Visual Analysis of Route Diversity

He Liu*  Yuan Gao†  Lu Lu ‡  Siyuan Liu§  Huamin Qu∥  Lionel M. Ni¶

The Hong Kong University of Science and Technology

ABSTRACT
Route suggestion is an important feature of GPS navigation systems. Recently, Microsoft T-drive has been enabled to suggest routes chosen by experienced taxi drivers for given source/destination pairs in given time periods, which often take less time than the routes calculated according to distance. However, in real environments, taxi drivers may use different routes to reach the same destination, which we call route diversity. In this paper we first propose a trajectory visualization method that examines the regions where the diversity exists and then develop several novel visualization techniques to display the high dimensional attributes and statistics associated with different routes to help users analyze diversity patterns. Our techniques have been applied to the real trajectory data of thousands of taxis and some interesting findings about route diversity have been obtained. We further demonstrate that our system can be used not only to suggest better routes for drivers but also to analyze traffic bottlenecks for transportation management.

1 INTRODUCTION
Driving route suggestion is a key feature of GPS navigation systems or online maps and is used by people every day. Once a user chooses the source and destination of a trip, and optionally) the departure time, one driving path can be automatically generated and suggested to the user based on various criteria such as distance and travel time. Recently Microsoft have developed a smart driving direction service called T-drive to suggest practically the fastest route to a destination at a given departure time. T-drive is based on the observation that taxi drivers are more likely to know a city and traffic best and thus are able to choose a route that can avoid congestion and reach a given destination in the shortest time. To discover the routes taken by taxi drivers, the trajectories of a large number of taxis are recorded and analyzed. The real system prototype of T-drive is built based on the trajectories of 30,000 taxis in Beijing over a period of 3 months. The results are quite promising and on average the suggested paths can save 16% of travel time.

However, in the real-world, taxi drivers may have multiple ways to reach the same destination, which we call route diversity. Drivers choose routes based on the traffic, road conditions, or customer preferences. The travel time also depends on a driver’s skills and familiarity with an area. Furthermore, traffic conditions may change over time. If more people take the routes suggested by T-drive, congestion may result, making those routes not optimal anymore. Therefore, by showing the trajectories with detailed statistic information about route diversity, each available route could help suggest better driving routes and help monitor the real traffic conditions.

Route diversity is also a key issue for transportation management and urban planning. More routes exist for some source/destination pairs than others. Some roads become bottlenecks because there are less alternatives for drivers. Knowing diverse routes and the importance of some roads for different source/destination pairs will greatly help transportation management and urban planning. Given the road network data, people can compute the road diversity for any given source/destination pairs. However, the results may be quite different from the real diversity derived from taxi trajectories. Some routes may be impractical or obstructed, therefore drivers familiar with an area would avoid them. Thus, it is important to analyze route diversity from real trajectory data.

In this paper we propose a trajectory visualization method that examines regions where diversity exists. Source/destination pairs with highly diverse routes are highlighted for further examinations. We also develop several novel techniques to visualize both real-world trajectories and statistics associated with different routes for diversity analysis. Through case studies, our visualization reveals trips of source/destination pairs of high diversity, and the different routing preferences of taxi drivers at different times. We further analyze the traffic conditions to discover what time periods certain bottleneck conditions occur.

The major contributions of the work are:
- We develop a comprehensive visual analytics system to study route diversity based on real trajectory data. To the best of our knowledge, it is the first system developed to visualize route diversity.
- We develop several novel visual encoding schemes to display the statistic information for different routes and reveal the importance of each road for different trips. Our system provides an intuitive way to compare and evaluate different routes.
- We demonstrate through case studies that our system can facilitate traffic analysis and route suggestion.

2 RELATED WORKS
In this section we summarize some relevant research work in navigation systems, geographical visualization, trajectory visualization, trajectory pattern analysis, and graph visualization.

Navigation systems Microsoft T-drive [29] makes recommendation of fastest paths taken by taxi drivers. The paths are computed based on historical trajectory data. It exploits the knowledge of taxi drivers who are more experienced and know a city’s road system and traffic situation well. The in-field evaluation have proved that over 60% of the paths given by T-drive are faster than those suggested by traditional methods. Our system can be used together with the T-drive system to better analyze different routes taken by taxi drivers.

