3
Collecting Wi-Fi Data

From the literature review in Chapter 2, we observed that out of all the technologies discussed, Wi-Fi seems to be the most promising one. We observed the advantages of Wi-Fi based data collection as,

- Universality as a standard technology globally,
- Independence from other types of data sources or infrastructure,
- High level of granularity both spatially and temporally,
- Possibility of passive data collection,
- Extreme ease of collection in terms of cost and effort and
- Scalability to cover study large areas.

Though it has its pitfalls in terms of intrusiveness resulting in risk to the privacy of the users, bias and uncertainty, Wi-Fi provides us with a strong base framework for fulfilling the opportunity to design and collect a large, long term and granular dataset which can be used for studying human activity.

In this chapter, we continue our research by looking at Wi-Fi technology closely to understand how it can be used to achieve the aforementioned goal. We start by looking at the Wi-Fi specification [1] and focus on the information available within the Wi-Fi probe requests. We then design and implement a series of data collection exercises which collect probe requests in various location with increasing level of complexity for analysis. We explore these datasets briefly to understand the usefulness of each set of information present in the probe requests along with the uncertainties in them. We also introduce 'Smart Street Sensor' project - a national scale effort for collecting Wi-Fi data at high streets across United Kingdom. Finally we summarise the data collection procedure with a detailed look at the each set of uncertainty in these datasets and draw conclusions on further lines of research into alleviating the uncertainty and noise so that they can be used to estimate human activity with confidence.

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

3.1 Wi-Fi as a Source of Data

Since the formation of 'Wi-Fi alliance' in 1999 to hold the trademark, Wi-Fi (Wireless Fidelity) has become synonymous to the IEEE 802.11 standard based internet connectivity. Today almost almost all devices use

this standard to create and connect to local area networks wirelessly. Due to its high fidelity and immense throughput up to 1 Gigabits per second, Wi-Fi has become the choice of technology for wirelessly transferring large amount data through networks. The adoption of 'smart' mobile devices - Smartphones, across the world has further cemented its position as on of the most ubiquitous technology which every one uses every day. In developed economies such as UK this has never been more true and having an infrastructure to serve and receive Wi-Fi signal greatly affects the ability to connect to internet in many areas. With close to 87%[2] of the population carrying one or more of these smart devices with Wi-Fi capability, provision of Wi-Fi as a service has become essential for any place making Wi-Fi one of the most used technology to access internet along with mobile networks.

[2] Deloitte. Mobile consumer survey - united kingdom, 2018. URL https://www.deloitte.co.uk/mobileuk/

Though the end goal of internet connectivity is the same, Wi-Fi greatly differs from internet connectivity through mobile networks such 3G/4G. The first difference is the range of the network, unlike mobile infrastructure where single tower can serve mobile phones for miles, Wi-Fi is designed to be and extension of the wired networking creating short range network with a range up to 20 meters. Due to this low-range and high throughput property, Wi-Fi is used primarily as a distributed infrastructure operated by owners of premises as a means to provide high speed connectivity to the users of these buildings unlike the large, national level, monolithic infrastructure that runs the mobile network. This creates a situation where urban areas are population by hundreds and thousands of these small area networks to which any mobile device can connect to. Unlike the mobile service providers and its customers, these Wi-Fi networks and mobile devices don't trust each other with specialised hardware. This creates a need for an introduction procedure - a sort of handshake between them where they exchange information about themselves. Moreover since these mobile devices constantly move across these Wi-Fi networks, it became necessary for them to carry out these 'handshake' process regularly and frequently so that they can traverse between the networks without loss of connectivity. This need for constant look out for new networks is solved by the 'Probe requests'.

3.1.1 Probe requests

There are numerous iterations and versions of the IEEE 802.11 standards but essentially all of them operate by exchanging packets of information called 'datagrams' or 'frames'. These frames have the information that is being exchanged along with the meta data about the information and information on the device that is sending them. Some of these frames have special purposes and one such purpose is the 'network discovery'. The frames used for this purpose by the mobile device and the Access Point (AP) are called the 'probe request' and 'probe response' respectively. Though the actual information exchanged between these devices are usually encrypted, these probe requests are unencrypted and are accessible to any device which is listening. The structure of a probe request is shown in Figure 3.1.

We can observe that the fixed part of the frames which has to be present in all of them are in the front showing the identity of the mobile device generating the frame along with one which is receiving the frame. There are additional meta data such as sequence number of the frame, controls denoting where the frame starts and ends. There is also a dearth of variable information which can be used to transfer data. For probe requests the destination device is set as 'broadcast' and the variable part usually contains the payload but for probe request frames, this contains 'information elements' which has data regarding the capabilities of the device organised in units known as 'tags' or 'parameter sets'. The significant information present in a probe request is detailed in Table 3.1 and full list of information available from a probe request is shown in the form of a sample probe request in appendix 7.4

Field	Notes
Source Address	Media Access Control (MAC) address
Time stamp	Precise time at which the frame is received
Signal Strength (RSSI)	The strength of the received signal
Length of the frame	Total length of the frame in bytes
Duration of transmission	Time it took to transmit the frame in milliseconds
Information Elements	List of various information about the device
Known SSIDs	Networks that are already known to the device

Table 3.1: Significant information included in a probe request

Essentially the above information is sent over and over by the mobile device expecting a reply from the APs around it so that it can keep a list of networks it can connect to. This process is usually carried out even when the Wi-Fi is switched off in the operating system so that the connection times are faster once it is switched on. Moreover operating systems use the replies they get for these probe requests and triangulate the device location with respect to the APs with known location information thus acting as a quick and easy localisation solution which along with the above makes this probing process almost non-stop.

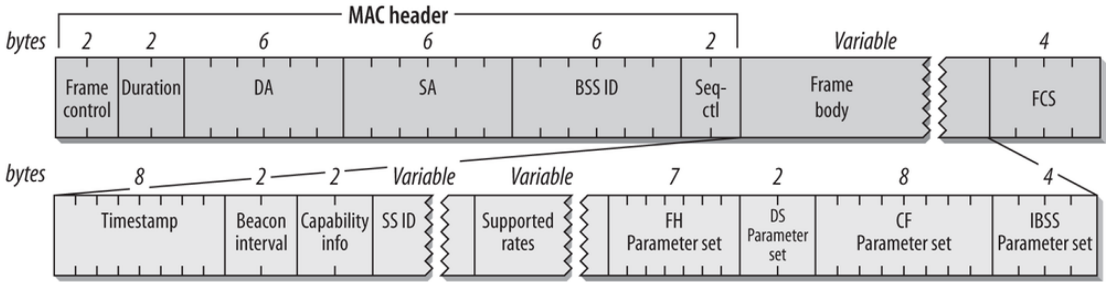

Figure 3.1: Structure of a probe request frame. *Source: IEEE 802.11 specification*

3.1.2 *MAC address*

Media Access Control (MAC) address is a 12bit unique identifier assigned to a device in a network, it is similar to the Internet Protocol (IP) address

but is assigned at the interface controller level and are usually assigned by the manufacturer of the device. Although the IP address of a mobile device might change regularly, the MAC address may remain just the same even for its life time making it akin to a unique identifier of a device and much more significant. It has two parts the first 6bits are known as Organisationally Unique Identifier (OUI) and gives us information about the manufacturer of the network card. Organisations need to register with IEEE to be assigned an OUI which they can use to generate a full MAC address. The second 6bit is the device part of the MAC address which is unique for the device from the manufacturer. Together both form the full MAC address which is unique to every device globally. The biggest draw for using Wi-Fi for mobility analysis comes from the fact that this globally unique ID is sent out regularly by mobile devices and can be collected passively through probe requests.

