TelcoFlow: Visual Exploration of Collective Behaviors Based on Telco Data

Yixian Zheng*, Wenchao Wu*¶, Haipeng Zeng*, Nan Cao†, Huamin Qu*, Mingxuan Yuan‡, Jia Zeng‡ and Lionel M. Ni§
*Department of Computer Science and Engineering, Hong Kong University of Science and Technology.
Email: {yzhengaj, wwuag, hzengac, huamin}@cse.ust.hk
†College of Design and Innovation, Tongji University. Email: nan.cao@gmail.com
‡Noah's Ark Lab, Huawei Technologies Investment Co. Ltd. Email: {Yuan.Mingxuan, Zeng.Jia}@huawei.com
§Department of Computer and Information Science, University of Macau. Email: ni@umac.mo
¶Corporate Technology, Siemens Ltd.

***Abstract*—Collective behavior is an important concept defined to capture behavioral patterns emerged among the crowd spontaneously. In social science, people's behaviors can be regarded as temporal transitions between a set of typical states (e.g., home and work) which are always associated with certain locations. This fact leads to an interesting research topic in developing ways to explore people's collective behavior patterns through movement analysis, which is our focus in this paper. In recent years, massive volumes of spatio-temporal data generated by mobile phones, called telco data, bring an unprecedented opportunity to study collective behaviors in terms of large coverage and fine-grained resolution. However, distilling valuable collective behavior patterns from the large scale of telco data is not an easy task. The challenge is rooted in two aspects, including the data uncertainty as well as the lack of methods to characterize, compare and understand dynamic crowd behaviors, which triggers the use of visual analytics to take full advantage of machines' computational power as well as human's domain knowledge and cognitive abilities. In this paper, we propose TelcoFlow, a comprehensive visual analytics system which incorporates advanced quantitative analyses (e.g., state-based behavior model) and intuitive visualizations (e.g., an extended flow view embedded with state glyphs) to support an efficient and in-depth analysis of collective behaviors based on telco data. Case studies with a real-world dataset and expert interviews are carried out to demonstrate the effectiveness of our system for analysts to gain insights into collective behaviors and facilitate various analytical tasks.**

***Keywords*-visual analytics; collective behavior; telco data; movement; spatio-temporal analysis**

I. Introduction

In social science, *collective behavior* refers to activities conducted by a temporary and unstructured group of people [1] [2]. It is an important concept defined to capture behavioral patterns emerged among the crowd in a spontaneous way. Mobile phones, as a kind of wearable sensors generating a large collection of spatio-temporal records of individuals (i.e., *telco data* [3] [4]), offer a proliferation of opportunities to study collective behaviors. Exploring collective behaviors in people's daily lives is of great value for many advanced intelligent applications. For example, analyzing time when different groups of people go to work and back home; or places where they work or shop, provides informative contexts that can be leveraged to infer personal habits and social identities, understand the dynamics of social system, as well as facilitate urban planning and location-based services.

However, discovery and analysis of collective behavior patterns from massive telco data is not a trivial task, which poses several challenges. First, how to characterize people's behaviors based on telco data for an effective collective behavior detection? Current studies [5] [6] point out that a person's daily behaviors can be regarded as transitions along a sequence of latent states. For instance, people tend to go to work in the morning and back home in the evening. Home and work are typical latent states while other states such as shopping are often included. Typically, for a person, these states are associated with certain locations [7]. Therefore, the basic intuition towards the goal of revealing people's collective behavior patterns is to model people's states by correlating spatial and temporal features captured by mobile phones and detect sequences of states shared by a group of people (i.e., sequential state clusters). Nonetheless, due to the uncertainty and sparsity of telco data, conventional methods, such as trajectory clustering, are not directly applicable, which would lead to overfitting and distorted results. Secondly, telco data are large in scale owing to a considerable number of mobile phone users and vast area of a city. Meanwhile, a single user's record may contain a long and complicated trajectory of movement without trip information. Then the emerging problem is how to extract a few most characteristic portions of data to identify meaningful collective behavior patterns. Due to these complexities, a fully automatic analysis is difficult, requiring considerable experience and profound knowledge in various fields. Therefore, analysts seek the help of visual analytics to take full advantage of both advanced computational power and human cognitive abilities to explore and interpret hidden patterns efficiently and effectively.

