CS 7311 Project Report
Texas State University
Fall 2018

Gentry Atkinson
gma23
gma23@txstate.edu

Abstract:

An analysis of of traffic flow patterns in the city of Austin, TX is presented with a focus on deducing the rate of public transportation usage. This analysis was conducted on a data-set produced with 147 BAT 433 Bluetooth device sensing stations around the city of Austin which is hosted at: data.austintexas.gov. These devices log the presence of a Bluetooth enabled device such as a phone or tablet at a particular time and place. This data has been aggregated by the city and analyzed to produce simple results such as the rate of travel along roads with these sensor devices at particular times. This paper extends that analysis by collecting "walks" over the data of sets of devices as represented by anonymized MAC addresses. An unsupervised clustering algorithm is then applied to these walks in order to detect a cluster of high length and high device number cluster which could reasonably represent groups of public transportation users.

1. Introduction:

Municipal sensor networks are becoming more common in the developed world. They allow city planners to respond to conditions within a city in real-time. But with the greater bulk of data generated by these sensor networks comes a greater burden to perform meaningful analysis of the results. Otherwise these modern and enormously useful tools are only contributing confusion rather than clarity. Austin, TX released the data collected by their Bluetooth sensing network to the public in the hopes that the innovative force of crowd collaboration would generate meaningful results in new and creative ways. Among the several specific questions that the city hoped to have answered by this public collaboration was this: what portion of our traffic is being generated by public transportation? This paper will approach this question by generating "walks" of sets of devices over a graph of the sensors and will then clusters
these walks to generate insight into the nature of traffic flow within the city of Austin. Random walks over graphs are a well established technique but are heretofore yet to be applied to this data set. [1]. A walk in this paper will refer to a set of devices being logged at the same series of sensors at the same rounded times. Rather than generating a small set of random walks on the graph of time-locations, this project generated the full set of trips taken by specific sets of devices on the data. This produced a more complete data-set on which to work at the expense of some computational power. Merely collecting these walks does not by itself tell us anything significant. After collection conventional techniques of comparison were used to find shallow distinctions between the city as a whole and the two, bus-heavy street which were selected. Furthermore, data clustering was performed using the BIRCH algorithm to find distinctions in the data which may not have been readily apparent from the conventional analysis.

2. Problem Statement:

While rapid growth can be economically beneficial for a city and its citizens, there is no way to deny that it can place a substantial strain on the infrastructure of a city. Cities in Texas face an unusual challenge in adapting to changes in population in that Texas does not collect income tax. Therefore the revenue available to the city is not proportional to the size of the population but only to the sum value of the land within the city. Deal with the increasing burden on its infrastructure without a proportional

increasein its revenue is forcing Austin to act more innovatively to do more with less. The city of Austin generated the “Hack the Traffic” (hosted at https://data.austintexas.gov/) transportation?” Solving this problem using the provided data-set presents several problems. The sensors only log the presence of a Bluetooth enabled device and the time that it was detected. There is no way to ascertain directly whether several devices which were logged were in a single vehicle or several.

In order to compensate for the difficulties presented by solving this particular problem with this particular data, this project chose to focus specifically on buses rather than including multi-occupancy cars. Buses are expected to be easier to detect within the data than cars for two reasons. The first is that they carry more people and therefore produce a larger set of devices traveling together. This larger set should be significantly easier to detect against the baseline of single occupancy vehicle. The second reason is that buses travel on predictable routes. This means that a pair of bus routes could be chosen as the focus of the project. Since it was certain that bus traffic is heavier on these lanes than elsewhere, it becomes reasonable to say that the distinctions found in the analyses are due to the presence of buses.

2.1 Related Work:

The devices being used to collect this project’s data are quite new. This is because Bluetooth is very young compared to many other municipal technologies. The sensors themselves also represent a cost that many cities have been reticent to adopt.

