A User Study on the Capability of Three Geo-based Features in Analyzing and Locating Trajectories

Xumeng Wang, Tianlong Gu, Xiaonan Luo, Xiwen Cai, Tianyi Lao, Wenlong Chen, Yingcai Wu, Jinhui Yu and Wei Chen*

***Abstract*—Visual analysis is widely applied to study human mobility due to the ability in integrating contextual information from multiple data sources. Analyzing trajectory data through visualization improves the efficiency and accuracy of the analysis, yet may induce exposure of the location privacy. To balance the location privacy and analysis effectiveness, this work focuses on the behaviors of different geo-based contexts in the process of trajectory interpretation. Three types of geo-based contexts are identified after surveying 94 related literatures. We further conduct experiments to investigate their capability by evaluating how they benefit the analysis, and whether they lead to the location privacy exposure. Finally, we report and discuss interesting findings, and provide guidelines to the design of privacy-preserving analysis approaches for human periodic trajectories.**

***Index Terms*—Periodic trajectory, location privacy, evaluation, geo-based context.**

I. Introduction

In recent years, human trajectory analysis has received extensive attention because of its numerous practical application scenarios [1], [2]. Urban sensors, such as smart phone applications, GPS navigators and surveillance cameras, keep track of everyone's movements. The collected trajectories are essential in many fields including urban planning, recommendation and so on. However, the risk of exposing location privacy is always our major concern considering human trajectory analysis, no matter the study targets are individuals or groups. Once the locations of the visits, especially the periodic ones, are detected, unscrupulous persons can accurately predict the trajectories of specific individuals or groups, and threaten their security or fraud their acquaintances. The location privacy is indeed a serious issue [3].

Duckham and Kulik [4] summarize location privacy-preserving approaches from the aspects of data acquisition and data processing. Controlling data acquisition is to control the sensitive location-based data uploaded to sensors. It is the most fundamental and efficient way to address privacy problems, but LBS (Location-Based Services) applications such as check-in services usually require precise locations. The common approaches regarding data processing are data modification (e.g. adding noise) and generalization [5] (e.g. data sampling). With regard to the former, the design of data modification operations is confronted with a trade-off consideration between privacy preservation and data utility [6]. As for data generalization, it maintains aggregated data instead of raw data. Such aggregation leads to loss of abnormal movements, which will restrict the scope of achievable analysis tasks. Besides, trajectories of a small population may also cause potential safety hazard. In some application scenarios, the individual trajectories are exposed with no choice. For example, some drivers can learn about the trajectory information of the passengers, who ask for ridesharing.

Existing methods fail to satisfy the demand of analysis and privacy preservation simultaneously. To solve this issue, we turn our attention to the analysis process. Visual analysis, as a universal approach, can facilitate analysts in exploring information efficiently. Targeted at visualization, we seek for a new perspective: geo-based contexts. While trajectories depict human movements, geo-based contexts describe the environments of the movements. The geo-based contexts are undoubtedly very helpful in geospatial visualization so that it is utilized in almost all location-based visual analytical applications. However, to the best of our knowledge, no study has ever researched the capability of the geo-based contexts in exposing location privacy. Thus, we decide to focus on the contribution of geo-based contexts to privacy exposure.

Varied features such as map, population density and POIs can be employed by location-based applications to facilitate exploration and understanding. However, specific choices are probably limited by the design space, schemes of privacy preservation and user preferences. As such, one immediate question is: *how to choose the appropriate features?* In this paper, we address the location privacy problem in trajectory analysis by summarizing three types of geo-based contexts. We build a trajectory analysis system to integrate the geo-based contexts. Based on the system, we conduct a user study to evaluate the capability of each context in both analysis and privacy exposure. Furthermore, we discuss the results and give guidelines for human periodic trajectory system design. The contributions of our work are summarized as follows:

- An innovative perspective that concerns both analysis and privacy preservation of trajectory data, which is to focus on the geo-based context;
- A state-of-the-art report on three types of geo-based features extracted from existed studies;
- A user study that explores the ability of geo-based features to facilitate analysts in analysis locating;
- Guidelines to the design of human periodic trajectory system with regard to location privacy preservation.

X. Wang, X. Cai, T. Lao, W. Chen, Y. Wu, J. Yu and W. Chen are with Zhejiang University, State Key Lab of CAD&CG. T. Gu and X. Luo are with Guilin University of Electronic Technology. E-mail: wangxumeng@zju.edu.cn; cctlgu@guet.edu.cn; luoxn@Huey.edu.cn; {xwcai, laotianyi, chenwenlong}@zju.edu.cn; ycwu@cad.zju.edu.cn; jhyu@cad.zju.edu.cn; chenwei@cad.zju.edu.cn. Wei Chen is corresponding author.

II. Related Work

Our objective is to investigate various geo-based contexts in trajectory analysis. In the process of the literature survey, we first categorize the features and their usages, then generalize the analytical tasks and the privacy-preserving methods respectively to understand the utility.

A. Geo-based Contexts

Geo-based contexts mainly fall into three categories: geographical features, statistical features and semantic features.

A large number of studies employed the entire map as one of supplementary information to facilitate analysis [7], [8]. The maps can be displayed with different scales depending on the level of details required in usage. For instance, at the national or provincial scale, analysts have the abilities of observing long distant travel trajectories [9] and analyzing from a macro perspective [10]. In contrast, at a city or street level, analysts are allowed to identify crossings and building blocks [11]. In some studies [12], [13], maps are simplified as road networks.

It's common that statistical features are overlaid on a map as dots [14], [15], [16], small charts [17], [18] and colors [19], [20], [21], [22]. In addition, statistical features can be visualized without the map. Tennekes and Jonge proposed a novel approach to not only reveal both the density and the composition of distribution, but also to adapt to zoom level [23]. From the illustrations they provided, we can distinguish the boundaries of building blocks based on statistical features without map.

The semantic features are the most sophisticated one among the three kinds of features. In accordance with twitter messages, Fuchs et al. [24] marked diverse personal places such as home, work and transport with different color on map. However, personal places have numerous categories, therefore it is a great challenge to find a large number of different colors or other visual channels that can be identified explicitly from each other. General POIs have the similar problem with personal places. Furthermore, because the specific POIs [25], [26] and social media data [27], [28], [29] are uncountable, they mostly appear in the form of text.

B. Location-based analysis

Zheng et al. [30] provided GeoLife to mine the correlation between users and locations by building a user-location graph, a location-location graph and a user-user graph based on trajectories. The reason why they emphasized on the correlation between users and locations is that the correlation between users and locations is one of the inferences bases for both the location correlation [31] and user correlation. The correlation between users and locations actually contains a large amount of information. To some extent, such correlation can reflect users' hobbies, lifestyles [30] and the diverse functions of locations, which are of great help to analyze other two correlations. In the follow-up research, this issue was extended to user-location-activity correlation [32] and widely applied in many fields like recommendation [7].

