Estimating Activity in Urban Areas Using Passively Collected Data - Case Study Using Smart Street Sensor Project.

Balamurugan Soundararaj
Department of Geography
University College London

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

Doctor of Philosophy

2019

Acknowledgements

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

Abstract

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

Impact Statement

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

List of Outputs

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

Contents

List of Figures

List of Tables

Chapter 1

Introduction

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

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

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

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

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

Chapter 2

Literature Search

2.1 Themes of Research

2.2 Technologies Used

2.3 Challenges Faced

2.4 Research Gaps

Chapter 3

Collecting the 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 smart-phones 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

Table 3.1: 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 $x10^6$ (per min)	Footfall No. (per min)
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)

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. 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.

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.

3.6 Discussion

Chapter 4

Data Processing

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 number of pitfalls as well. With developments in information technology, mobile communications and internet of things large assemblages of data are readily available leading to immense possibilities in terms of research but when we take up analysing these data at such scale, we encounter a equal amount of added complexity and cost as well. This makes it important that we are careful in choosing the methods and tools in dealing with such data. Moreover in several disciplines such as statistics and geography etc. the existing methods and tools are already developed for large scale data. These methods along with improvements in hardware recently has made the processing of big data possible to an extent in these disciplines without a major changes. At these scales, we need to devise the right methods and tools for the right problems. In this environment of constant change and growth, while trying to create a future proof big data approach, we cannot afford to lose the opportunity in extracting information from these large datasets . We need to have a pragmatic approach we look in to other disciplines and adopt best practices and solutions to develop a consistent philosophy rather than reinventing the wheel.

In the previous chapters we looked at the various methods devised to collect and process data from WiFi probe requests emitted by phones. Though we discussed the methods in detail, we left out the rationale behind the toolkit chosen to implement the methods. In this section we start by discussing the nature of 'Big Data', then discuss the previous research in the 'Big Data' space to understand the definition nature and challenges of big data. Then we look at our data-sets collected through the pilot studies and the Smart Street Sensor project closely and evaluate their nature in terms of the dimensions of the big data. We also discuss in detail the challenges faced in dealing with such dataset and the requirements expected from a toolkit to do so. Finally we look across disciplines, discuss and evaluate approaches and tools available in dealing with such datasets and their fitness of purpose to do so.

Big Data Discussion

Big data has become the new hot thing in research. Data has become the next oil and we are in a data deluge moving from a data sparse to data rich environment. As we saw in the previous chapters there are easy and quick ways to collect large amount of data. The technological advancements has enabled us to be able to think about utilising such large assemblages of data. Big data gives us numerous benefits, it can provide us insights which were not possible before. I can give comprehensive coverage. It can change the approach and methods in whole disciplines such as computer science where the shift towards machine learning and AI which is primarily fuelled by the explosion of data available. It also poses a lot of challenges because of its nature, in analysing and managing it. In addition to these, trying to solving 'normal' problem using big data tools and approaches can introduce enormous overhead and introduce additional complexity without providing much benefits. It can actually counter productive to use these methods when they are not necessary. It is absolutely necessary for any research to understand what is big data and how the subject at hand relates to big data.

The first and foremost challenge in big data is its definition. It is not clearly defined and the definition can vary widely based on the discipline and perspective. What is big data in one discipline may not be in another. Data also has various dimensions in which they can exhibit big data properties in limited number of dimensions. In general it is defined in contrast to traditional data. It is generally defined as the data which cannot be dealt with traditional desktop or server client computing methods and hardware and requires substantial change in the approach. Laney 2001 defined big data in three dimensions volume, velocity and variety - the three Vs of big data. These have been extended to include veracity - the reliability or truthfulness of the data, visualisation - the complexity in visual interpretation and presentation, and other additional dimensions such as visibility validity, variability, volatility and value. There have been other alternative dimensions proposed as well such as Cardinality, continuity and complexity - three Cs. However for our purposes we consider 5Vs of Big data, volume, velocity, variety, veracity and visualisation. It has to be noted that Not all data is 'Big' in all these dimensions and each dimension can lead to their own challenge in the analysis and management of the data. It is important to understand the dataset in terms of its Big data properties so that the corresponding challenges can be tackled with confidence.

The second set of challenges arise from the above mentioned nature of big data while processing it. The 'processing' of data can be categorised broadly as the following steps - data acquisition and recording, extraction cleaning and annotation, data integration and aggregation, modeling and analysis, and finally visualisation and interpretation. Each step in this processing poses its own challenges. The volume, velocity and variety poses the problem in data acquisition by introducing the need for distributed, crowdsourced collection of data, need for heavily parallelised computing and functional programming concepts while we clean and aggregate the data into valuable information.