Despite the rarity of these sensors several studies on their usefulness have been performed. One study [2] by the University of Maryland found that, in 2011, sampling rates for traffic using Bluetooth sampling varied between 2% and 8% of vehicles being detected. A report [3] by the Maryland highway administration in 2012 demonstrated that Bluetooth was as good as conventional sampling techniques for measuring vehicle speeds on roadways. This report also reveals that the 5-year operating cost for a single collection of data-sets with the hopes of finding creative and clever solutions to certain problems. Among these problems was this, “What portion of the traffic on our streets is generated by public sensor is $7200. Another group at Chulalongkorn University in Thailand[4] developed their own non-commercial Bluetooth detector and used it to track travel time through Bangkok in 2014. They demonstrated that their device was effective for its intended purpose but noted that not every device was detected by their sensor meaning that some trips were not properly recorded.

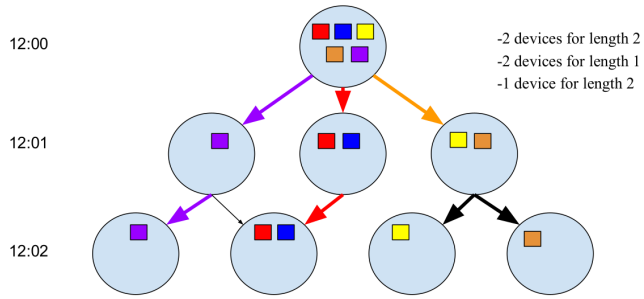
These studies all vary from the presented work in two ways. First, they made no attempt to analyze the variety of traffic that was present on roadways. So while they were all able to to make conclusions about traffic flow and speed, nothing was said about the nature of the vehicles that composed that traffic (e.g. single-occupancy vs. multi-occupancy vehicles). Second, they did not employ any powerful data-analysis techniques. Rather, their focus was primarily on data collection.

3. Methods:

The data provided was first converted to json. This semi-structured data format lends itself well to the analysis of large data sets. The collected results were then split into files representing a single day of data. This was done to allow parallel analysis of many days simultaneously. This could be done without degrading the results of the analysis because the device addresses were re-randomized everyday to help protect the identities and privacy of the drivers and passengers represented in the data. After the data was a graph in which a vertex represented a specific sensor at a time rounded to the minute. This rounding was performed to compensate for the slight “jitter” which can occur with real-world sensors based on the signal strength of individual devices, factors which might affect signal propagation between the device and the sensor, and processing delays within the sensors. Edges are added to this graph to represent one or more devices traveling from one sensor-time to another. Walks were then calculated on the graph

using the following algorithm:

Algorithm 1:
For each sensor-time:
  Set current-node to sensor-time
  For each neighboring node:
    Find the intersection of devices at current and next
    While intersection is still present:
      Check for intersection in neighbors
      If subset found, set current-node to neighbor
Record walk



A demonstration of 3 walks on a small graph.

A new json file was generated to store walks found by Algorithm 1. The specific information recorded for each walk was: number of devices, length of walk, time at start, time at end, speed at start, speed at end, the date, the location of the starting sensor, and the location of the ending sensor. In all about 50,000,000 walks were found in the data-set. Visualizations were created of the speeds for the walks, the times of day that they occurred, the number of devices, and the length of the walks.

Clustering was performed on this new data-set using the BIRCH clustering algorithm. The clusters generated by by analysis of the city as a whole was then compared to those on two bus-heavy routes, and one bus-light routes.

The specifc routes for these comparisons were chosen by comparing the map of all available bus routes provided by Cap Metro with a map of all available data sensors which was generated by processing of the data-set. The two bus-heavy routes identified by the process were Lamar Blvd and Riverside Dr. Both of these streets host both a high number of bus routes and Bluetooth sensors. The one bus-light routes identified was Red River St. More bus-light routes would have been useful but heavily trafficed streets attract both buses and sensors so relatively few streets have sensors but not buses.