Geographical visualization Geo-visualization provides interactive visual tools for exploration and analysis of data with geographical information. This is a broad and extensively studied field and we only summarize a few representative papers. Necklace Maps [24] project thematic mapping variables onto intervals on a curve that surrounds the map regions. Zhao et al. [30] presented Ringmap that visualizes multiple cyclic activities over

---

* e-mail: liuhe@cse.ust.hk
† e-mail: yuan@ug.gatech.edu
‡ email: ilu@ust.hk
§ email: siyliu@ust.hk
¶ email: huamin@ust.hk
∥ email: ni@ust.hk

IEEE Symposium on Visual Analytics Science and Technology
October 23 - 28, Providence, Rhode Island, USA
978-1-4673-0013-1/11/$26.00 ©2011 IEEE
time while preserving geographical information. Our system is inspired by the Ringmap and adopts a similar circular layout design to Ringmap [30] but encodes different attributes for a different application. Worldmapper [10] distorts the shape of the countries in a map such that the area is proportional to given scalar values. As we use a force-based method to distort the layout of graphs, the underlying map can be distorted accordingly using a scheme similar to Worldmapper. Butkiewicz et al. [5] used probes for selection and comparison of multiple regions. Fisher introduced hotmap [11] to represent aggregate activities and show users’ attention to the map. Hotmap is also used in our system to convey the diversity scores of different regions. Chang et al. [6] presented Legible cities to display large collections of data for urban context with different levels of abstractions. Wood et al. [28] discussed the geovisualization mashup techniques including tag clouds, tag maps, data dials, and multi-scale density surfaces for visual analysis of a large spatio-temporal dataset. We also integrate similar interactions and mashup techniques into our system to facilitate analysis tasks.

**Trajectory visualization** Andrienko et al. [3] summarized the approaches in visualizing movement data. Characteristics of movement data and methods to present dynamics, movements, and changes are discussed. Movement data is further classified into three types [2]: single object movement, multiple unrelated movements, and multiple related object movements. The corresponding approaches to visualizing the three types of movement are also discussed in [2]. GeoTime [17] displays a 2D path in a 3D space to provide a detailed view of the geographical and temporal changes in movement data. However, occlusion might become a big issue for a large number of paths in GeoTime. Guo et al. [13] presented a trajectory visualization tool that focuses on visualizing traffic at a road intersection. The spatial and temporal views are separated and the user can interactively explore the movement patterns of the trajectories. In our system, we adopt a 2D display similar to the one in [13]. The geographical attributes and temporal attributes are also shown separately to avoid occlusion. But two systems deal with different problems and data sets. In a proximity-based approach [7], the raw position is transformed into an abstract space such that the geographical information is transformed into meaningful multivariate data. Usually occlusion is a problem when displaying thousands of trajectories, therefore there are several ways to aggregate or cluster the trajectories to improve visual quality. One method is to aggregate the trajectories after dividing the temporal and spatial spaces into time intervals and compartments [1] [22]. Relevant data and features can be extracted from the database with both temporal and geographic aggregation, and similar trips with similar routes or start-end points are clustered, combined, and summarized to discover places of interest [4]. Hoferlin et al. [14] proposed schematic summaries as a novel approach to reduce visual clutter by trajectory bundling. In contrast, the goal of our system is to examine different routes taken at different time periods and thus it is important to show all possible routes instead of clustering them. However, if visual clutter becomes a serious issue, the above trajectory clustering and bundling approaches can be adopted to allow hierarchical exploration of route diversity at different granularity levels. Herter et al. [15] visualized aircraft trajectories. The system supports the display of multiple trails and the altitude of each aircraft. Willems et al. [27] visualized vessel movement as well as the vessel density along traces by convolving trajectories with a kernel moving with the speed of the vessel along the path. Nanni et al. [20] presented a density-based approach that has both high coverage and purity of achieved clusters, and is also resistant to noise and outliers. These methods can achieve impressive visual effects and can be used to enhance the trip views in our system. Besides geographical data, financial market data can also be plotted as trajectories [23], and those trajectories are clustered by unsupervised clustering algorithms for analysts to study.

