Population Mapping via Wi-Fi Network Analysis:
The Final Report

Sudipta Banerjee, Grant King, Glen Simon

CSE 824: Advanced Computer Networking and Communications

Abstract

In this paper we present an in-depth analysis on the usage of the Wi-Fi network located in the College of Engineering building at Michigan State University, with a focus on mapping the population distribution of the network users. The novelty in our work resides in the fact that we are not relying on GPS information from users' phones to provide localization. Instead, a dense collection of access points with known locations are used to identify the locality of connecting users. Our work provides statistical information about the behavior of congestion in the network, shows how the population distribution can be modeled using heatmaps, and offers a network usage prediction model that can represent how a population typically navigates through the building. We believe that this work is beneficial by allowing better understanding of how the current network is being used by the student population, helps identifying where highly congested locations are in the building, and can be used as a tool when determining the optimal deployment locations for future access points.

Keywords: Networking, Wi-Fi Tracking, Population Density, Population Movement Prediction

1. Motivation & Literature Study

Population mapping is commonly used in the context of mapping people in a particular geographical location to compute the 'population density'. It has a different connotation when it is associated with genomic projects and typically involves study of evolution patterns. In the current work, population mapping has been used to refer to analyzing patterns of mobility of a certain student population using data collected through Wi-Fi signals. We intend to compute statistical information from the Wi-Fi network to develop a predictive model which can provide important insight into mapping the students in a building and further provide diagnostic information about the network connectivity. Population mapping caters to a suite of applications, namely,

- Effective crowd-monitoring at social events.
- Time sensitive evacuation strategies.
- Hotspots for promotional events and advertising.
- Mobility tracking of a particular individual as part of surveillance operation.

Extensive research has been done in this regard [1, 2]. Most of the research is geared towards population tracking using Bluetooth signals, which comes with its own host of problems. Firstly, the devices need to be in *discoverable* mode for the traking to be initiated. A number of devices disable Bluetooth connectivity by default. Secondly, the range of Bluetooth communiction extends only to small distances ($\approx$ 30 ft.) compared to Wi-Fi which services typically upto 100 ft. A viable alternative is employing Wi-Fi signals, since, most devices (smartphones, laptops, smartwatches) accommodate connecton to Wi-Fi, but poses some privacy concerns since the GPS data may be integrated with the Wi-Fi information for localization of an individual, as done in [3]. Anonymized Wi-Fi signals have been widely used for analysis of wireless LAN technology in an academic setting [4]. The paper focuses on studying the patterns of network traffic from the context of Access Points (APs), Network Interface Cards (NICs) and protocols. However, the paper does not integrate the spatial and temporal characteristics to give geographic mobility of a user. Gross statistics are computed on the overall data but finer granularities are not exploited. The results obtained are generalized enough since the data is collected from diverse settings, spanning 161 buildings and involves close to 2000 users. However, involved research into nuances of the data is overlooked, for example, whether some access points exist which were continuously overloaded compared to others, or, if the network is down, how long does it usually take the network to be restored again, and so forth. The authors confirmed that the data collected through Syslog messages, Simple Network Management Protocol based polling and via tcpdump snifffers was not an accurate representation of the data, since some APs are not probably configured to transmit Syslog messages. The above work was essentially an extension of the WaveLAN study carried by Tang and Baker in [5]. Use of cellular data for analyzing the behavior of mobile users has been performed in [6]. Their work analyzes the specifics of the content and is not strictly applicable to population mapping. Similar research is conducted by Balachandran *et al.* in [7] but does not offer much insight into *how* the data can be used for understanding the moving patterns of a population. It focuses on a single day at a conference and the variability in the data is therefore restricted to homogeneous activity of the participants of the conference who followed a strict schedule. Our work has several contributions as follows:

1. The current work collects Wi-Fi signals from the College of Engineering building on Michigan State University campus over a timespan of $2\frac{1}{2}$ months for the Fall term 2017 with collaboration from DECS and analyzes the anonymized data for diagnostic purposes.
2. Analysis is done in the context of density of active access points, how does the activity change over the course of a day and spatial aspects correlating the location of the access points and the burst of netwrok activity is studied in great detail.
3. Temporal characteristics of the data is studied to present weekly pattern of network traffic and also track the movements of a particular user (volunteer) by integrating the spatial details with the timestamped Wi-Fi usage information.
4. Finally, a predictive model is proposed which closely approximates user mobility.

2. Data Set

For this study, all data was obtained through working in conjunction with the Division of Engineering Computer Services (DECS) at MSU. DECS is responsible for maintaining the Wi-Fi network within the engineering building on campus. To provide us with data for analysis, they agreed to deliver us log files containing information about the arrivals and departures for devices

connecting to each access point (AP) in the network. In these logs, devices were identified by their MAC address. However, since DECS also provides a network bouncer service, which corresponds MAC addresses of devices to student IDs, the MAC addresses in the logs were passed through a hash function before being delivered to us. This was necessary to protect students location information.

```
[time stamp]          [anonymized device ID] [arrival or departure] [location]
Sep 10 18:51:22    87bc88751a348ff47693c61e4785    UP    eb-ap-1-9-1307-east-hall
```

Figure 1: Example log entry from data logs used in this study.