4. Tools and Infrastructure:

Several Python standard libraries were employed in production and analysis of of our data. These included: json, numpy, matplotlib, gmplot, and itertools. The processing took place on several PC computers using the Linux operating system. All of the scripts used are available at: github.com/gentry-atkinson/traffic_analysis. Running these scripts on non-Linux computers may require some slight modification.

5. Data Processing Tools:

Data acquisition for this project was performed by the City of Austin using BAT 433 Bluetooth device sensors. This distributed array is logged on a central server hosted by the city of Austin which is not available for the public to access. Aggregated data has been made available by the city at https://data.austintexas.gov/. All processing was performed using Python scripts which can be viewed and downloaded at: github.com/gentry-atkinson/traffic_analysis.

5.1 Machine Learning Tools:

The BIRCH clustering algorithm was first described in 1996 [5]. Its intent was to provided fast and effective clustering on large, noisy data-sets using limited memory. It produces for its user a tree of clusters and sub-clusters. This means that an analyst can use it on some data and then choose the level of depth which best fits the specific data. This makes it very effective on data which may contain some unknown number of clusters.

BIRCH was chosen for this project based on its ability to compensate for noisy data with sacrificing its speed of execution. Large sensor networks are inherently noisy owing to several

factors including sensitivity, signal strength, network transmission speeds, and local interference. As noted in the Thai study [4] it is not uncommon for a sensor to fail to log the presence of a device although it should be noted that the sensors used in that study were built in-house rather than purchased off of the commercial market.

The implementation of the BIRCH algorithm used by this project is available at: freediscovery.io.

5.2 Visualization Tools:

The gmplot library was used to programmaticly plot points and trips on a map provided by Google Maps. This project can be found at: github.com/vgm64/gmplot.

Matplotlib was used to generate several charts which incorporated large amounts of data. This project is hosted at: matplotlib.org.

Finally Google Sheets was used to produce several charts which relied on smaller amounts of data.

6. Experiments:

As described in Section 3, two bus-heavy and one bus-light streets were first identified by comparing the Cap Metro bus map to the plot of all sensors used by the city. After the walks were calculated as described earlier they were divided into a full set of all walks throughout the city, those occurring entirely on Lamar, those occurring entirely on Riverside, and those occurring entirely on Red River.

These groups were first compared using conventional techniques. The speed, time of day, and size/length of the walk were compared for the full city group, the Lamar group, and the Riverside group.

Clustering was then performed on all four groups. Before clustering the data was cleaned to remove the walks with fewer than 3 devices or a length of less than two edges of the graph. These low-value data points represented 56% of all of the walks calculated and offered little of value to the final results. Therefore they were removed to prevent them from wholly subsuming the other data points and forcing a single data cluster which would eclipse the others.

6.1 Results:

It was found that the two bus-heavy routes skewed towards slightly slower travel. It would be irresponsible to though to settle on this distinction as being emblematic of the presence of a bus route. Rather, this is probably owing to the fact that the full-city group contains measurements taken from several highways, so its reasonable to think that a city surface street will generally move more slowly than a highway.

Whole of Austin:

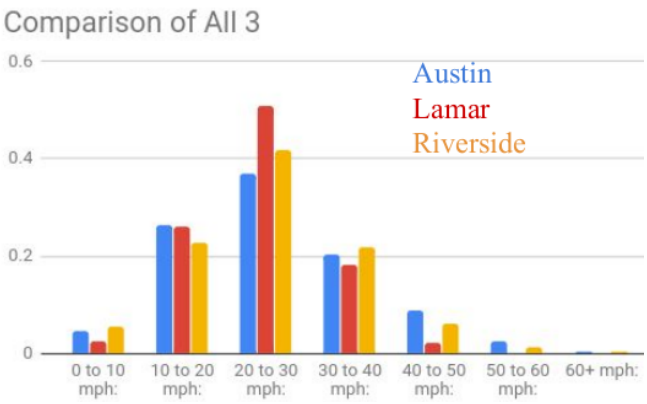
0 to 10 mph:	2256903	4.50%
10 to 20 mph:	13211872	26.36%
20 to 30 mph:	18475629	36.86%
30 to 40 mph:	10169677	20.29%
40 to 50 mph:	4483509	8.95%
50 to 60 mph:	1292444	2.58%
60+ mph:	231234	0.46%