As we saw in our literature review, this also creates an immense risk in terms of infringement of privacy both for the manufacturer and the user. Manufacturers of critical hardware components who do not want their unique MAC address to be publicised usually opt for registering a 'private' OUI which will not be given out to other manufacturers and will not be published publicly. Users (their mobile devices) who don't want to be tracked using their MAC addresses use a temporary MAC address which is unique only to the local network - 'local' OUI rather than using a 'global' OUI for unencrypted communications and switch to their original MAC address when a trusted encrypted connect has been established. This lack of uniqueness can be inferred from the second character of the MAC address being E, A, 2 or 6. Though this provides reasonably better privacy to the mobile users it also limits our ability to just use the MAC address from the probe requests as previous studies conducted with Wi-Fi. It is important to note that this is not a security measure rather an exception made available by IEEE 802.11 for assigning temporary addresses in ad hoc networks which has been used by most modern operating systems.

Essentially, there are two types of MAC addresses based on whether they have a public OUI or a private OUI. This distinction does not affect their uniqueness or usefulness in mobility research but hinders us from knowing about the device from the MAC address. There are also two types of MAC addresses based on them being unique globally or just in the local network. This distinction affects the feasibility of using them for device tracking or for studying movement of the users.

To summarise the above, we looked at the IEEE 802.11 standard to examine the significance and nature of the 'probe requests' which are broadcast by mobile devices constantly. We identified information present in these probe requests which are relevant to our study and examined the uniquely identifier MAC address field in detail. We see that though MAC address provides a way to globally identify a mobile device from probe requests it generates, this field can be often masked by using locally assigned addresses. We also observe that there are other relevant information which when combined can provide us an alternative to just using the MAC addresses.

3.2 *Initial Experiments*

With our theoretical understanding of the Wi-Fi standard and its capabilities, we move on to looking at the Wi-Fi landscape in real-world. We achieve this by designing small independent experiments where we record the Wi-Fi probe requests within controlled conditions along with the knowledge of the ambient population of the field of measurement. We then look at the collected probe requests, examine them in detail to look at their properties, aggregate them to footfall counts and compare them with the real-world counts to get a overall idea of how well they translate into real counts. The aim of these experiments to know more about the probe requests data and pick out the uncertainties and opportunities present in them. The objectives here are,

1. Design a simple method to collect probe requests.
2. Select locations with different levels of complexity.
3. Collect real-world data through manual counting.
4. Analyse the probe requests to extract useful information.

3.2.1 *Experiment Design*

The first step was to design a simple method to collect Wi-Fi probe requests. We accomplished this by using the open source, free software - *tshark* [3] on a regular laptop. First we put the Wi-Fi module of the laptop in 'Monitor mode' where it behaves as a wireless access point rather than a receiver. Then we invoke the command line interface of the Wireshark programme - tshark to collect the Wi-Fi probe requests received by the laptop in Character Separated File (CSV) format. The full shell script which collects the data is given below,

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

```
#! /bin/bash
tshark \
  -I -i en0 \
  -T fields \
  -E separator=, \
  -E quote=d \
    -e frame.time \
    -e frame.len \
    -e wlan_radio.signal_dbm \
    -e wlan_radio.duration \
    -e wlan.sa_resolved \
    -e wlan.seq \
    -e wlan.tag.length \
    -e wlan.ssid \
  type mgt subtype probe-req and broadcast
```

It is important to note that this script collects only particular data from the probe requests which we found to be relevant to us. The fields marked with *-e* are the ones which were collected and the correspond to the information in the probe requests as follows,

1. *frame.time* - Time stamp when the packet was received in microseconds.
2. *frame.len* - Total length of the packet in bytes.
3. *wlan_radio.signal_dbm* - Strength of the signal which delivered the probe request in dBm.
4. *wlan_radio.duration* - The duration for which the signal has been transmitted.
5. *wlan.sa_resolved* - The MAC address of the source device where the first part is resolved into a vendor name concatenated with 6 characters of the device part.
6. *wlan.seq* - Sequence number of the packet assigned by the source device.
7. *wlan.tag.length* - A list of lengths of the tags attached to the packet this acts a signature of the information contained within those tags and
8. *wlan.ssid* - The network for which the probe request is being sent for. This information is optional.

The name of the manufacturer/ vendor of the Wi-Fi module is extracted from the *wlan.sa_resolved* field into a separate column and the original field is hashed using the SHA256 algorithm implemented in R. In addition to this, the pedestrians next to the sensor were counted manually by the surveyor.

3.2.2 *Living room*

The first set of experiment was conducted with the laptop in the researcher's living room. The primary aim of this experiment to collect an initial set of probe requests to understand the information present in them in detail. The living room had 2 Wi-Fi enabled device - an Android phone manufactured by Motorola and an Android TV box manufactured by Remix. The house had 2 more phones - iPhone from Apple running iOS9 and an Android phone from Samsung in the rooms next door. The script was left running in the laptop on 15 Nov 2015 from 19:44 to 21:15 with an unexpected failure of 15 minutes in between from 19:55 to 20:10 approximately. In this duration, we collected around 3000 probe requests at the rate of 38 requests per minute.

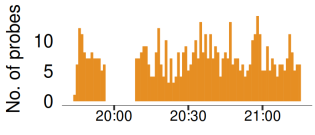

Figure 3.2: Number of probe requests collected every minute on 15 October 2017

The first thing we tried with the probe requests is to try and aggregate them based on the MAC addresses. Before the mobile devices started randomising their MAC addresses this should have accurately reflected the number of devices around the laptop. The data when aggregated showed that there were around 211 unique MAC addresses recorded. Being a residential area far away from traffic, these MACs are most likely not from unique devices. The high number must be the result of randomisation. More over since we know that there are only 2 - 4 devices in the house, there must be noise from significant distances beyond the house. The number of unique MACs recorded every minute are shown in Figure 3.2 We observe that on average we captured around 7 unique

MAC addresses every minute which quite far from the 2-4 range we were looking for.

Having established that just the MAC addresses are not enough to translate the probe requests into number of devices around the sensor, we started to look at other information we collected from the probe requests along with the MAC addresses. First we tried to isolate all the randomised, local MAC addresses by looking at the resolved vendor part. We aggregated the probe requests based on the vendor name present in the MAC addresses against all the other information present in them. The results are shown in Table 3.2 We looked how many unique values were present in these fields compared to the total number of probe requests. Initially we assumed that the randomised one won't have OUIs which are public and hence the ones which can be resolved should be the real addresses. But when we looked at the probes to MAC ratio of Google and Compex we realised that even the local MAC addresses could be registered public. This showed even when the OUI has been resolved into a vendor name, the original needs to be preserved for analysis. Samsung is special case since we know from the specification that their devices do not randomise the addresses yet they have so many unique addresses which needs to be looked in to closely.