An example of a log entry from the obtained data logs can be seen in Figure 1. Each log entry consists of 4 fields, including: a time stamp, anonymized device ID, a flag signaling either an arrival or departure, and the AP that this occurred on. Each field represents the following:

- timestamp : Datetime of when the action occurred
- anonymized device ID : 64 byte hashed MAC address
- arrival or departure : UP for arrival, DN for departure
- location : Approximate location of the AP within the building

Data logs were created for every day between September 10th 2017 to November 8th 2017. On average each log consisted of about 196,530 entries. To give this number better context, according to the MSU College of Engineering[1] , there were 6,770 engineering students (including undergraduate and graduate) enrolled for the Fall 2017 semester.

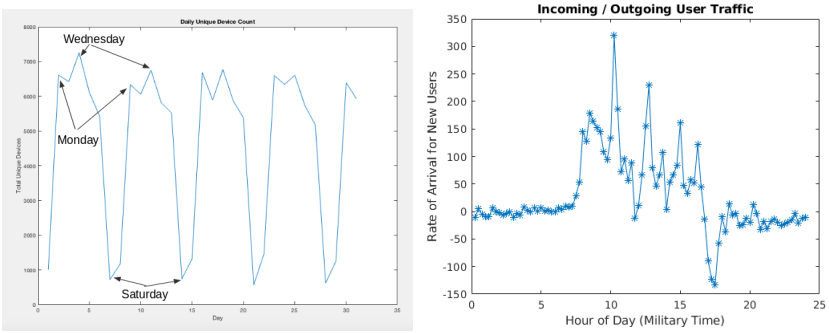
3. Preliminary Work

Upon receiving the data logs, initial analysis was preformed. Our primary interest at the start of the project was to identify if there were any trends in the day-to-day traffic and determine what the network usage looked like over the course of a typical week day. In Figure 2(a), it can be seen that weeks tend to show very similar usage patterns. The weekends always report a much lower count than week days, with Saturdays reporting lower numbers than Sundays and the least overall. Wednesdays almost always have the highest count of unique devices, followed closely by Mondays. This pattern makes sense due to how MSU classes are generally scheduled, being that if a course meets on Mondays, it also meets on Wednesdays.

Figure 2(b), shows the arrival rate of network users over the course of a sample day. These results are close to what you would expect from a typical building on a college campus. Most activity happens between 8am and 5pm, with a steady arrivals of users in the morning coming to a peak around 10am. This is not surprising, as many undergraduate classes are scheduled to start at the 10:20am time slot. Around 12pm, a decrease in arrivals is seen and for a brief period active users are actually departing the network faster than new ones can join. This observation coincides with the time that many often leave the building for lunch. Traffic than increases throughout the

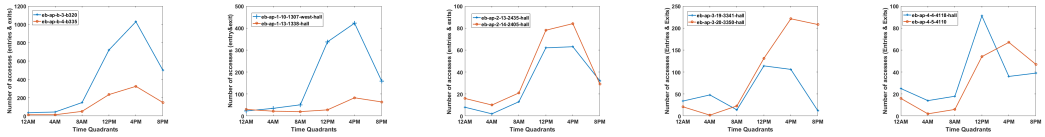
[1] https://www.egr.msu.edu/sites/default/files/content/UGS/fs17_collegeofengineering_enrollments_active.pdf

rest of the afternoon, until around 5pm when we observe a rapid departure of active users. Traffic after this point is slow for the rest of the night.



(a) (b)

Figure 2: (a) Shows the total number of unique device IDs that connected to the network at least once during the day over the course of a month. (b) Shows the rate of arrivals through a sampled day. Negative values represent users departing faster than they are arriving.



(a) (b) (c) (d) (e)

Figure 3: Arrivals and departures at two selected access points on September 10, 2017. (a) In the basement, (b) On the first floor, (c) On the second floor, (d) On the third floor and (e) On the fourth floor.

Another one of the primary objectives is to extract temporal characteristics from the data to gain a better understanding of *when* are the APs busiest on a given day? Are there some *discernible patterns* which can help identify densely populated areas at a given time? In this regard, we divided a day into 6 quadrants comprising of 4 hours each as follows, Q1: 12am - 4am; Q2: 4am - 8am; Q3: 8am - 12pm; Q4: 12pm - 4pm; Q5: 4pm - 8pm; Q6: 8pm - 12am. There is no overlap between consecutive quadrants, for example Q1$\in$ [12*am* $-$ 4*am*) and Q2$\in$ [4*am* $-$ 8*am*), and so on. Figure 3 depicts the arrivals and departures of two access points (indicated in the figure), having the two highest volume of traffic, located in the basement and on the four floors, on a given day. As anticipated, traffic is slow in the first two quadrants and gradually picks up around Q3 and peaks around Q4 and then slowly declines in the evening dipping around Q6. Similar trends are observed for the entire engineering building. However, this analysis does not integrate the spatial and temporal characteristics together, which may provide valuable insight. In addition, the preliminary work focuses on two APs carrying the maximum volume of traffic, but the analysis should be done on a larger scale carried out over a longer period of time, such that is generalizes to the entire building.

