How many people are at intersections in Boston?

Ben Lawson
Department of Computer Science
Boston University
Boston, Massachusetts 02215
balawson@bu.edu

Social media is a constant factor in many people's lives in current times. We explore the usefulness of data collected from public resources to derive estimates of a subsection of people at different intersections in Boston. Using mainly Twitter data, we are able to apply sampling methods to determine an estimate of the monthly visitation rate of almost 60 intersections. This information can be incorporated into other systems to improve traffic congestion estimation.

1 Introduction

In this project, we attempt to develop an understanding of human movement within the city of Boston. Using social media data from three companies, Brightkite and Gowalla, both of which are no longer active, and Twitter. To generate higher granular information, we used OpenStreetMap data to associate social media user's posts to specific intersections. Since not all people use social media, and thus are not represented in the datasets presented here, we must infer the actual amount of people that are present at these intersections in real life. Future work will attempt to cross-validate these methods with different types of observations, such as census population data and population counts at intersections derived from street cams and computer vision. Future work will also include using this data to solve classic problems, like max flow, in the pedestrian setting.

2 Data Resources

Many types of data were used in this project. Three sources of geosocial media data were used as well as geographical information from OpenStreetMap

2.1 Brightkite Dataset

This is a social media networking service that was acquired by a mobile social network, Limbo, in 2009. The dataset contains posts with a user id, geocoordinates, and a timestamp between 14 April 2008 and 18 October 2010.

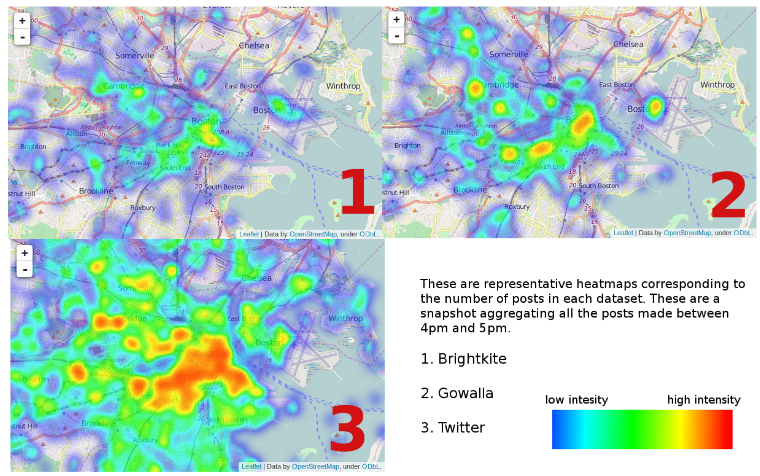

Fig. 1. Heatmap of tweets by direct count. Scaled proportionally total to each dataset.

2.1.1 Gowalla Dataset

This is a social media networking service that went out of business. Each post in the dataset has a user id, geocoordinates, and a timestamp, dating between 23 April 2009 and 22 October 2010.

2.1.2 Twitter Dataset

This is a micro blogging service that collects geological information about users's posts. This is still an active service and this data set was collected from 11 May 2015 until 2 April 2016. This data was collected via Twitter's streaming API, filtered by geolocation.

2.2 OpenStreetMap

This dataset was collected from OpenSteetMap, a collaborative, community based geographical open data resource. We use the labeled roads in the Boston area to create a chart of intersections.

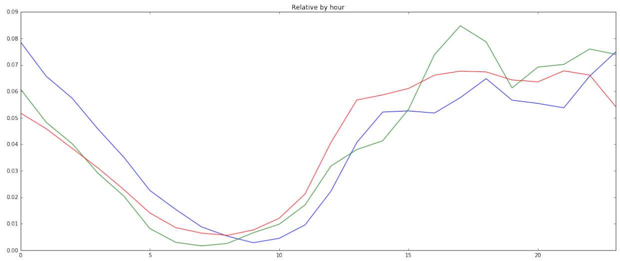

Fig. 2. Shows the percentage of tweets by hour. Demostrates that tweeting behavior mimics the human sleep cycle. Red is Brightkite, Green is Gowalla, and Blue is Twitter.