Vendor	No. of probes	MAC addr.	Signal (avg.)	Frame length	Duration	Tags	SSID	Seq. no.
AmazonTe	101	1	-80.53	4	4	5	3	101
Apple	77	7	-86.29	4	4	9	4	77
ArrisGro	7	1	-91.71	1	1	1	1	7
Azurewav	215	4	-87.82	3	3	12	10	213
CompexPt	75	28	-88.17	3	3	5	29	74
CompexUs	4	1	-87.25	3	3	3	4	4
Dedicate	2	1	-92.50	1	1	1	1	2
Fn-LinkT	64	1	-60.58	2	2	6	1	64
Google	1347	76	-69.14	4	5	41	6	1157
HuaweiTe	11	3	-87.91	3	3	3	1	11
IntelCor	25	2	-84.04	3	3	4	3	25
LenovoMo	1	1	-93.00	1	1	1	1	1
LgElectr	1	1	-90.00	1	1	1	1	1
Microsof	3	1	-90.00	1	1	1	2	3
Nvidia	65	1	-82.91	2	2	4	2	65
OneplusT	3	1	-86.67	2	2	2	2	3
Pepwave	4	4	-90.00	1	1	1	1	4
Sagemcom	3	1	-88.67	1	1	1	1	3
SamsungE	655	27	-83.81	26	26	54	23	621
SonyMobi	56	2	-78.66	2	2	2	1	56
TctMobil	1	1	-90.00	1	1	1	1	1
Tp-LinkT	31	1	-86.16	1	1	3	1	31
Wisol	143	3	-71.91	4	5	6	3	142
XiaomiCo	3	2	-88.67	2	2	3	2	3
Unknown	110	40	-88.86	19	18	21	5	90

Table 3.2: Number of unique values present in each field captured from the probe requests aggregated by the vendor names

We observe 24 different vendors in the data, even if we assume one device per vendor it is impossible for sensor to pick up 24 different devices without a significantly large area of measurement than we expect. We need a way to filter out this noise which are generated from the edge of the field of measurement. This is where the signal strength shows good promise. Looking at the Table 3.2 we can see that two of these vendors show significantly low average signal strength - Google and

Fn-LinkT which can easily correspond to the two devices present in the room. This can be explained by the decay of the signal as it passes through the walls. In our simple example, we can filter almost all the noise with just the signal strength of the probe requests.

We then look at all the other information we collected from the probe requests and see how they compare to the MAC address for aggregating. We observe that frame length and duration provides better aggregation into unique values than MAC address when they were randomised as seen with Compex and Google. This can be explained by the argument that since the devices were essentially sending the same information over and over with just changed MAC addresses, which is are of fixed length, same devices should be sending packets of same length. We also observe that duration of the transfer, being a function of the length of the signal, has almost the same amount of information in it as the frame length. We can confidently pick one of these fields and discard the other for further analysis. Though the tag lengths and SSID looked promising way to uniquely fingerprint the devices they don't have enough volume in them to offer substantial advantage. The set of tag lengths are not as unique as the lengths or duration while the SSID information is sparse for most of the vendors. For example, 66% of probes with local OUIs, 50% of the ones with Google and 38% with Samsung don't have any SSID information in them. This makes them very poor candidates for useful information in aid us with finger printing unique devices.

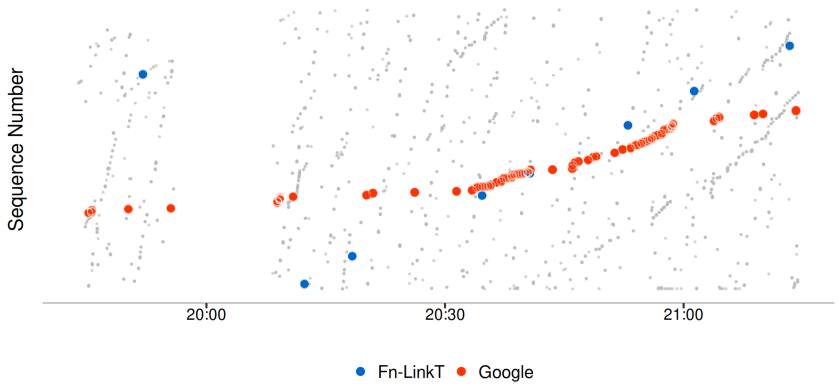

Figure 3.3: Sequence numbers plotted against timestamps showing clear patterns corresponding to unique devices.

Grey dots are probe requests with signal strengths lower than -70dBm.

Finally we found that the sequence numbers are the most interesting part of the data collected. Though they don't uniquely identify the devices directly through aggregation, along with timestamps they form visually discernible, interesting patterns that correspond to the mobile devices that generated them. In Figure 3.3 we have isolated the two vendors - Fn_LinkT and Google identified earlier and filtered only the probes requests with signal strength of more than -70dBm. We then plotted their sequence numbers against the precise time stamps when they were received. We can clearly see two devices which were present in the room and demonstrates the usefulness of sequence number in estimating the actual devices around the sensors. We need to devise a method for isolating this 'tracks' left by the devices in terms of their se-

quence numbers over time. We can also observe the rotation of sequence numbers at 4096 for the Fn_LinkT device which needs to be considered while devising such method. Figure 3.4 shows a similar exploration of Samsung devices. Though from the table 3.1 it looked as if Samsung devices are randomising their MAC addresses, we can clearly see in the figure that there are only two devices which were present for long time around the sensor and both did not randomise their addresses. The diversity of MAC addresses were indeed unique devices which must be present far from the sensor.

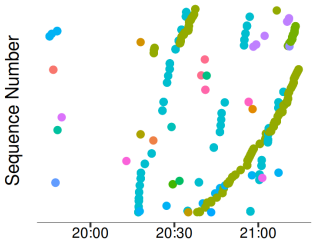

Figure 3.4: Sequence number patters in Samsung devices showing the diversity of MAC addresses showing that they are not randomised.
The colours show unique MAC addresses.

To summarise, we found that even when an unique non randomised MAC address is present when collecting Wi-Fi probe requests, we get significant noise from outside the perceived field of measurement. We also found that signal strength is a really good clue to filter out this noise. The frame length and duration looks proimising for the same purposes but they utlimately have the same information and can be used interchangeable with similar results. Finally we found that tag lengths and SSID are not useful information since they are either too varied or too sparse. Though the results of this exploratory analysis has been positive the main challenge is to make sure these methods are feasible when dealing with more real-time, real world data. We need to devise a more real world experiment to see frame lengths and signal strength work in a bigger data set for filtering out the noise.

3.2.3 *UCL South Cloisters*

This experiment was conducted collect a broader dataset from a real world setting so that we can examine the results from the previous experiments with further confidence. The specific goals were to validate the findings on signal strengths which correlated nicely to the distance from the sensor in the previous data and to further examine the usefulness of the frame length parameter. We also wanted this to be a standard test dataset on which we can test our methodologies before they can be applied to a broader project such as Smart Street Sensors. The data collection was conducted at one of the corridors in UCL - Southern cloisters which attracts a lot of pedestrian traffic during term time. This corridor also has substantial seating areas along the sides where students usually sit down for long periods of time to work. This provides us a source of devices which dwell near the sensors while constantly sending out probe requests. This area is also used heavily for lunch and for exhibitions/ events attracting large of visitors as well making it ideal for 'stress testing' our methods for cleaning and aggregation. The position of the sensor with respect to building is shown in **??**. The data was collected from 15:37 to 16:20 on 04 December 2017 and in that duration we collected around 14,750 probe requests using the scripts mentioned earlier and we counted 652 pedestrians walking directly in front of the sensors manually with pen and paper.