Lamar:

0 to 10 mph:	314944	2.56%
10 to 20 mph:	3211569	26.15%
20 to 30 mph:	6230284	50.74%
30 to 40 mph:	2252714	18.35%
40 to 50 mph:	255728	2.08%
50 to 60 mph:	12663	0.10%
60+ mph:	1258	0.01%

Riverside:

0 to 10 mph:	40575	5.63%
10 to 20 mph:	164743	22.85%
20 to 30 mph:	301767	41.85%
30 to 40 mph:	157833	21.89%
40 to 50 mph:	43694	6.06%
50 to 60 mph:	10387	1.44%
60+ mph:	2028	0.28%



The bus-heavy routes were also found to skew more towards early morning and later evening travel. This distinction is more difficult to dismiss. It is reasonable to expect more commuter traffic at this times. It is quite possible that bus commuters leave slightly earlier and return slightly later owing to the longer travel times of public transportation. It is promising then to see this distinction present itself in the data.

Whole of Austin:

0 to 1:	854242	1.70%
1 to 2:	620714	1.24%
2 to 3:	545207	1.09%
3 to 4:	372932	0.74%
4 to 5:	294156	0.59%
5 to 6:	683663	1.36%
6 to 7:	1586165	3.16%
7 to 8:	2603519	5.19%
8 to 9:	2863488	5.71%
9 to 10:	2753539	5.49%
10 to 11:	2681717	5.35%
11 to 12:	2860757	5.71%
12 to 13:	3029759	6.04%
13 to 14:	3074161	6.13%
14 to 15:	3074824	6.13%
15 to 16:	3213351	6.41%
16 to 17:	3329361	6.64%
17 to 18:	3259326	6.50%
18 to 19:	3050716	6.09%
19 to 20:	2546616	5.08%
20 to 21:	2123979	4.24%
21 to 22:	1925796	3.84%
22 to 23:	1606060	3.20%
23 to 24:	1167220	2.33%

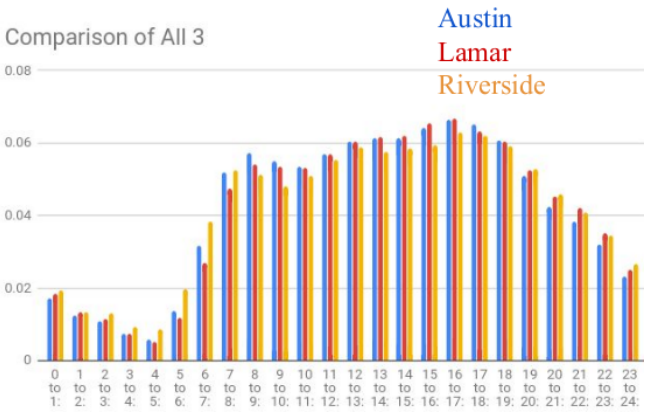
Lamar:

0 to 1:	227711	1.85%
1 to 2:	162681	1.32%
2 to 3:	140384	1.14%
3 to 4:	90572	0.74%
4 to 5:	62231	0.51%
5 to 6:	143480	1.17%
6 to 7:	331094	2.70%
7 to 8:	581809	4.74%
8 to 9:	664893	5.41%
9 to 10:	654655	5.33%
10 to 11:	653219	5.32%
11 to 12:	698403	5.69%
12 to 13:	741621	6.04%
13 to 14:	759069	6.18%
14 to 15:	762838	6.21%
15 to 16:	805579	6.56%
16 to 17:	820721	6.68%
17 to 18:	775234	6.31%
18 to 19:	742853	6.05%
19 to 20:	645560	5.26%
20 to 21:	555391	4.52%
21 to 22:	517505	4.21%
22 to 23:	433125	3.53%
23 to 24:	308532	2.51%