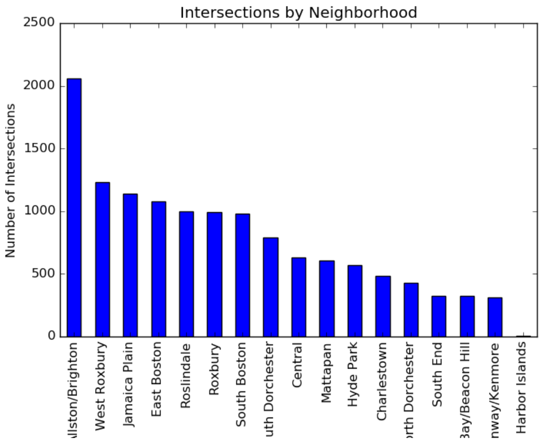

Fig. 3. The number of intersections per neighborhood

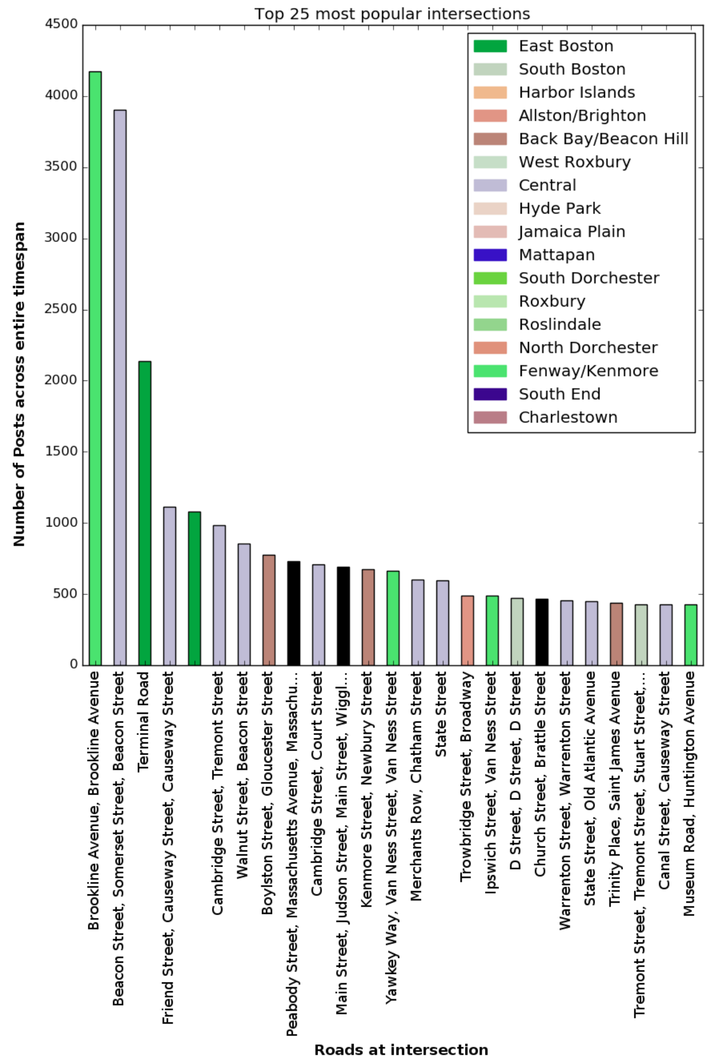

Fig. 4. The 25 most popular intersections overall (sheer volumne) colored by neighboorhood. Note missing road names and black color is due to missing data.

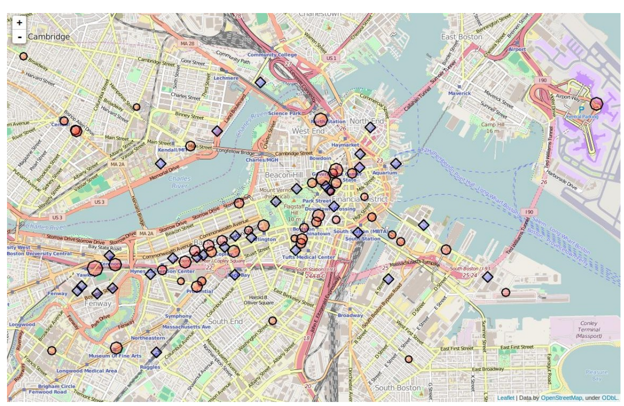

Fig. 5. Relative estimates for monthly social media users per intersection

2.3 BostonMaps: Open Data

This dataset contained bounding boxes and geometric shapes for each of the neighboorhoods in Boston. This was used to help give a general intuition for results.

3 Segmentation

Using the OpenStreetMap data, each Twitter post was associated with the closest intersection. This was done for each month in the Twitter dataset, so twelve months are represented from May 2015 to April 2016.