Besides, plenty of studies focus on group pattern mining [33], [11], [34], [35]. After the spatio-temporal aggregation of trajectories, Andrienko et al. [14] analyzed the place-related patterns through employing the synthesized approaches. However, only finding the patterns is not enough. Researchers also care about the significance of the patterns, thus it is necessary to not only characterize but also explain the patterns we detected.

C. Location Privacy

Individuals should have the right to control the extent of personal location known by others [4]. Regrettably, in order to enjoy the services, users have to allow their personal location to be collected and retained by variety of location-based services reluctantly [36]. When precise location is peeped by malicious person, his or her personal safety may be threatened [37]. It is necessary to lay emphasis on avoiding the leakage of individual location privacy. Yang et al. [38] proposed the urge of privacy preservation in the researches about individual life patterns. Actually, the studies about group pattern have the same urge as well in that outliers may expose just as the individual. Analyzing trajectories without limitation is an irresponsible behavior.

Massive researchers employed k-anonymity and other data-precessing techniques to address this problem. K-anonymity [39] eliminated some of the characteristics of the data to guarantee each "quasi-identifier" tuple occurs at least k times so that k or more individuals are indistinguishable. Somewhat similar to k-anonymity, data mining approaches sacrificed a part of the utility of the datasets for a little guarantee of privacy [40]. In addition to general data mining approaches, Parent et al. [37] gave a new concept called "sensitive stops". They proposed a semantic location cloaking paradigm for protection of sensitive information, such as visiting hospitals. However, this method also removed some data characteristics.

D. Privacy Preservation and Visualization

Nowadays, visual analysis is a popular data analysis approach. As people increasingly desire privacy preservation, combining visualization with privacy preservation is an inevitable trend. For example, Dasgupta and Kosara [40] proposed a privacy-preserving approach called "Screen-Space Sanitization" for parallel coordinates. With their approach, lines are blurred in different level based on k-anonymity so that the specific values are indistinguishable within ranges. In order to avoid location privacy disclosure from trajectory heatmap, Oksanen et al. [41] presented a privacy-preserving diversity method (ppDIV). Although their approach is superior in some respects to privacy preserving kernel density estimation (ppKDE) and privacy-preserving user count calculation (ppUCC), it may be ineffective in the areas with sparse population. Besides, Chou et al. [42], [43] designed interactive systems that allows users to process event sequence data and graph data in a variety of methods. From a different perspective, the visual approach [44] provided by Wang et al. set its goal as find the best balance between privacy and utility.

III. Background

We surveyed 94 related literatures that cover various research fields, including but not limited to geographical visualization, geospatial data mining, geographical information system (GIS), urban computing and location privacy. Based on the literatures we collected, we abstract an analysis task and three representative features.

A. Analysis Task

Classifying analysis tasks is complex because different literature generalize tasks from different perspectives. Task descriptions such as "recommendation", "correlation","pattern", and "urban planning" have overlaps between each other. Highlighting the tasks that may trigger privacy issues, we abstract a fundamental task that appears in most analytical processes: **explaining the trajectory**. For example, when the specific task is to recommend restaurants to users, analysts need to know which restaurants they or their friends have visited. Therefore, to solve that question, it is necessary to understand the information contained in the trajectory. Analyzing correlation of two individuals based on trajectories need this process as well. Such correlation is reflected in the overlap of their trajectories. To figure out which kind of correlations they have, analysts need explore the overlap trajectory. As for the observation of massive trajectories, the analysts need to explain the significance beneath the pattern, which is equivalent to the interpretation of a representative trajectory actually.

Aiming at explaining the trajectory, analysts start with characteristics of the trajectory. A trajectory contains a number of stationary parts (stay in the same places for a long time), we call them **trajectory points**. A sequence of trajectory points divide an entire trajectory into several **trajectory segments**. The trajectory segments next to a trajectory point are always explained as the process to achieve the goals,which is related to the trajectory points. For example, "students go to school every weekday", "patients go to see doctors", "friends go shopping", and so on. The characteristics of the trajectories can be replaced by inferring the user behavior at each trajectory point. Hence, we specify the analysis task as explaining the relationship between users and trajectory points.

B. Features

According to the literatures, a trajectory analytical system usually contains one or more geo-based context features. We find 16 kinds of auxiliary features from these systems, including entire map, building blocks, vehicle information, population distribution, social media information, POI data and so on. Based on the characteristics of the data types (figure or shape, numeric distribution, and texts), the features fall into three categories:

Geographical features are the most common features because they convey the basic geospatial concept to analysts. Tiled map data is the most common representation because it contains the most landscapes of an area, including roads, rivers, etc. Other features, such as district boundaries, road network and building blocks, are usually displayed at the corresponding scales, depending on the level of details required for analysis. City boundaries and provincial roads serve long-distance movements, while building blocks and street roads help with movements within a small neighborhood. What we care about is period daily trajectories. The range of daily trajectories is often within a city, so we focus on the city level. In practice, building blocks and road network are complementary information because the blank space between building blocks naturally form the shapes of road network.

Statistical features are the numeric distribution of measured objects such as population and temperature. In accordance with the specific tasks, a variety of statistical features are applied in the analysis process. For instance, the distribution of visitors' destinations reflects the most popular tourist attractions. In addition, the periodical changes of population over time indicate the major function hour of a region. The function hour of industrial districts is usually in the day time on weekdays, while that of residential districts is at night. The recreational districts are bustling in the evening and on weekends. However, statistical features are always collected by fixed facilities. Each of them is in charge of a certain area, so the granularity of this features may be great.

Semantic features, including POIs, social/news media, are the textual contents describe the functions of the regions. The extent of detail greatly affects the amount of information that can be extracted from the semantic features. Social media information [45] often includes long descriptions, which may result in more privacy exposure. Precise POIs contain the exact description of the place, which is a direct indication of location. A movement describing someone staying in a restaurant at 18:00 for about 15 minutes may indicate that the target had dinner in a fast-food restaurant. Once we know the name of the restaurant (based on the POIs), we are able to locate the position and find the target directly. In a movement-describing task, information like "having dinner at a fast-food restaurant" is sufficient to describe one's behavior. Thus, knowing the name of the restaurant would contribute nothing but privacy disclosure. Besides, comparing with statistical features, semantic features are more detailed.

IV. Experiment Design

We created a general analytical scenario for the user study. The visual encodings of geo-based contexts are the most widely adopted designs in the geographical analysis systems. In this section, we introduce our user study from the aspects of task, case and experiment process.

