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 ACTIVITY FROM PASSIVELY COLLECTED DATA

A CASE STUDY OF SMART STREET SENSOR PROJECT

UNIVERSITY COLLEGE LONDON

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, April 2019

Contents

List of Figures

List of Tables

Outputs From The Research

Book Chapter

1 *Consumer Data Research*

Journal Articles

1 *I.J.G.I.S paper*

2 *Medium data paper*

3 *Transfer Entropy Paper*

Conference Talks

1 *GISRUK 2017*

2 *GISRUK 2019*

3 *Retail futures Oxford*

4 *CCS 2016*

5 *CCS 2018*

6 *Italy 2018*

Other Talks

1 *Data natives 2018*

2 *Geo-data 2018*

3 *SUPF Workshop 2018*

CDRC Products

1 *C.D.R.C footfall dashboard*

2 *C.D.R.C footfall indicator*

3 *C.D.R.C footfall atlas*

4 *Clicker Android Application*

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)

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.

A Review of Literature

Themes of Research

Technologies Used

Challenges Faced

Research Gaps

Passively 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.

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.

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.

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 .

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.

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.

Processing the Data into Footfall

Medium 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.

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 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.

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

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 methods, 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.

[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

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.

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.

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

Table 1: Comparison of volume or size of the datasets of Wi-Fi probe requests.

** 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 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.

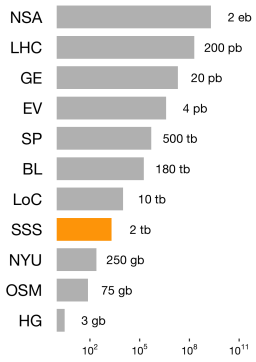

Figure 1: Comparison of volumes of datasets across various disciplines.

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 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 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 recieve upto 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 generates around a million requests every 15 mins. Compared to the LHC's 180 billon 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 which is updated every 10 years or survey based geographic data.

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.

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 Wi-Fi 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

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 Wi-Fi 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 Wi-Fi 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 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

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 3: Types and Use of Various Storage Solutions

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 be-

tween 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.

Visualisations and Applications

Footfall Indices

Events Detection

Pedestrian Flows

Discussion and Conclusions

Appendix

Source Code for Tools

Wi-Fi Sensor Stack

Footfall Data Toolkit

Footfall Indexer

Footfall Dashboard

Open Source Projects Used

Programming Languages

Libraries

Tools

Datasets

Bibliography

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

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.

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

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

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.

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.

Index