In this paper, we present a visual analytics system called TelcoFlow. Following a user-centered design process, we discuss specific design requirements to enable an effective analysis of collective behaviors based on telco data and describe how they have shaped the design of TelcoFlow. Moreover, we conduct case studies with domain experts demonstrating how our system can help to uncover stories behind the telco data and present information in an easy-to-digest way, thus facilitate domain experts to form hypotheses for subsequent analysis in real-world applications. The main contributions of this work are summarized as follows:

- The design and implementation of a comprehensive visual analytics system to investigate people's collective behavior patterns based on a state-based modeling;
- A suite of interactive visualization techniques enhanced with new features to support visually-assisted knowledge discovery and sense making from large-scale telco data;
- Case studies with a real-world dataset to exemplify the usefulness and effectiveness of our system in leading to interesting insights and facilitating analysis for various applications.

II. Related Work

This section provides an overview of prior studies closely related to our work.

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

Human Behavior Analysis via Mobile Phone Data With the popularization of mobile phones and rapid development of data gathering techniques, telecom operators collect enormous amounts of data every day, offering us unprecedented information resources to study people's behaviors in terms of large coverage and fine-grained resolution. In recent years, a few researchers attempt to model behavior patterns hidden in people's daily lives based on mobile phone data [8] [9] [10]. Eagle et al. [8] applied principal component analysis (PCA) to reveal underlying structures in each user's behaviors by extracting eigenvectors. However, this method employs PCA, a general-purpose dimension reduction technique, which extracts eigenvectors with no well-defined physical meanings. It fails to exploit the domain-specific knowledge that people's daily behaviors could be characterized as transitions along a sequence of typical states [7]. Considering such a state-based structure, Farrahi et al. [9] applied the latent topic model LDA to discover latent characteristic activities of human behaviors. Nevertheless, this is a supervised method requiring mobile cell stations to be labeled in advance with state semantics such as "home" or "work" for each user, which is not practical for real world applications. In addition, Zheng et al. [10] proposed a probabilistic framework to uncover people's latent behavior patterns and measure their similarity for cluster analysis, which is closest to our method. However, their model generates abstract latent variables that are difficult to understand and interpret.

Inspired by these state-of-the-art studies, TelcoFlow integrates interactive visualization techniques with analytical algorithms to provide an intuitive overview of extracted collective behaviors as well as an exploratory interface for underlying details in the data. To the best of our knowledge, TelcoFlow is among the first to provide a visual analytics procedure for a comprehensive analysis of collective behaviors based on mobile phone data.

Visual Analysis of Mobile Phone Data In recent years, mobile phone data has been widely studied for various applications. A systematic survey is provided by Calabrese et al. [11], and here we will mainly review some recent work on visual analytics. Di Lorenzo et al. [12] presented AllAboard, a visual analytics tool that analyzes mobile phone data to help city authorities visually explore urban mobility and optimize public transport. Wu et al. [4] developed a visual analytics system, called TelCoVis, to explore co-occurrence in human mobility. Andrienko et al. [13] proposed a suite of interactive visual analytics methods for reconstructing past events from mobile phone call records. Other studies include identifying "hot spots" of mobile phone usage [14] and geo-social analysis [15]. Most previous work [16] has concentrated on analyzing spatial distribution of population and extracted events, while in this paper, we address a different and more general issue which has not received much attention from researchers in the field of visualization.

Visual Analysis of Movement Visual analysis of collective behaviors falls into the general topic of movement analysis. The use of visual aids to explore movement data has become increasingly important and helpful in understanding and determining patterns, especially given large datasets and complex analytical tasks. Andrienko et al.'s summarized [17] previous work systematically and categorized related techniques into three major types, including direct depiction [18] [19], summarization [20] [21], and pattern extraction [4] [22]. In particular, the direct depiction techniques present paths of movement directly. However, uncertainties in telco data (to be discussed in Section IV-A) make this type of techniques hard to be applied directly, while visual clutter is another defect when handling large and complex datasets. Summarization techniques present movement based on statistical calculations, which could help to reduce uncertainties of movement data in spatial and temporal coverage [23]. Moreover, pattern extraction techniques support an intuitive discovery and analysis of various movement patterns. Our system integrates different types of techniques and enhances them with new features to provide a comprehensive solution for analyzing collective behaviors based on telco data.