A. Task Design

Playing the role of detectives, participants were asked to speculate on the life stories of the objects based on the information obtained from the system. To be specific, they were required to complete two tasks: to analyze the movement behaviors according to the corresponding trajectories, and to locate specific positions on the trajectories. The first task is adopted from [27]. The participants were asked to infer what the locations marked in our system mean to the trajectory

owner. This task is about the correlation between a person and a location, consisting two multiple-choice questions and the participants would choose an answer from options, like resting at home, night shift in the hospital, doing exercise in the park, etc..

We provide five options include a correct answer, three interference options, and an irrelevant option. To find appropriate interference options, we invite 21 participants to a pilot. Each participant need complete three cases in turn, using one feature at a time. In the pilot, participants are asked to describe the correlation in their own word. Without the limitation of options, they can give full play to imagination. The wrong answers with the highest frequency corresponding to each feature will be listed as the interference options of the corresponding question in the formal experiment. Besides, we randomly select an answer from all answers from other questions as an irrelevant option. Thus, there are totally six options, including the correct option, three interference options, an irrelevant option and a "not sure".

The first part concerns with the speculation on the life story belonging to the trajectory owner, while in the second part, we are interested in how the geographical features can expose the real position of trajectories. The participants were asked to answer the real positions of two locations farthest apart in the trajectory by observing the entire real map. In this task, they will play a police who wants to track an object. In order to compensate for the lack of background knowledge, they can see the entire city map with all three features here. However, the trajectory is still shown with only one feature.

B. Case Design

We designed the three cases based on three principles: meet the real situation; meet the common sense of participants; the difficulty difference should be as small as possible. Besides, the selections of the cases are representative trajectories that cover individual diverse life styles. We created three roles in the first attempt, which are a grandmother walking her grandson school every weekday and going public square dance every night, an office worker with a regular commuting life, a nurse with night shifts and day shifts in different weekdays. All these trajectories take place in the same city but with different movement patterns (areas, time, etc). We chose the trajectory point related to the most representative event of each case for users to analyze.

The correct rate of the questions about office worker is much higher than others. In addition, to our surprise, no one found all correct answers of the case of grandmother. To balance the difficulty, we decided to remove the case of office worker and split the daily life of grandmother into two parts: going exercise and walking a child school. These two parts will be separated to create two characters: an elder going exercise and a housewife walking a child school.

C. Experiment Process

We recruited 30 participants with ages ranging from 18-29 . At the beginning of the experiment, we invited participants to introduce themselves in at least two aspect, trajectory analysis experience and the familiarity with the target city. Among them, 6 participants have the experience of trajectory analysis. User information statistics.

We introduced the background of this user study before they starting the process. Then, our experiment mainly consists of a training phase and an experimental phase. We guided the participants how to accomplish our tasks through two cases. One is about the office worker, the other is about a tourist. In the first training case, we helped participants to understand the visual expression and showed them how to use the system. The second training case is about a tourist who has gone varied places, thus, participants can gain plenty of experience though analyzing and locating this case. During training, all of the geographical features are available while in each experimental case only one of the three features is available to participants. Then, the participants would be given an opportunity to get trained with the system by attempting to complete the two tasks based on the second training case. They were allowed to ask any question in the experimental phase. When they were sure to understand the entire process, they would enter the next phase.

In the experimental phase, participants need to repeat the process they have done in the second training case based on three new cases. As indicated in Table I, we adopted a 3×3 Latin Square to bind the features to the cases, which avoids the effect of sequence and the variance of the participants. To match the Latin Square, our participants were divided into groups of three, and each participants would take a row of experiment conditions in the Latin Square. During the experiment the participants wrote down their answers for each question. Note that they must start locating after they complete all analyzing tasks. Because they can get all information from the real map when they locating, which is not allowed during analyzing. Besides, to provide better participant experience for locating, we used dual screen systems to display two interfaces for trajectory range with one feature and entire real map with all three features respectively. Our system would help participants record the completion time for each task.

TABLE I
THE CASES AND FEATURES FOR THE THREE PARTICIPANTS IN THE SAME GROUP

	Case1	Case2	Case3
participant1	Road network	Heatmap	Point of interest
participant2	Heatmap	Point of interest	Road network
participant3	Point of interest	Road network	Heatmap

Next, participants would ask to evaluate three features by ordering them. The evaluation involves three issues: 1) This feature is useful when I analyze/locate; 2) It is easy to analyze/locate with this feature; 3) I have confidence in the result inferred with this feature. Finally, we interviewed participants and recorded their findings and perspectives about:

1) the process of completing the questionnaire with each feature.
2) more details explored from the trajectory.

The entire experimental process takes about one hour.

D. Visual System

We designed and implemented a web-based system. We used dual screen system to facilitate participants in locating by comparing the geo-contexts to real map. Both of the screen resolution are 1920*1080 pixels. The front end is developed by AngularJS. The back end server is implemented by JAVA and SpringMVC and deployed on a PC with Intel Core i5-6500 CPU, 16GB RAM, windows 10 (64bit). These settings ensure that participants can perform tasks smoothly.

1) Data Description: The data provided for the experiment consist of geo-based context data and trajectory data.

All of the geo-base context data are real data related to a medium-sized city with a population of nine million. For geographical data, we attained local road network data and map tiles from OpenStreetMap. We also collected POI data as semantic features. As is shown in Fig. 4 (a), the POI data has 12 categories such as transport, entertainment, and hospital. Each category has several subcategories, such as Karaoke and Leisure under the category of entertainment. The place name that directly reveal the precise locations are discarded. In addition, we employed population density data as statistical features [see Fig. 4 (b)]. We aggregated both statistical features every two hours to explore their time-varying patterns. Furthermore, the simulated trajectory data for experiment were adapted from the real cases, which characterize three life stories introduced in Section IV-B.

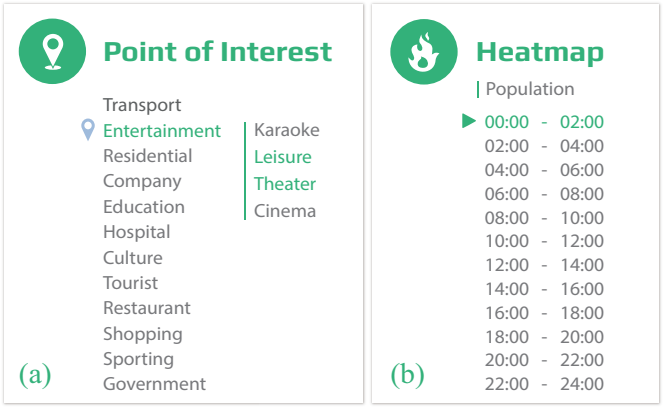

Fig. 1. A snapshot of the feature selection menu. (a) When the POI feature is activated, participants can filter the POI labels on the main view as is shown in Fig. 4. (b) The related population heatmap is shown in Fig. 3

TABLE II
TRAJECTORY DESCRIPTION (THE RELATED TRAJECTORY CAN BE FOUND IN FIG. 3.)