**Trajectory pattern analysis** Giannotti et al. [12] defined trajectory pattern as frequent behaviors in space and time, and discussed several approaches for mining trajectory patterns. Pelekis et al. [21] classified trajectory similarity into two major types: spatio-temporal similarity and temporal similarity, and distances between two trajectories are defined for the two major types and their variations for trajectory similarity search. Vlachos et al. [26] showed an interesting similarity measure which is based on Longest Common Subsequence, and the algorithm allows stretches in both space and time. Tietbohl et al. [25] proposed an approach to discover stops in trajectories. Trajectory location points are grouped by neighborhood and the time duration around the point is used to judge whether the object is stopped or on the move. In other words, the speed of the movement is used to find unknown stopping points. There have been some studies in traffic trajectories for urban planning. Liu et al. [19] studied inhabitants’ mobility behaviors through traffic records of buses and metros as well as trajectories of taxis. The temporal movement of metro rides reflects the geographical information and the pendulum of the people’s daily mobility. In another paper, Liu et al. [18] analyzed the operating behaviors of several top-income taxi drivers in order to find the underlying intelligence behind their high income. The trajectories and temporal changes to the operating regions of those selected drivers are analyzed. They discovered that top drivers move to regions with better traffic conditions in time, and plan their route route to make the maximum profit. In this paper, we study a quite different problem, trajectory diversity, which has not received much attention of researchers.

**Force-directed graph layouts** Force directed layouts are both flexible and easy to implement, and thus are widely used in practice. Kamada and Kawai [16] developed a popular algorithm which attempts to match the Euclidean distance to the length of the shortest path between the vertices. Davidson and Harel [9] took edge lengths, vertex distributions, and edge crossings into consideration and developed a better but rather costly graph layout algorithm. In our system, we also exploit force-directed layouts to solve occlusion and reduce visual clutter.

### 3 System and Data

In this section, we provide an overview of our system and also discuss the data used in the experiments.

#### 3.1 System Overview

Our system has three major components: the global view, the trip view and the road view. The global view shows the overall diversities of all important hotspots in the city, suggesting some interesting source/destination pairs and locations for further exploration. The trip view displays the trajectories with both geographical and statistical information for a selected source/destination pair, allowing users to analyze different routes and performances at different times. The road view visualizes the statistical information of all trips passing through a given road, including the speed, time and distance of each trip.

Figure 1 shows the flow chart of our system. Users first start with the global view. After feeding trajectories to be analyzed into the system, an overview of traffic flow and diversity is displayed in the global view. Users are free to explore any interesting spots and choose one as the source, and then all the destinations for trips starting from the source and the corresponding route diversity for each source/destination pair will be displayed. After that, users can select some interesting destinations for further investigation. When both the source and destination locations are fixed, all the trajectories from the source to the destination and the associated statistical information such as the travel speed, time, and distance of each route will be displayed in the trip view. If a road segment in the trajectory is interesting, users can open the road view to analyze all
Figure 1: The system overview. Our system consists of four components: 1) Global view which shows the overall route diversities of hotspots in a city; 2) Custom layer which distorts the global view to give more space to regions with high route diversities; 3) Trip view which displays the trajectories for a given source/destination pair as well as their associated spatial and temporal attributes; 4) Road view which visualizes the statistical information of all trips passing through a given road.

3.2 Data
The data used in this study is the trajectory data of over ten thousand taxis in Shanghai, China. Each GPS record contains the latitude and the longitude of the taxi, the date, the time of the day in seconds, the taxi’s status (occupied / vacant), and the speed of the taxi. After the data is sanitized by a map-matching algorithm, we are able to get the valid trajectory data of more than four thousand taxis. The diversity scores for each spot and each source/destination pair, as well as the statistical information for each route and road segment are computed as pre-processing.

4 Global View of Route Diversity
In this section we describe the global view that presents the route diversity of multiple source/destination pairs and provide visual cues to the user about where route diversities exist.

4.1 Design Rationale
In the global view we first want to show the hot spots of a city. Some locations may have more vehicles passing by than other locations. The aggregated traffic flow is used to compute the hotness (i.e., the number of vehicles) of each location and then a heat map is used to present the hotness of different locations. Then we need to display the route diversity score for each source/destination pair. For a given source/destination pair, the route diversity score measures the total number of different routes taken by drivers. In addition, we want to present aggregated diversity values for each location which measures the total number of different routes starting from or entering this location. Similar to the incoming/outgoing degrees of directed graphs, we use the incoming and outgoing diversities to measure the total number of routes leaving from or entering a location. The diversities are computed for a given time window. For some time windows, the incoming diversity and outgoing diversity might be different.