III. System Overview

TelcoFlow is designed for detecting, exploring and interpreting collective behaviors based on telco data, consisting of two phases: Data Modelling Phase and Visual Analysis Phase (Fig. 1).

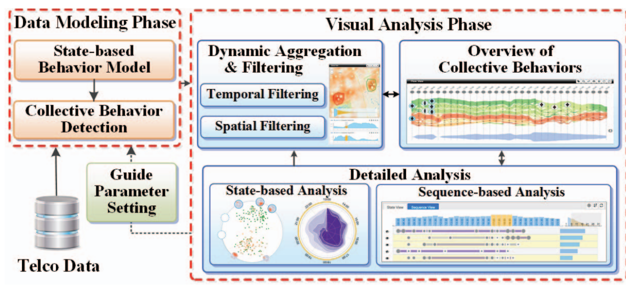


Figure 1: Overview of TelcoFlow system.

In the Data Modelling Phase, people's behaviors are characterized based on a state-based model and then collective behaviors are detected through biclustering. In the visual analysis phase, following the information seeking mantra *"Overview first, zoom and filter, then details on demand"* [24], TelcoFlow first enables dynamic aggregation and filtering in the Multi-facet Filter View, and then presents detected collective behaviors through an interactive timeline design (i.e., Flow View). After that, a more in-depth investigative analysis of the collective behaviors is supported by the State View and Sequence View.

IV. Modelling Collective Behaviors

In this section, we introduce the techniques of detecting collective behaviors based on telco data. We first analyze data characteristics and propose a state-based behavior model for characterizing people's status during a time interval. Then we show how to use this model for collective behavior detection.

A. Data Characteristics

The data involved in this research, called *telco data* [3] [4], is a type of all-in-one mobile phone data collected by telecom operators. The dataset contains a list of data exchange records between each mobile phone and cell stations when mobile phone users make phone calls, send messages or connect to the Internet. Basically, the data format is $< Timestamp, ID_{phone}, ID_{station}, Latitude_{station}, Longitude_{station} >$. Compared with call detail record (CDR) data widely studied previously, the popularity of smart phones with various mobile apps running both in the background and foreground has significantly increased the density of telco data, providing an unprecedented opportunity for an in-depth analysis of human behaviors.

In general, telco data can be viewed as a kind of movement data. For a better understanding, we further compare it with taxi GPS trajectories, a type of traditional movement data [19] [20] [25] widely studied by the community in recent years. Several unique characteristics are identified and summarized as follows: First, telco data has a much wider geographic (broad urban area not limited by road networks) and population (massive number of mobile phone users) coverage, which makes it a challenging task to extract valuable behavior patterns. Second, unlike taxi trajectories with pick-up/drop-off points, telco data does not contain clear trip information. The sampling rate may also change greatly from time to time and from person to person based on different habits of mobile phone users and running apps on phones. To make matters worse, the locations recorded in telco data are in the form of cell station IDs instead of exact GPS coordinates, which indicates the mobile phone user is within the range of a cell station. When a user stays in a region covered by signals from multiple cell stations, the mobile phone is likely to handoff among those stations frequently but not steadily connect to a single one, which is called ping-pong effect [15]. As a result, all these characteristics pose difficulties in applying conventional analytical approaches of movement to compare people's behaviors and exploit collective behavior patterns effectively based on telco data.