4. Results

In this section, we build upon the ideas presented in Section 3 to perform an insightful analysis of the data on a larger scale, with focus on mapping the APs throughput the building, have a visual representation of the regions in the building which are budding with Wi-Fi activity; followed by tracking mobility of individuals, analysis of diagnostics of the APs; and develop an analytical model which fits well to the real data and subsequently predict the probability of encountering an arrival at an AP succeeding the former arrival at that AP.

4.1. Access Point Mapping

As discussed in section 2, the Wi-Fi data logs only contained the access point (AP) name that arrivals and departures where seen on. To provide geographical context to this data a mapping had to be created to show the relationship between these AP names and the physical location in the building where the APs were installed. To create this mapping, first dimensionally accurate floor plans had to be obtained from the MSU engineering website at [2]. These obtained high resolution floor plans were than used as a basis to determine the spacing and relative locations between each of the APs.

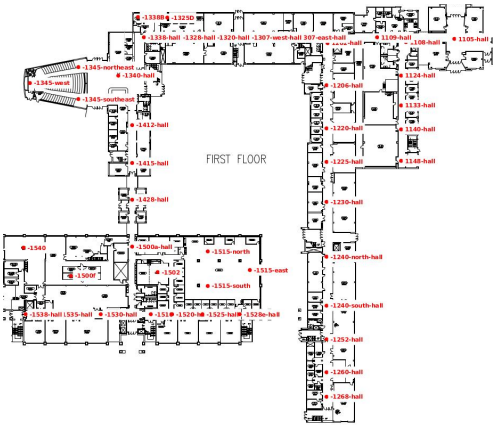


Figure 4: AP distribution for the first floor of the MSU engineering building.

To determine the spacing between APs and finish the geographical mapping of the locations, the engineering building had to be surveyed by our research group, looking for each physical AP in person, determining that AP's name, and finally translating it physical position to a X,Y pixel values on the floor plans. For this study, this mapping was done for 136 of the total 145 APs in the building including all APs on the first through third floors, but ignoring those in the basement or on the fourth floor. There were 45, 44, and 47 APs on the first, second, and third floor respectively.

Once the mapping of the AP names to locations on the floor plans were created and logged, a Matlab script was used to generate a plot to overlay the floor plan image to give user's a better idea of the AP distribution throughout the building. This AP overlay was used as a building block for many other aspects of this project, and an example showing the AP distribution for the first floor can be seen in Figure 4.

4.2. Access Point Temporal Analysis and Mobility Tracking

The preliminary work (see Section 3) initiated the analysis of the network traffic in different time quadrants through the course of the day. We suppose that the temporal analysis coupled with spatial information can be used towards delivering diagnostics of the APs and also, tracking an individual. The weekly traffic pattern of the APs is observed from the data over a 28 day period, September 10, 2017 - October 9, 2017. The APs considered in this analysis span the entire engineering building. In addition, we also illustrate the volume of the network traffic in different time quadrants.

[2] `https://www.egr.msu.edu/about/maps/engineering_building`

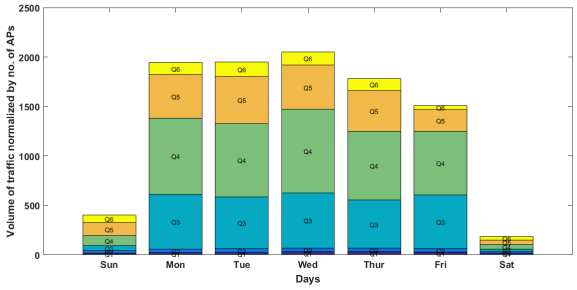


Figure 5: Analysis of network traffic based on a weekly pattern.

The key observations to be noted from Figure 5 are as follows:

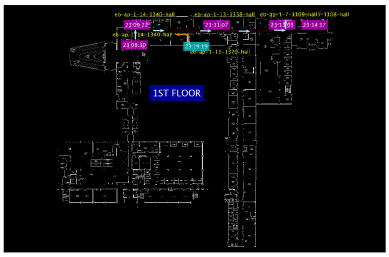
- There is a *sharp* contrast in the volume of traffic on weekdays (much higher) compared to weekends (typically lesser on Saturdays against Sundays).
- The largest proportion of the bar plot is occupied by the quadrant Q4, implying 12pm - 4pm is usually the busiest in the engineering building and the observations should be pertinent with any building on the campus. This is the time when people are moving around the building switching between classrooms, and also leave for their lunch break and then return to the building to resume classes.
- The smallest proportion of the stacked bar plot is occupied by Q1 and Q2. Usually the building is unoccupied during the 12am - 8am (the next day) interval when students retire for the night to their residences.
- Mondays and Tuesdays have similar network traffic, Wednesday reports the maximum peak and slowly declines on Thursday and Friday.