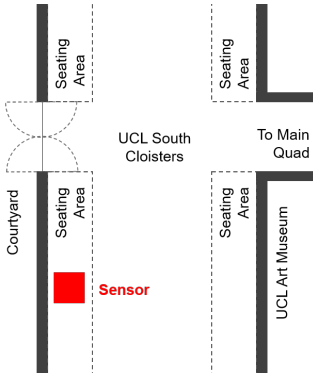

Figure 3.5: Illustration showing the configuration of the sensor at UCL south cloisters
* *Not to scale.*

Unlike the previously collected data in this experiment we made sure that the OUI information is preserved even after resolving them to vendor names. With this information we looked at the second character of the

OUI and categorised the probe requests as either 'local' - randomised or 'global' - non randomised. We then compared them to the vendor names to find if any manufacturers other than google have registered OUIs in the local range. Figure **??** shows the distribution of vendors within both local and global range of OUIs in terms of the number of probe requests collected.

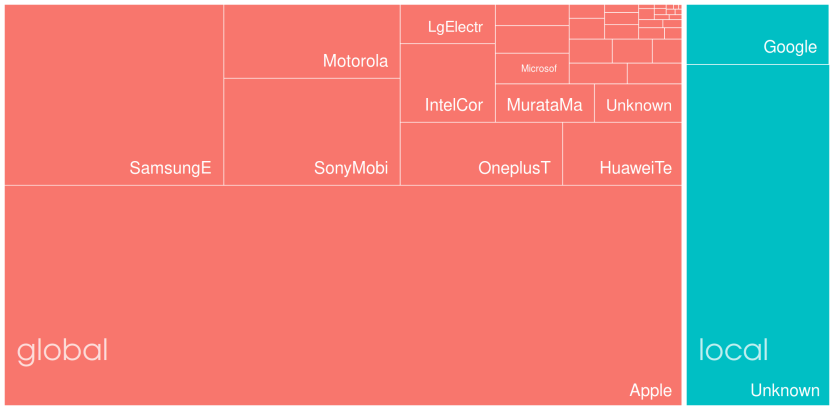

Figure 3.6: Composition of probe requests in terms of the vendor names and their type

Based on the number of probe requests

We observed that 'Google' is the only registered public OUI we find in the public range. We also noticed that the percentage of global MAC addresses collected was unusually large - 82%. This can be explained by the behaviour of the Apple devices while randomising the MAC addresses. Apple phones are known to randomise their addresses while probing for access points only when they are not connected to a Wi-Fi network already as documented by Vanhoef et al. [2016]. Since most of the members of UCL have access to 'eduroam' network and are connected to it at the campus, most of the apple devices we captured haven't randomised their addresses. This made this dataset heavily biased and not suitable for testing device finger printing methods but it does give us an opportunity to examine the nature of probe requests generated by Apple devices in particular.

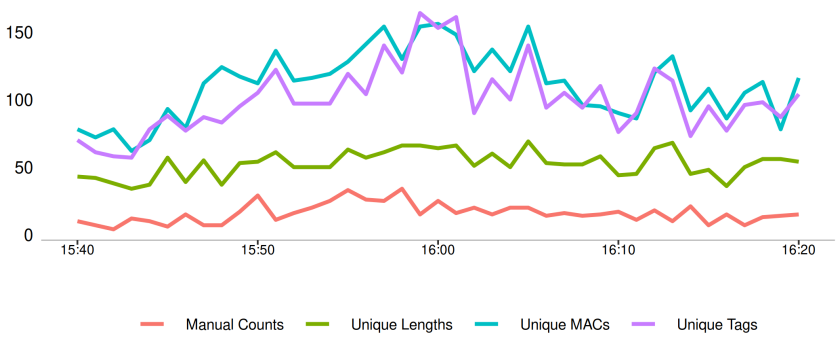

Figure 3.7: Comparison between the manual footfall count and aggregated counts from sensor collected data at UCL south cloisters.

The second step was to see how much the sensor collected counts differ from the manual counts. We aggregated the sensor counts for every minute in terms of the number of probe requests, unique MAC addresses

and the unique frame lengths and compared them to the manual counts done for each minute. The results are shown in Figure 3.7. We can observe that the original Mean Average Percentage Error (MAPE) when aggregated with MAC addresses is around 736% showing the immense amount of noise we can experience in a real world environment. This was reduced to 643% and 300% when aggregated by tag lengths and frame lengths but it is still far from being anywhere near accurate for being able to be used for estimating footfall. When we filter the probe requests for just the ones which have signal strengths more than -70dBm - the threshold which we got from the previous experiment, the MAPE for aggregating by MAC addressed, tag lengths and frame lengths is reduced to 80%, 87% and 67% respectively. The results after filtering with signal strength are shown in Figure 3.8

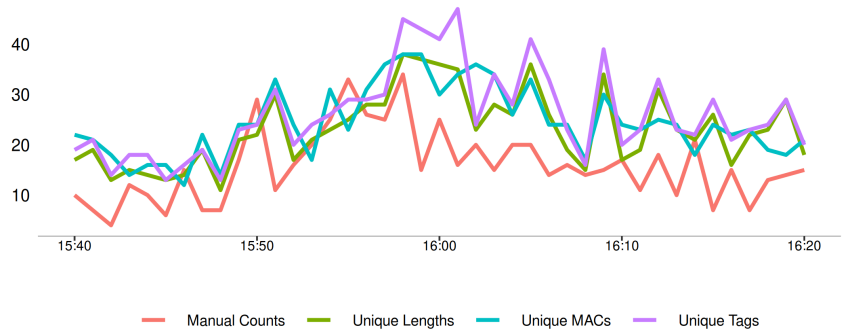

Figure 3.8: Comparison between the manual footfall count and aggregated counts from sensor collected data at UCL south cloisters after filtering probes with low signal strength

Though the signal strength filtering works to remove noise, we are still not clear about how this works or what is the most optimum cut off for filtering. We looked at the distribution of the signal strengths to find that they do exhibit patterns in terms of concentration around certain cutoffs as shown in Figure 3.9. These cut-offs can be detected dynamically from the data using one dimensional clustering methods such as k-means which are usually used to find the class intervals in one dimensional data. Figure 3.9 also shows the results of k-means clustering on the data to divide the data into 4 clusters.

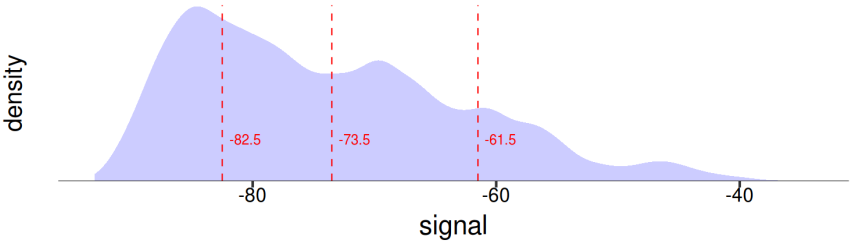

Figure 3.9: Density distribution of the signal strengths of the probe requests collected at UCL south cloisters along with class intervals.

