies, 1989. URL http://books.google.com/books?id=sNcRAQAAIAAJ.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.1177170.

Balamurugan Soundararaj. Clicker - an Android app for counting pedestrian footfalls with accuracy. https://play.google.com/store/apps/details?id=com.bala.manualcount, 2018.

John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, and Henk Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223–243, 2013. ISSN 0343-2521.

John Steenbruggen, Emmanouil Tranos, and Peter Nijkamp. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3-4):335–346, 2015. ISSN 03085961. DOI: 10.1016/j.telpol.2014.04.001.

Shan Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):70–73, 2014.

Paul Sutton. Modeling population density with night-time satellite imagery and gis. *Computers, Environment and Urban Systems*, 21(3):227–244, 1997. ISSN 0198-9715.

Karen P Tang, Pedram Keyani, James Fogarty, and Jason I Hong. Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 93–102. ACM, 2006.

Ole Tange. *GNU Parallel 2018*. Ole Tange, March 2018. ISBN 9781387509881. DOI: 10.5281/zenodo.1146014. URL https://doi.org/10.5281/zenodo.1146014.

Deborah S K Thomas. Data, data everywhere, but can we really use them. *American hazardscapes: The regionalization of hazards and disasters*, pages 61–76, 2001.

James J Thomas. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005. ISBN 0769523234.

Emmanouil Tranos. *The geography of the internet: Cities, regions and internet infrastructure in Europe*. Edward Elgar Publishing, 2013.

Emmanouil Tranos and Peter Nijkamp. Urban and regional analysis and the digital revolution: challenges and opportunities. *Hub cities in the knowledge economy: Ashgate, Forthcoming*, 2012.

Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: 10.1145/2897845.2897883.

Leah K VanWey, Ronald R Rindfuss, Myron P Gutmann, Barbara Entwisle, and Deborah L Balk. Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences*, 102(43):15337–15342, 2005. ISSN 0027-8424.

Tien Dang Vo-Huu, Triet Dang Vo-Huu, and Guevara Noubir. Fingerprinting Wi-Fi devices using software defined radios. In *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 3–14. ACM, 2016.

Stephen R. Walli. The posix family of standards. *StandardView*, 3(1):11–17, March 1995. ISSN 1067-9936. DOI: 10.1145/210308.210315. URL http://doi.acm.org/10.1145/210308.210315.

W. Wang, A.X. Liu, and M. Shahzad. Gait recognition using wifi signals. In *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016. ISBN 9781450344616. DOI: 10.1145/2971648.2971670.

Yan Wang, Jie Yang, Hongbo Liu, and Yingying Chen. Measuring human queues using wifi signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 235–237, 2013. ISBN 9781450319997. DOI: 10.1145/2500423.2504584. URL http://dl.acm.org/citation.cfm?doid=2500423.2504584{%}5Cnhttp://dl.acm.org/citation.cfm?id=2504584.

Wei Xi, Yuan He, Yunhao Liu, Jizhong Zhao, Lufeng Mo, Zheng Yang, Jiliang Wang, and Xiangyang Li. Locating sensors in the wild: pursuit of ranging quality. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 295–308. ACM, 2010. ISBN 1450303447.

Zhuliang Xu, Kumbesan Sandrasegaran, Xiaoying Kong, Xinning Zhu, Jingbin Zhao, Bin Hu, and Cheng-Chung Lin. Pedestrain monitoring system using Wi-Fi technology and rssi based localization. *International Journal of Wireless & Mobile Networks*, 5(4):17, 2013.

Yew Yuan, Richard M Smith, and W Fredrick Limp. Remodeling census population with spatial information from landsat tm imagery. *Computers, Environment and Urban Systems*, 21(3):245–258, 1997. ISSN 0198-9715.

Yihong Yuan and Martin Raubal. Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8):1594–1621, 2016. ISSN 1365-8816. DOI: 10.1080/13658816.2016.1143555. URL http://www.tandfonline.com/doi/full/10.1080/13658816.2016.1143555.

Robbert Zandvliet and Martin Dijst. Short-term dynamics in the use of places: A space-time typology of visitor populations in the netherlands. *Urban Studies*, 43(7):1159–1176, 2006.

Vasileios Zarimpas, Bahram Honary, and Mike Darnell. Indoor 802.1 x based location determination and real-time tracking. In *Wireless, Mobile and Multimedia Networks, 2006 IET International Conference on*, pages 1–4. IET, 2006. ISBN 0863416446.