B. State-based Behavior Model

A person's daily behaviors can be regarded as transitions along a sequence of states each associated with a certain time interval [5] [6]. In addition, considering the data characteristics, a reasonable behavior model for our system should 1) capture the most informative features in telco data to characterize a person's behaviors with consideration of uncertainty; 2) support effective measurement of behavior similarity of different people for collective behavior detection; 3) be easily interpreted for comprehension and analysis. Hence, we employ a state-based behavior model defined as follows:

*1) **Definition of State**:* Intuitively, a person's states (e.g., home, work) are usually characterized by different locations. Thus formally, the *state* of a person u_j for a time interval T_k can be defined with two components $< \mathscr{P}_{j,k}, \mathscr{C}_{j,k} >$:

- **Cell-based probabilistic distribution (CPD):**

  $$\mathscr{P}_{j,k} = \{P_{i,j,k} | i = 1,2,...,n\}$$

  where $P_{i,j,k}$ is the probability that the person u_j appears at the cell station c_i during the time interval T_k.
- **Certainty:** $\mathscr{C}_{j,k} \in [0,1]$

*2) **Computation of State**:* Based on the above definition, we compute the state of a person for a time interval in two steps:

① **Calculating CPD at a time point t:** Considering the discrete nature and irregular sampling rate of telco data, we start from calculating CPD $\mathscr{P}'_{j,t} = \{P'_{i,j,t} | i = 1,2,...,n\}$ for each time point, where $P'_{i,j,t}$ is the probability that a person u_j appears at cell station c_i at the exact time point t.

If there is a record at t with cell station q, then

$$P'_{i,j,t} = \begin{cases} 1, & \text{if } i = q \\ 0, & otherwise \end{cases}, (i = 1,2,...,n) \tag{1}$$

If there is no record at t, the CPD should be inferred based on two adjacent records before and after t (i.e., records at t_{prev} and t_{next}). The basic idea is that if a person appears at a cell station c_i at time t', there is a probability of P_Δ that he also appears at that cell station at $t' \pm \Delta$. Then we need to explicitly define P_Δ. Based on an extensive review of civil and transportation research literatures [26]–[28] and suggestions by domain experts, we employ one-dimensional Gaussian distribution $\mathscr{N}(t', \sigma^2)$, where σ^2 is inversely proportional to the average moving speed during the time interval between t_{prev} and t_{next} which can be estimated based on the geographical distance between the corresponding recorded cell stations (i.e., c_{prev} and c_{next}). The exponential nature of Gaussian distribution makes the subsequent process much more tractable and interpretable. Formally,

$$P'_{i,j,t} = \frac{P'_{i,j,t_{prev}} \cdot e^{-\frac{(t-t_{prev})^2}{2\sigma^2}} + P'_{i,j,t_{next}} \cdot e^{-\frac{(t-t_{next})^2}{2\sigma^2}}}{2}, (i = 1,2,...,n) \tag{2}$$

where $\sigma^2 \propto (t_{next} - t_{prev})/dist(c_{next}, c_{prev})$.

② **Aggregating CPD for a time interval T_k:** Based on the CPD at a single time point, the state $< \mathscr{P}_{j,k}, \mathscr{C}_{j,k} >$ of a person u_j for a time interval T_k can be calculated as follows:

$$\mathscr{P}_{j,k} = \left\{ P_{i,j,k} | P_{i,j,k} = \left(\int_{T_k} P'_{i,j,t} dt \right) / \sum_{i=1}^{n} \left(\int_{T_k} P'_{i,j,t} dt \right), i = 1,2,...,n \right\} \tag{3}$$

$$\mathscr{C}_{j,k} = \sum_{i=1}^{n} \left(\int_{T_k} P'_{i,j,t} dt / T_k \right) \tag{4}$$

Note that we normalize $\mathscr{P}_{j,k}$ to facilitate comparison of states for detecting collective behaviors, while certainty information is captured by $\mathscr{C}_{j,k}$ based on probability values before normalization.

*3) **Construction of State Cube**:* To support an efficient exploration, a *state cube* (Fig. 2a) is constructed to organize people's behaviors in a clear structure for subsequent analysis.

We first divide a day into regular time intervals, whose length can be manually selected to best fit the time scale of analysis for different applications. Then, we compute the sequence of a person's states as described above. The obtained state sequence is stored in a matrix (termed as *person matrix*), which is a slice of the state cube. In particular, in the matrix of person u_j, each column corresponds to a time interval T_k; each row corresponds to a cell station c_i; and the value $P_{i,j,k}$ in the cell (i, k) of the matrix indicates the probability that u_j appears at cell station c_i during T_k.