This is where the approaches in Programming languages such as Haskell, distributed storage and computing such as MapReduce, Hadoop and parallel processing approaches such as spark are introduced. There is a lot of bias arising from the unstructured way of collecting data which requires consideration on calibrating and weighting the data collection process to weed out any potential uncertainties and biases. The veracity of the big data introduces significant challenges in modelling and analysis of the data. The need for new ML algorithms which are data heavy and Deep learning methods such as CNN which focuses on processing heavy become prominent. The high performance computing methods with ultra parallelised techniques such as GPU processing are indispensable to be able to model and predict in such data sets. These properties also introduces significant challenges in presenting the knowledge mined with such data like visualisation. Visualisation is not only a presentation task supporting critical activity such as policy decision making support but also an indispensable tool in exploratory analysis. The volume and velocity of the data makes is hard to be able to visually digest the data, the variety especially in the time dimension can make the data too complex to interpret even when presented in small parts. GIS systems do a good job in visualising complex geographic datasets but struggle to maintain legibility and understandability when dealing with the temporal dimension. There is a need for interactivity and connected components to be able to visually keep track of information presented. There needs to be a balance between functionality, aesthetics and performance. Finally the variety of the data set poses the challenge of interoperability between various approaches and tools. This creates need for consistent standards in terms of dealing with data.

Apart from these processing challenges, we also have management challenges to consider such as privacy and security, Data governance and ownership, data and information sharing, and cost. Privacy and security is one of the basic consideration in any project dealing with big data. The approach, methods, tools should strictly have the capability comply with the legislation as well as the research ethics set up by all the stakeholders. At these scales there is no security by obscurity. The data collected itself may not be private but in conjunction with other datasets, it can be disclosive. There is also the ambiguity of the ownership of the data and risk associated with handling the data, it needs to comply with GDPR regulations. There also needs to be clear way to share the data securely with other projects along with publications of results so that the value of the all these processing is realised. The project management and issue tracking tools need to be capable of handling these data ownership and sharing concerns. This needs to be done in a timely, accessible manner as well. That leads us to the final challenge of cost of big data. Though most of the big data tools are developed openly and free there can be lot of hidden costs associated with collecting, processing and managing big data. There are the operational costs collecting data, network costs moving the data, server costs storing and processing the data, cost of specialised tools (or at least support for the tools) to use and the human resource cost in terms of training, time and money needed managing the data.

Study	Maximum	Minimum	Average
Pilot Study Main Study Smart Street Sensor			

Table 4.1: Volume of the data

This can make even a small project balloon in terms of cost when implemented in a scalable 'big data' way. This brings us to the need of us looking at the data at our hands closely so that we can make an informed decisions on how 'big' is our data and pick and choose the methods which are the most efficient in dealing with such dataset.

Is WiFi probe requests collected Big data?

In this section we take a detailed look at the WiFi datasets collected in the research using the 5Vs big data framework. Out aim is to understand the nature of the data thus the challenges we will face processing and managing it. We want to answer the following questions, Can the WiFi data be defined as Big data? What are the aspects where Wi-Fi data shows 'Big data' properties? This gives us good information when evaluating a viable and efficient toolkit to process and manage the data. First we look at the WiFi data in each dimension and then arrive at a general categorisation of the data.

Volume

We saw earlier that every location is bombarded with probe requests from phone in the area. Quantifying how much data is generated / arriving at each location is the first step of looking at the volume. We look at the absolute scale of the data rather than the rate at which it is being generated. For the sake of comparison we have used the measure of size on disk for a year's worth of data encoded in text format. We have to note that this can vary quite a lot between fields and tools. This can act as a quantification of volume of the data.

We can see that on average of data is generates at a place and project of the size of smart street sensor is estimated to generate around of data. This is not a trivial volume of data.. Desktop normal analysis of such volume is next to impossible. At the same time this is not truly what is meant by Big data as well. True big data sources in terms of volume can be of sizes. They cannot be even stored at a central single location. Within this interdisciplinary context, we can say that a national level sampled collection of Wi-Fi data is medium at best. A comparison of data sizes in various fields after being standardised is shown in the figure. We can see that data sources such as LIDAR, social media, internet sources are several orders of magnitude larger. Even if we assume the largest possible data possible - One sensor in every retail location in UK, we can assume something along the lines of This assuming data collected as we did in the main study and xxxx retail locations

Study	Maximum	Minimum	Average
Pilot Study Main Study Smart Street Sensor			

Table 4.2: velocity of the data

across UK in average. To summarise, in the volume dimension WiFi data is 'medium' in size and is expected to remain so in the future.

