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 dataset produced with 139 Bluetooth device sensing stations around the city of Austin. These devices log the presence of a Bluetooth enabled device 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. [[what algorithm? What parameters? What results? Is this too long?]]

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. Amongst 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. [[reference?]] 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.

Merely collecting these walks does not by itself tell us anything significant. Therefore the new dataset is then clustered using [[algorithm? DP-means? Hierarchical k-means? BIRCH? Is fuzzy clustering a possibility in this timeframe?]].

[[what do i assume the clusters will mean? What distinguishes a bus?]]

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. Land values within Austin are certainly increasing but much more slowly than land values [[source maybe? Mention the homesteader act which limits the rate that land values can grow?]]. Deal with the increasing burden on its infrastructure without a proportional increase in 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 datasets with the hopes of finding creative and clever solutions to certain problems. Amongst these problems was this, "What portion of the traffic on our streets is generated by public transportation?"

Solving this problem using the provided dataset 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.

[[mention contact with city?]]

[[some research on infrastructure costs to the city]]

[[limit yourself to busses? Is more possible]]

2.1 Related Work:

[[similar studies? Other results derived from Hack the Traffic? Previous uses of this clustering algorithm?]]

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. [[is this necessary? Did any important filtering or processing happen during the conversion?]] 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 was generated for each day, in which a vertex represents a 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 travelling 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
  Calculate the power set of devices
  For each device subset:
    For each neighboring node:
      Check for inclusion of subset in device list
        If subset found, set current-node to neighbor
          Repeat with subset until not found
          Record walk
```

A new json file was generated to store walks found by Algorithm 1. Clustering was performed on this new dataset using

[[whatever clustering algorithm]]. The clusters generated by this analysis were then compared to the expected [[whatever assumptions I'm making]]

[[Compare global data to our "bus heavy" routes on Riverside and Lamar? We still need quantifiable comparisons to identify a bus]]

4. Tools and Infrastructure:

[[How much to say about LEAP?]]

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]
[2]
[3]