If we want to show the diversity score for each source/destination pair, we will face a serious visual clutter problem. The most natural way to represent the diversity score is to use a graph with the nodes representing locations and edge thicknesses encoding the diversity scores. However, if all the pairs and their diversity scores are plotted, the display will be visually cluttered and users would hardly be able to identify any patterns. In this case, some clutter reduction would have to be done. Clutter reduction methods based on trajectory clustering and edge bundling might not work for our task as the goal of our system is to show all possible routes. To address this issue, we introduce an encoding scheme which provides different layers of information.

4.2 Visual Encoding Schemes
To avoid visual clutter and provide information at different levels of detail, we design an interactive visualization framework which consists of three information layers.

Heat Map Layer First, we need to identify hot spots in which higher percentages of vehicles travel around. We rasterize the city map into pixels. Then we compute the total number of departing and arriving vehicles for each pixel. Then we use the heat map to reveal the hotness (i.e. the total number of vehicles) of pixels over a 2D map. Figure 2 shows the heat map layer with red areas representing regions of high densities of vehicles while white areas represent regions of relatively low densities. Users can use the heat map to choose some interesting locations for further analysis.

Location Layer We further group similar pixels into locations. We adopt a single-link clustering algorithm that iteratively merges the two closest pixels with similar hotness values. We set a threshold (e.g., two kilometers) for the maximum size of each cluster. Each cluster is called a location. The total number of locations is then much less than the number of pixels in the heat map layer. For each location we calculate the outgoing diversity which is the total
Diversity score of all pairs starting from this location, and the incoming diversity as the total diversity score of all pairs ending at the location. Then we use a node to represent each location with the node size representing the total diversity score.

To calculate the diversity score for a source/destination pair, we first find all trajectories starting at the source and ending at the destination, and then group the trajectories into different routes. If the points from one trajectory are close enough to an existing route, we treat them as the same route. Otherwise, they will be treated as two different routes. The final diversity score $D$ can be represented using a statistical entropy formula:

$$D = -c \sum_{i} \left( \frac{N_i}{N} \ln \left( \frac{N_i}{N} \right) \right)$$

Where $N_i$ is the number of trajectories in the $i$th route and $N$ is the total trajectories number for the pair, and $c$ is a constant(e.g. 10).

Then we use a graph to encode locations and the route diversity scores for location pairs. The size of a node encodes the node’s aggregated diversity score. For route diversity analysis, edges with high diversity scores are more important, and thus they are rendered with high opacity to make them more visible. Figure 3 shows the location layer.

To further reduce edge overlapping, a force-based layout is employed. Let $F$ be the net-force exerted on each node representing a location. Then $F$ can be computed using the following formula:

$$F = F_e + F_{anchor} + F_c$$

Where $F_e$ is the elastic force keeping the original length of each edge. The second term $F_{anchor}$ keeps each node close to its original position. $F_c$ is the Coulomb repulsion between locations which prevents nodes moving together. The electronic charge of each location is proportional to its diversity score such that locations with high diversities will push other nodes further away to reduce overlapping. In addition to the net-force of each location, edges that cross one another will generate an extra repulsive force to prevent them from overlapping.

One drawback of the force-directed layout is that nodes may move far away from their original locations and users may have problem to know the locations they represent. There are two possible solutions. First, we only show the force-directed layout but insert a label near a node to show the location name of the node. The second solution is to distort the underlying map to provide extra geographic information. We can first triangulate the nodes with Delaunay triangulation. Each node will be attached to the underlying map. After the node positions are adjusted, the underlying map will be distorted accordingly using texture mapping. To avoid edge flipping, an extra force to penalize edge flipping can be added.[8].

Custom Layer After users choose some places in the location layer, more detailed information about the diversity scores and the traffic flow distribution can be displayed in the custom layer.

