Visual Exploration of Sparse Traffic Trajectory Data

Zuchao Wang, Tangzhi Ye, Min Lu, Xiaoru Yuan, *Member, IEEE*,
Huamin Qu, *Member, IEEE*, Jacky Yuan and Qianliang Wu

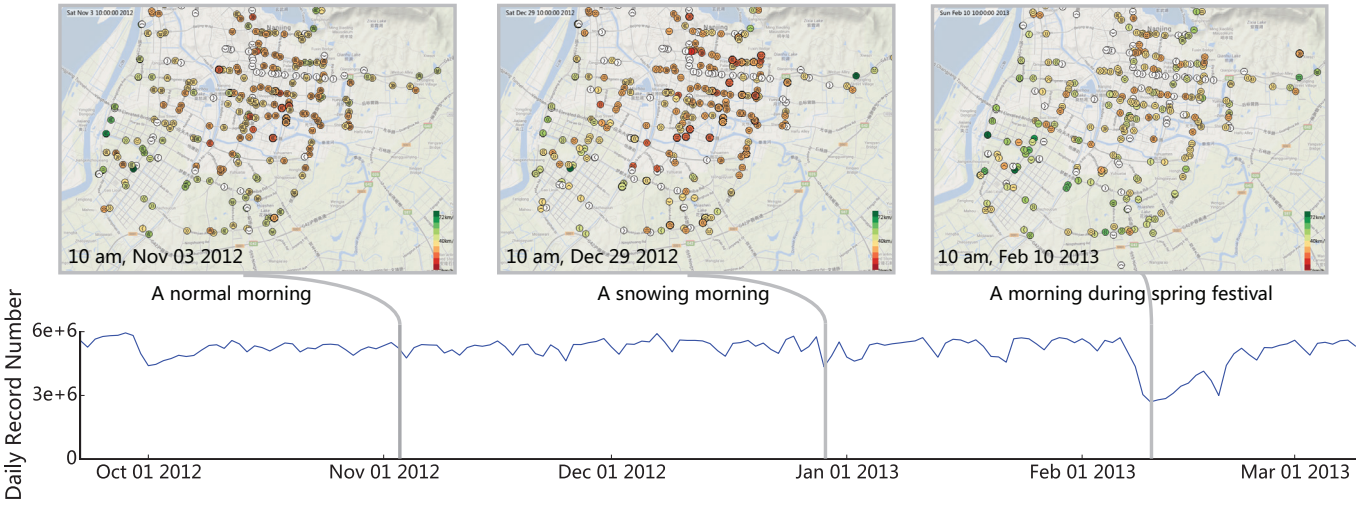

Fig. 1. Our sparse traffic trajectory dataset spans half a year. As shown by the line chart below, the daily traffic flow is rather stable, except during the spring festival. In a snowing morning, the transportation cells are redder, indicating high traffic load. In a spring festival morning, the transportation cells are greener, indicating low traffic load.

Abstract—In this paper, we present a visual analysis system to explore sparse traffic trajectory data recorded by transportation cells. Such data contains the movements of nearly all moving vehicles on the major roads of a city. Therefore it is very suitable for macro-traffic analysis. However, the vehicle movements are recorded only when they pass through the cells. The exact tracks between two consecutive cells are unknown. To deal with such uncertainties, we first design a local animation, showing the vehicle movements only in the vicinity of cells. Besides, we ignore the micro-behaviors of individual vehicles, and focus on the macro-traffic patterns. We apply existing trajectory aggregation techniques to the dataset, studying cell status pattern and inter-cell flow pattern. Beyond that, we propose to study the correlation between these two patterns with dynamic graph visualization techniques. It allows us to check how traffic congestion on one cell is correlated with traffic flows on neighbouring links, and with route selection in its neighbourhood. Case studies show the effectiveness of our system.

Index Terms—Sparse Traffic Trajectory, Traffic Visualization, Dynamic Graph Visualization, Traffic Congestion

1 Introduction

Transportation visualizations have been studied for many years. Researchers have designed methods to study various kinds of transportation data, including radar based vehicle counting data, taxi GPS data, subway IC card data, etc. Such visualizations help us gain insight into the complex transportation system.

In this paper, we focus on a new type of transportation data: sparse traffic trajectory data. Unlike traditional GPS data, our data encompasses the movements of nearly all vehicles in a city, not just taxis or buses. Therefore, it can provide accurate traffic statistics, such as flow volume on each cell and between each Origin-Destination (OD). These statistics are very precious in transportation modeling. However, our data is not recorded continuously as in GPS data. Rather, the data is recorded only when vehicles pass through transportation cells. Therefore, it is spatially sparse, and usually temporally sparse as well. The above differences make our data unique.

To analyse such data, our major challenge is to deal with the uncertainties caused by the sparsity. For example, the exact tracks on the links between two consecutive cells are unknown. The exact start and end locations of a trajectory are also unknown. Nevertheless, we consider movements in the vicinity of cells as of high certainty. Therefore, our first technique is to generate a local animation, which only shows movements near the cells. Following Andrienko et al.'s work [9], the second technique is to aggregate many trajectories, in order to compensate for the uncertainties in spatial and temporal coverage. In this way, we are able to study the cell status patterns (e.g. average speed pattern, congestion pattern) and inter-cell flow patterns.

Beyond studying these patterns separately, we propose to study their correlations. That corresponds to interesting domain questions such as how traffic congestion on one cell is correlated with traffic

- *Zuchao Wang is with Peking University. E-mail: zuchao.wang@pku.edu.cn.*
- *Tangzhi Ye is with Peking University. E-mail: yetangzhi@sjtu.edu.cn.*
- *Min Lu is with Peking University. E-mail: lumin.vis@gmail.com.*
- *Xiaoru Yuan is with Peking University. E-mail: xiaoru.yuan@pku.edu.cn.*
- *Huamin Qu is with Hong Kong University of Science and Technology. E-mail: huamin@cse.ust.hk.*
- *Jacky Yuan is with Nanjing Intelligent Transportation Systems Co., Ltd. E-mail: zipmagic@sina.com.*
- *Qianliang Wu is with Nanjing Intelligent Transportation Systems Co., Ltd. E-mail: wuqianliang@gmail.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346746

flows on neighbouring links, and with route selection in its neighbourhood. To study the correlations, we view the aggregated trajectory data as a dynamic graph, and apply dynamic graph visualization techniques.

The contributions of this work are:

- We present a visual analysis system to explore sparse traffic trajectory data, addressing the uncertainties with local animation and trajectory aggregation techniques.
- We study the correlation between cell pattern and link/route flow pattern with dynamic graph visualization techniques.

We will first review related work in Section 2. After that, we give an overview of our system in Section 3, followed by preprocessing in Section 4 and interface design in Section 5. Then we show the effectiveness of our system with case studies in Section 6. We discuss its potentials and limitations in Section 7, and conclude in Section 8.

2 Related Work

Our work is most related to traffic visualization, trajectory visualization and dynamic graph visualization. Besides, some analytic methods can be applied to our dataset.

2.1 Traffic Visualization

There are three major types of traffic data: event based, location based and movement based. Event based traffic data are usually log data collected manually. Each event usually has position, time and a set of attributes. Location based traffic data are collected by roadside detectors, including inductive loops, video cameras, etc. Such data are mainly used for traffic monitoring at predefined locations. For each location, it records a few statistical quantities, e.g. flow volume, occupancy or speed. Movement based traffic data are either directly collected from GPS devices, or reconstructed from images, videos or point clouds. It records the trajectories of a set of vehicles.

In this paper, we focus on sparse traffic trajectory data. It is a combination of location based and movement based traffic data, because it records the movement of vehicles only at predefined locations. No such traffic data has been studied in the visualization community.

For location based traffic data, Lu et al's HOMES [29] enables people to visually summarize inductive loop data in a city at different levels. Piringer et al.'s AlVis [33] dealt with video camera data within a tunnel. It is especially designed to enable situational awareness during emergency events. In above cases, traffic status are visualized in a stand-alone window. Alternatively, it can be embedded into the map as glyphs [27, 42, 38] for better situational awareness.