Figure comparing the typical size of the WiFi dataset to a small dataset and truly big data.

Velocity

Unlike volume, velocity is defined as the rate of collection data at short term. It is quantified similar to volume with size on disk per unit time but the time period is smaller. The Wi-fi probe requests are generated almost continuously at each location across various channels. The temporal precision is around micro seconds, but for convenience we have collected and aggregated data per minute for the pilot studies and 5 minutes for the Smart street sensor project. Every location generates around xx GB per minute and in total the project generates xx GB per minute. This again varies from xx to xx and on average we can say a moderately busy location is expected to generate xx in a minute. When we compare this to long term data sources such as census, slow data such as sample surveys it looks impressively fast and almost real time, but when we compare to actual real time data, Internet ad click through, Large Hadron Collider or Aviation, it is not as fast. This comparison is shown in figure. As we saw with volume, even in velocity the Wi-fi Data can be described as 'medium' data which is aggregated every 5 mins and mostly processed daily batches of xx GB. We can safely say that there are no real need for low latency real time solution while dealing with this data while at the same time we need to recognise that unlike traditional datasets, we have a steady stream of data which needs to be processed regularly and sustainably.

Figure comparing the typical size of the WiFi dataset to a small dataset and truly big data.

Variety

This aspect of the data cannot be quantified in absolute time but has to be subjectively discussed. Since the origin of the data is in the Wi-Fi standard, the core of the data is very very structured. Every probe request almost always have a core set of data which is highly structured and remains the same all over the world. This further propagates as well process the data further. There are two main sources of variability identified withing the WiFi probe requests

- Information elements : The standard specifies a good amount of flexible data which can be encoded in the request detailing the capabilities of the mobile device. These information element are implemented as per the manufacturer discretion.

- There is a lot of variability in the rate at which the probe requests are generated.

The former is slowly becoming an unimportant one since most of the manufacturers following apple have started to include almost no information elements in the probe request packet to protect the privacy of users by eliminating the possibility of detailed fingerprinting of devices. The rate of probe request generation still varies widely for different manufacturers but overall cannot constitute much of the variability. In terms of standardised footfall counts there is only one Ordinal data point along time intervals with a geographic unit and time. From the above, we can safely assume that the Wi-Fi data shows no Big-data characteristics in the variety dimension.

Veracity

This is the dimension in which the Wi-Fi data exhibits big data characteristics. This comes from the fact the data is collected in a unstructured way and passively. The first source of veracity originates from the unreliability of the data collection process. The data is collected through a network of 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 at best and at any given period of time about 10% of sensors fail to send back data regularly. More over the sensors are installed and uninstalled regularly as and when the data partners join 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. This poses immense challenge when we attempt to aggregate the data. There is a requirement for cleaning and filling in the gaps of the data.

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 the way it is installed and due to the context of the location as showed in the data cleaning procedures. Cite Karlo's work. This leads to a situation where the accuracy of the data collection varying quite widely across location and times. There is a question of weather 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 can increase the estimated footfall without any change in footfall at the location.

The final veracity issue is the changing mobile landscape. Though the wifi 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 the bias 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. From the above we can conclude that Wi-Fi data shows Big data characteristics in terms of its veracity and requires tools and methods when aggregating, analysing and modelling it.

Visualisation and reporting

Visualisation is closely related to volume, velocity and variety of the data. The Wi-Fi data due to its non-trivial size and velocity, exhibits similar characteristics and challenges in visualisation. Since there not much variety in the data-set when we process it into footfall all we are left with is time, location and footfall. Out of which location and footfall is easy to visualise but the time is the complicated one because of its volume - 2 to 3 years worth of data and granularity - 1/5 minute intervals. This is really hard to simplify and visualise. The key here is using approaches that show change efficiently and legibly. This shows the need for a dynamic, interactive visualisation tools which can deal with continuous change over long periods of time. There is also need for multiple linked dynamic visualisation platform for separating the scope of the visualisation into manageable units. The second challenge is the communicating the veracity of the data without distracting from the message. Finally the 'near real time' aspect of the data needs to taken into consideration while visualising it hence the need for always on, interactive, real time dashboards with geographic capabilities. Considering the above we can say that in terms of visualisation, Wi-Fi data partially shows big data characteristics.

