CSCI 491/591 P6

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

April 28, 2016

In this project our group is exploring a data sets comprised of GPS trajectories collected from taxi 'trips' around various major cities across the globe, such as Athens in Greece, Berlin in Germany, Chicago in US, and Beijing in China. In the award winning paper, "Efficient Map Reconstruction and Augmentation via Topological Methods" [**?**], the GPS trajectories of major cities were processed using two critical topological concepts, persistence [**?**] and Morse Theory [**?**], in order to construct an accurate image of the road networks of these major cities. The ultimate goal of our project is to deeply understand the topological concepts and techniques used in the paper to analyze GPS trajectories, and to add these data processing techniques to our skill sets, since they have been shown to be effective against a variety of noise that are common to practical and experimental data.

Our group selected the GPS trajectories data set of Berlin, among several others that we have explored in the project. The Data Set Selection section provides information on where the data set is from, and several reasons why the particular data set was selected.

In the second part, we show the several steps in the map reconstruction process, which includes the pre-processing of trajectory data, the density field calculation, ridge extraction, and simplification of the road network. Then, we illustrate the visualization of results we got in this project, including the visualization of trajectory data, the resulting map from extracting the ridge of the terrain, and the map after persistence simplification. Finally, we give the conclusion, and possible improvements to work on in the future.

Data Set Selection

In recent years, the amount of Volunteered Geographic Information (VGI) data that is publicly available has grown very fast, and the computational map construction from these GPS trajectories has attracted great attention. [**?**] The data set selected for this project was the VGI trajectory data of Berlin, Germany. The items below discuss why the Berlin trajectory data set was chosen.

First, GPS trajectory information has many very practical uses, therefore there is much value in knowing how to process and analyze data sets of this type.

Second, VGI data sets tend to be massive in size. The data set selected for Berlin contains a moderate number of trips and GPS samples per trip when compared to data sets of other cities. This makes it a good candidate for efficient processing given the time constraint, since the input data set size has a significant impact on the running time of the algorithm. Additionally, it contains enough densely overlapping trip data for the algorithm to produce a road network estimation with

a degree of confidence (this will be explained more thoroughly in the steps section).

Third, the data sets are all open source. The data set is accessible via the Internet, and can be downloaded from the open source website conveniently [**?**], and many researchers [**?**] use this data for research, hence it is easy to compare between the results of our experiment and the results of other people's work.

Details of the Data

The data set of Berlin GPS trajectories was downloaded from the Map Construction Portal [**?**]. The GPS trajectories of Berlin contain the trajectory data of taxi trips in Berlin. A trip is a collection of discretely sampled GPS trajectories that describe one path traveled through the city over some time period. The trajectory data has more than 27,000 trips in .txt format.

Each trip file consists of a list of trajectory data represented by an x-coordinate, y-coordinate and time stamp. Table 1 provides a partial sample of a trip file taken from the trajectory data of Berlin.

Table 1: An excerpt of trajectory data of Berlin

time-stamp	x	y
2585542.00	393742.586772	5821049.184616
2585604.00	393747.949682	5821296.284551
2585677.00	393883.091662	5821448.203015
2585738.00	393759.343945	5821821.259046

Figure 1: A visualization of discretely sampled points of the trajectory data of Berlin, as they are collected by taxi drivers' GPS systems. The points are actually physically far apart. We set the scale to 10^5 for x axis and 10^6 for y axis, which enable us to see the big picture of the points with the scale.

Steps to Reconstruct the Map

Our group uses the map reconstruction algorithm [**?**] to estimate the Road Network, which includes converting input trajectories to a density map, and extracting the ridge of the density map; the steps are as follows.

Density Field Construction: Discretize the region of city into a grid where each grid cell has side-length r. Each input trajectories is a sequence of discrete sampled points, let P $= \{p_1, p_2, p_3, ..., p_n\}$ be the collection of discrete sampled points from the trajectories. Compute the density $\rho(g)$ for each grid point g (g_x, g_y) based on the density function in paper [**?**].

Ridge Extraction: The set of points $\{(g_x, g_y, \rho(g))\}$ in $\mathbb{R}^3$ is a discrete terrain, compute the saddle points in the terrain, where saddle point of terrain is a stationary point but not an extreme point. Then we compute the manifold $S(p)$ for each saddle point p, where $S(p)$ is the collection of points of the terrain such that for each point q in $S(p)$, there is a monotone descending path from q

(a) The visualization of the density field in two-dimension.