Other diagnostic information can be leveraged from the data by looking at the busiest APs and then try to interpret the densely populated areas, maybe because of the location of the APs near a cafeteria (Sparty's) or, close to a lecture hall. Some prior information can be incorporated to predict future patterns on a day in the future. In this context, we focused on the first floor of the building and found that in the quadrants Q3 and Q4, the AP '*eb-ap-1-10-1307-west hall*' would appear as one of the top three busiest APs on every Tuesday. The floor plans of the first floor were examined and we hypothesized that the AP may be capturing the network traffic in one of the classrooms. Our above supposition was confirmed when similar patterns was observed on Thursdays as well. The class schedules in the engineering building usually occur in pair-wise fashion, such as Mondays and Wednesdays or, Tuesdays and Thursdays. Further examination of the Schedule of Courses[3] revealed that the course CSE 836: Prob. Models and Algos in Comp was taught from 10:20am-11:40am (Q3) and CSE 841: Artificial Intelligence was taught from 12:40pm-2pm (Q4), both in **EB 1300** on Tuesday and Thursday. The AP was closest to the particular classroom. This gives insight as to the classroom will be populated in those particular time quadrants.

[3] `https://schedule.msu.edu/`

(a)

(b)



(c)

Figure 6: Tracking mobility of the first volunteer. (a) On the third floor, (c) On the second floor and (c) On the first floor.

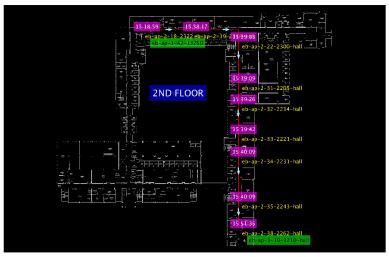


Figure 7: Tracking mobility of the second volunteer.

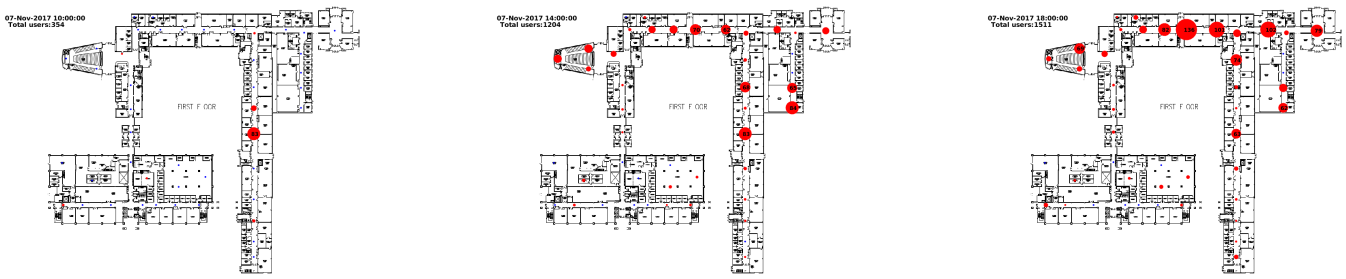
To track the mobility of an individual, two volunteers were tracked from the Wi-Fi data.

The motivation behind mobility tracking is two-fold: (a) Firstly, we would like to see if we can reliably locate someone from the data, and (b) Secondly, how does the tracking vary when a person briskly walks as compared to a stationary person who frequently stops during the course of his/her walk. The first volunteer moved through the building on October 24, 2017 from 11pm-11:20pm, covering the third floor, then second floor, first floor, then to the basement and exited the building through the first floor. The second volunteer moved on the second floor on October 27, 2017 from 3:18pm-3:50pm and stopped at locations multiple times for duration >10 mins. Figure 6 depicts the mobility pattern of the first volunteer with timestamps and some candidate APs (all the APs are not presented for the sake of brevity). It is clearly noted that the person is reliably tracked from the information gathered using the APs. However, Figure 7 indicates that at some locations the device carried by the volunteer connected to two APs on different floors (second and third floors), in spite of being physically located on the second floor only. This phenomenon is observed when the volunteer is stationary at one location for a considerable time, which reflects the unreliability of the Wi-Fi data to give a fairly *accurate* location of the individual. The primary reason for this is because the device is trying to aggressively look for a stronger signal and the state machines in the network interface card embedded in the device and the APs are out of sync causing the frequent roaming within a subnet. The same observation was noted by the authors of [4]. Thus, a person can be reliably tracked only if the person is moving continuously and does not pause for a long period of time.

4.3. Population Density Heatmap

Population density maps (heatmaps) are often useful tools when studying how a population utilizes or distributes through a space. Theses heatmaps can be even more useful if they are improved to show changes over time. Temporal heatmaps can show not only how a population distributes across a space, but also how that distribution changes over time. This can provide knowledge about trends in the movement of the population over the course of a day, week, month, or even longer.

