Erik Brakke, Tyler Waltze
CS 591: Data Mechanics
Lapets
5/4/16
Final Report

The aim of our project was to take some of the data that we gathered about the state of roads in Boston, and allow a user to analyze a route of their choice. In order to do this, we decided to simplify our dataset and just use the pothole data we gathered from the initial project we did. Other datasets, such as motor vehicle accidents and construction zones, were collected and normalized though and could easily be included.

In order to approach this, we wanted to come up with an idea of an average number of potholes on a route in Boston. To gather this information, we picked about 500 different start and end points (latitude, longitude) that fell within the boundaries of the pothole data we had. With these starting and ending points, we then query Google's map API to get bicycle directions. Google Maps returned between 1 and 3 different routes for each start and end point supplied. These routes were stored in a collection as a list of steps.

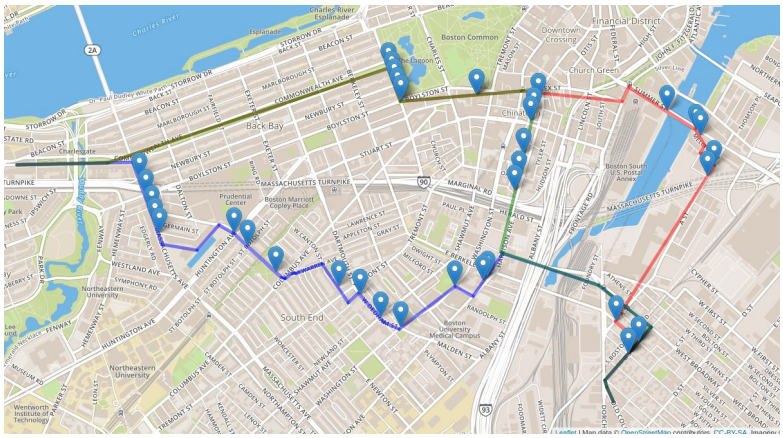
Then, for each step, we would query our own collection of potholes in the city and count how many potholes were within 3 meters of this line segment (using a formula to calculate the distances from a point to a line). Summing up all of these counts gave us the total number of reported potholes on a given route.

Afterwards, we calculated the average number of potholes per meter on a ride. Using this number as a baseline, we could then run our pothole finding algorithm on any bike route through Boston and in the end compare its average number of potholes to the average we calculated. The issue we had with this method is that we lacked data for Brookline and Cambridge. Given that some of the random routes would travel through those areas, the results of this were skewed. Given sufficient data for surrounding cities or a more intelligent random route selector, this issue would resolved.

We also implemented a heat map of different types of events around Boston. These events were things that we felt would be relevant to a biker (potholes, construction zones, vehicle accidents). This was easy to do as we had all of the latitude and longitude data for these events, so we just had to plot them on the Leaflet map. It was interesting to compare this with the Strava heatmap for bike rides around the Boston area. We would see some similarities between the concentrated areas of events we chose, and the cold spots on the heat map. This tells us that we can learn a lot about

what kinds of things bikers might care about through this city data. However, this is not an all telling data set. Bikers may not care about potholes, or construction zones, and may actually care about other events that we may not have taken into consideration. It would be nice to expand this project to create a customized heat map so bikers can get a better sense of which areas to avoid.

We implemented a simple web application for a user to input two location points and then run the suggested Google Maps routes against our algorithm and our average data set. This provides a user with a visual of where on the route all of the potholes are, as well as some indication of how much better this route is than the average route in Boston.



It is also important to note that our baseline dataset is run on completely random data point. It would be better to store each user's starting and ending points into our collection, as well and have our average continue to update as people used the app. Also, our baseline cannot take into account the fact that many people will just stick to the bike path, unless google maps suggested to use the bike path. This may weigh roads that are hardly traveled by bikers unfairly, but again, this could be solved by continuously updating the average.

Originally the idea was to assign absolute values to different routes, which we labeled as "danger levels". This value would be determined by the number of incidents on a given route (potholes, accidents, construction, etc). Each type of incident would be given a value. This value could fluctuate within a given incident type, for example an older pothole would be considered more dangerous than a newer pothole. We realized after the fact that ultimately these values were being arbitrary assigned by ourselves, and we had no hard data to claim that a pothole was deserving of a value of 5, while a construction zone would be 15. A seemingly better proposition would be to, given a

route traveled by an individual, take its start and end points and feed them into Google's map API. With the routes returned, compare it to the actual route the individual took. Assuming it is similar to one of the routes Google suggested, find points where the user's route differed from Google's. Then, at those points look for any possible incidents. From there you can attempt to calculate which incidents are more likely to cause a person to diverge from a route, and thus which incidents are more impactful relative to other incidents.

The web app is currently very simple, but it can easily be hooked up to any mongo data set. All one would have to do is use our program for finding events on a route to extend this project. It will be included in our project repo and is open for anyone to expand upon.

The takeaway from this project is that even with this simple data that the City of Boston has collected, we can begin to answer some very interesting and helpful questions. As more cities begin to install sensors into their infrastructure and make this data public for everyone, then we may have an even greater impact on people's day to day lives.