Riverside:

0 to 1:	13936	1.93%
1 to 2:	9684	1.34%
2 to 3:	9478	1.31%
3 to 4:	6746	0.94%
4 to 5:	6319	0.88%
5 to 6:	14256	1.98%
6 to 7:	27588	3.83%
7 to 8:	37967	5.27%
8 to 9:	36991	5.13%
9 to 10:	34680	4.81%
10 to 11:	36632	5.08%
11 to 12:	39991	5.55%
12 to 13:	42451	5.89%
13 to 14:	41562	5.76%
14 to 15:	42098	5.84%
15 to 16:	42950	5.96%
16 to 17:	45394	6.30%
17 to 18:	44757	6.21%

18 to 19:	42745	5.93%
19 to 20:	38155	5.29%
20 to 21:	33010	4.58%
21 to 22:	29553	4.10%
22 to 23:	24910	3.45%
23 to 24:	19174	2.66%



Charting the length of walks and the number of devices represented showed that most walks were low-length and low-device count regardless of the group. Lamar accounted for 12,279,128 of the 50,121,162 total walks found in the data. Riverside accounted for 721,027 of these walks. Lamar appears to present a more even distribution across devices numbers in it walks with 1-device and 2-device walks having nearly equal numbers but nothing was so profound as to produce a definite result.

Whole of Austin:

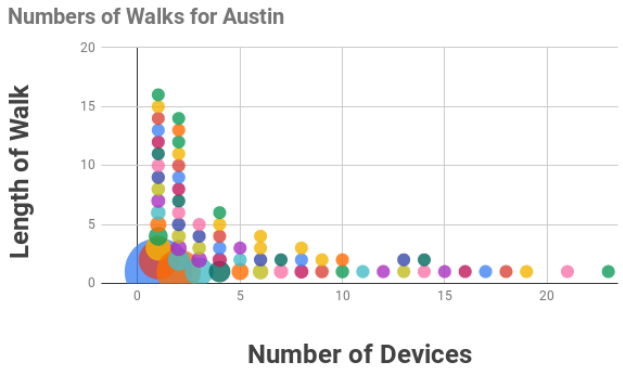
Number of walks with 1 devices:	36584301
Number of walks with 2 devices:	10218184
Number of walks with 3 devices:	2347977
Number of walks with 4 devices:	712916
Number of walks with 5 devices:	162350
Number of walks with 6 devices:	64764
Number of walks with 7 devices:	15561
Number of walks with 15 devices:	90
Number of walks with 10 devices:	1950
Number of walks with 8 devices:	8992
Number of walks with 9 devices:	2520
Number of walks with 11 devices:	462
Number of walks with 17 devices:	102
Number of walks with 12 devices:	456
Number of walks with 18 devices:	72
Number of walks with 19 devices:	38
Number of walks with 21 devices:	42
Number of walks with 23 devices:	46
Number of walks with 13 devices:	195
Number of walks with 16 devices:	96
Number of walks with 14 devices:	154

Lamar:

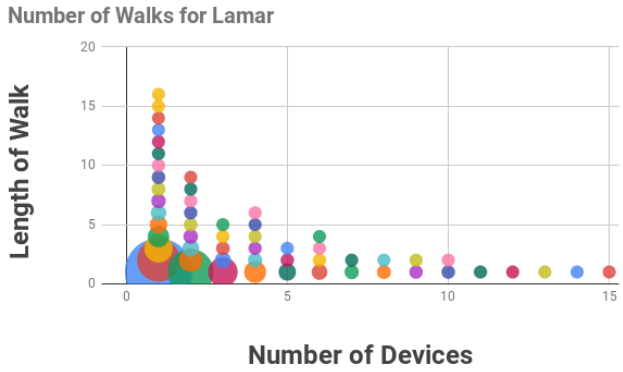
Number of walks with 1 devices:	8869887
Number of walks with 2 devices:	2563448
Number of walks with 3 devices:	604716
Number of walks with 4 devices:	176228
Number of walks with 6 devices:	16062
Number of walks with 5 devices:	41105
Number of walks with 15 devices:	30
Number of walks with 10 devices:	630
Number of walks with 8 devices:	2200
Number of walks with 7 devices:	3850
Number of walks with 9 devices:	765
Number of walks with 11 devices:	99
Number of walks with 13 devices:	52
Number of walks with 12 devices:	60
Number of walks with 14 devices:	28