Now more and more visualizations are for movement based traffic data. Various trajectory visualization techniques have been applied. At the local scale, Guo et al. designed TripVista [21] to study micro-traffic patterns at a road intersection. Zeng et al. designed interchange Circos [45] to study interchange pattern at subway stations. At the global scale, Liu et al. designed VAIT [27] to monitor city traffic. Ferreira et al. [20] and Chu et al. [17] studied city taxi datasets.

With our sparse traffic trajectory data, we would mainly explore the congestion patterns at each cell, and flow patterns at each link and route. Congestion pattern has been studied by Andrienko et al. [5]. They extracted traffic congestions from trajectory data, and visualized them in map and space time cube. Further, Wang et al. [42] studied the propagation of traffic congestions along the road network. Link pattern has been studied by Andrienko et al. [7]. They partitioned a city into regions and studied the traffic flows on the links between neighbouring regions. Route pattern has been studied by Liu et al. [26]. They studied the route diversity in a city at different levels. At the bottom level, they can compare different routes sharing the same origin and destination in terms of flow volume and attributes. Although each pattern has been studied, there's no work studying the correlations between these patterns. Correlation study is one focus of our paper.

2.2 Trajectory Visualization

Trajectory is a widely studied data type in visualization community. In the last two decades, many visual analysis techniques have been developed [3]. According to Andrienko et al.'s paper [4], these techniques can be categorized into three major types: direct depiction, summarization and pattern extraction. All three types of techniques can be applied to sparse trajectories.

With direct depiction techniques, trajectories are visualized in a direct way. That includes representing trajectories as an animation of moving objects [31], representing paths as polylines [30] or stacked bands [40], showing temporal information on a timeline [40, 39], and showing spatial and temporal information together with space time cube [22]. While applying direct depiction techniques to sparse trajectories, we need to address the uncertainties in trajectory reconstruction. For example, the exact tracks between two cells are virtually unknown, therefore visualizing sparse trajectories with animation and path polylines assuming linear movement with constant speed can be problematic. To address such issue, Stoll et al. [37] visualized the reconstruction uncertainties as colored band on top of animation and path polylines.

With summarization techniques, statistical calculations are performed on trajectory data. After that, these statistical summaries instead of the trajectories themselves are visualized. Trajectories can be summarized in many ways. Density map [43], spatial temporal aggregation [2, 7] and aggregated multivariate glyphs [35] summarize trajectories by location, time and attribute. Flow map and flow matrix [2], OD map [44] and Flowstrates [13] summarize trajectories by origin and destination. Various clustering algorithms [6, 19] summarize trajectories by routes. Andrienko et al. [9] argues that summarization techniques are suitable for sparse trajectories, because they reduce the uncertainties in spatial and temporal coverage. Bak et al.'s [11] also studied the aggregated patterns of sparse trajectories. In our paper, we applied summarization techniques. However, we not only study the aggregated patterns separately, but also their correlations.

With pattern extraction techniques, hidden patterns are extracted from trajectory data. After that, these patterns are visualized instead of the trajectories themselves. Typical patterns studied in visualization community include events [5], moving interactions [8] and movement semantics [25, 17]. We consider pattern extraction techniques suitable for sparse trajectories, but we do not study them in this paper.

2.3 Dynamic Graph Visualization

Existing spatial temporal aggregation techniques [2] can transform our sparse traffic trajectory data into flow map. This flow map can be considered as a dynamic graph, with cells as nodes, and inter-cell links as edges. Burch et al. [14] have already tested dynamic graph visualizations on eye tracking trajectories. Therefore, it is natural to also test it on sparse traffic trajectory data.

In their survey, Beck et al. [12] categorized existing dynamic graph visualization techniques into animation based techniques and time line based techniques. Animation based techniques visualize data as an animation of node-link diagrams [18]. They generally emphasize topological features and are more intuitive. In our paper, we would use animation to give an overview of city traffic, highlighting the backbone of the inter-cell links.

Time line based techniques map time to space, usually showing multiple node-link diagrams [15] or matrices [10] simultaneously. Some techniques integrate node-link diagram and time line in a much closer manner. For example, Massive sequence view [41] shows the existential dynamics of edges, while Flowstrate [13] shows the attribute dynamics of edges. Time series glyphs can be associated to each node [34], showing attribute dynamics of nodes. Shi et al. [36] combined time series glyph and node duplication, transforming dynamic route selections in a network into a tree style representation. Time line based techniques generally focus on temporal features and are more suitable for analysis. In our paper, we would use time line to analyze the dynamics of traffic status at each cell, and the dynamics of flow volumes on its related links and routes.

Ahn et al. [1] proposed a taxonomy for dynamic graph analysis tasks. They distinguished between low-level tasks and compound tasks. While low-level tasks are mostly addressed by existing works, compound tasks are not explicitly studied. Two most typical compound tasks are inferential task and comparative/correlational task. In our system, we would explicitly support two correlational tasks, which answer interesting domain questions.

2.4 Analytics

Many analytic methods are related to our work. In the trajectory mining community, researchers often down-sample vehicle trajectories to road or region resolution, aggregate them, and then study their general patterns or extract the outliers. For example, Pang et al. [32] counted the number of taxis in each region of the city at regular time intervals. Then they detected spatial temporal outliers based on these counts. Liu et al. [28] further structured the outlier events into outlier trees, therefore showing interactions between outliers in neighboring regions. In a later work [16], they used L1 optimization to infer the anomalous routes that cause these outlier events. Although above methods were originally designed for continuous GPS trajectories, skipping the down-sample step, they can be potentially useful to analyse our sparse traffic trajectory data. More trajectory computation methods are summarized in Zheng's book [46]. On the other hand, if we first aggregate the trajectories and view it as a dynamic graph, we can apply network analysis methods. Many of such methods are summarized in Kolaczyk et al's book [24]. These methods can be tested on our data. In this paper, we focus on visualization. Therefore we have not implemented these analytic methods, except for the minDistort algorithm [28] for abnormality calculation in each cell. However, we consider them complementary to our work, and would implement them if necessary.

3 Overview

In this section, we first describe the data we use. After that, we explain our design considerations and present the pipeline of our system.

3.1 Data Description

In the past few years, government in the city of Nanjing, China has been pushing a project on intelligent transportation system. In this project, several hundreds of transportation cells are set up on the roadside for traffic monitoring. The cells are usually installed at approx. 200 meters downstream a road intersection. Each cell is directed, meaning it is responsible for only one-direction of traffic flow. The cells have video cameras mounted in order to record vehicles passing through.

Our dataset contains two parts: trajectory data and cell data. The trajectory data are derived from the video cameras, in which vehicles are extracted from video streams, and identified via license plate recognition techniques. Basically, it contains a list of vehicle passing records, each corresponding to one vehicle passing through one cell. The data format is $\langle plate\_number, plate\_color, cell\_id, lane\_id, speed, timestamp \rangle$. A vehicle can be uniquely identified with its *plate_number* plus *plate_color*. For the cell data, it contains the names, spatial positions and directions of the transportation cells.

Our trajectory data spans 169 consecutive days, from Sep. 22nd 2012 to Mar. 9th 2013. The daily record number is shown in the line chart of Figure 1. All together there are 870 million records, with over 1 million vehicles. The data size is 39 GB. Besides, in the cell data, 472 cells are recorded.

Before visual design, we have made some preliminary analysis on the trajectory data of one day. We chose Dec 1.st, 2012. This was a Saturday, which contains 5,177,062 records. Figure 2 plots the positions of the cells. For each cell, we calculate its traffic flow volume as the number of records in trajectory data. This is mapped to the circle size on the map. A picture of the busiest cell is shown in the inset of this figure. As Figure 3(a) illustrates, on this day, one cell can have 0 to 100,000 records, with the average number being 14,000. There are 108 cells with less than 1000 records, 98 of which have 0 record.

Fig. 2. Locations of all transportation cells in Nanjing. Each circle represents one cell, with its area being proportional to flow volume on Dec. 1st, 2012. A picture of the busiest cell is shown in the inset.