(b) The visualization of the density field in three-dimension.

Figure 2: The figure demonstrates the density distribution of trajectories. It is the visualization of the density field after calculation, the very high density regions are colored in green or blue, the high level density regions are colored in red, and the low density regions are colored in dark red. The underlying main road networks are much clearer based on the density field map.

to p, and $S(p)$ is usually composed of discrete curves (each of them is from a local maximum point to p).

Persistence Simplification: The persistence is used to measure the scale or resolution of a topological feature. Compute the persistence [**?**] for each manifold $S(p)$ (whose projection is a piece of road), the persistence value is equal the difference between the density of local maximal point and the density of saddle point. We remove all the manifolds (unimportant roads) whose persistence is smaller than a given value. At last we project the remaining manifolds onto the plane, and obtain the road network of city.

Experimental Results

Trajectory Data Visualization

Each trip of the trajectory data is composed of a series of discrete points. Figure 1 shows the discrete points of the trips in Berlin. We can barely figure out the main roads from the spread out points in Figure 1, nor in the map of the city. In Figure 1, some regions have higher density of the discrete points while other parts have lower density of the discrete points. But even the higher density regions are not connected to each other, it is impossible to get the map of the city directly without further processing.

Density Field Visualization

In map reconstruction, the first step is density calculation. In P2, we gave the details to compute the density $\rho(g)$ for each grid point g based on the density function [**?**]. We set the grid size be $10m$, and the parameter t (an input variable that roughly indicates the noise level) be $\sqrt{t} = 20m$

Ridge Extraction

The details of ridge extraction is given in P2. The motivation of ridge extraction is to obtain the skeleton of the map for the following reasons.
First, even the high density regions in density filed map have multiplicities that span a wider region than the singular, ground truth network of city. Hence, the direct use of high density regions is not suitable to obtain the map. Second, ridge extraction can detect roads with few trajectories passing them even when they are next to heavily populated roads, which would otherwise be easy to eliminate as noise, introducing unwanted gaps in the resulting road network. Last but not least, it is convenient for further simplification on the map later. Since the 'ridge' is found by calculating

the maximum, that is critical, points given the distribution of overlapping trajectories on the z-axis, this maximal curve also corresponds to the most confident estimation of the underlying road.

Figure 3 shows the map obtained by extracting the ridge (we use a little extra operation to add the connection in the junction

Figure 3: The map is obtained by extracting the ridge of density terrain and projecting it in the plane. The underlying road network begins to emerge.

Persistence Simplification

Figure 4: The simplification of the map by removing all unimportant roads whose persistence are not greater than a threshold of 0.5. The main roads in the city are displayed.

We compute the persistence [**?**] for each road, and remove all unimportant roads whose persistence are not greater than a given value. By applying persistence filtration based on the estimated density, we obtain a simplified road network estimation. Since the previous step of ridge extraction identifies the regions of highest confidence, this simplification step acts to filter out the regions of lowest confidence. Figure 4 shows the important roads after simplification. It also shows the connectivity of the map is decreased after simplification. Our group will further improve the performance of simplification.

Conclusion and Future Work

Two important topological concepts used in this project are discrete morse theory and persistence. Morse theory enables one to analyze the topology of a manifold; by extracting the ridge, we can capture the skeleton of the density terrain of grid points. The main advantage of this is to obtain an estimated road with some degree of confidence based on the distribution of trajectories over a grid region. This identification of the most confident estimations is an effective way to use the high amount of overlapping trajectory data to create an estimation of the underlying road network that resembles a mode average of the overlapped trajectories, which smooths out some of the noise introduced by so many different trajectories. The purpose of persistence as applied in the algorithm is to further identify the features in the estimated road network that are the most 'true'. Given the degree of confidence as calculated by the density function, this step filters out the estimated roads that may not be as accurate as those which are associated with higher confidence scores.

There are several aspects that can be improved in this project in the future. First, the final visualization of the map can be improved by using plug-in components of Java Additionally, persistent homology is not the only factor in determining the accuracy of a particular road; we can combine the other features such as the length of road with the density of road, which was seen to improve the accuracy of the map after simplification in *et al.*. It would also be interesting to explore how to support more functionalities based on the current framework, such as returning a potential path for an input query region.

References