Riverside:

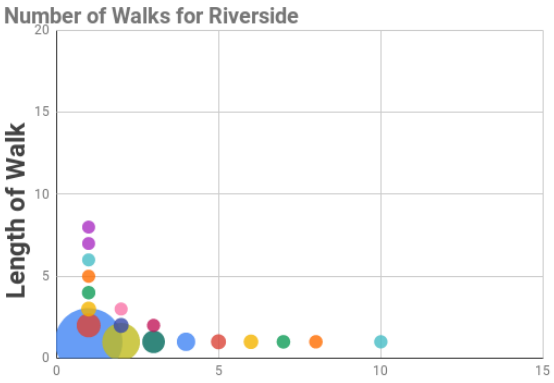
Number of walks with 1 devices:	580155
Number of walks with 2 devices:	116336
Number of walks with 3 devices:	17598
Number of walks with 4 devices:	5560
Number of walks with 5 devices:	765
Number of walks with 6 devices:	468
Number of walks with 7 devices:	77
Number of walks with 8 devices:	48
Number of walks with 10 devices:	20



A depiction of walks in Austin. Bubble size is proportional to the percentage of walks.



A depiction of walks on Lamar. Bubble size is proportional to the percentage of walks.



A depiction of walks on Riverside. Bubble size is proportional to the percentage of walks.

Clustering these walks produced a greater distinction amongst the groups than did traditional comparison. When BIRCH was applied to all 4 groups with appropriate filtering, it was found that Lamar and Riverside consistently produced a much smaller number of cluster as compared to both Red River and the city as a whole. This suggests that the data in those two groups either holds more tightly to a small number of regions of is dispersed enough that the algorithm seperates the clusters more broadly. Either way, clustering produced a noticeable distinction between the bus and none bus groups that could be built upon in future work.

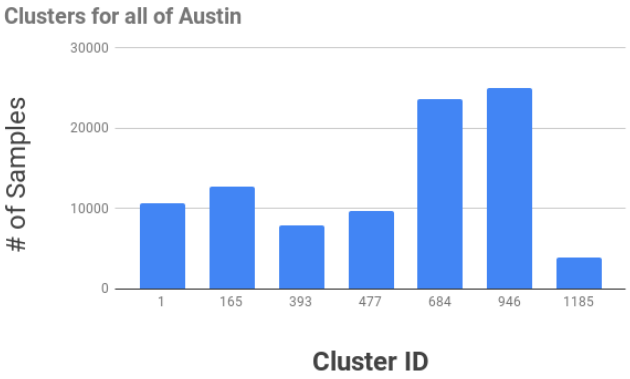
	Cluster ID	Samples
Austin	1	10668
	165	12809
	393	7946
	477	9754
	684	23598
	946	25092
	1185	3952
	Total:	93819

	Cluster ID	Samples
Lamar	1	19595
	226	8660
	Total:	28255

	Cluster ID	Samples
Riverside	1	434
	18	325
	Total:	759

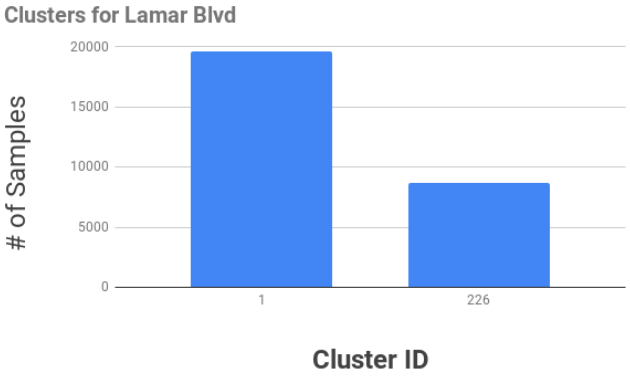
	Cluster ID	Samples
Red River	1	2347
	16	607
	27	1660
	39	232
	50	2476
	64	621
	Total:	7943

Notice that in each of these groups the number of samples is much smaller than in previous examples owing to the filtering that was done before clustering.