These cells may be obsolete, malfunctioned or under construction. As Figure 3(b) shows, the mean speed in this day is around 30 km/h. We can see that the main body of the speed distribution seems to obey a Gaussian distribution. Besides, there are two outlier peaks at 0 km/h and 10 km/h. This distribution also has a long tail, with maximum speed being 500 km/h.

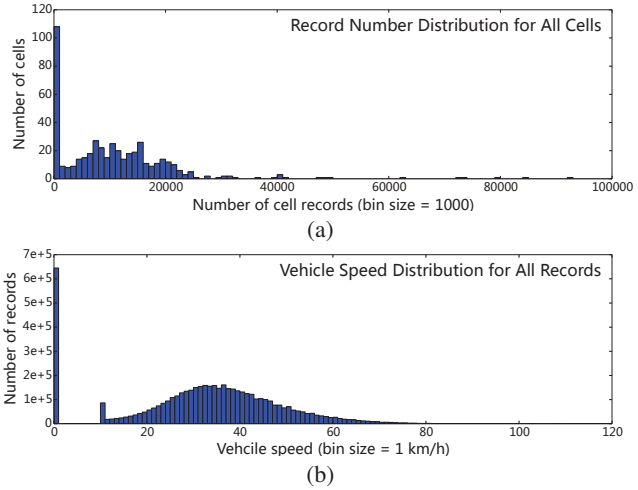

Fig. 3. Statistics of all cells in Nanjing, on Dec. 1st, 2012. (a) The distribution of record number in each cell. (b) The distribution of vehicle speeds in all cells.

We further focus on that busiest cell. Figure 4(a) shows its traffic flow in this day. We can see that the flow volume reached a high level at 8 am, and kept high until 4 pm. The maximum traffic flow per 10 minutes is 1174, at 9:50 am. From Figure 4(b), we can see the traffic speed began to drop at 7 am, and began to recover at 6 pm. At noon, there's a small increase of traffic speed, but it dropped back quickly.

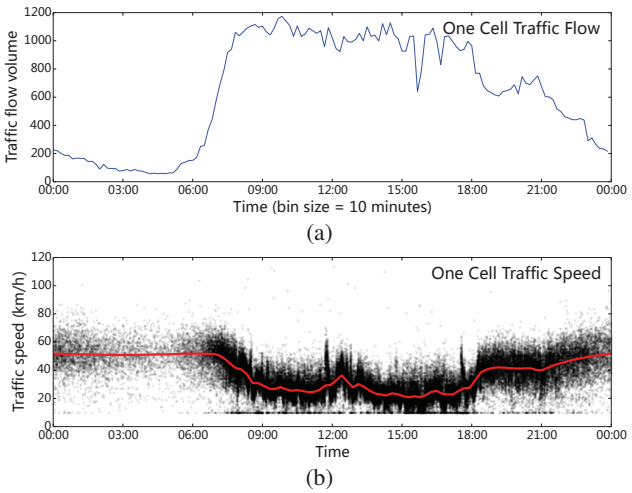

Fig. 4. Traffic condition of one cell in Nanjing, on Dec. 1st, 2012. This cell is highlighted in Figure 2. (a) One day traffic flow. (b) One day traffic speed, where each black dot is a record in the trajectory data, with y position representing speed value. The red line is a LOWESS smoothing of these speed values.

3.2 Design Considerations

We compare our cell-based sparse trajectory data with taxi GPS trajectory data. Their differences are summarized below:

- **Sparse in terms of location:** In our data, vehicle movements are recorded only when they pass through one of the cells. The exact start/end locations and the tracks between two consecutive cells are uncertain. This spatial sparsity also results in temporal sparsity. In contrast, in taxi GPS trajectory data, movements are recorded in a quasi-continuous manner.

- **Dense in terms of population:** In our data, vehicles are sampled very densely. It covers almost all vehicles running on the major roads of Nanjing. In contrast, GPS data are usually restricted to taxis, which are just a small subset of vehicles.

- **Accurate traffic flow volume:** Because of the dense sampling in population, traffic flow volume can be calculated accurately. It is usually not possible with taxi GPS data.

- **Accurate traffic speed:** Traffic speed can be estimated at high accuracy. In contrast, estimation based on taxi GPS data can be biased, because taxi drivers sometimes drive in low speed to search for passengers.

- **Suitable for network analysis:** Network analysis can be performed more accurately, due to accurate estimations of flow volume and speed. However, this can be problematic with taxi GPS data.

- **Not suitable for vehicle tracking:** Vehicles can not be tracked accurately all over the map, because their positions are only recorded at the cells. However, in taxi GPS data, vehicles can be tracked in high precision, with the help of map matching algorithms.

Although our data is sparse in terms of location and time, it still has much richer information than the OD data. When using OD data, only origin and destination location/time are known. It is impossible to get accurate traffic flow volume and speed information at each location, on each link and along each route. As a result, many network analysis tasks can not be performed. Vehicle tracking now becomes impossible, even locally.

In conclusion, compared with traditional GPS data and OD data, our sparse traffic trajectory data can give much more accurate traffic statistics, and therefore is the ideal data for network analysis. However, we have to consider the uncertainties.

Regarding the uncertainties, we make use of two techniques. The first technique is local animation. We believe showing an intuitive animation of vehicle movements is important in data exploration. However, due to spatial and temporal sparsity, we can not accurately track the vehicles. In our data, vehicles move on the complex city road network. Their start/end locations, route selection and travel speed between consecutive cells are all unknown. A reasonable trajectory reconstruction is very difficult. Therefore, we only reconstruct the movements in the vicinity of cells, where we believe the movements are of high certainty. Then we visualized such local tracking with local animation.

Following Andrienko et al.'s work [9], the second technique is to aggregate many trajectories, in order to compensate for the uncertainties in spatial and temporal coverage. In this way, we transform our data into flow map, which is essentially a dynamic graph. Therefore, we can perform network analysis to study the macro-traffic patterns.

According to Kelly's formalization [23], traffic flow can be modeled mathematically, as illustrated on the left of Figure 5. There are several core concepts in this model: *node*, *link*, *route* and *OD*. In our case, each cell C_i is a node. If vehicle moves from cell C_0 directly to cell C_1, without passing any intermediate cells, then $C_0 \rightarrow C_1$ forms a link. Each cell is connected to multiple links. Some are *upstream links* which end at this cell, while some are *downstream links* which start from this cell. A series of links $C_0 \rightarrow C_1 \rightarrow ... \rightarrow C_N$ forms a route. The start and end cells of a route forms an Origin Destination pair (abbrev. as OD), e.g. $C_0 \Rightarrow C_N$. Each route consists of a sequence of links, and a link can be shared by multiple routes. Each route corresponds to one OD, and for one OD there can be multiple routes.

Our network analysis consists of three steps, as shown on the right of Figure 5. Basically, users first get an overview of the traffic status at city scale. In this process, they find some cells interesting, select them and examine their local traffic patterns. Once they discover traffic congestions or abnormalities on a specific cell, they can check whether they are correlated with flow patterns on its upstream/downstream links or vehicles' route selection. The three exploration steps are detailed below:

- **Global Exploration** focuses on presenting an intuitive overview of the city traffic. Users can check the traffic status of all cells and the flow volumes on major links at a specific time. Congested cells and backbone links can be discovered.

- **Cell Exploration** focuses on revealing the traffic patterns at each cell. Users can select one *central cell* each time. They check how its traffic speed and flow volume change with time, and try to discover trend and periodicity. Users can also see when traffic congestions occur on the central cell, and when the traffic status looks abnormal. Users can partly reproduce the traffic scenario with local animation.

- **Correlation Exploration** focuses on testing correlations between cell patterns and link/route patterns. That corresponds to interesting domain questions such as how traffic congestion on one cell is correlated with traffic flows on neighbouring links, and with route selection in its neighbourhood. Given a central cell, users first select its major upstream/upstream links. Then they check which of these links have an increasing or decreasing flow volume when the central cell starts getting congested. Alternatively, users can select a route passing by the central cell. Then users check whether the flow volumes on this route and its alternative routes increase or decrease when the central cell starts getting congested. If correlation is detected, users can try to explain it by searching for news on the Internet.

