

Programmatically Predicting [Bus] Occupancy

Undergraduate Independent Research Report

ECE 492: Special Problems – Spring 2014

Daniel Bordak, Erin Corrado, Revan Sopher, and Ashley Weaver

Abstract—In this study we address the problem of programmatically predicting bus occupancy. We approach this in two ways: by using publicly available GPS positions to calculate stop time, and by using promiscuous Wi-Fi to count devices. While the stop time approach proved unfruitful due to the significant time requirement in establishing a benchmark by which to interpret the data, the Wi-Fi approach yielded an accurate model once we compensated for noise. We conclude by suggesting modifications to the experiment, and the application of this methodology to the determination of occupancy of other locales.

1 MOTIVATION

Due to the periodic nature of class schedules, students at Rutgers tend to move between campuses during certain known time intervals during the day. This leads to predictable migratory patterns. However, the bus schedule does a poor job of taking student schedules into account — for example, on Livingston campus, buses on the route to College Avenue can often be found lined up at the stops consecutively, while the buses on the route to Busch are much rarer. To make matters worse, the Busch buses are usually packed to capacity between class periods. The aim of this project is to develop a system for better predicting the demand for buses, such that an empirical argument for the optimization of the bus schedule might be made.

2 APPROACH

In order to facilitate simultaneous contributions from all four team members, we approached the problem in two ways:

- by directly collecting data about number of riders vs. stop time, and
- by measuring number of devices on the bus.

Two students were assigned to work on each approach. Ultimately, we planned to combine results into a single prediction method.

2.1 Stop Length Inference

In this approach, we planned to establish a correlation between the length of time a bus spends at a stop and the number of passengers currently aboard. Our initial theory was that a bus nearing capacity would take longer at stops due to the larger number of riders entering and exiting. Every Rutgers bus periodically reports its position, velocity, and destination to an online service — nextbus.com’s Public API. This service is free to use, and its information is publicly available [1]. As such, using it would not require any hardware or observers on the bus, allowing for cheap, large-scale execution. We first tried to measure the strength of this correlation; if we could prove a strong positive correlation, we could use the bus travel data to measure length of stops, and therefore predict occupancy. By running the analysis on a log of travel data, we could predict occupancy across a significantly larger domain and for theoretically no cost.

In order to establish a correlation between stop length and bus occupancy, we followed a bus on its circuit, recording data about its stops. Specifically, we recorded the number of people boarding or departing the bus, the length of the stop, and what time it was.

```

13:55:42.381750 273707991us tsft short preamble
    11.0 Mb/s 2462 MHz 11b -54dB signal antenna 1
    RA:XX:XX:XX:XX:XX (oui Unknown)
    Acknowledgment
13:55:42.381761 273708457us tsft 11.0 Mb/s
    2462 MHz 11b -52dB signal antenna 1
    RA:XX:XX:XX:XX:XX (oui Unknown)
    Clear-To-Send
13:55:42.384772 273711771us tsft 11.0 Mb/s
    2462 MHz 11b -52dB signal antenna 1
    RA:XX:XX:XX:XX:XX (oui Unknown)
    TA:XX:XX:XX:XX:XX (oui Unknown)
    Request-To-Send
13:55:42.385499 273712406us tsft 11.0 Mb/s
    2462 MHz 11b -60dB signal antenna 1
    BSSID:XX:XX:XX:XX:XX:XX (oui Unknown)
    DA:Broadcast SA:XX:XX:XX:XX:XX:XX
    (oui Unknown) Beacon (RUWireless_Secure)
    [6.0 9.0 11.0* 12.0 18.0 24.0 36.0 48.0 Mbit]
    ESS CH: 11, PRIVACY

```

Listing 1: Example tcpdump output. This log-file has been modified from its original version. It has been formatted to fit this page; the original output had one line per packet. Additionally, MAC addresses have been censored to protect the privacy of those we observed.

2.2 Promiscuous WiFi

In this approach, we used the inherent correlation between the number of smartphones on the bus and the number of passengers; it is reasonable to assume that more devices means more passengers, since the devices travel with the passengers. Nonetheless, we would try to verify this correlation by comparing collected network information with logged bus occupancy.

Data collection was performed with a laptop running Linux, equipped with a typical 802.11n wireless card. The computer ran the wireless card in monitor mode. In normal operation, a wireless card will receive every packet in the vicinity, but discard those that are not specifically addressed to it; in monitor mode all received packets are transmitted to the operating system regardless of intended recipient. With this setup, we ran the following command:

```
tcpdump -e
```

This command logs the headers of every packet that the wireless card sees, thus uniquely iden-

TABLE 1: Bus stop key

Label	Name of Stop
1	Buell Apartments
2	Werblin Recreation Center
3	Rutgers Student Center
4	Scott Hall
5	Student Activity Center
6	Visitor's Center
7	Stadium
8	Werblin Recreation Center
9	Hill Center
10	Allison Road Classrooms Building
11	Library of Science & Medicine
12	Busch Suites
13	Busch Campus Center

tifying every active device in the area [2]. A sample output of tcpdump is shown in Listing 1. The strength of signals is also recorded, so we can choose to ignore packets below a certain decibel strength. We initially planned to add a GPS device to the computer, so as to automate the logging of position along with the network data, but we found it easier to manually record approximate physical locations; we rode a bus along its route, logging packets and stop times.

3 BUS ROUTE

The Rutgers bus system encompasses several routes between its campuses. For the purpose of our study, we needed only to monitor one; conducting all collections on the same path is essential to producing comparable data. We decided to monitor the *A* route, one of the two routes connecting the Busch and College Avenue campuses [3]. The other of the two routes, the *H* route, was not selected due to its somewhat lower number of buses; it was easier for us to find *A* buses whenever we needed to collect data. The particular advantages of this route included its variety of locations: Busch and College Avenue offer a good contrast in terms of population, and thus device, density. Additionally, the regularity of short buses improved our ability to collect data — some Rutgers buses are of double length, making them harder for a single person to identify the number of passengers and the quantity of traffic on and off the bus.

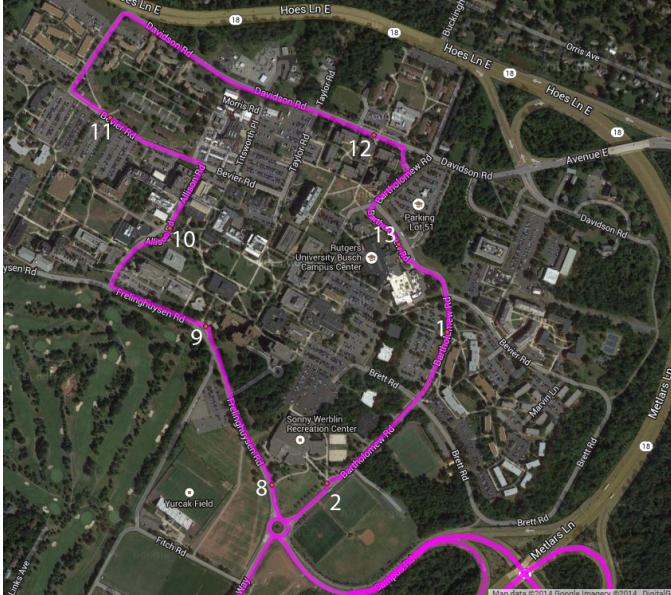


Fig. 1: A Route: Busch Classrooms. The bus route consists of two campuses, connected via bridge.