Start Time	Trajectory Description	End Time
	...	
Thu 13:05	A(FGH)I	Thu 13:20
Thu 13:20	I	Thu 15:42
Thu 15:42	I(HGF)A	Thu 15:57
Thu 15:57	A	Fri 07:47
Fri 07:47	A(BCD)E	Fri 07:54
	...	
Frequent trajectory segments: A(BCD)E(DCB)A		
Special trajectory segments: A(FGH)I(HGF)A		

2) System Design: Our experimental system consists of a task list, a trajectory description panel, a feature selection menu, and a main view. The task list shows participants how to complete tasks in order as shown at the top of Fig. 4. When they finish a task, they can submit the task here to switch to the scene of the next one or take a break. System will automatically record the completion time from participants receiving tasks to they submitting them.

The main view has two different versions for analyzing and locating, respectively. For the analyzing task, the main view shows the trajectory (see Fig. 2, Fig. 3 and Fig. 4). To avoid exposing redundant information, we restricted the range and the scales of the display space to fit the scope of trajectories. We applied a transformation of Dodeca-Rings [46] to emphasize the important trajectory points. Because the representation of trajectories is irrelevant to our user study, we translated the trajectories into textual descriptions in the trajectory description panel to simplify the learning process. The trajectory description panel (see Table II) lists the movements of the object in the main view. The descriptions contain the origins, destinations and the routes of the movements or the long-stay position as well as the corresponding time. Combining the trajectory on the map and the description, participants can grasp the whereabouts of the objects in any periods conveniently.

For the locating task, participants can see all the geo-based context features (but not the trajectory) and the OpenStreetMap of the entire city (see Fig. 5). Participants could see the exact locations with longitude and latitude by click on the real map. In the version of main view for analyzing, only one feature is allowed, while for locating, all of them can be activated.

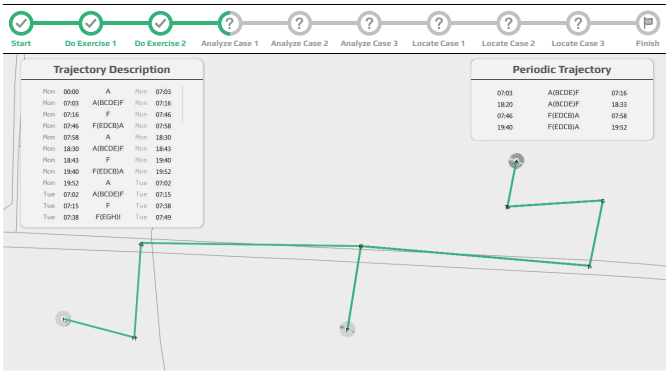

Fig. 2. The main view for the analyzing task displays the case of the elder with a geographical feature: road network.

V. EVALUATION

We summarized the score and time consuming of 30 participants in the analyzing part and locating part respectively. In this section, we introduce the detailed statistical results and related analysis. Note that we removed an outlier (a completion time of analyzing with heatmap), of which value is more than three standard deviations (3SD) away from the mean.

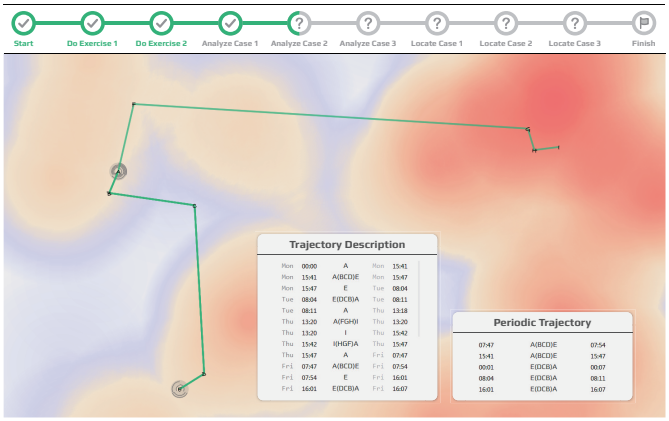

Fig. 3. The main view for the analyzing task displays the case of the nurse with statistical features: population density data. The statistical features are shown in the form of heatmap animation.

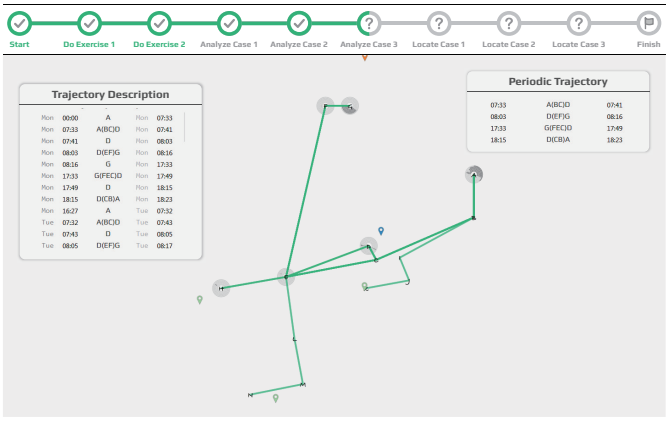

Fig. 4. The main view for the analyzing task displays the case of the housewife with the semantic feature: POI. Only the semantic located within one block along the trajectory route is displayed.

A. Trajectory Analysis

We only designed one question for each case. As is shown in Fig. 6 (a), the correct rate of analyzing based on POI is slightly better than road network and much better than heatmap ($\chi(1) = 4.022, p = .0449 < .05$). More than a half participants miss the correct answer when analyzing the case of nurse with night shift based on heatmap. They thought the target might take a break at home or visit his/her friends. Actually, the population fluctuation around a hospital is different from a residential area in that numerous people visit hospital in day time instead of night. This pattern is indeed demonstrated in this case, but may be it has less ability to convince participants. The analyzing error based on poi mainly happens in the case of elder who going exercise in a park. In this case, participants could only find a few POIs, because the park is relatively capacious, moreover, the entire park is labeled by only one marker rather than an area. Thus, participants were unable to determine whether the target was in the park or not. Compared with heatmap and POI, analyzing with road network reflected similar correct rate in all three cases.

The similar order also can be observed in the average completion time of correct answers (see Fig. 6 (b)). Semantic features like POI can provide specific explanation to locations. As long as participants find the right POI, they can find the answer with no hesitation. Exploring with other two features can only reach vague conclusions, therefore, participants need more time to scrutinize again and again. Furthermore, the information contained in road network is more intuitive than those in heatmap. Participants might take longer time to understand heatmap during analyzing.

From ranking results, we found that the advantage of POI is more obvious. Majority of participants rank POI in the first place in all three aspects (usefulness, ease and confidence), as shown in Fig. 6 (c). Especially for confidence, almost all participants had the greatest confidence in the results analyzed with POI. The ranking results indicate that there exists little difference between heatmap and road network.