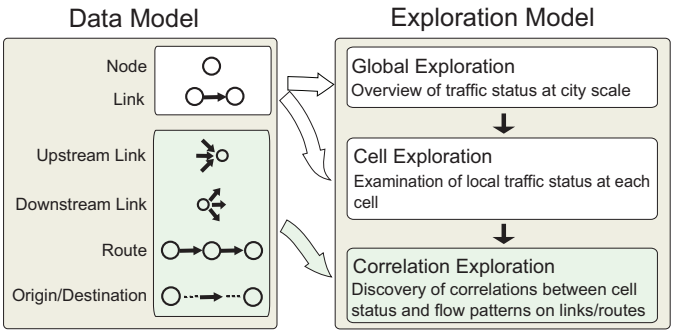

Fig. 5. Conceptual model of our network analysis: based on a formal network data model, we propose three exploration steps: global exploration, cell exploration and correlation exploration.

3.3 Pipeline

Our system pipeline contains a preprocessing phase and a visual exploration phase, as shown in Figure 6. In the preprocessing phase, we first clean the data. Then we derive statistics data for cells and links. The cell statistics data records the traffic flow volume, average speed, congestion status and abnormality status of each cell at regular time intervals. The link statistics data records the flow volume along each link at identical time intervals. The details of the preprocessing phase are presented in Section 4.

In the visual exploration phase, three kinds of explorations are supported (see Section 3.2). For global exploration, we plot the cells and links as a node-link diagram on the map. For cell exploration, we use pixel tables to show the speed, traffic flow, congestion status as well as the abnormality status of a cell over days. We also provide local animation view for each cell. For correlation exploration, we design route/link filtering view, allowing users to select major upstream/downstream links of and major routes passing by a central cell. Then we provide link/route flow view, allowing users to visually correlate the temporal pattern of cell status with flow volumes on selected links/routes. The details of visual interface are presented in Section 5.

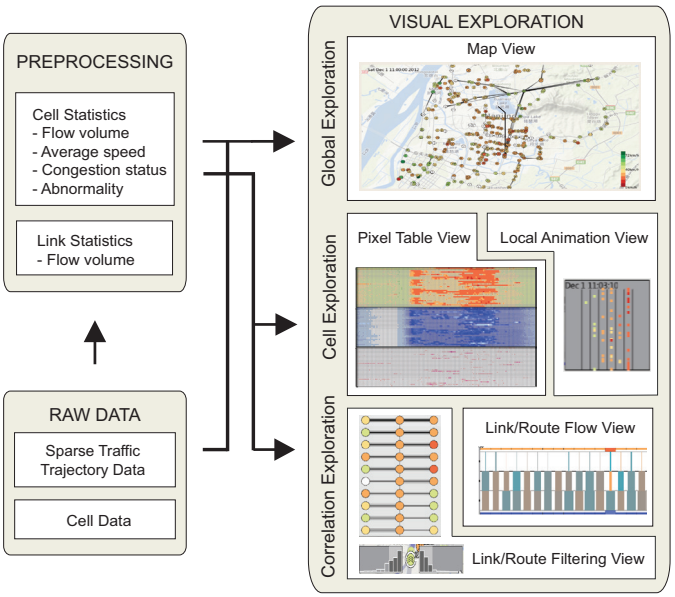

Fig. 6. Our system pipeline contains two phases: preprocessing and visual exploration. In the preprocessing phase, we calculate basic traffic statistics. In the visual exploration phase, we allow users to perform three kinds of explorations with various visual designs.

4 Preprocessing

In the preprocessing phase, we calculate the cell statistics data and link statistics data from raw sparse trajectory dataset. Before that, we first cleaned the data, by removing records with undefined cells or duplicated timestamps. In this step, 10% of the trajectory data are removed.

Then for each cell, we calculate its average speed, flow volume, congestion status and abnormality at regular time intervals. In this paper, we set the time interval to be 10 minutes. Basically, the flow volume can be calculated as the number of records passing the cell, while the average speed can be calculated by averaging the speeds in these records. However, in our data, some cells can be malfunctioned, and mistakenly record many moving vehicles with zero speed. Therefore, we remove all records with zero speed when calculating the average speed. If there's no record with non-zero speed, then the average speed is unknown. Although this strategy can potentially remove vehicle records in congestion, our domain experts consider it as acceptable.

The congestion status is a boolean value, indicating smooth traffic or congestion. It is estimated from the average speed value, following Wang et al's methods [42]. For each cell, we first estimate the *free flow* speed based on all average speed values for this cell. After that, we assume the cell is congested if the current speed is significantly lower than the free flow speed.

The abnormality is a float value, indicating whether the cell behaves abnormally, in terms of its speed and flow volume. We use an algorithm similar to the minDistort algorithm proposed by Liu et al [28]. That is, to calculate the abnormality, we would compute difference values between traffic status at the current time and that at the same time of the same weekday in the same month. For each time, the traffic status is represented as a ten dimensional vector. That is, the speed and flow volume for five time intervals: the current time interval, two previous ones and two following ones. The difference value is calculated as the Euclidean distance between these vectors. We define the minimum of these difference values as abnormality.

For each link, we calculate its flow volume also at 10 minutes intervals. This is stored as series of matrices, each for 10 minutes.

5 Visual Interface

Our visual interface consists of five components: map view for global exploration, pixel table view and local animation view for cell exploration, then link/route filtering view and link/route flow view for correlation exploration.

5.1 Map View

Map view presents an intuitive overview of the city traffic network, as shown in Figure 7. We follow the flow map technique [2], visualizing cells as nodes, and inter-cell links as edges. The flow map can be animated once users change the current time. This is equivalent to dynamic graph animation [18], with predefined node positions.

For each cell, we encode the traffic speed by color, where red represents low speed and green for high speed. If the speed is unknown, we use white. The "⟩" sign inside the cell indicates its direction, for example, west to east. The number of "⟩" signs indicates the relative flow volume passing through this cell. Each link is represented as a gray line connecting two cells, with saturation indicating direction: high saturation (black) end starts the link. The width of the link is proportional to its flow volume. To avoid visual clutter, users can filter out links with low flow volumes, therefore keeping only the major links.

5.2 Pixel Table View

Pixel table view shows the temporal patterns of one cell. As shown in Figure 8(a), it includes three kinds of tables: traffic speed table, flow volume table and abnormality table. Take the speed table for example. It shows the average traffic speed at one cell in Dec. 2012. Each row of the table represents one day, and each column represents 10 minutes of a day. The speed is encoded as a pixel in red-yellow-green color scale, located at respective row and column. If the speed is unknown, we use gray. Once the traffic is congested, we make the pixel larger.

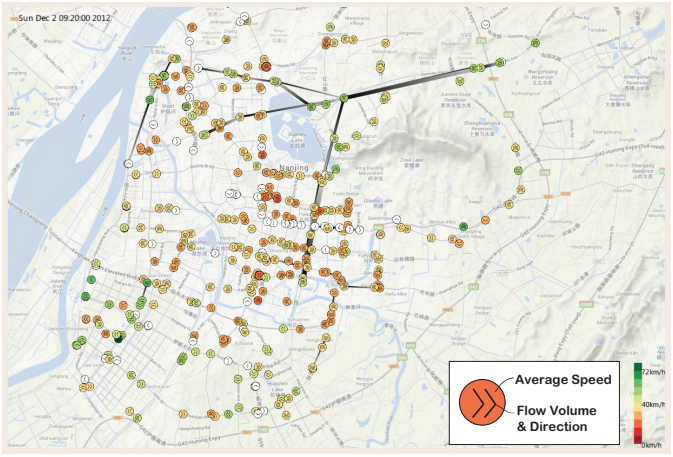

Fig. 7. Map view: overview of city traffic network, visualizing cells as nodes, and inter-cell links as edges. To avoid visual clutter, only links with 10-minutes flow volumes above 100 are shown.