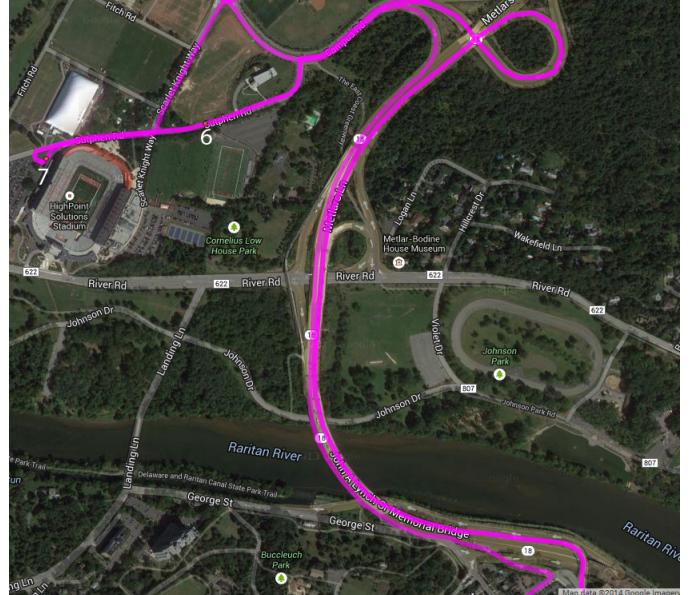


Fig. 3: A Route: Busch Stadium

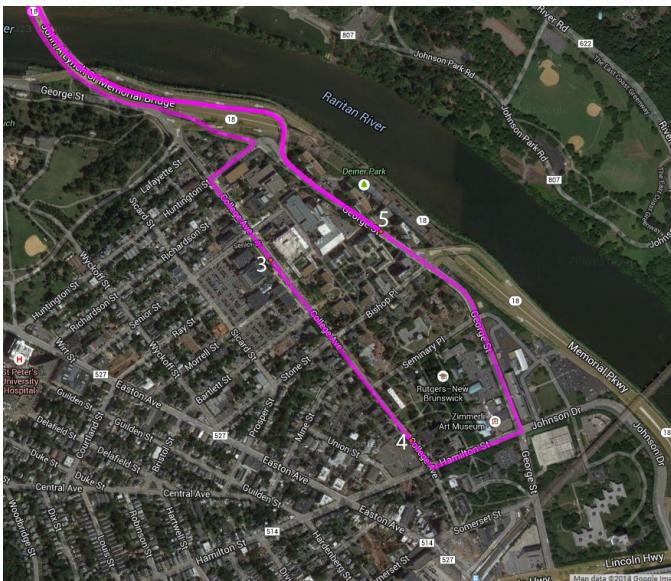


Fig. 2: A Route: College Ave

Figures 1-3 are maps of the stops serviced by the *A* bus route. We began riding the route on Busch Campus (Fig. 1), at the Busch Campus Center. It was too early to begin data collection there, however, so the first stop recognized in our dataset is the Buell Apartment stop, also on Busch.

4 GRAPH PRODUCTION

4.1 Preprocessing

The raw output from running `tcpdump` contains many irrelevant details — fields like the type of packet, or the antenna used. As such, our first step in graph production was to preprocess this data, preserving only the information relevant to our goal. The main data of interest was the list of MAC addresses seen in a given packet. Each of these addresses can take one of four roles: Source, Destination, Transmitter, or Receiver. Every packet will have some, but not all, of these types of addresses. Since each line of `tcpdump` raw output represents a packet, our log files are first parsed line by line, each line individually stored as a Python Dict — otherwise known as a Hash Map. These Dicts hold the four potential addresses, the packet's strength, and what time it was sent at (in microseconds). All the other information was disregarded; ignoring information during preprocessing is the simplest way of removing it from the data.

A few special considerations needed to be made during this step:

- Packets may not have a strength field. To make data handling easier later on, these are assigned a strength of zero. This is easier since it can be filtered out later

```
{
  "num_macs": 4414,
  "routers": [
    1, 3, 6, 7, 11, 14, 22, 24, 31, 43, 57, 64,
    73, 94, 103, 146, 152, 153, 189, 221, 228,
    252, 266, 284, (...),
  ],
  "packets": [
    { "str": -54, "adds": { "R": 0 }, "time": 0 },
    { "str": -52, "adds": { "R": 1 }, "time": 466 },
    { "str": -52, "adds": { "R": 2, "T": 1 }, "time": 3780 },
    { "str": -63, "adds": { "R": 1 }, "time": 11200 },
  (...),
  ],
  "initial_time": "13:55:42.381750"
}
```

Listing 2: Example json output, abbreviated: we construct this structured representation of the packet capture for plotting.

simply by checking that the strength field is nonzero; a zero strength appearing naturally is impossible.

- Some entries contain only the address of a router, with no particular destination in mind. These are broadcast packets, sent to notify devices of the router's existence. Since we are monitoring personal devices, not routers, these entries are ignored, and the address identified as a router is flagged for future filtering.
- Since MAC addresses are usually unique device identifiers, they could potentially be used to identify device owners. This kind of risk could subject our study to human subject regulations. In order to avoid this risk, addresses are saved into the Dicts as index numbers (starting from zero) rather than the actual address. This way, we can still differentiate between multiple devices without having to worry about our data being used for morally questionable purposes.

At the end of the parsing process, the Dicts are dumped into a JSON output file. JSON is an open standard data interchange format [4]. Using JSON made utilizing the data much eas-

ier when it came to producing the graphs; most languages, Python included, have a standard library implementation for JSON, eliminating the effort spent in re-parsing the data.

4.2 Filtering

The preprocessing done above is only performed once per dataset. Therefore, any sort of filtering that might change between two graphs must be done immediately before the construction of the graph. We built four filters to be run at this stage:

- The Strength filter removes packets below a given signal strength threshold. For example, a strength filter of -50 would filter out packets of -51dB.
- The Zero-Strength filter removes packets which have a strength field of zero. These packets correspond to tcpdump output lines missing a strength field, as mentioned above in the preprocessing section.
- The Router filter removes any addresses which were flagged as routers during the parsing step. It also removes any packets which would no longer contain any addresses after the previous is performed.
- The End Time filter removes any packets past a certain time. This is desirable when you only want to look at a part of a dataset. We never produced a corresponding Start Time filter since the need never arose.