The custom layer is shown in Figure 4. Nodes can be represented by circles or squares. Each node consists of an inner part and an outer part. The size of the inner part (shown in white) represents the incoming diversity while the size of the outer part (shown in black) represents the outgoing diversity. Though circular nodes might be more aesthetic, square nodes allow more precise measurement and comparison of the incoming and outgoing diversities. The edge with the greater diversity will drag the inner ring towards it. The offset of the inner ring from the center is calculated as following:

$$Offset = (\text{Radius}_{\text{outer}} - \text{Radius}_{\text{inner}})(\sum_{A} \text{Diversity}_{A} / \sum_{B} \text{Diversity}_{B})$$

Where Diversity is a vector starting at the location and pointing to each destination with a length proportional to the diversity score of the pair.

The diversity of each pair is represented by the width of the yellow lines. The width of the gray part of each edge represents the traffic flow. There is an indicator $C$ in each edge. The ratio of $|AC|$ to $|BC|$ encodes the ratio of the diversity score from $B$ to $A$ over the diversity score from $A$ to $B$.

4.3 Interactions

For effective diversity exploration, the global view supports rich user interactions such as selecting, dragging, and position binding.

Selecting In the location layer, the user can select a node or an edge by simply clicking on it. When the user presses the mouse and brushes an area, all the nodes inside that area are selected. When the user selects a node, a custom layer is generated to display the selected node, all adjacent nodes, and edges. If the user selects an edge, all routes between the nodes of the edge will be computed and displayed in the trip view.

Dragging The user can drag a node to change its position. This function is especially useful when there are too many nodes and the visual clutter problem cannot be solved with the default force directed model.
Figure 5: The trip view shows all the major routes for a given source/destination pair with the trajectories shown in the inside central area and the temporal statistical information encoded in the circular outside area. The numbers show the times of a day. The bars encode the total numbers of trajectories recorded at different time periods. Each trajectory is shown as a circular trace with two endpoints corresponding to the start/finish times of the trajectory. The color map for speed is also used in other figures.

Position Binding The user can choose to bind a node to its original position by pressing a key. As the force-directed layouts change the locations of nodes, their geographical information may be lost. This function helps users to observe the original position of a node.

5 TRIP VIEW OF DIVERSITY

In this section we describe the trip view that visualizes the trajectory information between a source/destination pair.

5.1 Design Rationale

We need to further analyze each route taken by drivers, thus, it is desirable to show both geographical and statistical information in one view. For each trajectory, we would like to display the departure time, the arrival time, the duration of the trip, the instantaneous speed, the average speed, and the geographical information. Since the traveling distance is indicated by the length of a trajectory, it is not explicitly displayed in our system. For each time period, we would like to display the number of taxis running in that time period, and also the speed of the taxis.

5.2 Visual Encoding Schemes

Figure 5 shows the trip view. In our visualization we have a circle with geographical information in the center, and statistical information around the outside of the central geographical circle in a 24 hour scale. As most of the statistical information is associated with time, we adopt a design similar to [30] and use a clock-like radial layout to show the statistical information distributed over time. The day starts from top, points to bottom at noon, and returns back to top at midnight when the day ends. Thus the time is shown on a circular axis, and variables such as trip speed and duration are plotted around the circle. We call the circle representing the durations of trips the duration circle, and the circle representing the speed of trips the speed circle. As speed is another very important attribute, we use color to encode speed. We use a continuous red-green colormap to encode speed. Red represents low speed while green represents high speed. The encoding scheme is intuitive since the red color will remind users of red light while green will remind users of green light.

Time The time of the day is displayed on a circle around the central circle like a clock. Each degree on the circle represents 240 seconds, while the total 360 degrees come to 24 hours.

Trajectory The trajectories between a given source/destination pair are plotted inside the central circle by their geographical locations. The default color for the trajectories is blue. Once a trajectory is selected, it is then colored by the speed of the trajectory.

Duration The duration of each trip is drawn in an arc between the speed circle and the time circle. The arc starts at the degree representing the corresponding time when the trip starts and ends at the time the trip ends. The arc is colored to show the average speed of the trip. As stated before, red means low speed and green means high speed.

Number of trajectories The numbers of trajectories in different time periods of a day can be represented by the bars on the speed circle. The longer the bar is during a time period, the more trips there are, which is the same as in typical bar charts.

Speed The instantaneous velocity is presented by colored round dots in the bar chart of the outer circle by time of the day in the form of 24 hours.

5.3 Interactions

Various user interactions are supported in our system to allow users to select interesting trajectories and study their patterns. There are two ways to select trajectories.