C. Detecting Collective Behaviors

Based on the state-based behavior model, if a group of people share a similar sequence of states, we define it as a collective behavior. Thus, the problem of detecting collective behaviors can be cast as sequential state-cluster detection, consisting of two steps:

Step 1: Clustering people with similar states

We extract a *time interval matrix* (Fig. 2b) from the state cube for each time interval, based on which people are clustered into groups according to the similarity of their states. Intuitively, people with similar states should be spatially close to each other, and in the meantime, have similar CPDs. Therefore, we employ Wasserstein metric, also called as Earth Mover's Distance (EMD) [29], which is defined as a minimal cost to be paid to transform one distribution into the other. In order to fully capture spatial information, we calculate such a cost based on geographical distance. In this way, the similarity of two states can be thought of as the average distance between two sets of cell stations with consideration of the corresponding probabilistic distributions. In addition, we choose to adopt density-based clustering due to its insensitivity to noises and cluster shapes [30]. The entropy theory and simulated annealing technique [31] are employed to automatically determine the most suitable parameters for clustering.

Step 2: Biclustering for collective behavior detection

In this step, we adopt biclustering for sequential state-cluster detection, which is a popular data mining technique to identify coordinated relationships between groups of entities in bioinformatics [32]. First, we build a binary matrix based on the extracted clusters. Let $\mathscr{G} = \{G_{k,h} | k = 1,2,...,l; h = 1,2,...,s_k\}$ be the set of clusters in all time intervals, where s_k is the number of extracted clusters for time interval T_k. Then a binary matrix is generated as shown in Fig. 2c, where each row corresponds to a person $u_j (j = 1,2,...,n)$ and each column corresponds to a cluster $G_{k,h}$. A cell of the matrix is marked as 1 if the corresponding person in the row belongs to the corresponding cluster in the column, and marked as 0 otherwise. In this way, the biclustering algorithm can be easily applied to mine collective behaviors efficiently, each in the form of a set of state clusters. Note that we focus on closed biclusters [32] to avoid double counting.

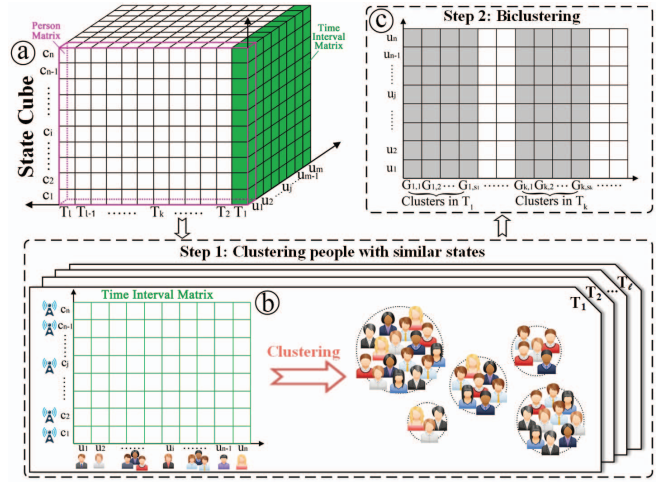


Figure 2: Process of collective behavior detection. (a) A *state cube* is constructed to characterize and organize people's behaviors in a clear structure. (b) A *time interval matrix* is extracted from the state cube for each time interval, used to cluster people with similar states. (c) A binary matrix is generated based on the extracted clusters for detecting collective behaviors through biclustering.

V. Visualizing Collective Behaviors

In this section, we first discuss the rationale inspiring the design of our system, then introduce a set of visualization designs for an in-depth analysis of collective behaviors based on telco data.

A. Design Rationale

We followed a user-centered design process to develop and improve our visual analytics system iteratively. We worked closely with three domain experts from a research consortium focusing on human behavior analysis to better understand the problem domain and identify challenges in their work. Two of them specialize in data mining and machine learning approaches for analyzing large-scale movement data, and the third one has been long engaged in the research on sociology and media studies. The research consortium also held regular meetings with end-users, including government analysts and industry practitioners. We conducted interviews with these experts at different design and implementation stages to present prototypes and collect their feedback. The extensive discussions gradually led to the following design requirements for developing visual analytics system to support an effective analysis:

R.1 **Enabling dynamic aggregation and filtering.** Due to the large scale of telco data, the first challenge analysts encounter is how to find people with collective behavior patterns. To tackle this challenge, in TelcoFlow, aggregation is needed to provide broad awareness of the entire dataset in different aspects. Then analysts can probe the data by filtering to discover subsets of data of their interest for further exploration.