4.3 Graphing

For the construction of the graphs, a few additional libraries were needed: Numpy, Pandas, and Matplotlib. Numpy is an advanced mathematical library for Python [5], and Pandas is a library built on top of Numpy to provide high-performance data structures [6]. We used these for data analysis. Matplotlib is a 2D plotting library, which we used to generate the graphs [7]. In order to simplify the graph production process, we made individual scripts for each graph type, plus a meta-script to make a united interface. This meta-script reads a JSON file obtained from the preprocessing step, imports the data into a Pandas DataFrame (which can be thought of as a complicated 2D array), and

sends this off to be filtered, as explained in the filtering step. After the data has been filtered, it is handed off to one of the six plotting scripts:

Normal Plots

The Packet and Unique plots were our easiest plots to generate. This is because they utilized the data in a fashion similar to how it was stored. They were created by generating an array of 3-second duration bins and adding the appropriate statistic to the bins. In the case of the packet plot, this was simply the number of packets found over those three seconds; the bin is simply an integer. In the case of the unique plot, however, the bin is a mathematical set holding the MAC addresses seen over the three second period. After the entirety of the three seconds are parsed, the bin is assigned the length of the set.

Inverted Data Plots

The First-Last Grid, Packet Histogram, and Vector plot were created by analyzing our data in a different fashion. Rather than looking at each packet over time, we instead looked at the data as a list of addresses, each containing a list of times at which it appeared — an inverted data index. The first time each MAC address is encountered, it is added to a list with an initial value of the time it was seen as both its first and last time seen, and seen-count of one. Each additional time an address is seen, its last time seen value is updated to the current time, and its seen-count is incremented. In the Grid, the addresses are graphed using the first time seen as the x-value and the last time seen as the y-value. In the packet histogram, packets which fulfill the requirement of

$$t_{\text{first}} + t_{\text{max}} - t_{\text{last}} < 120$$

are selected, where t_{max} is the total observation time. The threshold of 120, or two minutes, was chosen to allow for some leeway in case a device was not detected immediately. These selected packets are then graphed with their y-values set as their seen-count. To make the graph easier to interpret, the y-axis was made to be logarithmic, and the addresses are sorted by their y-values. For the Vector plot, first

an array of zeros was initialized with length equal to the length of observation in seconds. For each packet, every value between t_{first} and t_{last} was incremented. The resulting array was plotted as a step function, with the x-axis being its indices.

Segment Plot

The segment plot is the most complicated plot we generated. First, a list of mathematical sets are generated, one for each stop. These sets are called the stop-sets. The stop-sets are populated with the MAC addresses seen during the corresponding stop. Next, a second list of mathematical sets are generated, again, one for each stop. These sets are called the intersect-sets. The intersect-sets are filled with the intersect between the stop-set of the same index value, and every other stop-set, combined by union. For instance, intersect-set 2 would contain

$$(stop_2 \cap stop_0) \cup (stop_2 \cap stop_1) \cup \\ (stop_2 \cap stop_3) \cup \dots \cup (stop_2 \cap stop_n)$$

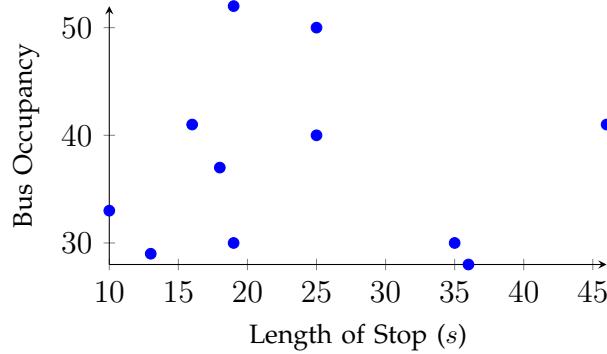
The length of each intersect-set is considered to be the number of people on the bus at the stop. A special exception is made for the intersect of stop-set 0 and stop-set n. This set represents addresses present both at the beginning of the ride and the end. No normal student would do this — this would consist only of our own devices and possibly devices present around this stop. As such, this set is instead subtracted from all other intersect sets; it is likely garbage. Afterward, we add a flat value of 2 to every stop, representing our two cellphones which we filtered out.

Some graphs also had the actual number of people on the bus, as observed, overlaid. This overlay was generated with a simple step function, using the value at $n + 1$ as the value for the line between n and $n + 1$.

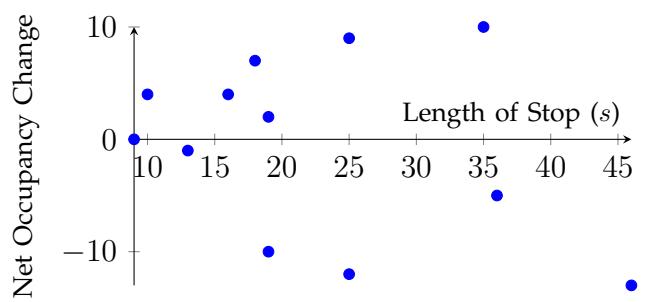
5 RESULTS

5.1 Stop Length Inference

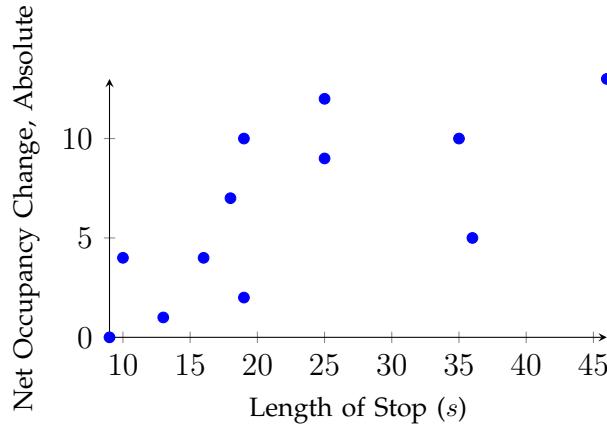
Based on the experimental procedure described in Section 2.1, we produced Figure 4. Figure 4a depicts a lack of significant correlation between stop length and occupancy. This makes sense,



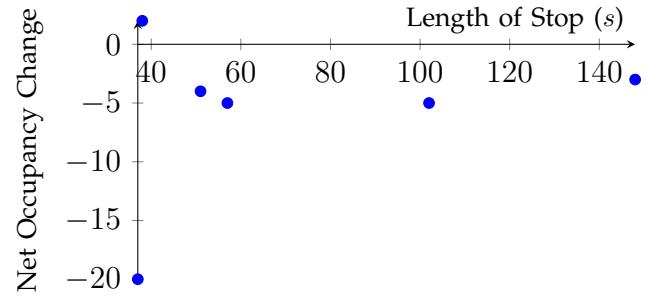
(a) Bus Occupancy as a function of stop length. Occupancy is measured after each passenger transaction (e.g. a transaction taking 10s and adding 5 passengers to an existing pool of 20 would be charted as (10,25)). Stop length has no apparent correlation to Occupancy.



(b) Net Delta Occupancy as a function of stop length. No immediate relationship visible, however negative occupancy changes seem to be exclusive to larger stop lengths.



(c) Absolute Delta Occupancy as a function of stop length. Strong correlation between magnitude of change and stop length. We can predict changes if we can show certain bus stops tend to positive or negative change.



(d) Net Delta Occupancy as a function of Stop Length at a Single Stop. This supports the idea that we can characterize individual bus stops as being prone to positive or negative change.