This table like design helps summarize daily traffic pattern, and has proven effective in Wang et al.'s work [42]. In our system, we allow users to switch on/off each of the three tables. We further allow users to make periodic filtering. For example, in Figure 8(b), we only show the traffic status during the daytime of weekends.

5.3 Local Animation View

In addition to pixel tables, we also design local animation view for cell exploration. We believe an intuitive animation will be important in giving users a first impression of the data. It can also help validate discoveries made by statistic or data mining methods, as shown in Wang et al.'s work [42]. However, as mentioned in Section 3.2, traditional animation is not applicable due to the uncertainties caused by sparsity. Therefore, we only reconstruct the traffic within 200 meter interval downstream of each cell. In the reconstruction, we assume that the vehicles move with constant speed and never change lanes. Then we are able to show a local animation, as in Figure 8(c). In the local animation view, we draw black lines to separate the lanes, and dots to show the vehicles. The dot color represents the speed of vehicle.

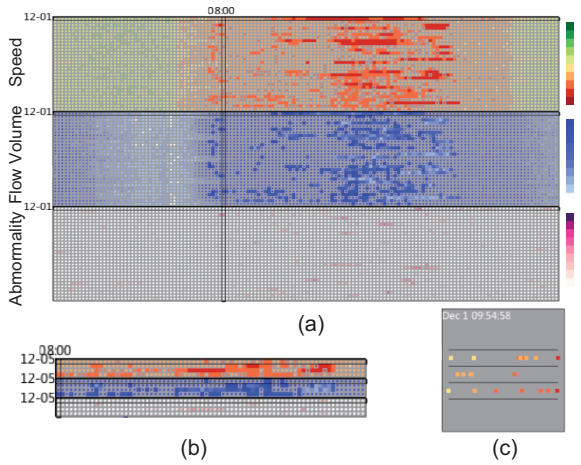

Fig. 8. (a) Pixel table view shows the temporal patterns of one cell, including traffic speed, flow volume, abnormality and congestion status in Dec. 2012. (b) Pixel table supports periodic filtering. This table shows the traffic status during the daytime of weekends. (c) Local animation view shows the traffic animation at one cell.

5.4 Link/Route Filtering View

Given a central cell, the first step of correlation exploration is to select its related links/routes. This is supported by link/route filtering view.

Link filtering view supports the selection of most related upstream/downstream links, as shown in Figure 9(a). It consists of two histograms. Given the central cell, the left histogram displays the top ten highest flow volumes of its upstream links, while the right histogram is for downstream links. Users can select links directly on these histograms, for example, the top three upstream links and top five downstream links. Selected links are highlighted on the map, with their width proportional to the flow volumes at the current time. We can color these links in gray, identical to that in the map view. However, in order to show whether the flow volume is increasing or decreasing, now we prefer to color them in a yellow-gray-blue color scale, as shown in Figure 10. Yellow indicates decreasing flow volumes, while blue indicates increasing flow volumes. Again, saturation indicates direction, with high saturation end starting the link.

Route filtering view supports the selection of related routes. That includes one *central route* R passing the central cell, and route R's multiple alternative routes. For simplification, in our system we only consider routes consisting of three cells. So the central route R is in the form $C_{start} \rightarrow C_{central} \rightarrow C_{end}$, where $C_{central}$ is the central cell. C_{start} and C_{end} are the start and end cells of the route. The alternative routes of R share the same OD $C_{start} \Rightarrow C_{end}$ with R. However, they bypass the central cell, so each alternative route is in the form $C_{start} \rightarrow C_{alternative} \rightarrow C_{end}$, where $C_{alternative}$ is a cell different from the central cell $C_{central}$. Our system first tries to recommend the central route R. As shown in Figure 9(b), our system would choose the top ten routes passing the central cell with highest flow volumes. These ten routes are arranged vertically and aligned horizontally. For each route, we have three circles. From left to right, they represent the start cell, central cell and end cell. The cell color indicates its traffic speed at the current time. Two lines connecting the circles represent links between cells. The width of the line is proportional to the flow volume on route R. Each line also has a gray background, whose width is proportional to the total flow volume on the link. This gives some contextual information. As we consider routes with high flow volumes more relevant to the central cell, we suggest users to choose route R with thick black lines, which are arranged on top. Once users select route R, its top ten alternative routes with highest flow volumes will be automatically selected. On the map, the central route R will be highlighted, with an "S" sign besides the start cell, and an "E" sign besides the end cell. However, the alternative routes will not be highlighted automatically. Users can highlight them manually on the route flow view, which will be mentioned later.

Fig. 9. (a) Link filtering view supports the selection of major upstream/downstream links of the central cell. These links will be used for cell-link correlation discovery. (a) Route filtering view supports the selection of one route passing the central cell, and its alternative routes. These routes will be used for cell-route correlation discovery.

5.5 Link/Route Flow View

After filtering the related links and routes, users try to discover the correlation between cell status pattern and link/route pattern. That is supported by link/route flow view.

With link flow view, we compare flow volumes on links, and correlate them with traffic status on the central cell. This would be a dynamic graph exploration task. Among various visualization techniques for dynamic graph, we think Flowstrates [13] will be a good starting point for our design. It is a time line based method specially designed to compare temporal patterns on links. As shown in Figure 10, the link flow view has a horizontal time axis. In the middle of this view, it's the link region, showing the flow volumes of four links on four rows. Their names are labelled on the left, in Chinese character. At each time interval, the flow volume of each link is represented as a rectangle. Instead of purely using color encodings in Flowstrates, we use both color and width of rectangle. This is because we want to show two properties simultaneously. We use the width of rectangle to encode the magnitude of flow volume, and color to encode its increasing/decreasing rate. The color scale is yellow-gray-blue, identical to that used in the link/route filtering view: yellow for decreasing and blue for increasing. The upstream links are drawn at the lower half of the link region, while the downstream links at the upper half. Then the upstream links and downstream links are ordered by total flow volume separately. On top of the link region is the cell speed band, where we show the speed pattern of the central cell. Below the link region is the cell flow band, where we show the flow volume pattern of the central cell. At some time interval, the band appears larger than usual. That indicates traffic congestion. These two bands are aligned with the link region, enabling visual correlation of cell status pattern and link flow patterns.

The route flow view reuses the above design, but shows cell-route correlation instead of cell-link correlation. The only difference in our design is that now the central link region becomes the route region, where each row represents one route. One of these routes will be the central route R. It passes through the central cell. Others are the alternative routes of R. These routes are vertically ordered by total flow volume.

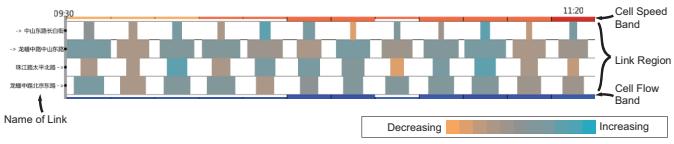

Fig. 10. Link flow view supports visual comparison among flow patterns on upstream/downstream links of a central cell, and visual correlation between those flow patterns and traffic status on the central cell.

6 Case Studies

In this section, we present four studies, demonstrating the three kinds of explorations supported by our system. The first case is for global exploration, the second for cell exploration, while the last two cases are for correlation exploration.

6.1 Case 1: City Traffic Network Exploration

In this case, we study the city level traffic condition. We choose four typical times on Dec. 1st, 2012, and compare the network patterns at those times. In Figure 11, we can see that the traffic during morning peak and evening peak are very different. At 8 am in the morning, the traffic is relatively smooth at all cells. There are many high flow volume links, indicating the traffic load is high. However, at 6 pm in the evening, the traffic is rather congested at all cells. At the mean time, there are much fewer high flow volume links. It perhaps indicates that traffic flow drops due to congestions during the evening peak.

From the figures, we can clearly see the backbone of Nanjing's traffic. Position A seems to be the most important hub. It is the railway station in Nanjing. Major traffic routes in the city are between A and B, A and C, A and D, in both directions. Position B is on the intersection of two highways, and Position C is the commercial center of Nanjing. Position D is a high-tech enterprise zone. We postulate that the ODs $A \Leftrightarrow B$ are for traffic entering and leaving Nanjing. ODs $A \Leftrightarrow C$ and $A \Leftrightarrow D$ are on the inner express way of Nanjing, connecting different functional regions in the city.