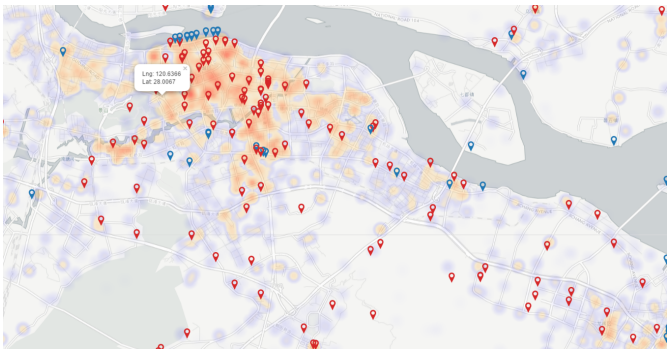

Fig. 5. The main view for the locating task. The entire map with all the information that can be found in the road network, heatmap and POI features are shown.

B. Position Locating

The criterion of locating is whether the error is greater than the threshold, the approximate distance of a block 200 meters). In addition, the range of trajectory may affect the accuracy of location, so we multiplied the threshold with a coefficient that is proportional to the trajectory range. Regarding position locating, our participants were required to locate two trajectory points on the real map for each case. We chose two locations instead of one because two points can determine a plane coordinate system. Accordingly, their answers had three results: two hits, one hit or no hit. The no hit situation can be simply inferred as privacy preservation. The two hits situation indicates privacy exposure because it determinately define the map plane. We defined the one hit situation as uncertain privacy exposure, which suggests a potential location exposure.

It can be seen that the results of the two statistical method are similar (see Fig. 7 (a)). By chi-square tests, we calculate the correlation between features and the accuracy based on two criteria: two hits and at least one hit. The feature employed is proved to have significant effect on the accuracy based on the results of both at least one hit ($\chi(2) = 20.09, p = .000 < .01$) and two hits ($\chi(2) = 21.11, p = .000 < .01$). Then, we calculated the test statistic of each two of the features (see

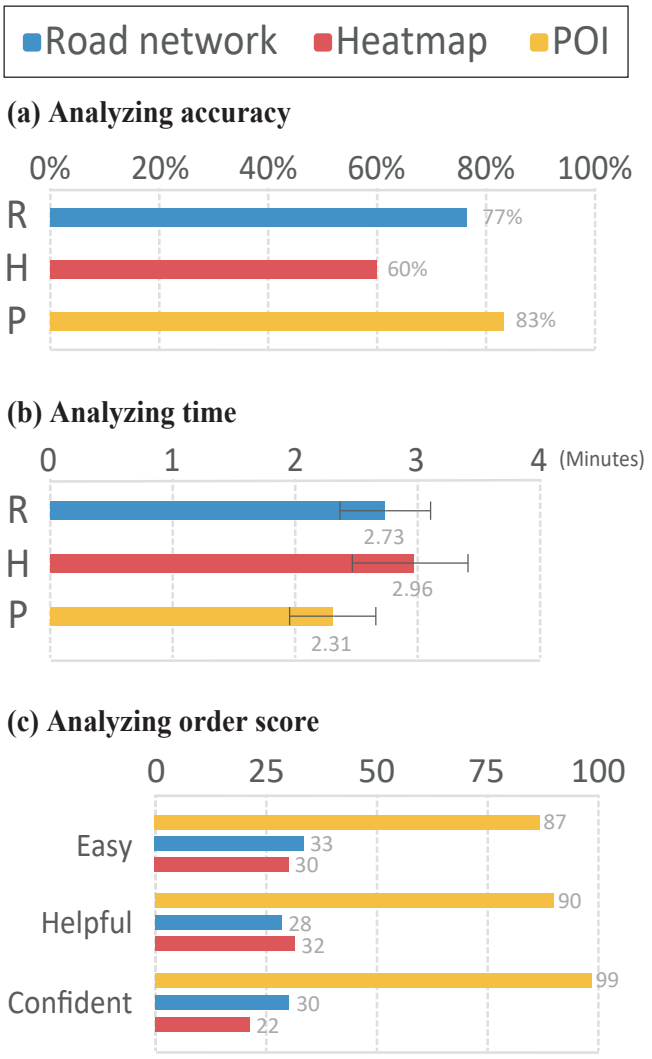

Fig. 6. Analyzing results statistics. a) The total accuracy of 30 results. b) We calculated the 95% confidence intervals for completion time. c) We converted the ranking results to the corresponding score (the first gets 100, the second gets 50 and the third gets 0) and calculated the average.

TABLE III
THE CHI-SQUARE TEST RESULTS OF THE CORRELATION BETWEEN FEATURES AND LOCATING ACCURACY. THE TOP-RIGHT TRIANGLE AND THE BOTTOM-LEFT ONE SHOW THE P-VALUE OF AT LEAST ONE HIT AND TWO HITS, RESPECTIVELY. WE LABEL THE SIGNIFICANCE AT THE .01 LEVELS BY *.

	Road network	Heatmap	Poi
Road network		0.000 *	1.000
Heatmap	0.000 *		0.000 *
Poi	0.602	0.000 *	

Table III). It turns out that when locating with heatmap is less accurate than POI data and road network separately, while the accuracies of POI data and road network have no significant difference. After locating one position, via the continuity of the road network, most participants can locate another one correctly. Although semantic features like POI are discrete markers, they sometimes are dense or unique, that's the reason why they have almost as much ability as road network. Compared to the POI, heatmap is relatively blurry and sparse so that locating two position precisely is not so easy.

Due to the low correct rate of locating by heatmap (only two participants got two hits), it make little sense to analyze its average completion time. As is shown in Fig. 7 (b), participants take shorter time to locate based on road network than POI. On the one hand, viewing multiple categories of POI distribution requires interactions. On the other hand, POI provides a wealth of information, which increases the difficulty of identifying the patterns that can be collated to real locations.

Finally, we turn to the ranking evaluation. It indicates the same assessment with correct rate and completion time. Unsurprisingly, heatmap is considered to be the worst feature in locating, and road network is slightly better than POI, as shown in Fig. 7 (c). However, we noted a phenomenon that three participants have most confidence on their locating results based on POI, which is wrong unfortunately. Instead, their results based on other features are correct. We think the reason is that participants were misled by POI. For instance, there may exist more than one shopping malls in a city. Most shopping malls, taking up a large area and sharing similar POI distribution characteristics, is hard to differentiate.