Fig. 4: Plots created with data from visual occupancy observation. Aside from d, which was created from the data from the *H* route single-stop observation, these plots use the data from observation of the *A* route.

because the number of sedentary or otherwise intransient riders does not cause slowdowns; rather, the relevant metric is the net occupancy change; the number of people entering or exiting the bus at any given stop. Plotting the net occupancy change as a function of stop length (Fig. 4b), however, was also inconclusive, as the longer stops could signify either positive or negative changes. From stop timings alone, we would not be able to predict the direction of change.

Plotting absolute change (Fig. 4c), however,

shows a much stronger correlation than either of the previous graphs. Absolute change is not a useful metric on its own, but if it could be proven that certain bus stops tended to have a generally positive or negative net change, we could predict the sign of the change and thus the occupancy changes over time. In other words, if more people tend to get on at a given stop than get off, a longer stop here means a large positive change to the number of passengers. Conversely, a stop which is associated mostly with disembarking, a long stop would

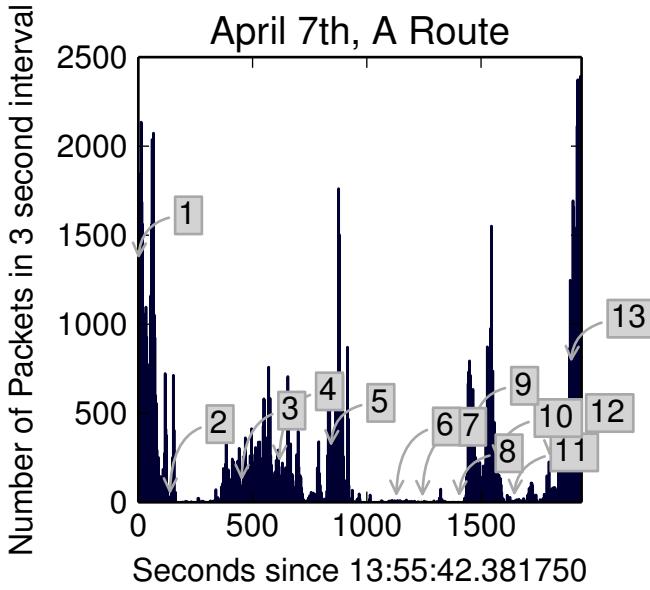


Fig. 5: Packets detected per 3-second bin: there are visible differences between stops, and frequencies are higher on busier stops.

mean a large negative change.

To test this, we waited at a single bus stop and counted the number of passengers entering and leaving each bus on a certain route, and the length of the stops (Fig. 4d). Data for this experiment was slower to collect, as the buses were spaced out. This data suggests that this particular bus stop is predominantly a destination, rather than a starting point; longer stops here, then, would imply a net decrease in number of riders.

However, the data collected was insufficient to draw these conclusions; there can be significant variation depending on time and day of the week. This data collection would need to be repeated at every bus stop, on a variety of days and times, in order to provide the required basis of prediction. This clearly increases the time requirement to well beyond the scope of this study.

Thus, while the preliminary foray suggests that this approach could be successful, it is not feasible for this study. We move on to our next approach.

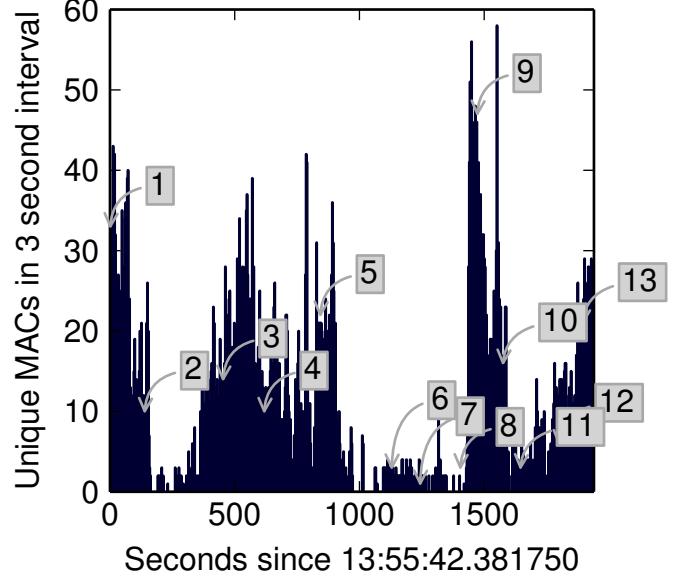


Fig. 6: Unique MAC addresses detected per 3-second bin: by plotting devices instead of packets, we get cleaner frequencies.

5.2 Promiscuous WiFi

Packet Plot

In figure 5, we plot the number of packets in three second bins. Analyzing this figure, we see that there are visible differences between stops. This roughly lines up to the expected traffic around each stop — stops like the BCC, Hill, the ARC, and the entirety of College Avenue have a large amount of traffic, while stops near the Visitor’s Center and Werblin have very little.

Unique Devices Plot

Network traffic is not a good indicator of occupancy, however, because it represents the amount of activity, rather than the number of actors; a single person transmitting a large amount of data may be interpreted as multiple people with the Packet Plot. As such, our next step was to switch out “number of packets” for “number of unique MAC addresses” (Fig. 6). The Router filter would also be used here, since we don’t want to count routers among the number of riders. The values on this plot differ noticeably between stops. However, this is not due to occupancy; there are clearly not nearly that many people on the bus. Instead, it