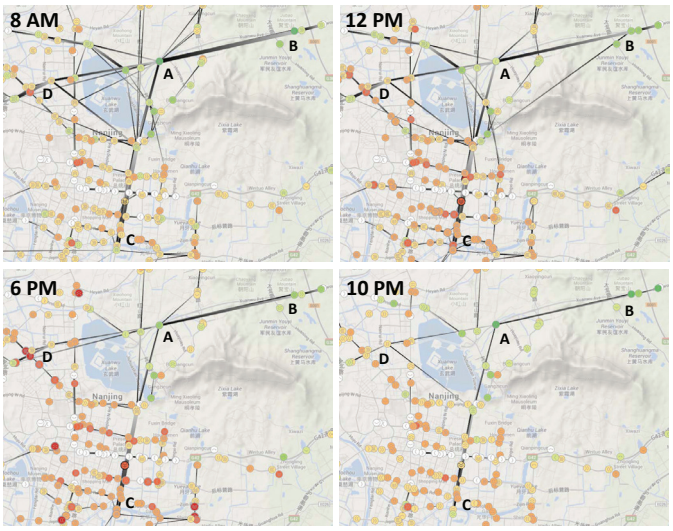

Fig. 11. Case 1: City level traffic network conditions at different times on Dec. 1st, 2012. Links with 10-minute flow volume larger than 50 are shown.

6.2 Case 2: Cell Traffic Event Exploration

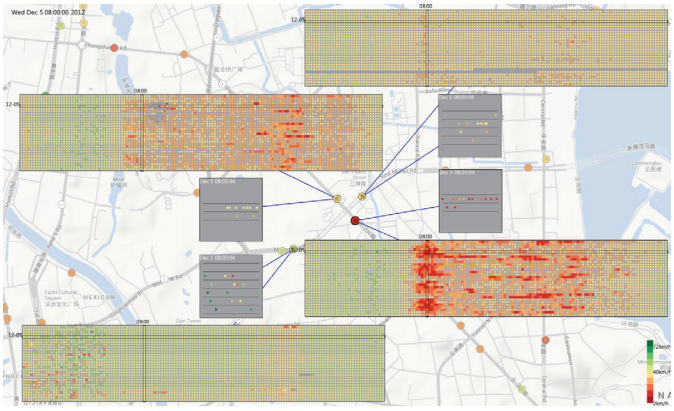

Fig. 12. Case 2: Cell traffic event exploration at the high-tech enterprise zone.

In this case, we explore the cell traffic events. We look at position D, the high-tech enterprise zone, as shown in Figure 12. We select the four cells at the downstream of a road intersection. We can see from the pixel tables, that the two cells on the north have some speed values unknown. That is indicated by the continuous gray color. The northeast and southwest cells have more lanes, and the traffic is usually smooth. However, the northwest and southeast cells have only two lanes, and are congested frequently. The congestions have clear periodic patterns. Take the southeast cell for example. On weekdays, it mainly congested during the morning peak, from 7:30 am to 8:30 am. On weekends, it mainly congested in the afternoon and evening, from 1:30 pm to 3:30 pm, and from 5:30 pm to 6:30 pm.

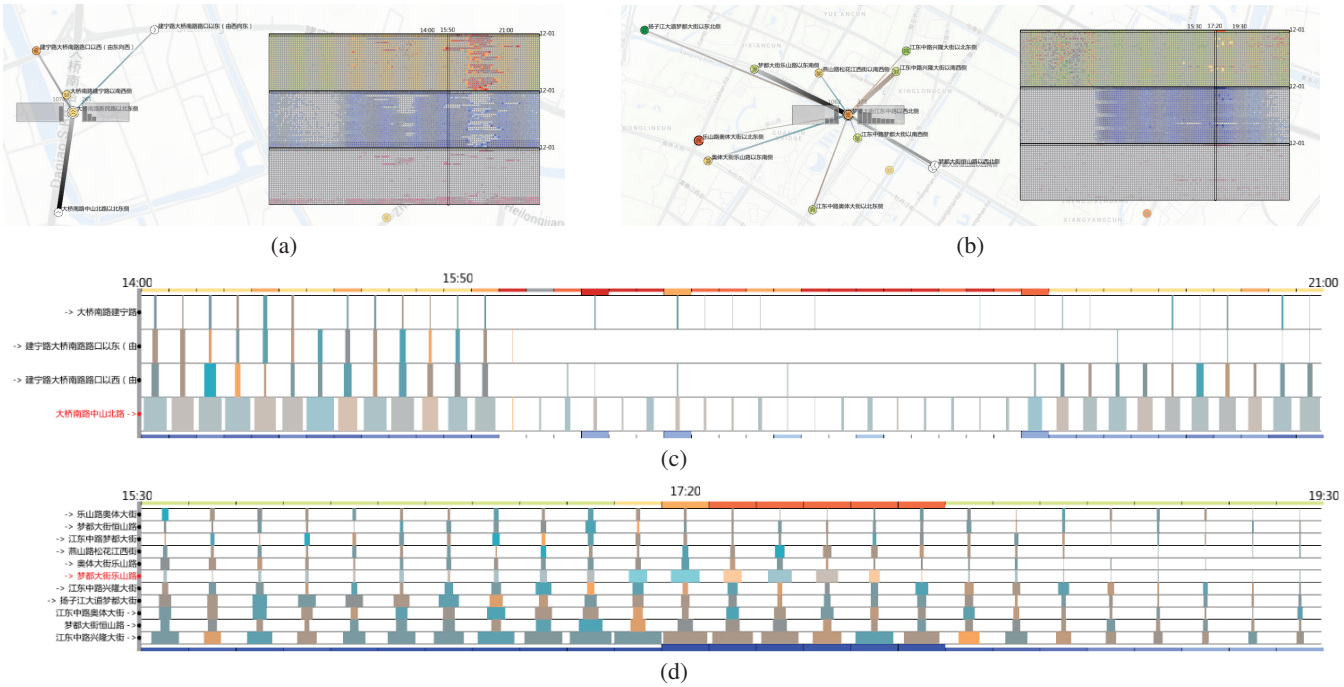

Fig. 13. Case 3: Correlation studies between the congestions on one cell, and flow volumes on its upstream/downstream links. When cell (a) is congested, flow volumes on its major upstream/downstream links drop significantly (c). When cell (b) is congested, flow volumes on some of its major upstream/downstream links increase significantly (d).

6.3 Case 3: Correlating Congestions with Link Flow

In this case, we study the correlations between the congestions on one cell, and flow volumes on its upstream/downstream links. By common sense, people would expect that flow volumes on these links change during congestion. Now we are able to study whether it is the case. We have checked many cells. For most of them, the traffic flows drop significantly. For example, in Figure 13(a), we select one cell. We can see from the pixel tables that it usually gets congested during the evening peaks. Therefore, we continue to select a time range from 2 pm to 9 pm, on Dec. 1st, 2012. Then from the histograms of link filtering view, we can see that it only has one major upstream link and three major downstream links. We select these links, and generate the link flow view, as shown in Figure 13(c). From the cell speed band on top, we can see that the cell is congested from 4:20 pm to 7:20 pm. In the central link region, there are the four links. The bottom row is for the only upstream link, while the top three rows are for downstream links. We can see that the flow volumes drop considerably during congestion, as indicated by the emptiness in the link region. Besides, we can also see that the bottom link has a much larger flow volume than other links. The name of this link is highlighted in red in the link flow view, while the link itself is highlighted in black on the map.