Brushing by location Inside the central geographical circle, users can press and drag the mouse to select an area. All trajectories passing through that area are selected. Selected trajectories are highlighted using special colors that represent the speed of the trips. Meanwhile the durations of the trips and the speed of the trips are also shown in the trip view. This part of brushing is for spatial analysis. For example, if there are two routes reaching the same destination, users may select the trajectories associated with...
one route and compare the patterns of the trajectories. Figure 6 illustrates the brushing by location feature.

**Brushing by time span** In the outer statistical circle, users can select an arc to specify the start time and the finish times. All the trip information within that specified time period is highlighted, including the trajectories in the geographical circle, the speed information in the speed circle, and the durations in the duration circle. This is for temporal analysis. For example, users may select a time period and see the routes taken in that time period.

**Union selection** The brushing by location also supports union operations. Users can add a group of trajectories to previous selections. If users find an outlier trajectory that should be removed from selection, they can erase that trajectory by union operation.

### 6 Road View of Diversity

The road view of the system shows the importance of the road and traffic conditions. The roads with less alternatives will be more important. For each road, we compute the route diversity scores for all location pairs which have routes passing through this road. If the diversity score is low for a certain pair, then the road is important for this pair, and vice versa. Figure 7(a) shows the road view.

#### 6.1 Design Rationale

Since the road view shows mainly two aspects of the road: importance and traffic conditions, we split the view into two parts, the upper part showing the importance of the road and statistical information associated with each route; the lower part showing the speed distribution which reflects the traffic conditions.

The upper part shows important information such as the distances between each source/destination pair, the route diversity value of each source/destination pairs and the percentage of trajectories from one source to one destination that use this road. If there are multiple routes from a source to a destination which means the route diversity score is high, and only a small portion of the trajectories use this road, this road is not that important. If only short distance trips use this road, the road may not be a major road and may not be suitable for long-distance trips as drivers may not be familiar with it.

The lower part shows the speed distribution of all vehicles passing through this road. The height of the bars represents the volume of traffic at that time on that road, which helps users to infer the peak times when the number of trajectories is relatively high. The color of the bars represents the speed of each trajectory, which helps users to identify the time when the traffic moves slowly. For example, if all traffic is moving at low speed in a certain time slot, we can infer that a traffic jam exists at that time. If the traffic is moving at low speed and the diversities of the source/destination pairs are low but the percentage of trajectories using that road is high, it usually means that this road is a traffic bottleneck.

#### 6.2 Visual Encoding Schemes

**Distances** The left and right axes represent source locations and destination locations respectively. The source and destination names of a route will be displayed if users use mouse to hover the route. The locations are sorted by the distances to the road. The locations in the upper part of the axis represent trips of short distances while the lower part shows long distances. This would help users identify the importance of the road for different types of trips (i.e., short-distance trips or long-distance trips).

**Diversity** The diversity of each source/destination pair is plotted in the center from bottom to top. The lower the polylines are, the fewer the routes drivers can use. The more evenly distributed the trajectories are on different routes, the higher the diversity scores. A number is displayed near each polyline to show the diversity score of this polyline. Therefore, if the diversity is low, the drivers have a limited choice of routes from the source to destination.

**Percentage of trajectories** Our system can also show the percentage of trajectories using that road from a source to a destination. For example, for a source/destination pair, the percentage of trajectories passing through a given road indicates the importance of the road. If 100% of the trajectories pass through the road, then this road is the only option for that source/destination pair. If most of the trajectories use the road but the diversity is high, the given road is popular and important but there are other alternatives.

**Time** Same as the trip view, the time of day is divided into 360 parts. The x-axis represents the time, starting from midnight to morning, and then to noon, afternoon, and finally midnight.

**Number of trajectories** The height of the bars below the parallel coordinates represents the number of trajectories using that road in a given time slot. Each trajectory is counted only once in each time slot.
Speed The color of the bar below the parallel coordinates encodes speed with red representing low speed and green representing high speed. From the height and the color of the bars, we can infer the distribution of speed and the traffic changes. Meanwhile, the polylines connecting a source and a destination are also colored by the average speed of the trips.