Figure: Example of too much data in time dimension

Summarising the above discussion, we can say that the Wi-Fi data collected from probe requests is at best a 'medium' size data which shows big data characteristics in terms of its veracity. Any toolkit devised to be used with it need to be able to deal with its medium volume, velocity and visualisation needs and at the same time need to able to deal with the huge veracity of it. This leads us to devising a 'medium data toolkit' which can be used in such cases so that not to introduce the cost and complexity introduced by broader big data tools.

Figure spider graph showing the profile of Wi-Fi data.

A Survey of Methods and Tools

In this section, we survey the tools and methods available at various stages of the data processing and management process we discusst he tools with respect to the performance (throughput), flexibility, complexity and cost. We finally try to devise a toolkit which best suits our data needs.

Collection

There are numerous tools available for data collection with network of sensors under the umbrella of internet of things. The primary consideration in the data collection is the scale of the infrastructure and the cost associated with it. The smartstreetsensor project uses its own proprietary sensor system which collects data at the location. The tooling decisions were made with the commercial application in mind and is not entirely relevant to our discussion, but for the research conducted with the data, it is necessary to understand the data collection process and how the toolkit integrates with the rest of the setup. We start by looking at different approaches in the data collection tools and try to reason the most appropriate solution for the WiFi data. At the hardware level, the lowest level of tool can be a micro-controller such as audrino with a dedicated hardware module with custom software to collect the exact data needed. This is time consuming, cumbersome takes a lot of cost to develop, but is very 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 set is collected through multiple sources and syndicated into a cleaned source by a thirdparty. This can be costly and inflexible but quick and easy. The middle ground on this to deal with a complexity as much as the WiFi data, is to use a single board computer with external modules and use general purpose, tools to build a data collection device.

The toolkit we have adopted consists of RaspberryPi, Linux, tcpdump/tshark and nodejs. The RaspberryPi and the linux platform it provides is one of the most diverse and general purpose systems available. On top of which we can build our data collection system with specialised, opensource and free Wi-Fi sniffing tool such as tcpdump, tshark along with a general purpose runtime such as nodejs which provides other functions such as scheduling, personal data obfuscation and data transmission. This system has a capacity to sniff and transfer large amounts of data and with a 3G module is very versatile in terms of location.

Storage

This is one of the most diverse set of area in terms of both methods and tools available. It has been constantly in development since the beginning of computer systems and is one of the fastest changing landscapes. The aspects to be considered while choosing the data storage solutions are,

1. Speed
2. Redundancy
3. Reliability
4. Cost & complexity

One of the spectrum is just using file systems for storage. Though it seems to be primitive, this has a lot of advantages. Operating systems usually are really good at managing - reading, writing and searching filesystems, They usually have no overhead involved and are efficient. Hierarchical organisations can be pre-indexed for hierarchical data and finally is very reliable. But the primary disadvantage is the inability to handle complexity or variety in data.

On the other end of the spectrum is the highly distributed big data systems such as Hadoop HDFS which are built for ¿ petabyte datasets and query them without loss of speed. There are

Type	Comment
Filesystem	for hierarchical data around 10TB range
Cloud Storage	$<$ 10TB, can add hdfs stuff, more reliability
Relational DB	1-5TB, Good for relational Data, Row-wise, Partitioning
Document DB	-10TB, Good for unstructured data, column wise,Clustering
HDFS	$>$ 10 TB, Good for scale and structure
Hive, Hbase	on top of HDFS, bring DB to HDFS

Table 4.3: Types and Use of Various Storage Solutions

hybrid file systems which are hadoop compatible as well, Azure blob storage, Amazon S3 cloud storage which can be used a storage/ dump for a large amount of data.

In the middle there are databases, which are built prioritise and balance the database needs. The two major approaches are the relational databases which are optimised for structured data which are related to each other in tabular format. They are row heavy databases and are good for high volume, low veracity data which has need for consistency. SQL databases PostgreSQL, Mysql, SQLserver are examples. The other approach is the document store databases which are column heavy databases which are optimised for high variety data which doesnot need immediate consistency. These can be as simple as key-value based databases and as complicated as graph databases. Mongodb, couchdb, cassandra as examples. Both these approaches can be scaled/distributed for less redundancy and increased throughput. The former tend to scale veritcally and the latter scale horizontally. Some like cassandra are built to be highly distributable.