However, for a few cells, the flow volumes on their upstream/downstream links would increase during congestion. In Figure 13(b), we select such a cell, and choose a time range from 3:30 pm to 7:30 pm, on Dec. 1st, 2012. From the right histogram of link filtering view, we can see that the flow volume distribution on its downstream links is rather flat. In the end, we select three upstream links and eight downstream links, and generate the link flow view in Figure 13(d). From the cell speed band on top, we can see that the cell is congested from 5:20 pm to 6:20 pm. In the central link region, the bottom three rows are for upstream links, while the top eight rows are for downstream links. We can see that the upstream link on the very bottom has the largest flow volume among all links. Before congestion, its flow volume continues to increase, indicated by blue color. However, during congestion, its flow volume continues to decrease, indicated by yellow color. In contrast, the sixth row counting from bottom has much larger flow volumes during congestion. The name of this link is highlighted in red in the link flow view, while the link itself is highlighted in black on the map. Besides, the eighth row also has larger flow volumes during congestion. These two links are both downstream links and correspond to vehicles turning around. It may indicate that some events happened there, where many vehicles previously parking nearby were leaving. We searches on the Internet, and find this cell to be on the north of Nanjing Olympic Center. On Dec. 1st, an exhibition ended there just at 5 pm, then a concert started at 6 pm. It is conceivable that there could be high traffic load.

6.4 Case 4: Correlating Congestions with Route Selection

In this case, we study the correlation between the congestions on one cell, and vehicles' route selection in its neighbourhood. As shown in Figure 14(b), we select the cell in the middle of the map. Then this cell get an additional circle around it on the map. We choose a time range from 5 am to 11 am, on Dec. 1st, 2012. As shown in Figure 14(a), our system recommends ten central routes, and we select the fourth route. This central route is shown by the black lines in Figure 14(b), where the start cell and end cell are highlighted with a "S" sign and a "E" sign respectively. Its five alternative routes are automatically selected by our system, which share the start and end cell. Their flow dynamics together with the central route's dynamics are shown in the route flow view in Figure 14(e). From the Figure, we can see that there are three major routes, on the top three rows. From top to bottom, the first route is shown in Figure 14(d), the second in Figure 14(c). The third route with red label is the central route in Figure 14(b). We can see that the flow volume on the central route is exceptionally large during congestion. However, when the traffic is smooth, this route is seldom travelled. It seems that in this case, vehicles do not avoid the congested central cell. It is more likely that it's the high flow volume in the central route that causes the congestion. Unfortunately, we can not confirm this discovery.

7 Discussion

In this paper, we have studied a new kind of transportation data, namely sparse traffic trajectory data. Such data contains almost all vehicles on the major roads of Nanjing. Therefore it gives accurate

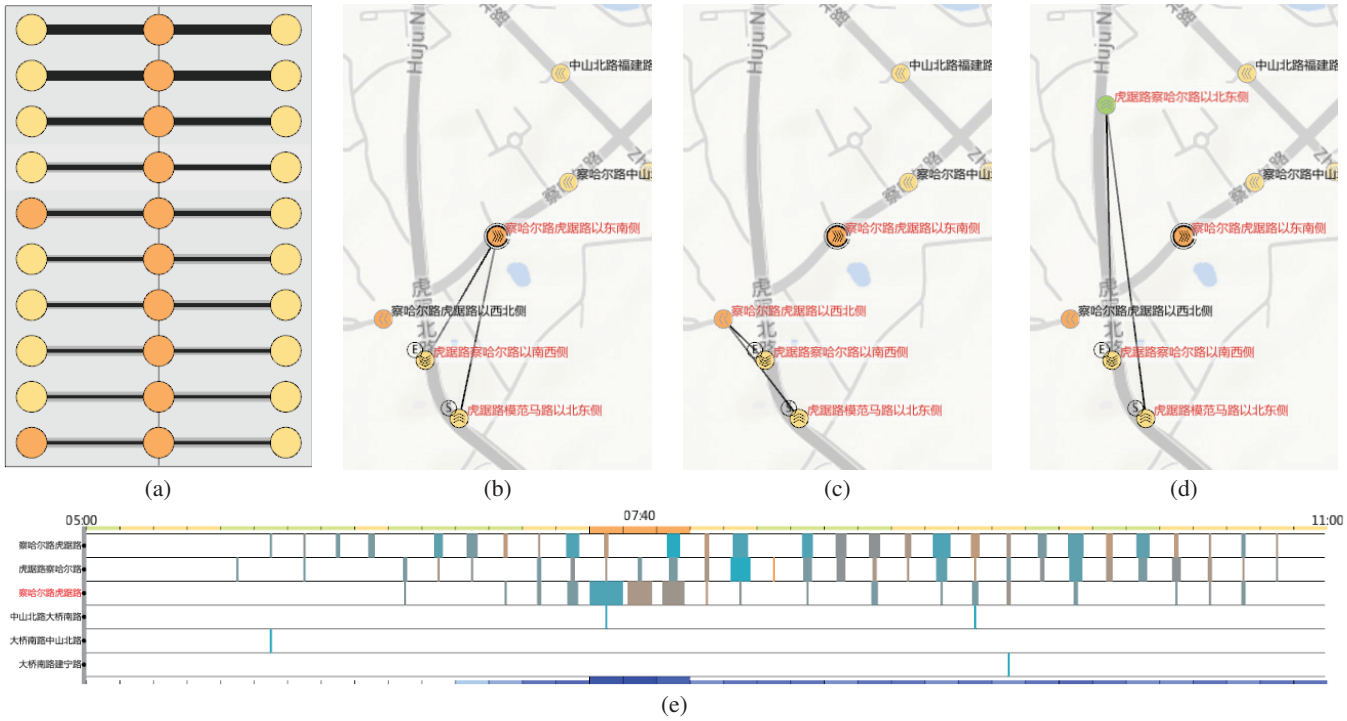

Fig. 14. Case 4: Correlation studies between the congestions on one central cell and vehicles' route selection in its neighbourhood. (a) Route filtering view recommends ten central routes passing the central cell. (b) The central routes, which passes the central cell. (c,d) Two alternative routes bypassing the central cell. (e) Route flow view helps compare route flows and correlate them with the traffic status on the central cell.

traffic statistics and is very precious for macro-traffic analysis. We analyse such data from the angle of network analysis. We have studied the global network patterns, cell patterns and correlation patterns. The value of such patterns and the effectiveness of our system in analysing such patterns have been demonstrated in the case studies. Our domain experts confirm that: *"This system is nicely designed according to the characteristics of our data. It has produced many useful analysis results."*

Performing network analysis on our data is a natural choice, because city traffic is a network-constrained movement. Andrienko et al.'s flow map [2] can be already seen as a dynamic graph. While they address the patterns on node/link separately, we further study the correlations between them. Such correlation explorations are important to answer many interesting domain questions, and we study them explicitly with dynamic graph visualization techniques. Our domain experts consider correlation exploration as very valuable. They comments that: *"The correlation exploration analyses our data from a new perspective."*

From the research perspective, our system is limited in several aspects. First of all, it does not support all major network analysis tasks systematically. Currently there's no separate link exploration, route exploration and OD exploration. Secondly, it focuses on visualization methods, and lacks sufficient support for automatic analysis. In many cases, we have to scan the data manually to discover patterns. It is like finding a needle in a haystack. It would be much more powerful if we have automatic algorithms to search for patterns and visualization methods to validate and explain them. Besides, in the cell-route correlation exploration, we constrain that the central route and its alternative routes have three cells. This is not realistic, because most re-routings would relate to more than three cells. However, if we consider longer routes, we would need more complicated strategies for central route recommendation. A mere flow volume comparison may not work, because longer routes will systematically have less flow volume.

From the application perspective, our system has some other limitations. First of all, our domain experts find the direct manipulations in our system too fancy. One of them said: *"I can see that direct selections on pixel tables and histograms are very advanced interactions, but we are more comfortable with standard menus and dialogues."* We consider it as a general issue in user preference. Besides, they find the link/route flow view not intuitive, even after we have greatly simplified the visual design in the current version of our system. One of our domain experts said: *"Although I can understand it, my boss can't."* We consider it as a general problem for time line based dynamic graph visualization techniques, which focus on analysis but are less intuitive. Finally, our domain experts hope that the system can support their daily workflow, and directly address specific application questions. For example, they would like it if our system can show how the traffic changes if trucks are not allowed to pass through the city center. They also wish our system can discover illegal taxi operations. Currently our system is mainly exploratory. It is not able to answer such specific questions.