In this study, temporal population density heatmaps were created to show the population trends in the MSU engineering building for each day of the week. The population trends are based off of the number of users that are connected to each AP throughout the building and is based off the assumption that a high number of visitors to the engineering building are carrying a device that connects to the public Wi-Fi when in range. These temporal heatmaps were produced by first generating a series of snapshot heatmaps, each showing the population distribution at a time step **Ts**. **Ts** was varied, at a configurable rate with a granularity of minute lengths, from midnight at a particular night to midnight of the following night to create a series displaying one day of traffic. Each of these snapshot heatmaps were than compiled into a video to show a smooth transition of what the number of user's on each AP throughout the day looks like. Example snapshots from one produced temporal heatmap can be seen in Figure 8. It can be seen in this figure that the total number of users increases as expected throughout the day, often coming to a peak around 6pm (behavior varies on weekends). It can also be seen that the population is much more dense on the North side of the building (North corresponds to the top of the image). This is true on all mapped floors of the building and for every week day (Monday - Friday).



(a) (b) (c)

Figure 8: Population density heatmap for the MSU engineering building at various periods during the day on a sample Tuesday. (a) First floor at 10am. (b) First floor at 2pm. (c) First floor at 6pm.

Heatmaps at each time step **Ts** were created in Matlab. At a high level this was done by comparing the number of arrivals to the number of departures for each AP, but more details will be discussed.

```
0 3 5 3
1 7 10 12
4 8 6 8
```

Figure 9: Sample *user_counts* matrix.

The heatmaps were generated from the matrix *user_counts* with a height equal to the number of APs on the floor and a width equal to the number of *minute_edges*. The number of *minute_edges* was calculated by

$$(time\_length/\mathbf{Ts}) + 1$$

where the *time_length* defines how many hours of data the temporal heatmap is to display. One extra edge must be added based on how the binning algorithm works to handle any users present before the first time step. Each value within the *user_counts* matrix represents the total number of users connected to a particular AP at a given time of day. A sample of this can be seen in Figure 9. This example shows the number of users for three APs (represented by rows) over the course of four different times (represented by columns). So at **Ts** = 2 (using 0 to represent the first index), the first AP has 5 current users, but when **Ts** increases to 3, the number of users connected to this AP drops to 3.

The change of users connected to an AP overtime can be calculated by subtracting the number of departures from the number of arrivals over the course of **Ts** for the AP. These values are calculated before the creation of the *user_counts* matrix and stored in a *traffic* array for each AP. The *traffic* array for the first AP in the example would like this:

```
[3 2 -2]
```

Note that the length of the *traffic* array is shorter by one index, because it corresponds to events occurring between values in a a row of the *user_counts* matrix and can have negative values representing more departures from this AP than arrivals.

The *user_counts* matrix is then used in conjunction with the AP overlay function described in section 4.1 to produce a graphical representation of the distribution of network users. Here the size of the circle representing AP locations is directly correlated to the number of users currently connected to the AP. The number of current users connected to an AP can optionally be displayed

when greater than a configurable threshold to provide better context to what the size represents. In Figure 8 this threshold is set to display the number of current users when above 60. Also the total number of users on the displayed floor is shown in the top left corner underneath the date and time.

4.4. Probabilistic Model

There are three types of events that need to be captured and modeled in order to represent the connection logs. Arrivals, Departures, and Transitions. Arrivals are defined as being logged arrivals with no associated departure, and vice versa for departures. Transitions are defined by a logged departure simultaneously followed by an arrival at the same timestep for the same user. In order to generate these statistics, the data logs were divided into device bins, so that each series represented the path of a single device to be traced.

To understand the patterns in which devices were arriving, we first ran a small analysis of the arrival data. For each AP the delay between each subsequent Arrival was recorded. From this list, we binned the total number of delays into bins with one second increments. Plotting the sum of these bins for every AP, shown in Figure 10, shows us data that roughly follows an exponential distribution with a dip at the 0 region. Finding the best fit of an exponential distribution appears to work well for modeling Arrivals, as tested on APs 1 and 6 in Figure 11, as these arrival delays represent a continuous distribution. Saving the lambda parameters from these exponential distributions gives us a set of exponential distributions that can be used to create random delays when simulating AP Arrivals.

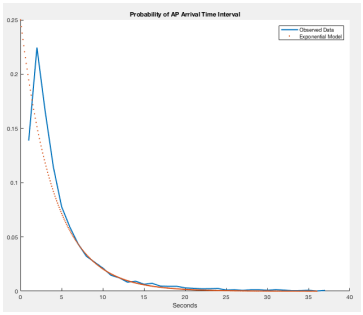
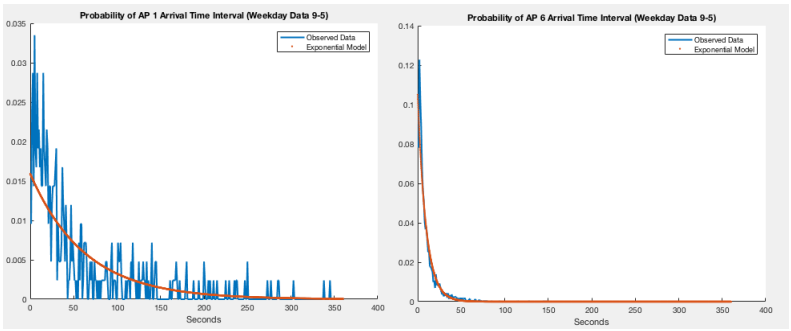


Figure 10: Arrival delay totals for all APs.