C. Interview

Some interesting opinions are mentioned in the pilot, hence we also summarize them in this section. In the interview, some participants mentioned that they inferred some simple information by observing the trajectory itself. For example, the trajectory points with the longest residence time may be home and the trajectory point with long residence time in daytime may be work place. However, this idea is not feasible to some unusual trajectories like the case of nurse with night shift. About the second question, participants with trajectory analysis experience are better at speculating the entire stories. Besides, in the tips of POI, participants can get more specific speculations. For the three features, they expressed perspectives as follows:

Road network: One of the participants with trajectory analysis experience said that he distinguished the main road form small-scale road networks and identify the location of the city center. Notwithstanding, such information does not contribute to the completion of the analysis tasks. Some participants without analyzing experience thought that it is confused to interpret useful information from the outline of the road network in the analysis process. About locating with road networks, our participants expressed different opinions, which is depending on the range of the object trajectory. Finding the same small-scale road network from an entire map is like fishing for a needle in the ocean, but the problem becomes much easier as the range of the road network expands.

Heatmap: Some statistical features can only bring their superiority into corresponding analysis tasks, however, the extra information will assist more or less for locating the position. However, for most of the participants, it is very challenging to locate by heatmap. Because it is too sparse

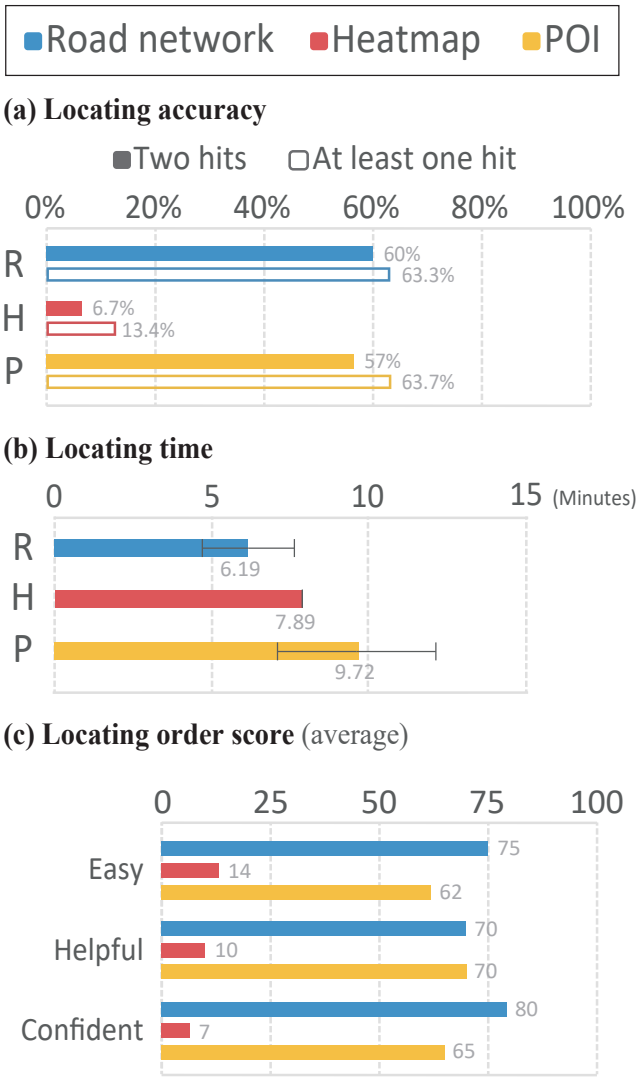

Fig. 7. Locating results statistics. a) The total accuracy of 30 results. b) We calculated the 95% confidence intervals for completion time. c) We converted the ranking results to the corresponding score (the first gets 100, the second gets 50 and the third gets 0) and calculated the average.

in some areas, participants can hardly get useful information. The dramatic changes shown in the heatmap is the main basis of locating.

POI: Most participants agreed with the opinion that POI is of great help to identify the behaviors. Rest of them reckoned that massive POI data interfere their judgment in that they can only check all the POIs nearby to analyze the trajectory. Furthermore, POIs like restaurants are located throughout the city, hence it is necessary to analyze with other information like the time period of stay. Filtering out the irrelevant POI may reduce the uncertainty of speculation, which leads to the improvement of efficiency and accuracy. About locating, some smart participants utilized the specific POI that have sparse distribution. For instance, there are only two train station in this city, if we find that the trajectory pass one of them, locating is just a piece of cake.

VI. Discussion

This paper focuses on a small topic: the role of three categories geo-based features in analyzing trajectories and identifying locations. We discuss four limitations of our work and propose some guidelines to the privacy-preserving design of trajectory visual analytical systems.

A. Limitations

Our trajectory analysis task do not cover all analytical scenarios. For some specific scenarios, some geo-based features are essential. For instance, road network is indispensable in traffic congestion analysis [47] and road planning [22]. The analysts may even have access to the road names, which reveal the exact locations. In such cases, privacy can only be protected by data transformation like aggregation and reduction. Also, POI data are inevitable in analyzing the functions of a local neighborhood. One example usage of POI is to distinguish the origins of noises in the city [48].

Different visualization and interaction designs expose information from different perspectives. POI data reflect two attributes: location and type of places. In our system, we emphasized the overall distributions of the selected POI types. Alternatively, analysts can focus on all types of POIs within a local area. The difference between the two approaches is that the former exposes the relative locations between POIs and trajectories, while the latter hinders the cognition of the entire map.

We also found that capability of the features in analysis and privacy exposure may differ from various trajectories. Locating based on road network or heatmap takes a longer time for a trajectory in a small area than that in a large area. Besides, heatmap is better at assisting the analysis of ordinary periodic trajectories (the office worker case) than abnormal trajectories (the tourist case). For this issue, the results of POI have no significant difference.

A visual analysis system usually contains more than one geo-based features. Perhaps one feature alone can not produce beneficial content. But the combination of multiple features may generate new knowledge, which increases the risk of privacy leaks. To figure out the effectiveness of the combinations, an extended study is required.

B. Guidelines

We get some enlightenments from the experimental results. In this subsection, we share our findings in designing trajectory analysis systems with the three features.

Remove the geographical features from the system if they are irrelevant with the tasks. Although participants have a relatively satisfying correct rate of analyzing, we demonstrate in the Section IV that original road network is beneficial for locating the two specific positions. Furthermore, it provides the most confidence in locating and relatively less confidence in analyzing. Thus, with road network, analysts have a great opportunity to speculate the positions of the entire trajectory. Some road network transformation approaches were proposed, such as simplification [49] and segmentation [50].

We are not sure about how such transformation approaches contribute to location privacy preservation. Due to the limited contribution, we recommend to remove irrelevant geographical features directly. Suppose that analysts need to study modern lifestyle, they need no knowledge about the specific street names, but behavior summarization, like traveling.

Statistics features need more intuitive expressions. In our experiment, heatmap exert tiny effect on both analyzing and locating. Changes in data or tasks may generate different results. Nevertheless, the visual expression of dynamic heatmap does hinder participants to understand the statistic information. If what analysts concerned about are mainly a few important trajectory points, then the system can only provide them with statistics of the corresponding locations. By using a set of locations, more visualization options are given. In the meantime, the information redundancy is reduced, and therefore the location privacy leakage is also avoided.