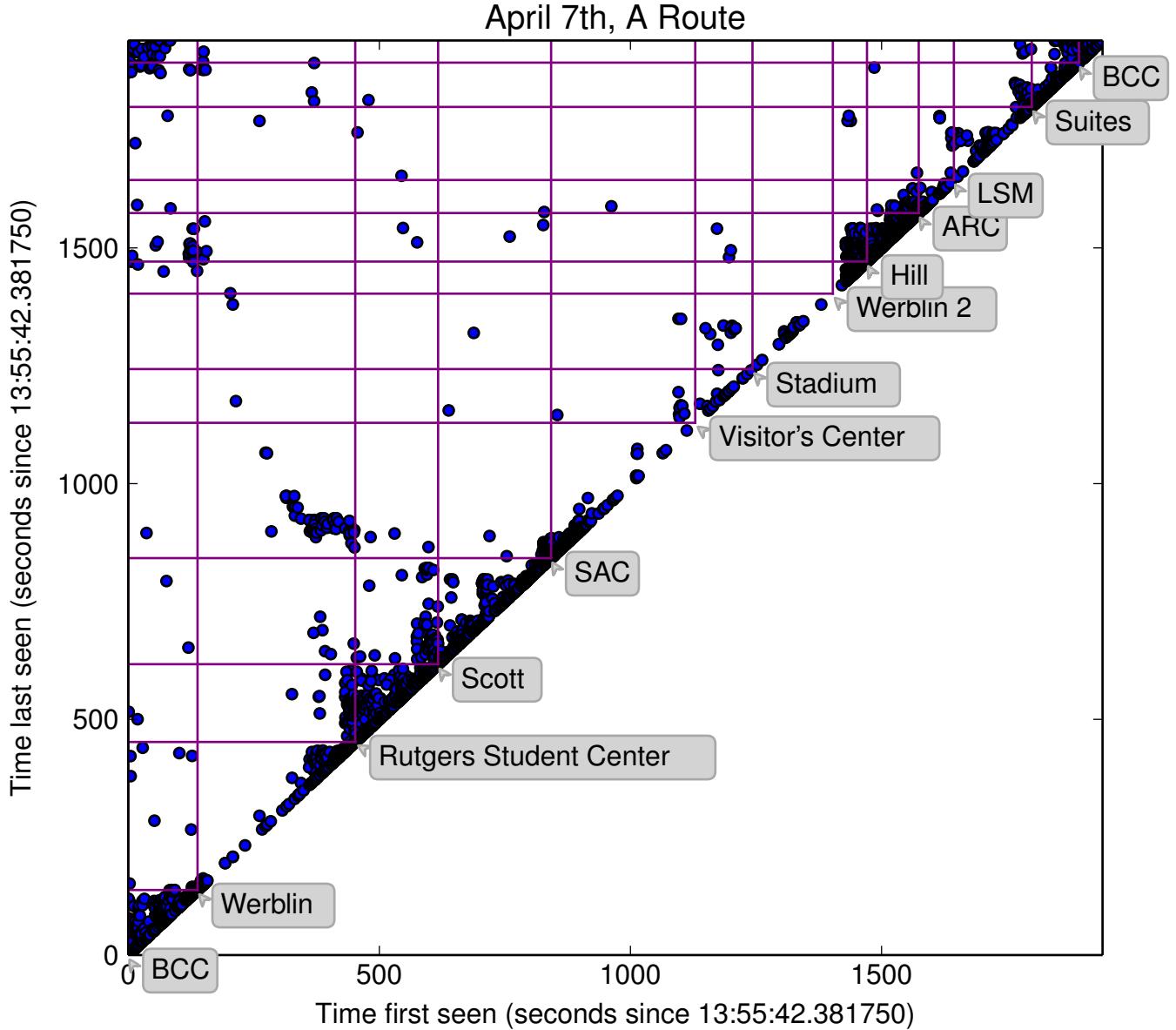


Fig. 7: First and Last times seen for each MAC address observed while on the bus route: clumps along the line $y = x$ are noise. Points further from the line represent devices present for longer durations.

seems that the shape of the plot is primarily influenced by the proximity to routers — the portions of the bus loop by athletic fields and between campuses has nearly no data.

First-Last Grid Plot

In order to establish which devices were on the bus, we plotted the time of the first and last occurrence of each device (Fig. 7). Each point here represents a unique device, where the x-coordinate is the time of the first sighting, and

the y-coordinate is the time of the last sighting. Thus, the high concentration of data points along the line $y = x$ represents the transient devices, most likely people the bus was passing by. The clump of data points in the top left corner of the plot represents devices which were seen throughout the journey. These are most likely our own devices, or stationary devices around the geographically identical first and last stops.

As a variation on this, we also plotted the

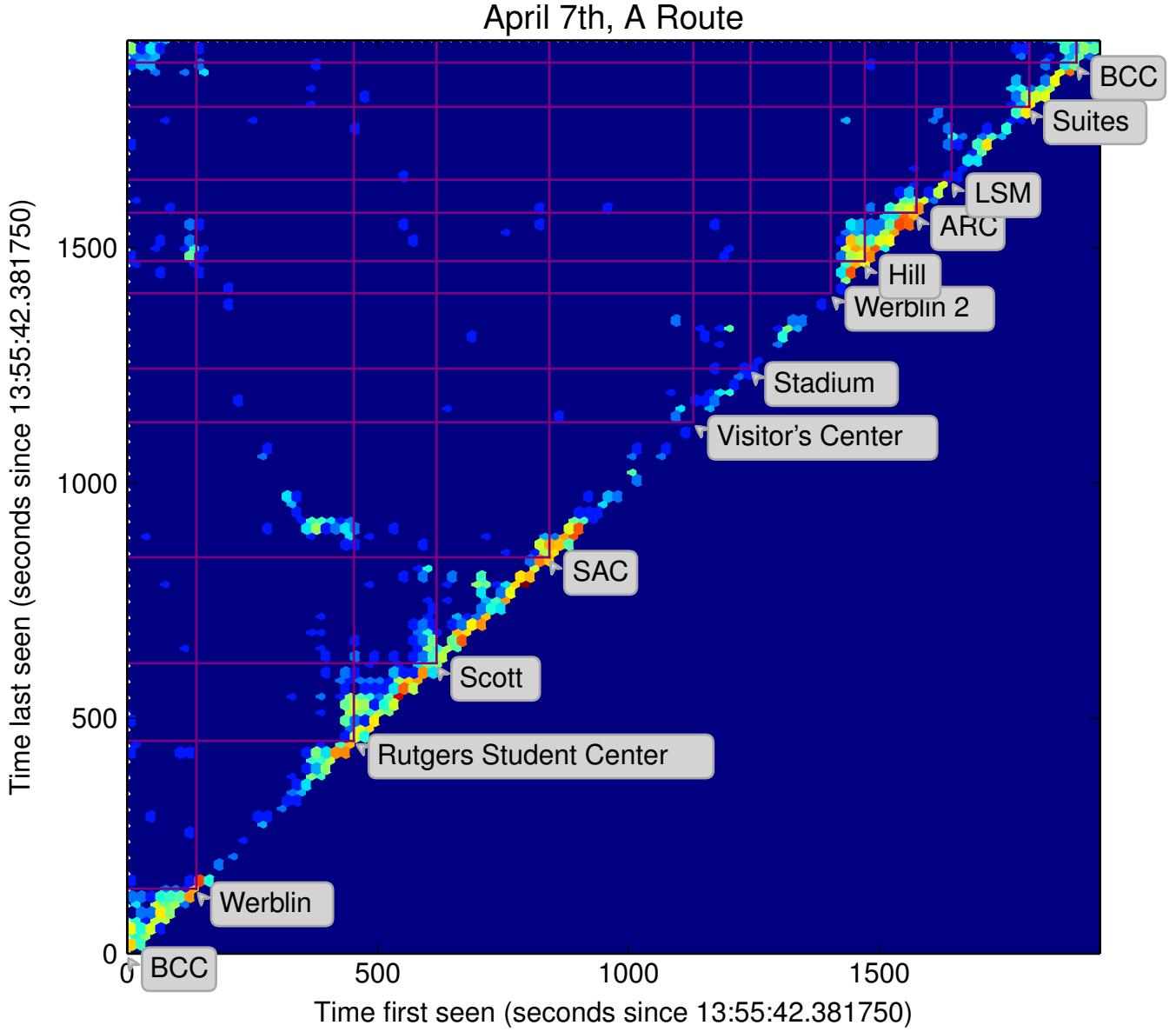


Fig. 8: First and Last times seen for each MAC address observed while on the bus route: similar to previous plot, but clumps are easier to identify.

same data as a heat map (Fig. 8), such that higher concentrations of points are depicted by more vivid colors. In the original grid plot, any number of datapoints at the same coordinates would be shown the same as a single datapoint. This method allows us to see more easily how many points are in each area of the plot (between Rutgers Student Center and SAC, for example).