R.2 **Providing overview of collective behaviors.** With the filtered data, the domain experts first want to obtain an intuitive overview of collective behaviors via state-based modelling (R.2.1). In particular, what is the overall distribution of people's states at each time interval (i.e., are there any clusters or outliers)? How stable are these clusters (i.e., do they share a common or diverse state evolutionary path)? In addition, a general understanding about the "state" is also desired by the experts (R.2.2). At this stage, they want to know whether a certain cluster of "states" indicates that people are staying together at a fixed place or moving along a similar route.

R.3 **Facilitating detailed analysis.** Based on the overview, detailed analysis in two aspects needs to be further enabled for a better understanding.

a) State-based analysis. First, analysts have difficulties in understanding semantics about why those people are clustered together at a certain time interval. Hence, a visualization is needed to present people's states contained in a cluster with the awareness of spatial information and support an intuitive comparison of different clusters (R.3.1). Another interesting state-based task highlighted by the experts is to compare detailed state evolutionary trends of different clusters in both space and time domain (R.3.2).

b) Sequence-based analysis. Collective behaviors could be described by a sequence of state clusters. Thus, one of the most important tasks for detailed analysis is to extract and visualize those sequences in an intuitive way and navigate analysts to find ones of their interest (R.3.3). To this end, as suggested by the experts, a joint analysis from three aspects needs to be supported: 1) the *length* of sequence indicating how long a collective behavior lasts; 2) the *number of people involved* indicating the scale of a collective behavior, and 3) the specific *time intervals* when a collective behavior occurs.

B. Visualization Design

In line with the design rationale, we developed the interface of TelcoFlow (Fig. 3), consisting of five interactively coordinated UI components that serve different analytical purposes, including Multi-facet Filter View, Flow View, Sequence View, and State View with two subviews: Radial Distribution Map and Radial Distance Map.

