**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 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/) 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

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 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 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.

**5. Data Processing Tools:**

Data acquisition for this project was

performed by the City of Austin using the

[[sensor model]] Bluetooth device sensor

from [[manufacturer]]. 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

incorporating [[whatever libraries]] which can

be viewed and downloaded at:

git.txstate.edu/gma23.

[[Is Spark still necessary?]]

[[Can we work out fuzzy clustering?]]

**5.1 Machine Learning Tools:**

[[describe clustering algorithm]]

**5.2 Visualization Tools:**

[[describe gmplot]]

[[how else do we want to view this data?]]

**6. Experiments:**

[[describe the identified “bus heavy” lanes]]

[[cluster all the walks and then cluster walks

from known bus lanes. Is there a more

prevalent high device number cluster in the

bus heavy lanes?]]

[[what variables can i control for and what am

i testing for? How much of a distinction is a

positive result?]]

**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.

[[I found something, positive or negative]]

**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 SENSORS: 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