Segments Plot

The segments plot was created using the procedure for segments plot described in the Graph-

ing section. The plot (Fig. 9) shows the actual number of bus passengers as we observed (the purple function) overlaid with the function that we calculated using the lengths of each intersect-set (the blue function). One can see that the calculated function is accurate for the stops after Werblin. The discrepancy between the actual and calculated number of bus passengers at Buell (label 1) may be influenced by the fact that we had started observation and recording at this location. In this method, more routers equals more stop-to-stop change.

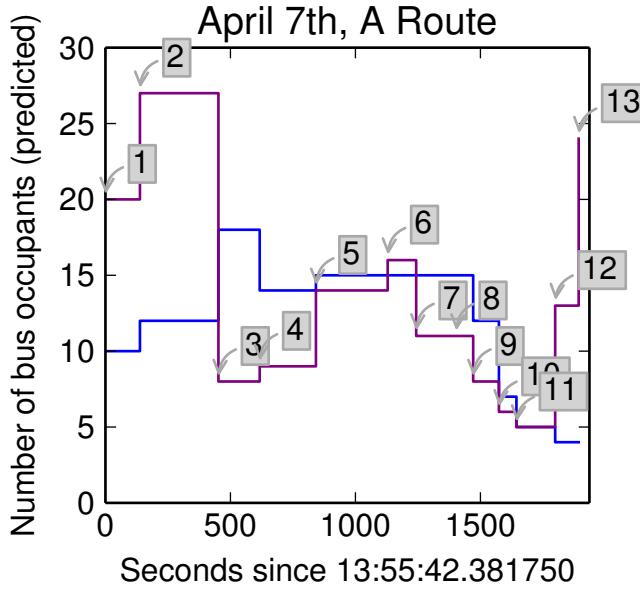


Fig. 9: Predicted number of passengers using the Segments approximation (blue) versus actual number of passengers (purple). Segments approximation calculated from the intersection of sets of observed devices at each stop.

This fact means the calculated function will appear more accurate in areas with more devices. There are more routers around the College Avenue areas, so the calculated function appears more accurate at these stops in comparison to stops like Werblin, which have fewer surrounding devices and routers. Due to the limitations of this method, we progressed to creating the vector plot.

Presence Vector Plot

In an attempt to minimize noise from transient devices and to compensate for the lack of traffic in uninhabited areas, we constructed a plot (Fig. 10) similar to the previous. For every address i we encountered in the collection, we create a vector $X^{(i)} = [X_1 \cdots X_T]$ such that

$$X_t^{(i)} = \begin{cases} 1 & t_{\text{firstSeen}}^{(i)} \leq t \leq t_{\text{lastSeen}}^{(i)} \\ 0 & \text{otherwise} \end{cases}$$

with the sum of these vectors represented by X , such that

$$X = \sum_i X^{(i)}$$

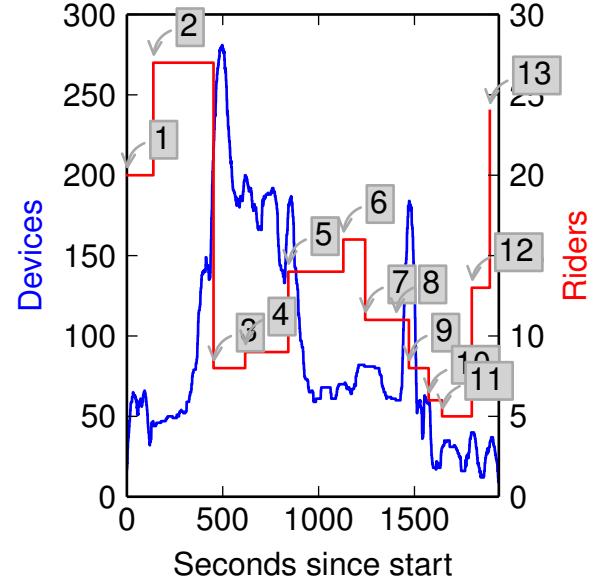


Fig. 10: Predicted number of passengers using the Vector approximation (blue) versus actual number of passengers (purple). Verctor approximation calculated by summing number of devices between first and last sightings.

The value at any point of X indicates how many devices we can infer are presently on the bus. Overlaid on this plot is the step function corresponding to the observed number of riders. It is scaled differently, as there is approximately a three time difference between the devices seen and the people riding.

We see that there appears to be an excellent correlation here, with the exception of the 400-700 second segment. This segment corresponds to the entrance to the College Avenue campus, a highly concentrated area that appears to have contributed considerable noise. We do see, however, that the two changes in occupancy in this region seem to correspond to peaks in the traffic — this means that the observed traffic is probably consistently higher than the actual over that area, which we can account for in prediction.

6 ALTERNATIVE APPLICATION

The methods we developed for the purpose of measuring bus occupancy can also be used for similar measurements in other settings. To

prove this, we measured occupancy in a classroom (Fig. 11). Just as with the bus data, the points around $y = x$ consist of transient devices. The points in the clump to the top right is probably due to a buildup of students outside the lecture hall awaiting the following class. The points in the clump to the top left represent the devices that were present for the duration of the capture. Other scattered points may be noise, but could also be sleeping devices which only report themselves occasionally.

There were exactly 106 students in the lecture hall during this collection. While judging exact cutoffs for the measurement area is of a complexity outside the scope of this project, a rough estimate of the clump in the top left gives about 120 addresses. In order to judge how reliable this count is, we created a new type of plot, counting the packets including each of these long-present devices (Fig. 12). This plot shows that most devices measured transmitted or received a reasonable quantity of packets — only a few are particularly low on packet count. This suggests that this approach does generate an accurate count of the lecture hall occupancy.

We can also observe that some devices are involved in a significantly larger quantity of packets. This graph is generated with router filtering on, meaning it is unlikely that these are routers.

7 BUS OCCUPANCY: DISCUSSION

The Stop Length Inference approach supported correlations between the number of passengers on a bus and the length of time the bus stops. Unfortunately, this method is impractical to utilize on a mass scale since collecting enough data to make conclusions about each stop would take an exorbitant amount of time and resources. This approach may be possible with more people and time, but the time constraints make this method beyond the scope of this project.

The Promiscuous WiFi approach is much more efficient but has the problems of noise filtering and determining whether a device is actually present on the bus. The methods that we used to analyze the WiFi data each had

certain advantages and disadvantages. The best way to use the data is to count the number of unique MAC addresses; the number of packets is not a reliable metric to determine the number of bus riders. The First and Last time seen grid graphs reduce noise, but they do not represent the data in a fashion that is particularly easy to comprehend. The Vector and Segment graphs are the most accurate in determining the number of bus occupants. Both of these two graphs have specific sections where one is more accurate than the other. Overall, the Vector and Segment approximations are the best way to analyze the data from the Promiscuous WiFi; therefore, these are the best approaches for predicting bus occupancy.

8 FUTURE WORK

8.1 Continuation of Benchmark

As mentioned earlier, the Stop Length Inference approach seemed promising, when looking at the simple quantity of traffic at a stop. In order to utilize this, however, we would need to determine the tendency of each particular stop. This means we would need to extensively analyze each stop individually over long periods of time. Our preliminary research indicates this approach offers great opportunity for future research. This research would, unfortunately, require far too much time and resources for four college students.