Class intervals calculated using k-means clustering with the number of clusters defined as 4.

To summarise, in this experiment conducted at UCL south cloisters we collected a bigger set of data over a longer period of time to validate the previous finding and serve as test dataset for further research. We found that signal strength is one of the key information with which we can remove the external noise from the dataset. We also found that the

tag lengths and frame lengths though looked promising does not gives us any significant advantages. Unfortunately the data was also found to have major bias towards non-randomised probe requests because of the availability of the campus Wi-Fi. This requires us to collect a more representative data set for further research into using sequence numbers to finger print devices. Finally we also found that accurate measurement of real footfall is challenging to be done manually we need a better method to collect data for the surveyors to maintain accuracy and precision.

Figure 3.10: Location and configuration of Wi-Fi data collection carried out in Oxford Street, London.

3.2.4 *Oxford Street*

From the results of the previous two experiments we arrived at the task of devising a final 'real world' experiment collecting probe requests at a high street with high volume of footfall. Similar to the previous data, the aim here was to generate a dataset which can be used to test and validate signal strength based filtering and sequence number based clustering methodology against the scale and complexity of a busy, open public area such as a retail high street. The location chosen was Oxford street, London - one of the busiest retail streets in the world. The data was collected from 12:30 - 13:04 hrs on 20 December 2017 using the same methodology as above from a laptop in a backpack. The surveyor positioned himself at the front of a store while carrying the 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. The location where the data was collected and the configuration of the sensor with respect to the street is shown in Figure 3.10.

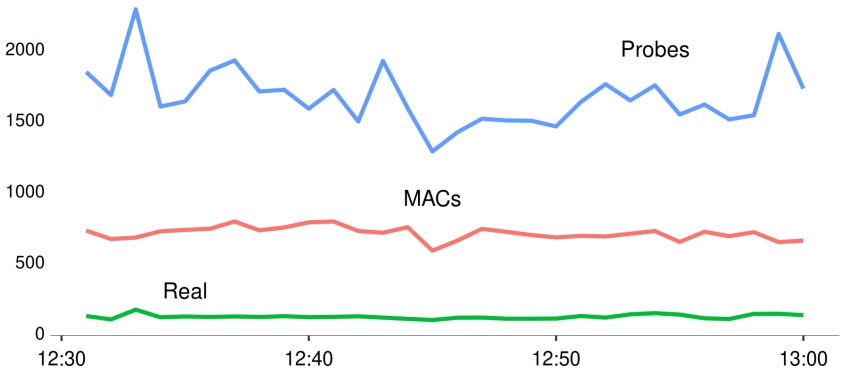

Figure 3.11: Comparison of the counts from aggregated probe requests and MAC addresses with manual counts at Oxford street, London.

The manual counting was done using a node-js base command line app running under Termux on an android phone. The application is detailed in section 7.1.1 which counts the number of times a key has been pressed on the phone. This has an additional advantage as the phone used is kept unconnected to any Wi-Fi and with the screen on for counting, emits probe requests at regular intervals. Moreover we

know the phone to be of the vendor 'Google' and randomises the MAC address, giving us a good base line to compare our results to.

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. Initial exploration of the data showed in Figure 3.11 where we compared the sensor aggregated counts to the manual counts of footfall. It shows that the data has large amount of noise giving a suitable candidate for testing. Moreover, with 55% of local mac addresses, it is free from high concentration of global mac addresses as we saw in the data from UCL corridors. This dataset is extensively used in the development of the filtering and cleaning methods and which are discussed in detail on Section 4.2

In this section, we saw the design, implementation and initial results of small experiments we conducted to understand the nature of the probe requests and the opportunities they provide us. We identified useful information in the probe requests and discarded the ones which were not useful The major conclusions arising from these experiments are,

1. MAC address on its own is not enough to aggregate probe requests into devices or footfall.
2. Signal strength is crucial to removing noise from outside the field of measurement
3. Sequence number is promising in isolating devices when their MAC addresses are randomised.
4. Frame lengths, duration, tag lengths and SSID information do not add additional value in cleaning the data.

We finally collected a fairly representative Wi-Fi dataset from a high volume retail location for use in further research on methods to clean the data.

3.3 *Pilot Study*

As we see later in Section 4.2 the efficiency of the methods to clean and aggregate data not only depend on the noise and bias in the data itself but also on external factors such as, the configuration of the sensor in relation to the environment, the day of the week etc. Thus the dataset captured in our initial experiments, though acts as a good starting point, cannot enable us to generalise our findings to all possible configurations. This necessitates an even larger data collected over longer durations in challenging situations we usually find in real world conditions. This was our primary motivation in conducting a pilot study collecting data at 5 locations across London. The aim was to collect probe requests with information we found relevant in the initial experiments for every location surveyed for at least a full week so that we can understand any artifacts caused by the periodicity of the data. We also wanted to collect data at all these locations in parallel for at least a week so that they can be compared to one other.

3.3.1 *Methodology*

The hardware setup for the sensors is illustrated in Figure 3.12. It design of the hardware is not original as it is heavily influenced by the proprietary technology of the data partner for the Smart Street Sensor project albeit a much simpler form. The core of the hardware is the general purpose single board computer - Raspberry Pi Model B running Linux Operating system. Two communication modules - 3G and Wi-Fi were connected to this machine via Universal Serial Bus interface. 3G modem was equipped with a SIM card which it uses to connect to the internet while the Wi-Fi modem is set to 'Monitor' mode. The board takes power from an outlet and the software is pre installed with the operating system which resides in a Memory card.

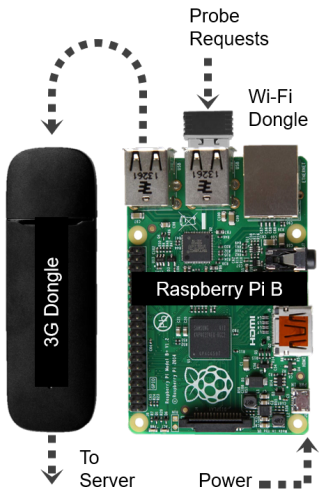

Figure 3.12: Hardware setup used to collect data in the pilot studies.

The software used for the sensors consists of two parts - sensor software and server software. The sensor software was written as a mix of Bash script and NodeJS. Essentially these scripts use wireshark program to capture packets, parse them, anonymises the MAC address fields, adds the location information, encodes them into JavaScript Object Notation format and finally sends it to a server through Web-Socket protocol. The code used at the sensor side is detailed in Appendix **??**. At the server side we have a similar NodeJS application which listens to the data sent over web sockets, parse them and saves them to a PostgreSQL database. The server side code is detailed in Appendix **??** and schematic diagram for the whole process is shown in Figure 3.13. The information collected from each probe request at these locations are,

1. Time stamp at which it was received
2. MAC address of the source device.
3. Signal Strength of the packet.
4. Total length of the packet.
5. Sequence number of the packet.
6. OUI part of MAC address.
7. Location at which it is collected.