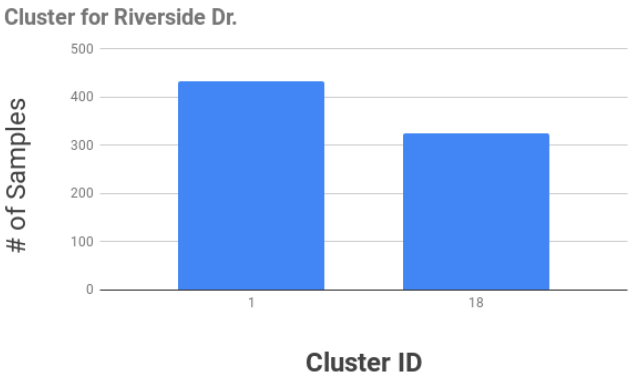


Bar graph of the number of samples per cluster for Austin.

Bar graph of the number of samples per cluster for Red River St.



Bar graph of the number of samples per cluster for Lamar Blvd.



Bar graph of the number of samples per cluster for Riverside Dr.

7. Conclusion:

This paper has described a technique which has been applied to data collected by Bluetooth device sensors arranged around the city of Austin, TX. Devices were followed across these sensors over time in order to build a set of walks which were then clustered in order to determine what portion of this traffic was composed of high occupancy public transportation and what portion of it was low or single occupancy vehicles.

This project has concluded that there is a good chance that some street can eventually be identified as containing a higher portion of public transportation. There is no indication that any data point within the set can be identified as belonging to a single- or multi-occupancy vehicle, not any one walk as calculated by this project. Finally, it does not appear to be possible to calculate the exact percentage of trips on a street which is being produced by public transportation. If it is possible to make any distinction it will be only to determine which street contains more multi-occupant traffic as compared to another.

7.1 Future Work:

This project was sufficient to demonstrate within limits that some distinctions can be drawn from unsupervised learning techniques regarding the presence of busses (and possibly multi-occupant vehicles in general) in streets monitored by Bluetooth sensing devices. In order to confirm and expand upon this data it will be useful to find more streets for comparison. Although the Hack the Traffic data-set is very large in terms of the

sheer number of entries, there are not so many roadways represented in the data. Most problematicaly for these finding, there are very few streets in Austin which have sensors but do not have Bluetooth sensors. Expanding the base of this projects analysis to other cities could yield more useful results than those that have been found by this limited analysis.

References:

[1] L. Lovasz, "Random Walks on Graphs: A survey," *Bolyai Society Mathematical Studies. 2*, pp. 1-46, 1993.

[2] Elham Sharifi, Masoud Hamedi, Ali Haghani, Hadi Sadrsadat, "Analysis of Vehicle Detection Rate for Bluetooth Traffic Sendors: A Case Study in Maryland and Delaware," 18th World Congress on Intelligent Transport Systems, October 2011

[3] Stanley E. Young, "Bluetooth Traffic Detectors for Use as Permanently Installed Travel Time Instruments," State Highway Administration, February 6, 2012

[4] Ladawan Klinkusoom, Chaodit Aswakul, Panuwat Janpugdee, "Bluetooth Sensors for Vehicular Traffic Monitoring," *I29th International Technical Conference on Circuit/Systems Computers and Communications*, July 1-4, 2014

[5] Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," SIGMOD ’96 6/96 Montreal, Canada, 1996.