8 Conclusion

In this paper, we have presented a visual analysis system to explore a special kind of transportation data, i.e. sparse traffic trajectory data. We use local animation and aggregation techniques to deal with the uncertainties in such data. After trajectory aggregation, we are able to study the macro traffic patterns from the angle of network analysis, and perform three steps of explorations. We starts from the city scale network status, then drill down to each cell. Finally, we choose some congested cells, and study the correlation between patterns on these cells and the traffic flows on related links and routes. For each of the exploration tasks, we produce real case studies with our system.

Acknowledgments

The authors wish to thank the anomynous reviewers for their valuable comments. This work is supported by NSFC No. 61170204 and HKUST grant No. SRFI11EG15PG. This work is also partially supported by NSFC Key Project No. 61232012.

REFERENCES

[1] J.-w. Ahn, C. Plaisant, and B. Shneiderman. A task taxonomy for network evolution analysis. *IEEE Trans. Vis. Comput. Graph.*, 20(3):365–376, 2014.

[2] G. Andrienko and N. Andrienko. Spatio-temporal aggregation for visual analysis of movements. In *Proc. IEEE VAST*, pages 51–58, 2008.

[3] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[4] G. Andrienko, N. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. *Information Visualization*, 7(3):173–180, 2008.

[5] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *Proc. IEEE VAST*, pages 161– 170, 2011.

[6] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *Proc. IEEE VAST*, pages 3–10, 2009.

[7] N. Andrienko and G. Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Trans. Vis. Comput. Graph.*, 17(2):205–219, 2011.

[8] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi. Space transformation for understanding group movement. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2169–2178, 2013.

[9] N. Andrienko, G. Andrienko, H. Stange, T. Liebig, and D. Hecker. Visual analytics for understanding spatial situations from episodic movement data. *Knstliche Intelligenz*, 26(3):241–251, 2012.

[10] B. Bach, E. Pietriga, and J.-D. Fekete. Visualizing dynamic networks with matrix cubes. In *Proc. ACM SIGCHI*, pages 877–886, 2014.

[11] P. Bak, F. Mansmann, H. Janetzko, and D. A. Keim. Spatiotemporal analysis of sensor logs using growth ring maps. *IEEE Trans. Vis. Comput. Graph.*, 15(6):913–920, 2009.

[12] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. The state of the art in visualizing dynamic graphs. In *Proc. EuroVis STAR*, 2014.

[13] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne. Flowstrates: An approach for visual exploration of temporal origin-destination data. *Comput. Graph. Forum*, 30(3):971–980, 2011.

[14] M. Burch, F. Beck, M. Raschke, T. Blascheck, and D. Weiskopf. A dynamic graph visualization perspective on eye movement data. In *Proc. Eye Tracking Research and Applications*, pages 151–158, 2014.

[15] M. Burch, C. Vehlow, F. Beck, S. Diehl, and D. Weiskopf. Parallel edge splatting for scalable dynamic graph visualization. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2344–2353, 2011.

[16] S. Chawla, Y. Zheng, and J. Hu. Inferring the root cause in road traffic anomalies. In *Proc. IEEE International Conference on Data Mining*, pages 141–150, 2012.

[17] D. Chu, D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen. Visualizing hidden themes of trajectories with semantic transformation. In *Proc. IEEE PacificVis*, pages 137–144, 2014.

[18] K.-C. Feng, C. Wang, H.-W. Shen, and T.-Y. Lee. Coherent time-varying graph drawing with multifocus+context interaction. *IEEE Trans. Vis. Comput. Graph.*, 18(8):1330–1342, 2012.

[19] N. Ferreira, J. T. Klosowski, C. E. Scheidegger, and C. T. Silva. Vector field k-means: Clustering trajectories by fitting multiple vector fields. *Comput. Graph. Forum*, 32(3):201–210, 2013.

[20] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2149–2158, 2013.

[21] H. Guo, Z. Wang, B. Yu, H. Zhao, and X. Yuan. Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. In *Proc. IEEE PacificVis*, pages 163–170, 2011.

[22] T. Kapler and W. Wright. Geotime information visualization. In *Proc. IEEE InfoVis*, pages 25–32, 2004.

[23] F. Kelly. *The Princeton Companion to Mathematics*, chapter The Mathematics of Traffic in Networks, pages 862–870. Princeton University Press, 2008.

[24] C. G. Kolaczyk, Eric D. *Statistical Analysis of Network Data with R*. Springer, 2014.

[25] R. Krueger, D. Thom, and T. Ertl. Visual analysis of movement behavior using web data for context enrichment. In *Proc. IEEE PacificVis*, pages 193–200, 2014.

[26] H. Liu, Y. Gao, L. Lu, S. Liu, H. Qu, and L. M. Ni. Visual analysis of route diversity. In *Proc. IEEE VAST*, pages 171–180, 2011.

[27] S. Liu, J. Pu, Q. Luo, H. Qu, L. Ni, and R. Krishnan. Vait: A visual analytics system for metropolitan transportation. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1586–1596, 2013.

[28] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proc. ACM SIGKDD*, pages 1010–1018, 2011.

[29] C.-T. Lu, A. P. Boedihardjo, J. Dai, and F. Chen. Homes: highway operation monitoring and evaluation system. In *Proc. ACM SIGSPATIAL GIS*, pages 85:1–85:2, 2008.

[30] P. Lundblad, O. Eurenius, and T. Heldring. Interactive visualization of weather and ship data. In *Proc. International Conference Information Visualisation*, pages 379–386, 2009.

[31] OpenDataCity. Visitor flow analysis by public wireless. http://apps.opendatacity.de/relog/, 2013.

[32] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng. On detection of emerging anomalous traffic patterns using gps data. *Data & Knowledge Engineering*, 87:357–373, 2013.

[33] H. Piringer, M. Buchetics, and R. Benedik. Alvis: Situation awareness in the surveillance of road tunnels. In *Proc. IEEE VAST*, pages 153–162, 2012.

[34] P. Saraiya, P. Lee, and C. North. Visualization of graphs with associated timeseries data. In *Proc. IEEE InfoVis*, pages 225–232, 2005.

[35] R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *Proc. IEEE PacificVis*, pages 17–24, 2014.

[36] L. Shi, Q. Liao, Y. He, R. Li, A. Striegel, and Z. Su. Save: Sensor anomaly visualization engine. In *Proc. IEEE VAST*, pages 201–210, 2011.

[37] M. Stoll, R. Kruger, T. Ertl, and A. Bruhn. Racecar tracking and its visualization using sparse data. In *Proc. Workshop on Sports Data Visualization*, 2013.

[38] G. Sun, Y. Liu, W. Wu, R. Liang, and H. Qu. Embedding temporal display into maps for occlusion-free visualization of spatio-temporal data. In *Proc. IEEE PacificVis*, pages 185–192, 2014.

[39] A. Thudt, D. Baur, and S. Carpendale. Visits: A spatiotemporal visualization of location histories. In *Proc. EuroVis (Short Papers)*, 2013.

[40] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2565–2574, 2012.

[41] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk. Reordering massive sequence views: Enabling temporal and structural analysis of dynamic networks. In *Proc. IEEE PacificVis*, pages 33–40, 2013.

[42] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. van de Wetering. Visual traffic jam analysis based on trajectory data. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2159–2168, 2013.

[43] N. Willems, H. van de Wetering, and J. J. van Wijk. Visualization of vessel movements. *Comput. Graph. Forum*, 28(3):959–966, 2009.

[44] J. Wood, J. Dykes, and A. Slingsby. Visualization of origins, destinations and flows with od maps. *The Cartographic Journal*, 47(2):117–129, 2010.

[45] W. Zeng, C.-W. Fu, S. M. Arisona, and H. Qu. Visualizing interchange patterns in massive movement data. *Comput. Graph. Forum*, 32(3):271–280, 2013.

[46] Y. Zheng and X. Zhou, editors. *Computing with spatial trajectories*. Springer, 2011.