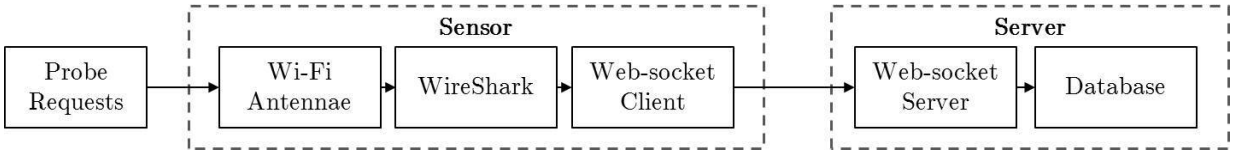

Figure 3.13: Schematic diagram showing the data collection process in the pilot study.

The manual counting at these locations were done using a custom application Soundararaj [2018]. The application was built for recording pedestrian footfall with precision and accuracy which was not possible when counted manually. The app records precise time stamp of every footfall with the precision of micro seconds which can be aggregated later at different time intervals. The code for the app is detailed in Section 7.1.2.

3.3.2 *Locations*

Five retail locations were chosen in consultation with the data partner for the pilot study keeping in mind their complexity and volume of footfall. The sensors were installed at the locations in a phased manner and multiple manual counts were conducted at each location for 30 minute intervals. The locations and their descriptions are summarised in Table 3.3.

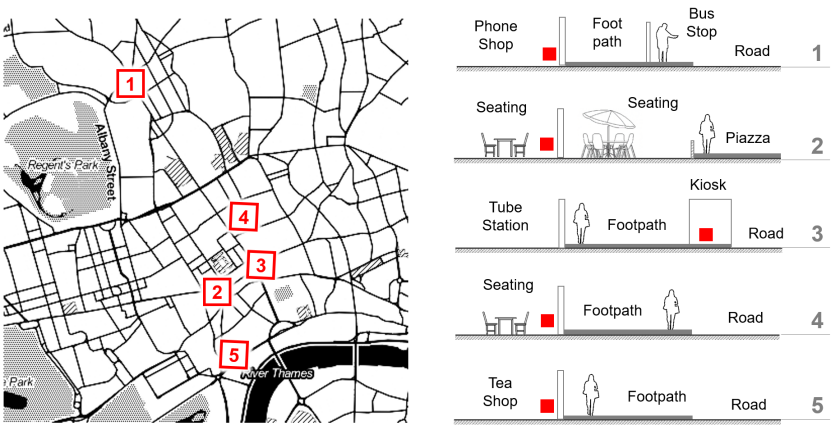

Figure 3.14: Pilot study locations in London along with their corresponding sensor installation configurations.

- *Location 1* is at the Camden high street in front of a mobile shop behind a bus stop. This location was chosen specifically because of the large amount of dwelling population at the bus stop and the stationary mobile devices inside the shop which is expected to create a large amount of noise along with the high footfall in the high street. The challenge here is to isolate the footfall from two sources of noises which are at equal distance from the sensor.
- *Location 2* is at a square with a very low footfall but has a large amount of seating of the restaurant all around it. The challenge here is similar to that of the previous location but just that the volume of footfall is low which makes it one of the hardest locations for accurately estimating footfall.
- *Location 3* is in front of Holborn station entrance in an information kiosk. This location was chosen for the really high volume from the station which is expected to cause noise. The challenge here is to be able to isolate the crowd inside the station from the footfall in the pavement.
- *Location 4* is at a fast-food restaurant at a shopping center. The sensor has restaurant seating at one side and a pedestrian footfall at the other. The challenge here is that the stationary noise and the footfall are equidistant from the sensor.
- *Location 5* is at the frontage of a shop at Strand with a mobile shop next door. This is the 'cleanest' locations of all with only

one clear source of noise which is at different distance from the footfall.

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. The schedule of the data collection and the days at which the manual counting were done are shown in Figure 3.15.

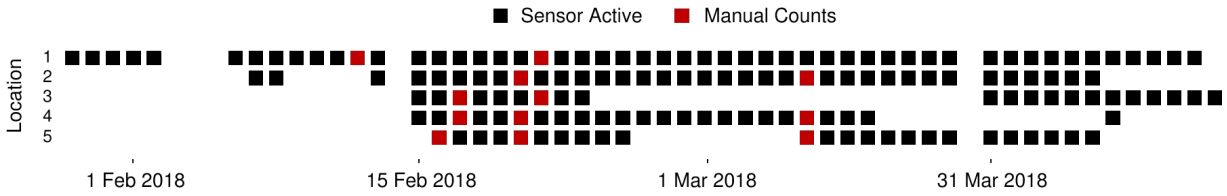

Figure 3.15: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

The survey was conducted for almost 2.5 months and about 33.5 million records were recorded which takes up to 1.8 GB of space on disk when encoded as text. During the manual counts around 10,000 people were counted walking past these sensors. A detailed account of the volume and velocity of data collected at these locations were given in Table 3.3. The dataset collected was used extensively to develop and test the signal strength based filtering and sequence number based clustering methodology which are detailed in the Section 4.2.

Table 3.3: Locations of data collection in the pilot study and the amount of data collected at each locations

** Total probe requests in $\times 10^6$(per minute) ** Total footfall (per minute)*

Id	Location	Type	Installation notes	Probes*	Footfall**
1	Camden St.	Phone Shop	Bus stop in front	9.9 (297)	3683 (33)
2	St.Giles	Restaurant	Seating on both sides	3.9 (169)	0346 (05)
3	Holborn Stn.	Info. Kiosk	Front of station entrance	4.3 (303)	2956 (46)
4	Brunswick	Fast Food	Seating on one side	3.4 (210)	0960 (12)
5	The Strand	Tea Shop	Phone shop next door	8.4 (382)	1969 (21)

3.4 *Smart Street Sensor Project*

The Smart Street Sensor 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 and Consumer Data Research Centre, University College London (CDRC, UCL) and Local Data Company (LDC). The project was designed to serve as the first and unique comprehensive research into the patterns of retail activity in UK high streets by measuring their real-time footfall from collecting Wi-Fi probe requests. The data for the project was collected through sensors installed at around 1000 retail locations across UK.

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 is 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 the occupiers of the retail premises who want to improve revenues, landlords who want to increase the value of the property, local authorities who want to improve the vibrancy of the retail economy and finally investors and consumers within the retail industry. The project also aims to facilitate decision making by these stakeholders in addition to the tremendous opportunities for academic research.

Figure 3.16: Hardware setup used to collect data in the pilot studies.

3.4.1 *Methodology*

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. Figure 3.17 shows all the locations in United Kingdom city-wise and Table 3.4 shows the region-wise distribution of locations.

Region	Locations
Greater London	479
Scotland	118
Yorkshire and the Humber	114
South East	103
North West	98
South West	87
East Midlands	68
East Of England	49
West Midlands	39
North East	26
Wales	17
Northern Ireland	2

Table 3.4: Regional distribution of Smart Street Sensor locations across UK

We can see that the project has a strong London bias which along with other retail centers in Greater London area accounts for almost half of the locations. We also need to note that the locations are retail and any insight from the data needs to be looked at with a retail point of view.

A custom footfall counting technology using Wi-Fi based sensors (Figure 3.16) was also designed, developed by LDC and the sensors were installed the identified locations. The sensor employs proprietary hardware and software, monitors and records probe requests sent by Wi-Fi enabled mobile devices present in its range. In addition to this, the number of people walking by the sensor was counted manually for short time periods during the installation of the sensors at the corresponding locations. The project aimed 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 are converted into a unique cryptographic hash at the 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.