7 EXPERIMENTS
The data used in this study is the trajectory data of ten thousands taxis collected in Shanghai, China during an eight-month period. Each GPS record contains the latitude and longitude of the taxi, the date, the time of the day in seconds, the taxi’s status (loaded/vacant), and the speed of the taxi. The data is sanitized by removing erroneous trajectories that register impossible speeds, immediate location changes or large off-road distance errors. Valid trajectories are then stored in binary files to reduce storage space and processing time.

Our system is implemented using JAVA. The experiments are conducted on an Intel(R) Core(TM)2 2.66GHz Laptop with 4GB RAM and a NVIDIA Geforce GT330M GPU with 512MB RAM. The preprocessing to sanitize data of thousands of taxis in one week took about two hours. We were able to get the valid trajectory data of more than four thousands taxis. After the preprocessing, our system supports interactive real-time visual displays and user interactions.

In the following, we demonstrate the usability of our visualization through three case studies: diversity exploration, traffic monitoring, and route suggestion.

7.1 Case Study 1: Diversity Exploration
We first used our system to explore the diversity distribution and identify source/destination pairs with high diversity scores.

We generated a global view of the data. After observing the heat map layer in Figure 2, we found the high departure/arrival density areas that are displayed in red and yellow. After checking the geographical information of these areas, we found that they are in the downtown area. Then we added the location layer with global route diversity information to the heat map layer (see Figure 3) and checked the diversity distributions by observing location sizes and edge opacities. However, in the downtown area, the location nodes are too close and edges overlap too much. To reduce the visual clutter, we rearranged the layout using our force based model and the layout topology becomes more discernable (see Figure 8(a)). We closely examined the large location nodes with solid edges, which represent the locations with high diversities according to our encoding scheme.

We found that the location with the greatest diversity score has similar outgoing and incoming diversity scores. Figure 8(b) shows that the outgoing diversity scores are extremely unevenly distributed. The thickest edge appears to have a high traffic flow and high diversity score. Since the indicator on the edge is closer to the selected node, the diversity score is higher when the selected location acts as a source.

For comparison, we selected a source/destination pair with high diversity and another source/destination pair with lower diversity, and displayed their trip views respectively (see Figure 8(c)). The trip view highlighted in the red rectangle in Figure 8(c) shows that there are four routes existing for this pair. While the trip view highlighted in the blue rectangle shows that only two routes are available for the pair. For source/destination pairs with high diversity, there are more routes to choose. It poses challenges for GPS navigation systems and drivers. With the real trajectory data, we can further analyze each route and discover its advantages and disadvantages. There may exist multiple quality routes. Thus it might not be appropriate to always recommend the same route every time. Thus, route diversity exploration can serve as a start point for route suggestion.

For source/destination pairs with low diversity, there are not many choices left for drivers. But transportation bottlenecks may exist. We can use our system to analyze the traffic and identify potential bottlenecks. The findings can facilitate transportation management and urban planning.

7.2 Case Study 2: Traffic Monitoring
Our next case study shows that our visualization system can reveal traffic congestion and transportation bottleneck in some area.

Traffic congestion We explored the traffic conditions with the trip view. Figure 9 shows the trajectories in a small part of the downtown area. We found that except one outlier trajectory, all vehicles were at low speed on average. Very few trajectories had relatively higher instantaneous speed, but the instantaneous speed was still low compared with the speed in other areas. This indicates serious traffic congestion in this area. Meanwhile, no taxi driver took longer routes to get to their destination, so all of them endured the low speed. This may be due to the short distance between the source and the destination. Another possibility is that in the downtown area all other roads had similar low speeds at the same time period, and therefore it was better to wait than to detour.

Traffic bottleneck We used the road view to examine the traffic conditions of one road. The leftmost axis shows the sources of the trip, sorted by the distance to the road. The rightmost axis shows the destination of the trip, sorted by the distance from the
road. In the center, the trips are grouped by the diversity score of each source/destination pair.

From Figure 7(a) we inferred that this road was very important since for most source/destination pairs this road was the only choice as most trajectories in the center mostly showed low diversity. In addition, after examining the percentage of trajectories passing through the road, we found that for the same source/destination pair the percentage of all trajectories using this road is very high. Thus, this road was important for sources and destinations in both short and long distances. However, from the speed distribution graph below we found that almost all the trips during all the time periods were at very low speed, suggesting the existence of a traffic bottleneck. Most trips were taken during the night but the velocities were still low. Figure 7(b) shows the location of the road on the map (marked with a red circle) while all trajectories passing through the road are plotted in blue.