4 Sampling

Capture & Recapture. This type of sampling was first used when measuring animals in traps. Trappers would mark animals and then count how many of these animals returned to derive an estimate of the total population. We discovered only 94 intersections, of the almost 25,000, had five or more visitors each month. Only 57 of these intersections had visitors that returned during the capture/recapture period. The red markers show the intersections that estimates could be computed, scaled with the $\log_2$ function. The blue markers show the intersections that did not have returning users. Fenway park had that max estimate with approximately 12,000 visitors per month. This calculation was only done with the

Twitter dataset. Future work will explore the best way to intergrate the Brightkite and Gowalla datasets.

$$\hat{N} = \frac{Kn}{k} \tag{1}$$

$\hat{N}$ is the estimation of the total population of social media users, n is the number of users during the first month, K is the number of users during the second month, and k is the number of users from the first month that are observed during the second month. With this formulation, we can derive estimations for the number of social media users that post near an intersection per month. It will be important to discover what percentage the social media users are of the total population. This can be done by observing a true population through web cams and utilizing popular methods in computer vision, like object dectection and background subtraction. We can also obtain an estimate by comparing all the social media activity in a neighborhood to the total population take from the census, however this would be less accurate and would be difficult to argue as the true population.

$$\frac{|A \cap B|}{|B|} = \frac{|A \cap S|}{|S|} = \frac{|A|}{|S|} \tag{2}$$

Once we know the relationship between the number of social media users and the true total population, we can use proportional sampling to determine the most accurate estimate for the number of people passing through an intersection. For example, if we knew that $\frac{1}{10}$ people used social media, we can use Equation(2) to determine the estimate of people at each intersection, instead of an estimate of the number of social media users at each intersection. If we take the most popular intersection, Brookline Ave @ Brookline Ave, our estimate is 12,835.2. Plugging this into the formula gives us: $\frac{1}{10} = \frac{12835.2}{\hat{N}}$ resulting in an appromation of $128,352$ total people passing through that intersection each month.

5 Conclusions

Although the social media data consisted of many posts, only a fraction of the intersections had data spanning the entire collection period. Of these intersections, only a fraction had entire data to compete the estimate via the capture & recapture method. Future work will be to include the Brightkite and Gowalla datasets in the estimation of intersection occupancy and analyzing flow problems associated with pedestrian traffic.

Acknowledgements

Special thanks to Andrei Lapets who supported this project and helped flush out ideas about general project direction. Special thanks to Evimaria Terzi for help with the development of the Twitter dataset and Sofia Maria Nikolakaki for her work with tweet segmentation by intersection.

Appendix A: Most Popular Intersections (monthly estimates)

Road Intersection	Monthly Estimate
Brookline Ave, Brookline Ave	12835.2
Beacon St, Somerset St	9226.3
Friend St, Causeway St	3685.0
Cambridge St, Court St	3596.0
Main St, Judson St, Wigglesworth St	2928.0
Terminal Road	2135.2
Kenmore St, Newbury St	1840.0
Avenue De Lafayette, Washington St	1392.0
Trinity Place, Saint James Avenue	1368.0
Boylston St, Gloucester St	1357.3
Museum Road, Huntington Avenue	1323.0
Arlington St, Newbury St	1020.0
Massachusetts Ave, Douglass St	783.0
Newbury St, Exeter St	726.0
Belvidere St, Huntington Ave	704.0
Tremont St, Stuart St	640.0
Fairfield St, Newbury St	580.0
Boylston St, Exeter St	528.0
Dartmouth St, Newbury St	455.0
School St, Province St	448.0
Church St, Brattle St	387.8
Tremont St, Seaver Place	320.0
Back St, Berkeley St	320.0
Washington St, Hayward Place	312.0
Main St	310.5
World Trade Center Road	300.0
State St, Congress St	286.0
Cambria St, Boylston St	270.0
Massachusetts Ave, Brookline St	250.8
Park St, Beacon St	250.0
Brattle St, Massachusetts Ave, JFK St	243.8
Chester St, Elm St	238.0
Wadsworth St, Madison St	237.6
Cambridge St, Tremont St	203.1
Boylston St, Tremont St	180.9
Huntington Avenue	156.0
Atlantic Avenue, Congress St	135.0
Berkeley St, Newbury St	130.0
Massachusetts Ave, Essex St	116.0
Blackfan St, Longwood Ave	115.0
Farnsworth St, Congress St	114.0
Drydock Ave, Tide St	102.0
Congress St	91.0
Bedford St, Kingston St	84.0
Washington St, Union Park St	80.0
Highland Ave, Central St	78.0
Franklin St, Pearl St	65.0
Ellery St, Broadway	64.0
Florence St, Louise Road	57.8
Province St, Bromfield St	56.0
Stuart St, Warrenton St	52.5