Figure 3.17: City wise distribution of locations with Smart Street Sensors installed.

The sensors are usually installed on partnering retailer's shop windows so that its range covers the pavement in front of the shops. A typical configuration of a sensor at a location with respect to the premise

the pavement in front of it is illustrated in Figure 3.18 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 Azure container at regular intervals of 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 project began on July 2015 with the first sensor installation and has grown to an average of 650 daily active sensors as of January 2019 with a total of 1200 locations been involved in the project since the beginning. We have collected around 2TB of data comprising of around 73 Billion probe requests.

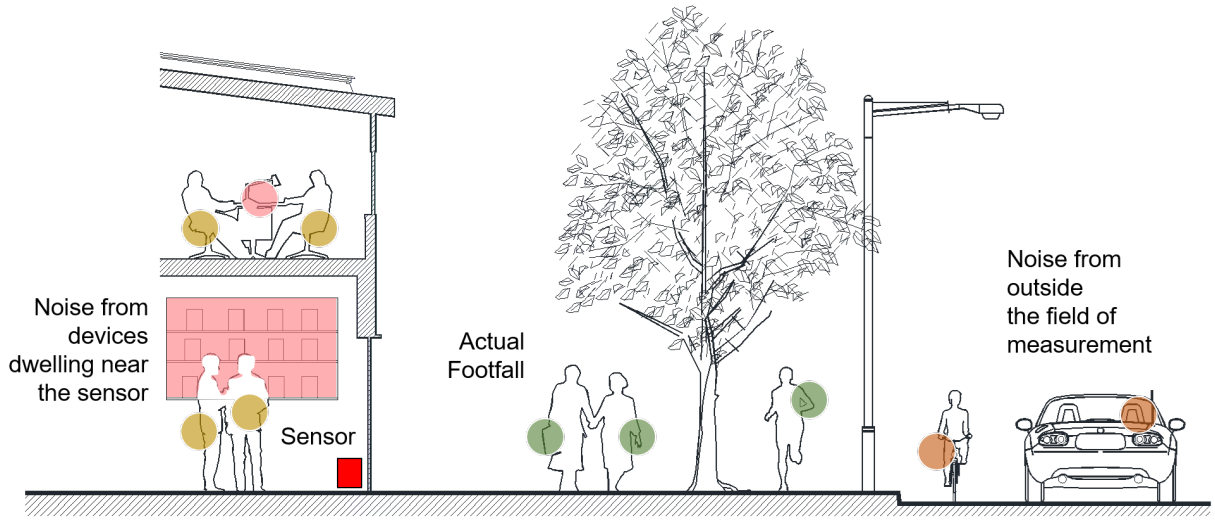

Figure 3.18: Cross section showing a typical installation of Smart Street Sensor in a retail frontage.

Due to the scale and the commercial nature of the project the sensors collect fewer than the previous experiments. The information collected by the Smart Street Sensors are the 5 minute interval when the probe request was collected, hashed MAC address and signal strength. The probe requests are aggregated by the MAC addresses within the same 5 minute intervals and hence the signal strengths are aggregated to the minimum signal strength reported. Due to the longitudinal nature of the project, the data collection methods have changed over time as well. The hardware has been upgraded with more capabilities in the early 2016, the interval they reboot has been adjusted several times in 2017 and finally with the MAC randomisation problem accentuated in the later part of 2017 the signal strength aggregation was changed from minimum to maximum in March 2018. Essentially, the data has changed over time and we need to consider the changes while devising the methodologies for cleaning the data.

3.5 Discussion

In the previous sections we designed and implemented data collection processes to arrive at 3 different datasets - Small experiments, Pilot Study and Smart Street Sensor project. The small experiments were designed as way to collect as much data as possible from the probe requests for short periods of time aiming to collect small sets of comprehensive data under controlled conditions for exploratory purposes. The pilot study extended this further by collecting data for a longer time in real world conditions aiming to validate the insights collected with the experiments and the methodologies we devise for the research. The Smart Street Sensor is the most comprehensive study which collects very small focussed set of data at a national level for very long periods of time. These datasets give us a well rounded set of data to set up our toolkit and devise out methodologies. The summary of the datasets in terms of their characteristics is show in table 3.5.

Before we move on to develop methods to process the data into information on footfall in these locations, the crucial action is to look at the possible biases and uncertainties in these datasets arising due to the data collection methodology and from the broader context. These form the framework on which built our data processing pipeline where we propose to solve each of these uncertainties in each step.

From our understanding of the data, we observe that the major sources of uncertainties are regarding the range of the sensor, the frequency at which mobile devices generate probe requests, the way and rate at which the mobile devices randomise their MAC addresses, the collisions caused due to the hashing of the MAC addresses and finally the gaps introduced by the failure of the sensors. There is an inherent bias to these data caused by the mobile phone ownership in the population which varies across time, location and demography. We discuss each of these uncertainties and biases in detail below,

Table 3.5: Summary of the collected data-sets.

Dataset	Locations	Time	Detail	Purpose
Small Experiments	3	30 - 60 mins	High	Exploratory analysis
Pilot Study	5	6 weeks*	Medium	Devising and calibrating methodologies
Smart Street Sensor	1000*	4 years*	Low	Real world insights

*approximate

3.5.1 Range of the sensor

The first and foremost uncertainty we face with wireless sensors such as Wi-Fi and Bluetooth is the delineation of the field of view of the sensor. Though the Wi-Fi signals can be partly managed or restricted by manipulating their strength, there is no reliable way to precisely delineate a survey area for these sensors. The manipulation of signal strength can be done by installing metal shields around the sensors to block certain directions and prioritise others but the method cannot block out all the

signals and will leave some uncertainty about where the probe requests are coming from. Moreover, strength of the signal received from a mobile device by the Wi-Fi access point depends on numerous factors such as,

1. Distance between the mobile device and the Access Point.
2. Thickness of the objects present in between them.
3. Nature of obstructions such for e.g. metal vs glass
4. Interference from other wireless devices.
5. Power level of the transponder of the Access Point.
6. Power level of the transponder of the Mobile device.
7. Even atmospheric conditions such as humidity, temperature etc.

The signal strength drops non-linearly when moving away from the access point as shown in Figure 3.19 and there is a *close-range non-monotonicity* as well - where within 10 feet, A closer device can report lower signal strength than a farther one [Cisco, 2008]. The relationship between the two is given by the equation [4],

$$\log_{10} d = \frac{(P_o - F_m - P_r - (10 \times n \times \log_{10} f) + (30 \times n - 32.44))}{10 \times n} \quad (3.1)$$

Where,
d = distance - Sensitivity of the receiver
F_m = Fade Margin - Sensitivity of the receiver
n = Path-Loss Exponent, ranges from 2.7 to 4.3
P_o = Signal power (dBm) at zero distance - Measured by testing
P_r = Signal power (dBm) at distance - Measured by testing
f = signal frequency in MHz - Specific to the hardware

[4] Zengrzengr, Andy, and Cabral. Calculate distance from rssi, 2017. URL https://bit.ly/2Ohskf9

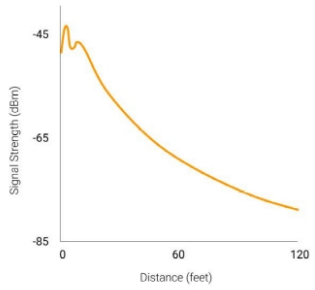

Figure 3.19: The decay of signal strength (RSSI) with respect to distance.