7.3 Case Study 3: Route suggestion

Finally, we demonstrate how to use our system to explore different routes between a given source/destination pair. Figure 10(a) and (b) show two routes available for a source/destination pair. We selected one route and then all the important information for this route would be displayed using our system.

Upper route We first selected the upper route in Figure 10(a). We found that all the trajectories were from the afternoon to midnight. We also observed that no one used the upper route before 4pm, and all trajectories were at high speed.

Lower route We then selected the lower route as shown in Figure 10(b). We found that from morning to midnight there were almost always taxis on this route. However, the speeds of the trajectories were normally not high. We also found there are more trajectories on this route than on the upper route, which shows the taxi divers preferences.

Morning After spatial analysis, we moved to temporal analysis. We selected the time interval from 7am in the morning to 2pm in the afternoon, observing taxis only using the lower route. From Figure 10(c), it is clear that most trajectories were at low speed. However, some trajectories showed relatively high speed at the end of the trip, showing that there was congestion on this route.

Evening We then selected the time period from evening to midnight to analyze those trajectories (see Figure 10(d)). We gained the insight that taxis used both the upper route and the lower route as shown in Figure 10(d). Most of the trajectories were at high speed, which shows that there was no congestions in the evening. However, there were still some trips at low speed, which may be due to taxi drivers’ driving skills and experiences.

Figure 10(e) and 10(f) show routes for another source/destination pair. The upper routes highlighted in 10(e) have longer distance but the average travel speed is faster. Most routes highlighted in 10(f) have shorter distance but the average speed is low. However, we also noticed that there are some trajectories (that zig-zag yellow ones in 10(f)) that travel longer distance at low speed. Some drivers may feel these detour routes could avoid traffic. But with the traffic flow analysis, we found they did not result in faster speed. Thus, these trajectories should not be recommended to drivers.

8 Conclusions

In this paper we have studied the problem of route diversity based on real taxi trajectory data and presented our visual analytics system to reveal the global diversities in a city, the route diversity for any given source/destination pairs, and the statistic information for any trip or road taken by taxi drivers. Some well established visualization techniques such as heatmap and force-directed graph layouts are integrated into our system. We also proposed several novel visual encoding schemes such as the trip view and the road view to reveal important spatial and temporal information associated with a trip or a road. We have demonstrated the usage and usefulness of our system for diversity exploration, traffic monitoring, and route suggestion by several case studies. The advanced visual analytics techniques combined with intuitive user interactions allow users to interactively explore the data and analyze the spatial and temporal patterns of route diversity. The novel visualization methods can also be used to analyze other spatial temporal data.

There are multiple avenues for future work. Our visualization system may not scale well for very large datasets. When there are too many trajectories in an area, visual clutter becomes a serious problem. Thus, we plan to integrate some advanced clutter reduction methods into our system. Currently we only show the time of day for each trip. However, there may exists correlation between route diversity and other factors such as seasons and weather. For example, some routes may not be preferred in bad weather or in winter. Thus, we will also show the information of the date and weather condition for each trip. Traffic analysis is a very complicated problem. It may be strongly influenced by events like social gatherings (e.g., sport games), road construction, and car accidents. We plan to take these factors into consideration and suggest routes for unusual or urgent situations. Finally, we want to conduct a formal user study and seek feedback from domain experts to further improve our system.

Acknowledgements

We would like to thank Baining Guo at Microsoft Research Asia who introduced the Microsoft T-Drive to us and inspired us to work on the project. We also thank the anonymous reviewers for their constructive and valuable comments. This work was supported in part by grant HK RGC GRF 619309 and an IBM Faculty Award.
Figure 10: Route analysis: (a) The upper route is selected. This route is mainly taken between afternoon and midnight. (b) The lower route is selected. This road is used from morning to night. However, the speeds of the trajectories were normally not high. (c) The trajectories from morning to afternoon are selected. Most taxis used the lower route but were at low speed. (d) The trajectories in the evening are selected. Both routes are used and most of the trajectories are at high speed. (e)(f) Routes for another source/destination pair.
REFERENCES