Finally there are solutions such as Hive and hbase which are database like functionality built on top of distributed file systems combining power of both concepts. This behaves like a hyper large scale database system and works in conjunction with other big data tools

Raw wifi data has temporal hierarchy and is of medium size hence a normal filesystem is sufficiently suitable for its needs. When the same data is aggregated it loses its scale and is highly strucutred so a relation DBMS is sufficient for it. In case the project runs longer and more longitudinal analysis had to be done on raw data HDFS needs to be used and if the aggregated data scales to ¿10TB we can handle it with a timescale db should be suitable. PostgreSQL is more suitable than other databases because of its better support to geographic data.

Processing

The primary considerations while surveying are the volume, velocity and veracity of the data. We should be careful to choose the tools which are right for the size. The perfect tools for a medium size data can be as much as 230x faster than big data tools (ref). At one end there are Big Data analysis tools such as Hadoop based impementations such as Mapreduce and Spark, Business toosl such as skytree, realtime tools such as storm and samoa, cleaning tools such as Openrefine. All these tools are optimised for the cluster/grid computing and the processing is heavily parallelised across the clusters. There is also a lot of overhead asssociated with moving data across clusters and we won't be making up for these overheads until we hit certain size of the data. As we know Wi-Fi data is not at the scale these tools operate, we can look into how large streams of data are handles in computer science/ systems engineering. Data processing is done in two stages, the first one is the filtering, cleaning and aggregation of the raw Wi-Fi data and the second step is the analysis and modelling of the aggregated data. As we saw in (ref) the system tools in combination with parallel processing across CPU cores, can be used and can be actually faster for medium sized data. The data transfer format is text since it is standardised with utf8 and is easily understood and shared between UNIX tools. This also helps us in the data sharing and management which is discussed later. For the first part of the processing - filtering & cleaning we use the following tools,

1. **sed** - streaming text editor. A fully featured text editor which works on stream of text. The stream is processed usually by each line and is the most commonly used to search and replace (translating) text streams using regular expressions.

2. **Grep** - grep (global regex print) is a special case of sed where we search the stream for regular expression and print the result. This is usually used for searching and filtering text streams.

3. **awk** - This is a turing complete special purpose higher level programming language which is optimised for sorting, validating and transforming text streams. It is full featured enough to be able to manage a small text based database by itself. This is usually used to transform tabular delimited data.

4. **jq** - This is similar to awk, has a emcascript based scripting language for transforming text data which is in the JavaScript Object Notation format. These four tools form a core toolkit for tranforming, translating and filtering data. All these tools are single threaded and need an external tool to parallelise the processes. For this we can use gnu-parallel.

5. **parallel** - This is a tool built with perl (citation) which parallelises the any operation across CPU cores and even across multiple nodes through secure shell (ssh). This gives us a medium sized cluster which is well suited dealing with text data stored in a file system.

Bash completes the toolkit to provide a overarching high-level scripting interface to combine all the smaller tools and managing data transfer between them as text streams using the 'pipe' operator. This along with core bash tools such as sort, uniq can give us a basic data filtering, transformation and aggregation toolkit with a reasonable throughput. Example, For a normal word count problem, this toolkit can give us a through put of 540MB per minute without parallelisation and with parallelisation this can be improved to 2.5GB per minute.

For complex data cleaning techniques such as filling in the gaps, we can use higher level languages such as R or Python through their scripting environments and linking them to our pipelines using bash. Security in terms of obfuscation can be done through hashing algorithms implemented by openssl, nodejs and R and for encryption, we can use the gnupg. The toolkit being open source free software has the added advantage of being secure as well.

Visualisation

Tableu, Omniscope.

Conclusions

To summarise we have done a survey of tools and arrived at the following toolkit

Figure of the data toolkit.

4.2 Signal Strength

4.3 Sequence Numbers

4.4 Phone to Probes Ratio

4.5 Data Completeness

4.6 Discussion

Chapter 5

Applications

5.1 Footfall Indices

5.2 Events Detection

5.3 Pedestrian Flows

5.4 Discussion

Chapter 6

Conclusions

Appendix A

Source Code for Tools

A.1 Wi-Fi Sensor Stack

A.2 Footfall Data Toolkit

A.3 Footfall Indexer

A.4 Footfall Dashboard

Appendix B

Open Source Projects Used

B.1 Programming Languages

B.2 Libraries

B.3 Tools

B.4 Datasets

Bibliography

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