Source: Wi-Fi Location-Based Services, Cisco

All these factors vary widely in real world conditions at each location depending on where and how the sensors are installed. They also vary widely over time due to changes in the context and vary across different directions at each location as well. This makes it extremely difficult to model the distance between mobile device and the access point as a function of the received signal strength. The equation 3.1 can be approximated and simplified as,

$$R = (-10 \times \log_{10} d) + A \quad (3.2)$$

Where R is the reported signal strength and A is the signal strength at 1 meter. Though equation 3.2 can help us to infer the distance of the mobile device roughly, the location wise uncertainty of this method makes it lose the meaning when compared across locations.

From the above we can conclude that it is almost impossible to delineate the field of measurement precisely and accurately by simple methods using the information present in the probe requests. This leads to uncertainty in the data collected which needs to be resolved with explicit assumptions or specific methods to reduce the resulting noise. We require a method to isolate this noise from data generated by devices within the field of measurement. This methods needs to be independent of the micro site configuration, temporal changes in the context.

3.5.2 *Probe request frequency*

The second uncertainty we face is the frequency at which mobile devices generated the probe requests which varies wildly. The number of probe requests generated from a mobile device depends on,

1. Manufacturer of the device. E.g. Samsung vs Apple
2. Version of the software running on the device. E.g iOS 7 vs iOS 8
3. State of the device. E.g. If it is already connected to internet? Has the location services been switched off?
4. The number of access points already known to the device.

Studies done by Freudiger [2015][5] has shown that the number of probe requests generated by a mobile device vary widely across manufacturers such as Samsung, Apple and LG, across different state the devices are in such as charging, screen being on, Airplane mode being on etc and depends heavily on the number of access points known to the device. It is also seen from our initial experiments that these probe requests are generated in short bursts rather than being generated at regular intervals. This makes predicting a base factor for calculating a number of mobile devices based on the number of probe requests received much more complex. The variety of device models available and the pace of change in software that run these models further complicates this. Though we can simply aggregate these probe requests based on the unique information on them, in absence of such information it becomes extremely critical. We to consider this uncertainty in detail while making any simple assumptions on the relationship between number of probe requests and the number of mobile devices that generated them.

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

3.5.3 *MAC address randomisation*

Randomisation of MAC address is one of the recent uncertainty introduced in the data. As we saw in section 3.1 MAC address is the unique identifier for each mobile device and we aggregate the footfall numbers based on this. Since the probe requests are transmitted unencrypted and can be received by any access point, this is one of the biggest leak of personal data which occurs in the Wi-Fi based communications. Modern mobile devices solve this problem by using a randomised MAC address for the probe requests which can result in large over estimation of number of mobile devices in the vicinity.

The method of randomisation and the frequency varies widely between device manufacturers and also changes as new versions of the software are released. This seriously affects the usefulness of the data long term where methods designed to overcome this randomisation can be rendered inefficient in the future. Figure 3.20 shows the increase in the share of randomised MAC addresses since 2015. We can observe that in addition to the overall upward trend there are bursts of increase around late 2016 and 2017 which coincides with the release of new mobile operating systems. This makes it necessary for devising a method to overcome MAC randomisation problem to be able to uniquely fingerprint devices so that they can be aggregated together. As we saw in our literature

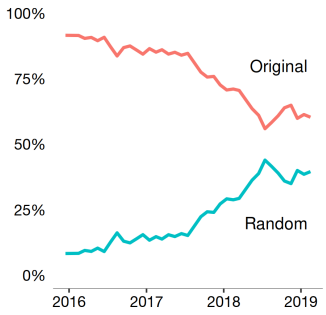

Figure 3.20: Increase in the share of randomised MAC addresses compared to non-randomised original ones over the years.

From data collected at Regent Street, Cambridge.

review, this is also one of the major opportunities in research on human mobility using Wi-Fi data.

3.5.4 *Mobile Device Ownership*

One of the major external bias in all the datasets collected from mobile devices is the overall volume and nature of the ownership of these devices. The ownership of mobile devices, specifically Wi-Fi enabled ones, have been on the rise since 2005. Though mobile ownership has reached unprecedented level in recent years, there is still an underlying increasing trend present in the ownership of these devices which manifests itself in the collected data. Moreover the mobile ownership varies widely between demography of age and geography as well. Figure 3.21 shows the mobile ownership across age groups in UK from 2012 to 2018. We can observe that the older age groups are under-represented in our data. This needs to be taken into consideration while using this data for extrapolating demographic conclusions out of it. In addition to this, the overall upward trend needs to be adjusted assuming 1% increase monthly and 0.2% weekly when using this data across long periods of time.

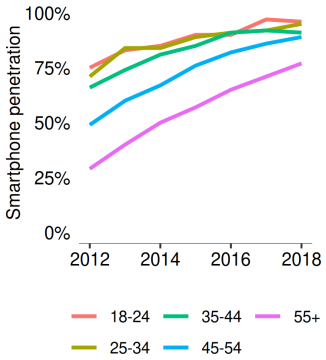

Figure 3.21: Smartphone penetration by age group in United Kingdom (2012-18)

Source: UK edition, Deloitte Global Mobile Consumer Survey,

3.5.5 *Missing Data*

Being collected by distributed set of sensors located in busy real-world scenarios, the data has a large number of gaps as well. These gaps are caused by various reasons such as, failure of the sensor hardware and software, failure in internet connectivity to send the data back, External factors such as store closures which can cause power loss, regular disruptions such as software update and maintenance and finally other factors such as unauthorised tampering and unplugging of the sensor. This is leads to a dataset which contains several small and medium sized gaps as shown in Figure 4.2. More over, the Smart Street Sensor project is implemented and managed with commercial motive, the sensors are installed and uninstalled at locations as retail partners join and leave the project. This leads to an uneven availability of data across locations over longer time periods which creates challenges while aggregating data across locations. We need to implement a methodology to fill in these gaps which considers the periodic patterns in the data. We also need to devise a measure for aggregating the counts across locations which removes the bias introduced by long term gaps in data.

3.5.6 *MAC address collisions*

Finally, from the initial analysis we have observed that there are few instances of collisions occurring in the hashed MAC addresses. This has been observed as unique hashed MAC addresses appear at different locations within a short period of time which cannot be explained by the physical travel by the user between these locations. These collisions are caused by the limitation of the hashing algorithm used and exist only in very large amounts of data. It is important to note that this

collision are specific to non-randomised MAC addresses as we don't expect any consistency within the randomised ones. Even though this is an inevitable side effect of the hashing process, the probability of such occurrence is very low and is calculated as 2^{-n}, where n is the number of bits in the output of the hashing algorithm. The total number of estimated collisions between m unique values is given by, $2^{-n} \times \binom{m}{2}$[6]. This translates to around 100 collisions across a million unique devices with a 32 bit algorithm and 2 collisions across 10 Billion devices when using a 64 bit algorithm. Though these collisions might cause issues in granular mobility models, for long term and broad studies where we don't track individual devices, they can be safely ignored.

[6] mikeazo and poncho. Formula for the number of expected collisions, 2015. URL https://bit.ly/2YS6zYl