(a) (b)

Figure 11: Normalized arrival delay and model fit for AP1 and AP6.

An initial design for capturing the rate at which AP Transitions should occur was to record counts and waiting times for every single Transition and generate a matrix of average seconds-to-transition for each AP pairing. Testing this method on a small set of data yielded a scaled probability matrix that showed approximately appropriate local patterns. For a better picture of the generalization of this approach, we applied our algorithm to a larger set of data, about 300

thousand data points from a single weekday. This time, the matrix had appeared to converge to a small set of high probabilities with far fewer local trends visible, as seen in Figure 12. After seeing these updated results, and upon looking at some numbers by hand, it became evident that there were outliers affecting the results of the data and our previous assumptions were incorrect. For Transitions that had fewer data points, it became much more likely that short Transition times would skew the reported Transition per second frequency. For example, a very frequently transitioned route from AP 2 to AP 45 had 405 devices Transitions with a total wait time of 92550 seconds , giving us a Transition probability of 0.438% per second. However, the route from AP 2 to AP 105 only had 7 device Transitions, but a total wait time of 156 seconds, giving a much higher Transition probability 4.49% per second.

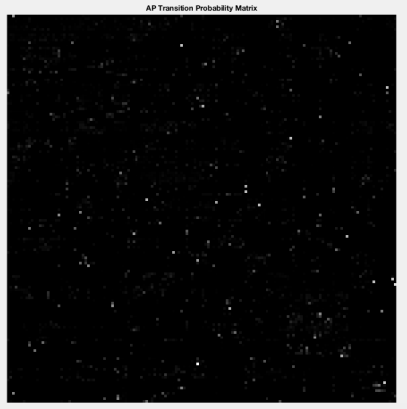


Figure 12: Poorly structured seconds-to-transition probabilities matrix.

Though there are ways to mathematically account for these outliers, we also wished to analyze both of these trends independently. This indicated the need to separate the concepts of location transition probability and transition time. With this in mind, we updated our script to maintain a separate transition count matrix, as well as tracking the mean and standard deviation of time-until-transition for every AP pairing. Time-until-transition statistics allow us more finely estimate how long any device with be at a given AP until it moves to another. In order to avoid the added space complexity of storing this entire stream of numbers, a recurrence method for calculating a running mean and variance was used, as described in Donald Knuths The Art of Computer Programming[8]. The final result is a matrix that represents the likelihood of transferring to any given AP starting from any other, and corresponding parameters for the mean and variance of the time spent at the prior AP before transitioning. The transition probability matrix generated by this simpler method contains a number of local features that indicates it is likely capturing meaningful data, as seen in Figure13. Departures are handled as a special case of Transitions where a device transfers from any given AP to the Departure AP based on Departure data.

Looking at departure counts and frequencies, the vast majority of departures are accounted for by APs on the first floor, with higher concentrations around entrances to the Engineering Building. A few notable exceptions include APs on the second and third floor that are near the transition from the Engineering Building to Anthony, which our data does not cover. There are a very small number of APs not near any building entrances with high departure rates, and this

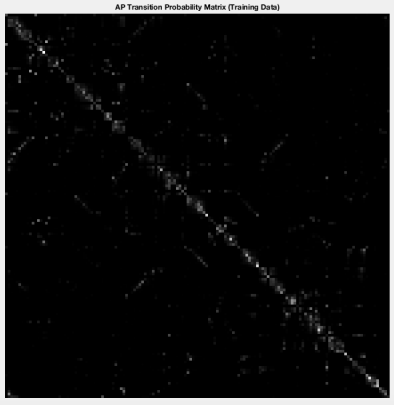


Figure 13: New time invariant transition probability matrix (time-to-transition information is stored in separate matrices)