8.2 Alternative Approach

The largest problem with the interception of packets is the distance between the bus and access points. Many, possibly most, of our detected packets were from passersby due to their relative proximity to routers; without a router to transfer a packet to, there cannot be a packet to detect.

One possible solution to this issue would be the installation of WiFi access points on Rutgers buses, using the standard Campus WiFi Access Point SSID. Using the same SSID would cause student smartphones to automatically connect, without requiring any effort on the part of the students. This would limit detection to Rutgers students, as well; no others would be able to

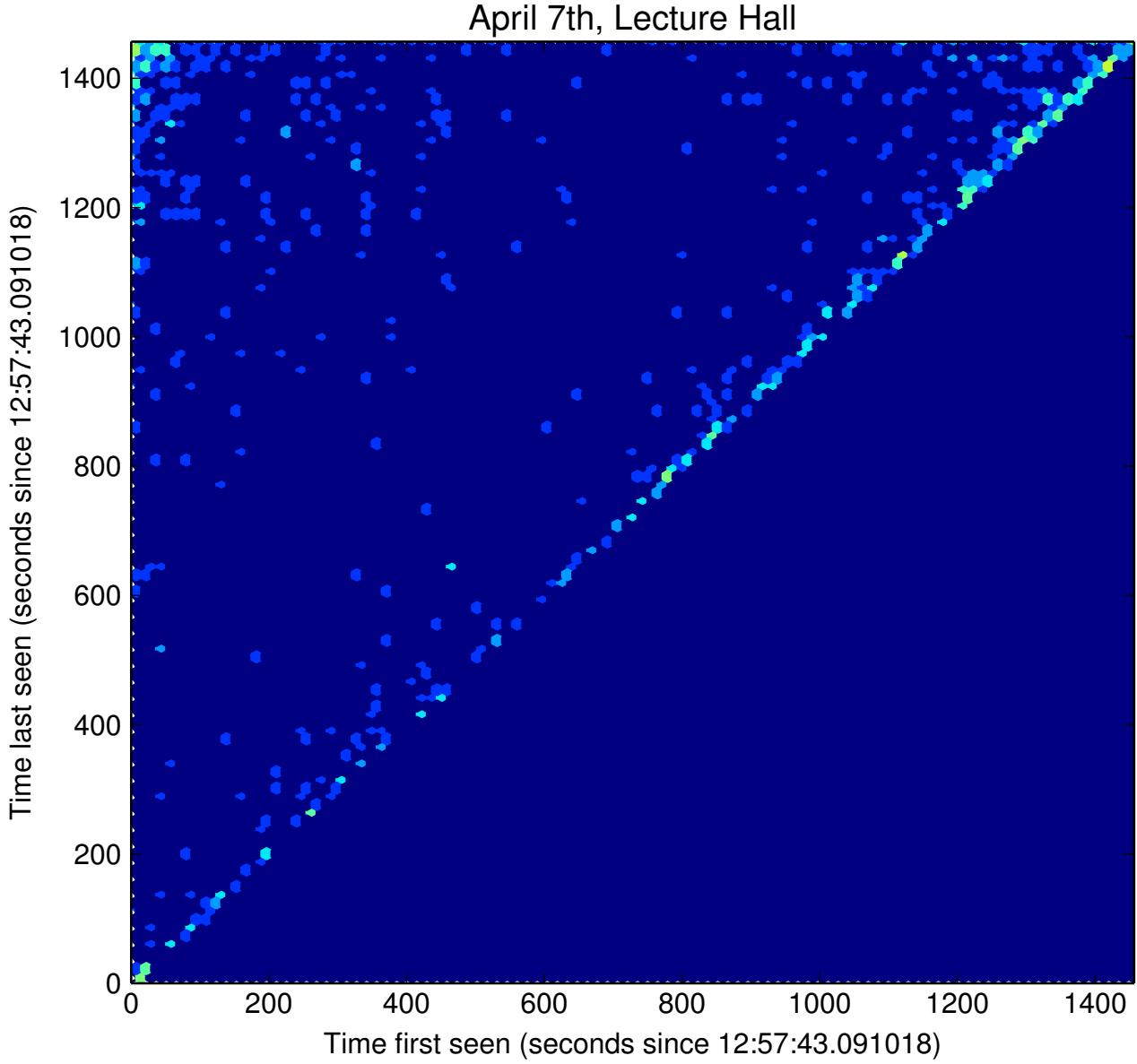


Fig. 11: First and Last times seen for each MAC addresses in a lecture hall: transient noise along the diagonal, and a high concentration that were visible throughout the lecture.

connect to this AP. This might, in fact, reduce accuracy on the College Avenue campus, but the overall impact would most likely be an increase in accuracy. It is unlikely that Rutgers or First Transit would install these devices strictly for the purpose of detection, since our methods are still unproven. Adding functionality in the way of internet, a router's expected purpose, could likely be done only through mobile data due to the bus routes going far outside the range of campuses. Mobile data is very expensive, however, making this unlikely.

Any access points installed unofficially would be an explicit violation of the terms of use of the Rutgers networks, and, as such, even more infeasible.

Putting this burden of observation onto other students rather than officials would be much more manageable in terms of dealing with bureaucracy. To do this, we could replace the laptop in our setup with a smaller, dedicated unit to be installed in the buses, to collect data for longer periods of time. Running several of these units would allow us to collect enough