Simplified semantic features are favorable to trajectory analysis. On the one hand, detailed semantic information, such as the specific name of a restaurant and the event happening at 10:00 AM on Monday, will expose the location immediately. We reckon that generalizing the semantic features to reduce the specificity of location can convey appropriate information to analysts. For example, McDonald's can be generalized as a fast food restaurant, or even a restaurant. Actually, anonymity approaches, like k-anonymity achieve the individual privacy preservation by using the same idea. If some semantic features are so significant that they can not be generalized, analysts need to weigh between the risk of location privacy exposure and the importance of this kind of features to analysis. On the other hand, the distribution of semantic features can also be abstracted. As illustrated in the experimental results, semantic information is of conduciveness to trajectory analysis. However, displaying too much locations of semantic features may also cause privacy exposure. Many participants located the two positions based on the distribution of POIs. The unique arrangement of POI in an area can lead to exact matches. In fact, during the analysis process, analysts are not sure which POI is relevant to the behavior behind the trajectory. In general, they can not avoid to check all POI types near the trajectory points. Thus, we are more inclined to summarize all the semantic features nearby together than to label the locations individually in the entire region. For instance, it could be described as "there are five restaurants, one school and two companies nearby" so that others have less chance to identify the specific locations.

Take features off the map as far as possible. No matter what type of geo-based features appear on the map, there is a risk of exposing location privacy. In accordance with the interview, we found that the feature distribution patterns do contribute to the location identification. Participants tend to locate the real position by finding similar context pattern with the real map. In order to avoid such situation, our advice is to visualize information off the map. As noted above, the analysts may have no needs to learn all features within the entire region. Instead, they focus on one area a time.

VII. Conclusion

With the purpose of analyzing trajectory effectively without privacy exposure, we present a new perspective to deal with the privacy issue in geographical visual analysis systems, which is to select geo-based contexts reasonably. We first classify geo-based contexts into three features. We then evaluate their contributions on assisting analysis and exposing privacy with a user study. Finally, we summarize five guidelines based on the results attained from the experiment.

For future work, we would like to separate the statistical features and semantic features from the map. We will design more user studies to verify the effectiveness; Moreover, we will take more attention to the details of the analysis process instead of the accuracy analysis.

Acknowledgment

This research has been sponsored in part by the National 973 Program of China (2015CB352503), National Natural Science Foundation of China (61772456, U1609217, 61761136020).

References

[1] W. Chen, F. Guo, and F.-Y. Wang, "A survey of traffic data visualization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 2970–2984, 2015.

[2] W. Chen, Z. Huang, F. Wu, M. Zhu, H. Guan, and R. Maciejewski, "Vaud: A visual analysis approach for exploring spatio-temporal urban data," *IEEE Transactions on Visualization & Computer Graphics*, no. 1, pp. 1–1, 2017.

[3] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, no. 6, pp. 391–399, 2009.

[4] M. Duckham and L. Kulik, "Location privacy and location-aware computing," *Dynamic & mobile GIS: investigating change in space and time*, vol. 3, pp. 35–51, 2006.

[5] G. Andrienko and N. Andrienko, "Privacy issues in geospatial visual analytics," in *Advances in Location-Based Services*. Springer, 2012, pp. 239–246.

[6] R. Chen, B. C. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Information Sciences*, vol. 231, pp. 83–97, 2013.

[7] F. Zhang, D. Wilkie, Y. Zheng, and X. Xie, "Sensing the pulse of urban refueling behavior," *Acm Transactions on Intelligent Systems & Technology*, vol. 6, no. 3, pp. 13–22, 2013.

[8] Z. Wang, T. Ye, M. Lu, and X. Yuan, "Visual exploration of sparse traffic trajectory data," *IEEE Transactions on Visualization & Computer Graphics*, vol. 20, no. 12, pp. 1813–1822, 2014.

[9] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel, "A conceptual framework and taxonomy of techniques for analyzing movement ," *Journal of Visual Languages & Computing*, vol. 22, no. 3, pp. 213–232, 2011.

[10] A. Slingsby and E. Van Loon, "Exploratory visual analysis for animal movement ecology," in *Computer Graphics Forum*, 2016.

[11] G. Andrienko, N. Andrienko, S. Rinzivillo, and M. Nanni, "Interactive visual clustering of large collections of trajectories," in *IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 3–10.

[12] B. Jiang and Y. Fei, "Vehicle speed prediction by two-level data driven models in vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1793–1801, 2017.

[13] S. Wang, S. Djahel, Z. Zhang, and J. McManis, "Next road rerouting: A multiagent system for mitigating unexpected urban traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2888–2899, 2016.

[14] 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 *IEEE Conference on Visual Analytics Science and Technology*, 2011, pp. 161–170.

[15] K. Mohamed, E. Côme, L. Oukhellou, and M. Verleysen, "Clustering smart card data for urban mobility analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 712–728, 2017.