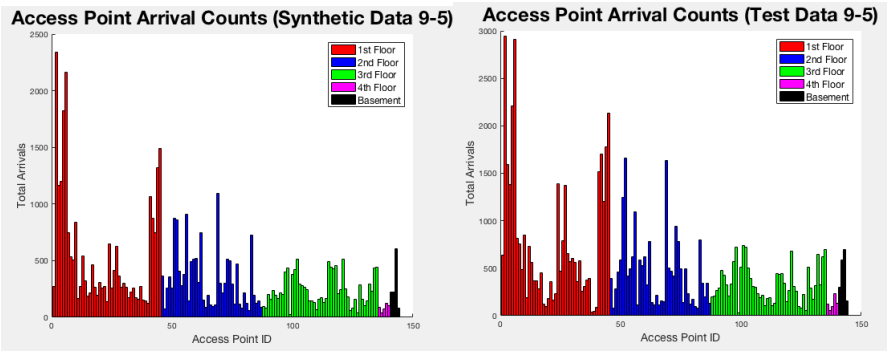
may be indicative of poor service from these APs or locations in general. This is anecdotally supported by one authors experience working in a basement lab, where there are a high number of departures that cannot easily be explained by leaving the building. In some cases, Departure counts were so low in some APs, that there was little to no chance of a departure actually occurring, leading to unsustainable device population growth in simulations. This was accounted for by adding a small, constant chance of a random disconnect occurring, based on an empirical rate of about 0.7 Departures per second in the data used for training.

4.5. Synthetic Data Experiment

Using these models, we wrote a script to generate a population of devices and simulate how the mass of the population would move around and the data that would be generated. A simulation clock ran in one second increments, updating all appropriate variables at every tick. The array of exponential distribution lambda parameters was used to maintain a queue of arrivals. When initializing, and every time there was an arrival at a given AP, a random arrival delay was generated using that APs respective arrival distribution. For devices that were in the population, a next destination AP and transition delay were generated using the transition probability matrix and time-to-transition data respectively. Arrivals feed the population size, while Departures reduce it. This simulation was run for 24 hours of simulation time and the generated logs were processed for comparison against both the training data that generated it and an unseen test day. The plot for arrival counts were generated, as well as the transition probability matrices for all cases. Noting that sparse probability vectors have a higher likelihood of orthogonality if they are randomly compared, we decided on computing the cosine similarity between matching rows of the probability matrices.

Synthetic Data Cosine Similarity		
	Mean	Standard Deviation
Training Data	0.8495	0.2418
Test Data	0.8136	0.2301

Table 1: Synthetic Data Transition Model Comparison



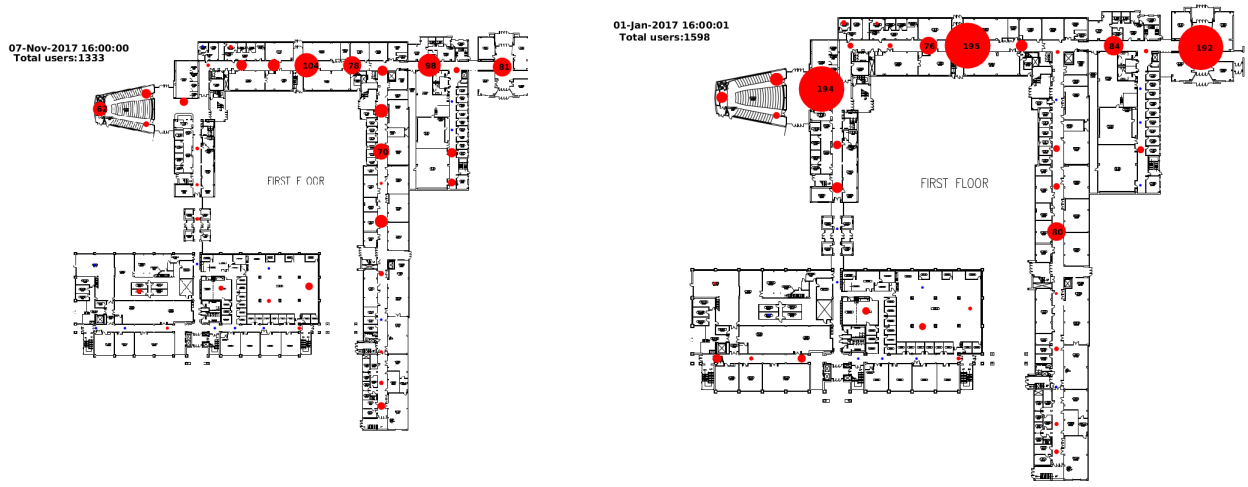
(a) (b)

Figure 14: Arrival counts from 9am to 5pm for synthetic and testing data.

The plots for arrival counts in Figure 14 show fairly similar visual trends across the different APs when compared to unseen data. This would indicate both that our exponential probability arrival model is capturing the underlying distribution of AP arrivals and that it generalizes well across the course of a single weekday. When comparing synthetic data transition probabilities against the training data, the first 10 AP transition probabilities have an average cosine similarity of 0.9665. These 10 APs account for most of the initial traffic that exists in the building starting on the first floor and many of the common transitions and departures. This is evidence that our model is generating synthetic data that accurately represents the distribution observed in our training data, and much of the erroneous data comes from APs and transitions with smaller amounts of training data and more uncertainty in reconstruction. The results of these comparisons across the training and testing data for an entire 9-5 period are shown in Table 1