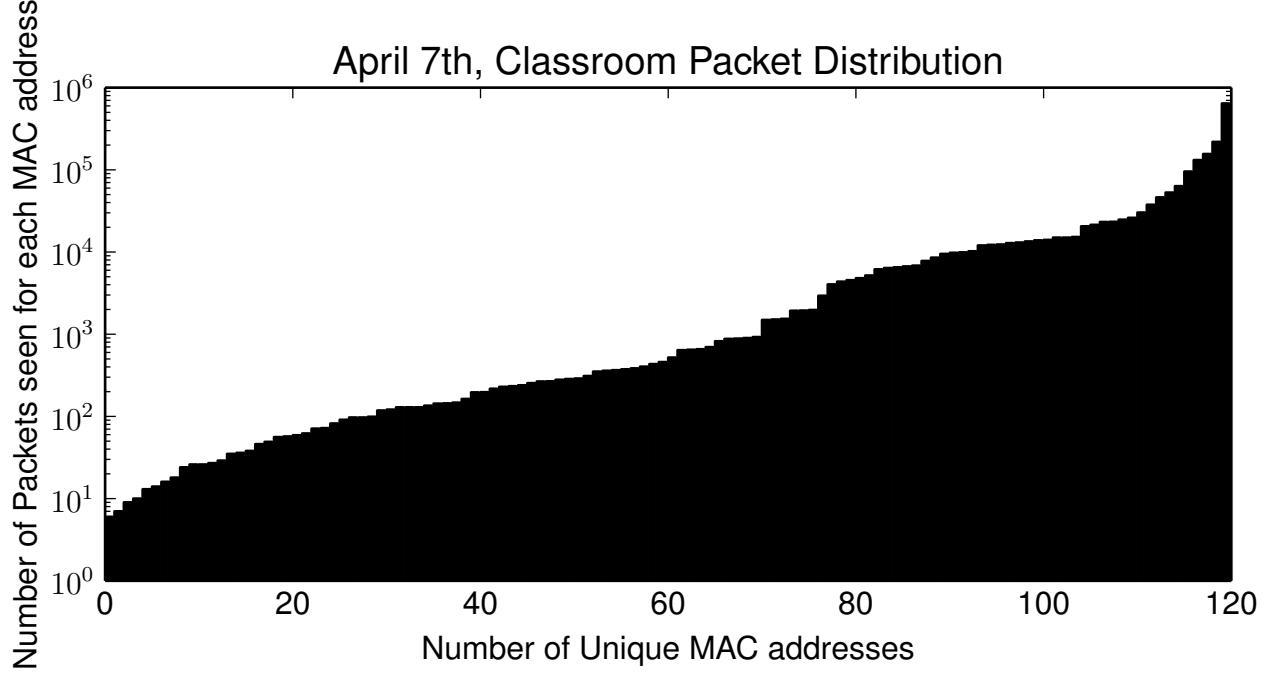


Fig. 12: Number of packets sent by each device during the entirety of the class period, sorted by activity: the logarithmic scale reveals that the median device sent several hundred packets.

data to make occupancy predictions. However, the manpower required for this also makes this an improbable option.

8.3 Potential Upcoming Problems

Any of our occupancy graphing methods depend upon MAC addresses corresponding to passengers. Similar techniques are utilized in the real world to determine traffic patterns. For example, the city of Houston uses the MAC addresses from Bluetooth devices to figure out the flow of traffic on city streets in their TranStar monitoring system [8]. Unfortunately, this technique may not be effective for much longer. The next update for iOS, iOS 8, is expected to be released in Fall 2014 [9]. This update will change how Apple devices broadcast their packets when searching for networks. The device will randomly generate a MAC address during each round of scanning, rather than display the hardware MAC address. This is to deter tracking by companies for the purpose of targeted advertisements and to provide higher security for consumers [8]. Though we have

no malicious intent, this privacy feature will still put an end to our tracking — rather than seeing an iPhone moving along a bus route, we will instead see many fake devices, with no apparent correlation. Unfortunately for us, Apple iPhones make up over 40% of the smartphones owned in the US, so this update would greatly reduce the effectiveness of our program [10].

REFERENCES

- [1] (2014) Nextbus rutgers university. [Online]. Available: <http://www.nextbus.com/predictor/stopSelector.jsp?a=rutgers>
- [2] V. Jacobson, C. Leres, and S. McCanne. (2014) Tcpdump documentation. [Online]. Available: <http://www.tcpdump.org/manpages/tcpdump.1.html#lBAI>
- [3] R. D. of Transportation. (2014) Rutgers campus buses. [Online]. Available: <http://rudots.rutgers.edu/campusbuses.shtml>
- [4] (2013) Introducing json. [Online]. Available: <http://json.org/>
- [5] T. Oliphant. (2014) Numpy. [Online]. Available: <http://www.numpy.org/>
- [6] (2014) Python data analysis library. [Online]. Available: <http://pandas.pydata.org/>
- [7] J. Hunter, D. Dale, E. Firing, and M. Droettboom. (2013) Matplotlib. [Online]. Available: <http://matplotlib.org/>

- [8] S. Oliver. (2014) MAC address randomization joins apple's heap of iOS 8 privacy improvements. [Online]. Available: <http://appleinsider.com/articles/14/06/09/mac-address-randomization-joins-apples-heap-of-ios-8-privacy-improvements>
- [9] (2014) Apple - iOS 8 - overview. [Online]. Available: <http://www.apple.com/ios/ios8/?cid=wwatus-kwg-features-com>
- [10] N. Hughes. (2014) Apple's iPhone now represents 42% of smartphones owned in the US - NPD. [Online]. Available: <http://appleinsider.com/articles/14/01/16/apples-iphone-now-represents-42-of-smartphones-owned-in-the-us---npd>

APPENDIX

TEAM MEMBER CONTRIBUTIONS

Daniel Bordak

Collected tcpdump and classroom data

Created tcpdump output parser

Current implementation's design came from privacy concerns and speed requirements. Encountered difficulties in matching arbitrary MAC addresses and determining their type, as well as the aforementioned privacy concerns.

Optimized graph plotting functions

First implementation of original plotting functions were painfully slow and had large amounts of redundant code. Refactoring solved the redundancy, but more importantly the usage of external libraries solved the speed issue. Integration of these libraries was particularly difficult, however, due to their demand of vectorized code.

Cowrote First-Last Grid Plot.

Solved issues with axes.

Cowrote Segments Plot.

Rewrote to better utilize Pandas features, dramatically increasing speed and accuracy.

Wrote script to generate stop annotations

The desire to reduce redundancy led to the implementation of a common parser between both stop data and packet data. This script was applied to all graphs based on bus data.

Wrote all plot generation sections.

Difficulty encountered in remembering everything I did.

Formatted report.

Difficulty figuring out a whole lot of L^AT_EX stuff.

Rewrote report.

Original draft was inaccurate as to the purpose of some graphs, etc. Difficulty in choosing what needed to be rewritten, and of course rewriting large sections of the report.

Added Bibliography.

Difficulty in figuring out modern best practices for L^AT_EX bibliographies.

Erin Corrado

Collected bus and classroom data.

Cowrote First-Last Grid Plot.

Difficulties in learning more advanced Python and learning how to utilize the Python libraries Numpy and Matplotlib.

Wrote length/occupancy correlation plot.

Data was difficult to interpret at times. Data was time consuming to collect and limited.

Cowrote poster.

Difficulty with trying to get the poster printed and transported to the proper place in time.

Cowrote Packet Histogram.

Encountered difficulties with utilizing the new libraries used such as Pandas.

Proofread and edited report and graphs.

Hardest part was keeping everyone's contributions in sync with one another. Also the group had to decide on style and formatting.

Wrote Conclusion

Figuring out exactly what to write and conclude.

Revan Sopher

Wrote Proposal

Designed two-method approach, laying out plan for study.

Collected tcpdump data

First data collection of study, so needed to figure out which tool to use and how to use it, and how to put a wireless card in monitor mode.

Wrote Raw Packet Plot

First plot of study, so needed to choose and learn tools (matplotlib, numpy). Also needed to write annotation script and parser for tcpdump output, including a regex for matching MAC, as this was before Dan's rewrite.

Wrote Unique Device Plot

First data processing of study. Later plots drew from this one.

Cowrote Segments Plot

This was after Dan's refactoring, so needed to situate myself and learn the more efficient looping techniques.

Wrote Vector Plot

Need to create and sum vectors from the data caused this plot to be significantly different.

Wrote Report

Wrote entirety of report as first submitted. Expanded content and wrangled with LaTeX for later revisions.

Ashley Weaver

Cowrote report

Editing first drafts. Research for future works section.

Cowrote Poster

Difficulty designing layout to fit all graphs and descriptions.

Collected bus data

Spent a long time simply waiting for the bus. Bus was often delayed due to traffic.