[16] Z. Li, D. P. Filev, I. Kolmanovsky, E. Atkins, and J. Lu, "A new clustering algorithm for processing gps-based road anomaly reports with a mahalanobis distance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1980–1988, 2017.
[17] G. Andrienko and N. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in *IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 51–62.
[18] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko, "Visually driven analysis of movement data by progressive clustering," *Information Visualization*, vol. 7, no. 3, pp. 225–239, 2008.
[19] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie, "We know how you live: exploring the spectrum of urban lifestyles," in *ACM Conference on Online Social Networks*, 2013, pp. 3–14.
[20] R. Walker, A. Slingsby, J. Dykes, K. Xu, J. Wood, P. H. Nguyen, D. Stephens, B. L. W. Wong, and Y. Zheng, "An extensible framework for provenance in human terrain visual analytics." *IEEE Transactions on Visualization & Computer Graphics*, vol. 19, no. 12, pp. 2139–2148, 2013.
[21] A. M. Maceachren, A. Jaiswal, A. C. Robinson, and S. Pezanowski, "Senseplace2: Geotwitter analytics support for situational awareness," in *IEEE Conference on Visual Analytics Science and Technology, Vast 2011, Providence, Rhode Island, Usa, October*, 2011, pp. 181–190.
[22] 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 Transactions on Visualization & Computer Graphics*, vol. 19, no. 12, pp. 2149–58, 2013.
[23] M. Tennekes and E. d. Jonge, "Coloring Interactive Compositional Dot Maps," in *EuroVis 2016 - Posters*, T. Isenberg and F. Sadlo, Eds. The Eurographics Association, 2016.
[24] G. Fuchs, G. Andrienko, N. Andrienko, and P. Jankowski, "Extracting personal behavioral patterns from geo-referenced tweets," 2013.
[25] Y. Chen, K. Jiang, Y. Zheng, C. Li, and N. Yu, "Trajectory simplification method for location-based social networking services," in *International Workshop on Location Based Social Networks*, 2009, pp. 33–40.
[26] D. Schürmann, J. Timpner, and L. Wolf, "Cooperative charging in residential areas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 834–846, 2017.
[27] I. Liccardi, A. Abdul-Rahman, and M. Chen, "I know where you live: Inferring details of people's lives by visualizing publicly shared location data," in *CHI Conference on Human Factors in Computing Systems*, 2016.
[28] T. Fujisaka, R. Lee, and K. Sumiya, "Discovery of user behavior patterns from geo-tagged micro-blogs," in *International Conference on Uniquitous Information Management and Communication*, 2010, pp. 1–10.
[29] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1623–1632, 2017.
[30] Y. Zheng, X. Xie, and W. Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *Bulletin of the Technical Committee on Data Engineering*, vol. 33, no. 2, pp. 32–39, 2010.
[31] Y. Zheng, L. Zhang, X. Xie, and W. Y. Ma, "Mining correlation between locations using human location history," in *The Workshop on Advances in Geographic Information Systems*, 2009, pp. 472–475.
[32] V. W. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang, "Collaborative filtering meets mobile recommendation: A user-centered approach," *Aaai*, 2010.
[33] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 330–339.
[34] K. Zheng, Y. Zheng, N. J. Yuan, S. Shang, and X. Zhou, "Online discovery of gathering patterns over trajectories," in *IEEE International Conference on Data Engineering*, 2014, pp. 242–253.
[35] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang, "Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data." *IEEE Transactions on Visualization & Computer Graphics*, vol. 22, no. 1, pp. 160–169, 2016.
[36] N. Ozer, C. Conley, D. H. O'Connell, T. R. Gubins, and E. Ginsburg, "Location-based services: time for a privacy check-in," *ACLU of Northern California*, 2010.
[37] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, and Z. Yan, "Semantic trajectories modeling and analysis," *ACM Computing Surveys*, vol. 45, no. 4, p. 42, 2013.
[38] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie, "Mining individual life pattern based on location history," in *Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*. IEEE, 2009, pp. 1–10.
[39] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2012.
[40] A. Dasgupta and R. Kosara, "Adaptive privacy-preserving visualization using parallel coordinates," *Proceedings of the IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2241–2248, 2011.
[41] J. Oksanen, C. Bergman, J. Sainio, and J. Westerholm, "Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data," *Journal of Transport Geography*, vol. 48, pp. 135–144, 2015.
[42] J.-K. Chou, Y. Wang, and K.-L. Ma, "Privacy preserving event sequence data visualization using a sankey diagram-like representation," in *Proceedings of the SIGGRAPH ASIA Symposium on Visualization*, 2016.
[43] J.-K. Chou, C. Bryan, and K.-L. Ma, "Privacy preserving visualization for social network data with ontology information," in *IEEE Pacific Visualization Symposium*, 2017.
[44] X. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K.-L. Ma, "A utility-aware visual approach for anonymizing multi-attribute tabular data," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 351–360, 2018.
[45] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.
[46] C. Guo, J. Xia, J. Yu, J. Zhao, J. Zhang, Q. Wang, Z. C. Qian, Y. V. Chen, C. Wang, and D. Ebert, "Annotatedtimetree, dodeca-rings map & smart: A geo-temporal analysis of criminal events," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2014, pp. 303–304.
[47] H. Guo, Z. Wang, B. Yu, and H. Zhao, "Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection," vol. 18, no. 1, pp. 163–170, 2015.
[48] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang, "Diagnosing new york city's noises with ubiquitous data," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014, pp. 715–725.
[49] J. Kopf, M. Agrawala, D. Bargeron, D. Salesin, and M. Cohen, "Automatic generation of destination maps," in *ACM Transactions on Graphics)*, vol. 29, no. 6. ACM, 2010, p. 158.
[50] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 712–725, 2015.

Xumeng Wang is a Ph.D. student in the State Key Lab of CAD&CG at Zhejiang University, Hangzhou. She earned the B.S. degree in information and computing science from Zhejiang University in 2016. Her research interests are visual analytics and privacy preservation.

Tianlong Gu received the M.Eng. degree from Xidian University, China, in 1987, and the Ph.D. degree from Zhejiang University, China, in 1996.

From 1998 to 2002, he was a Research Fellow with the School of Electrical and Computer Engineering, Curtin University of Techonology, Australia, and a Post-Doctoral Fellow with the School of Engineering, Murdoch University, Australia. He is currently a Professor with the School of Computer Science and Information Security, Guilin University of Electronic Technology, China. His research interests include formal methods, data and knowledge engineering, software engineering, and information security protocol.

Xiaonan Luo is a professor with the School of Computer Science and Information Security, Guilin University of Electronic Technology, China. He is the director of National Engineering Research Center of Digital Life and the director of Digital Home Standards Committee on Interactive Applications of China Electronics Standardization Association. He won the National Science Fund for Distinguished Young Scholars granted by the National Nature Science Foundation of China. His research interests include computer graphics, CAD, image processing and mobile computing.

Xiwen Cai received his B.Sc. in Psycology from the Zhejiang University, China in 2014. He is currently a master student at the State Key Lab of CAD&CG, Zhejiang University. His research interests include human computer interaction and visualization.

Tianyi Lao is currently working toward the M.S. degree at Zhejiang University, Hangzhou, China. Her current research focuses are information visualization and visual analytics, especially visual analytics of sports data.

Wenlong Chen is currently working toward the Master degree with the State Key Laboratory of Computer Aided Design and Computer Graphics, Zhejiang University, Hangzhou, China. His research interests include visualization and visual analytics.

Yingcai Wu is a ZJU100 Young Professor at the State Key Lab of CAD&CG, Zhejiang University. His main research interests are in visual analytics and information visualization, with focuses on user behavior analysis, urban informatics, social media analysis, and text visualization. He received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology. Prior to his current position, Dr. Wu was a postdoctoral researcher in the University of California, Davis from 2010 to 2012, and a researcher in Microsoft Research Asia from 2012 to 2015. For more information, please visit http://www.ycwu.org.

Jinhui Yu is a professor in State Key Lab of CAD&CG at Zhejiang University, P.R.China. Ph.D, Computing Science Department of Glasgow University in 1999. M.Eng and B.Eng, Harbin Engineering University (Former Harbin Shipbuilding Engineering Institute) in 1987 and 1982 respectively.

Wei chen is a professor in the State Key Lab of CAD&CG, Zhejiang University. His research interests include visualization and visual analysis, and has published more than 30 IEEE/ACM Transactions and IEEE VIS papers. He actively served as guest or associate editors of the IEEE Transactions on Visualization and Computer Graphics, the IEEE Transactions on Intelligent Transportation Systems, and Journal of Visualization. For more information, please refer to http://www.cad.zju.edu.cn/home/chenwei/.