CSCI 491/591 P2

Group 4
Chenglin Fan, Jici Huang, Angelica Davis, Peng Zou

March 1, 2016

In this project our team is exploring a series of data sets comprised of GPS trajectories collected from taxi 'trips' around various major cities across the globe. In the award winning paper, "Efficient Map Reconstruction and Augmentation via Topological Methods" by Wang et. all, these GPS trajectories were processed using two critical topological techniques in order to construct an accurate image of the road networks of these major cities. Much of the strength of this paper comes from the reconstruction process's robustness against noise. The end goal of our project is to deeply understand the topological concepts and techniques used in the paper, and to add to our skill sets a particularly powerful data processing technique against multiple kinds of noise that are common in real data sets.

For this deliverable, we have selected an interesting data set, the GPS trajectories of Berlin, among several others that we would like to explore in the project. In this document we will provide the information on where the data set is from and explain the meaning of the collected data. We will also conjecture what we might find in the exploration of the data set in addition to digging into the map reconstruction processes.

Data Set Selection

As mentioned above, the data set that will be explored in depth for this project is the collection of GPS trajectories taken from the German city, Berlin. The team decided it was a suitable representative data set since the size of the data is about midway between that of Chicago and Athens (two of the other trajectory data sets). The size of data set is a reasonable characteristic for comparison because each of the data sets represent overlapping trajectories as well as noisy trajectories. Thus, a mid-sized data set may well represent the noise ratio (either due to noisy trajectories or the lack of sampling uniformity).

Details About the Data

The Berlin GPS trajectories dataset was downloaded from the Map Construction Portal (`http://www.mapconstruction.org/data_downloads/`). The GPS trajectories of Berlin data set contains the trajectory data of taxi trips in Berlin. Here the trip is a collection of GPS trajectories that describe one path traveled over some time period. The trajectory data has more than 27,000 trips in .txt format.

Each trip file consists of a series trajectories each represented by an x-coordinate, y-coordinate and a time stamp corresponding to the x, y-coordinate location at that particular time. Table 1 provides a small sample of a trip file taken from the trajectory data of Berlin.

Table 1: An excerpt of trajectory data of Berlin in a trip file

x	y	Time stamp
393742.586772	5821049.184616	2585542.00
393747.949682	5821296.284551	2585604.00
393883.091662	5821448.203015	2585677.00
393759.343945	5821821.259046	2585738.00
⋮	⋮	⋮

Exploration

The items we will explore in this data set are as follows:

- Estimate the underlying road network by using the trajectory data.
  As described above, the trajectory data is a compilation of taxi trips around a major city. As a real world event the complete path of a taxi trip is unpredictable, and the complete data set is comprised of overlapping trajectories. With this in mind, the raw data could be imagined as a monochromatic drip painting of the city. Since the web of trajectories stretches out in any direction with virtually indistinguishable overlaps, it will not be immediately clear from the trajectory data where the distinct roads occur. Therefore it will be necessary to apply some simplification to introduce more clarity into the image. This will be done by analyzing the most topologically persistent features, and removing the data not associated with these features.

- Find the roads with heavy traffic, namely the roads with high density of trajectory. We also want to analyze the topological structure of the heavy roads in the city.

Conclusion

In this project, our group will explore the GPS trajectory data of Berlin in Germany. We aim to construct an image of the road network using this data. Since the collected data is real, its raw form has little organization that would be immediately helpful in identifying distinct roads. Therefore we will be applying the topological techniques used in the map reconstruction paper. From this process, our group hopes to gain a deeper understanding of these topological concepts and at the same time learn how to process real, noisy data sets.