Another method used to verify our synthetic data was to use it to generate the same population density heatmaps that were discussed in section 4.3. Results from this can be seen in Figure 15(b). For clarity, Figure 15(a) shows a heatmap generated from real data corresponding to the same day of the week as the synthetic data represents. It can be seen that while the synthetic data is close in many respects, it overly crowds certain APs. We have a few theories on to way this is occurring. First, Figure 15(a) is displaying a particular time of a sample day, which may have had a different traffic pattern than the dataset that our probabilistic model was trained on. Second, our probabilistic model heavily relies on making decisions based on the number of arrivals and departures from APs. This means APs that have many of short connections to them appear busier than APs that have fewer persistent connections. This results in APs near popular entrances and exits of the building, which have many users passing through making short connections, but few actually staying in proximity to the AP, appear to have a larger user base in the synthetic data

over which actually appears in the real data. This effect is further magnified by the fact that the probabilistic model assumes that there are no users initially in the building, when in the real data there are a considerable number of devices present in the building past mid-night at any given day. It therefore floods the common entrances to the building with new arrivals to increase the total number of users to levels seen in the real data. Lastly, the displayed results are just snapshots from the produced heatmaps and the synthetic data more closely captures the real distribution of users at less heavily congested times during the day.



(a) (b)

Figure 15: Population density heatmaps showing they population distrubtion throughout the MSU engineering building for (a) real logged data and (b) synthetically generated data created from the probabilistic model.

While this model exhibits reasonable performance on generalized statistics of a population across a given day, it is lacking the detail of tracking how individual users and devices travel through the network. A potential future extension to address this would be to have probabilistic transition model augmented to include the previous and current AP as two prior dimensions to make the prediction of the next AP a device will connect to. This would much more accurately simulate the behavior of a single device, but would involve much more overhead in model generation. Give the number of unique APs in the building, a three dimensional transition model with prior knowledge of two APs would require over 3 million parameters as opposed to the just over 20 thousand in the current model. It is questionable whether there is enough data in the form of 3-grams to fully populate this model, and this may exacerbate some of the issues that may be associated with limited data previously mentioned. Hidden Markov Models may be a more viable candidate solution for future probabilistic transition models to track individual paths.

5. Conclusion

In this paper, we have discussed how our work provides statistical information about how a campus population, specifically within the engineering building on Michigan State University's campus, can be mapped by analyzing network usage logs. Spatial-temporal characteristics of the population can be gleaned from the network traffic to procure a heatmap representing the distribution of users across the active access points. It was shown that mobility patterns of individuals could be extracted from the logged data to provide accurate position tracking while carrying at least one device connected to the network. A novel predictive model was developed which yields the probability of arrivals at any given access point; with an approximate probability of transition between access points. This model can be used to make accurate assumptions about the distribution of users across the network for any day of the week. Future work is directed towards utilizing the information extracted from the raw data to design feasible solutions promoting development in infrastructure and other user driven diagnostic operations. We believe that this work can be used both as a foundation for many future applications and as a tool that caters to: creating effective evacuation strategies, providing crowd monitoring at events, identifying potential hot spots for recruiters and advertisements within the building, assessing the effectiveness of the current access point deployment, and assist in determining optimal locations for the installation of new access points.

6. Acknowledgement

We would like o express our gratitude to Michigan State University engineering building system administrator Adam McDougall and Computer Science and Engineering department administrator Kelly Climer for their assistance in securing the data. We would like to thank Professor Guoliang Xing for his feedback which contributed towards the experimental design.

References

[1] F. M. Naini, O. Dousse, P. Thiran, and M. Vetterli. Population size estimation using a few individuals as agents. pages 2499–2503, July 2011.

[2] Goudeseune S. van Bossche F. van de Weghe N. Versichele M., Neutens T. Mobile mapping of sporting event spectators using bluetooth sensors: Tour of flanders 2011. *Sensors (Basel, Switzerland)*, 12(10), 2012.

[3] Gatej R Lehmann S Sapiezynski P, Stopczynski A. Tracking human mobility using wifi signals. *PLoS ONE*, 10(7), 2015.

[4] David Kotz and Kobby Essien. Analysis of a Campus-wide Wireless Network. *Wirel. Netw.*, 11(1-2):115–133, January 2005.

[5] Diane Tang and Mary Baker. Analysis of a local-area wireless network. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking*, MobiCom '00, pages 1–10. ACM, 2000.

[6] Atul Adya, Paramvir Bahl, and Lili Qiu. Characterizing alert and browse services of mobile clients. In *Proceedings of the General Track of the Annual Conference on USENIX Annual Technical Conference*, ATEC '02, pages 343–356, Berkeley, CA, USA, 2002. USENIX Association.

[7] Anand Balachandran, Geoffrey M. Voelker, Paramvir Bahl, and P. Venkat Rangan. Characterizing user behavior and network performance in a public wireless lan. In *Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '02, pages 195–205, New York, NY, USA, 2002. ACM.

[8] Donald E. Knuth. *The Art of Computer Programming, Volume 1 (3rd Ed.): Fundamental Algorithms*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1997.

23