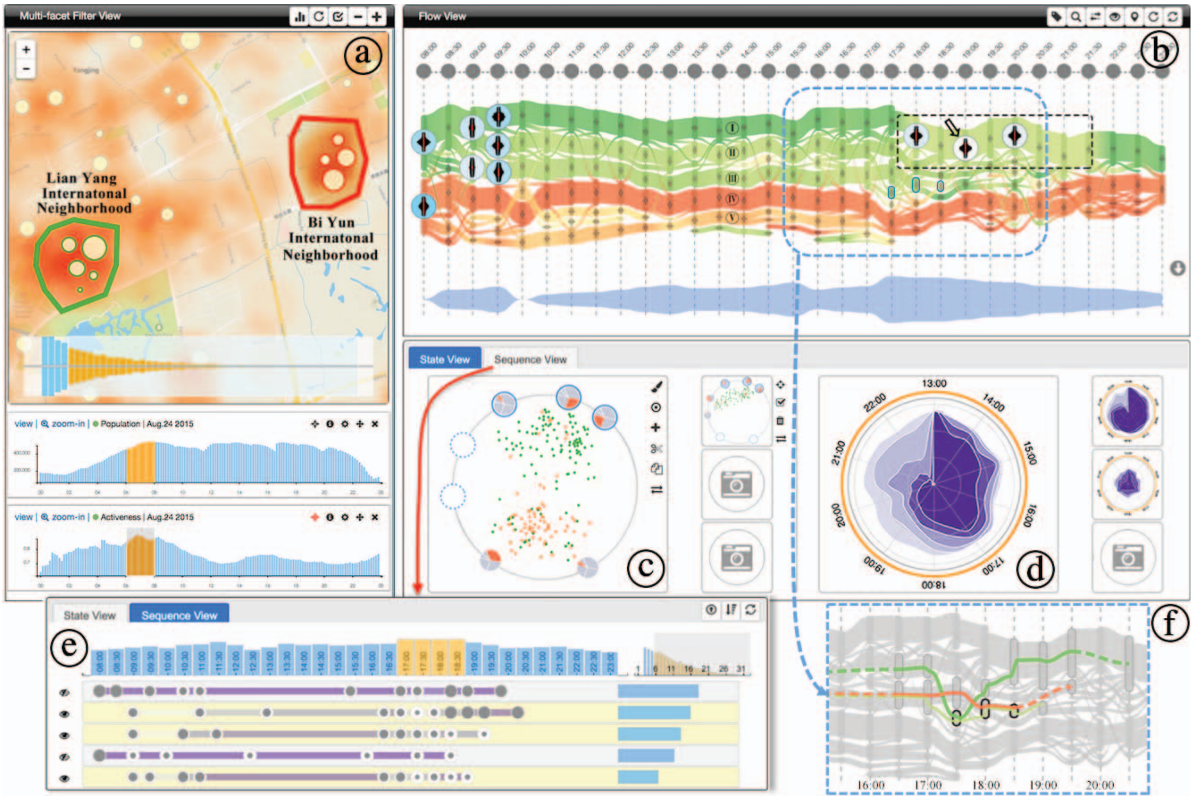


Figure 3: The TelcoFlow visual interface contains five coordinated UI components: (a) Multi-facet Filter View that supports dynamic aggregation and filtering in multiple aspects; (b) Flow View that provides an overview of collective behaviors through state-based modeling; State View, containing (c) Radial Distribution Map and (d) Radial Distance Map, that conveys the details of a state cluster to facilitate understanding and comparative analysis; (e) Sequence View that visualizes the extracted sequences of state clusters representing different collective behaviors and navigates analysts to find ones of their interest, which can be highlighted on the flow (f) for further exploration.

*1) **Multi-facet Filter View:*** TelcoFlow provides the Multi-facet Filter View (Fig. 3a) to support dynamic aggregation and filtering of data (**R.1**) within spatial and temporal context. Well established visualization techniques are preferred by our collaborators for better interpretability and scalability. Thus, we employ a multi-layer map with interactive histograms to help analysts make a dynamic query of data from three aspects, including population, activeness and correlations, which are summarized based on the observation and practical needs by the experts.

Analysis based on population: If analysts want to identify time intervals and places with a fair amount of recorded population, they can brush on the *population histogram* showing the temporal distribution of recorded population. Then a *heat map layer* will be generated on the map revealing the corresponding spatial distribution (i.e., the number of people recorded at different places). Analysts can select places by poly-brush to filter subsets of people at certain places during selected period for further exploration.

Analysis based on activeness: Secondly, analysts are more interested in people who move frequently and significantly rather than those who always stay at the same place. Thus, we employ a statistical entropy formula to quantify the activeness of each recorded person based on the state-based model (Section IV-B):

$$E_{j,k} = -\mathscr{C}_{j,k} \cdot \sum_{i=1}^{n} \left(P_{i,j,k} \cdot ln(P_{i,j,k})\right) \qquad (5)$$

Intuitively, a more diverse CPD implies a more active person, and $\mathscr{C}_{j,k}$ penalizes persons with states of low certainty. To facilitate visual exploration and filtering, the *activeness histogram* is generated to show the varying activeness of people recorded during different time intervals. Moreover, the *activeness layer* can be activated on the map with nodes representing cell stations and the node size encoding the average activeness of people appearing at each cell station. To avoid visual clutter, we adopt a force directed layout similar to Liu et al.'s work [33] and also enable analysts to extract top N (set by users) cell stations with the most active people.

Analysis based on correlations: After choosing a subset of people, the *correlation histogram* can be activated at the bottom of the map to show correlations of these people with others. We define correlation between two persons based on the number of time intervals during which their states are in the same cluster. Thus, in this histogram, the height of each bar encodes the number of people who share similar states with the chosen subset of people for a corresponding number of time intervals indicated on the horizontal axis. Analysts can brush to filter people with certain correlations.

In summary, the temporal query on the population histogram and activeness histogram, the spatial query on the multi-layer map, and the range query on the correlation histogram enable an efficient cross filtering of the large telco data for subsequent analysis.

*2) **Flow View:*** The design goal of this view is to allow users to capture a general picture of collective behaviors from different aspects (R.2). Thus, a composite visual design is adopted to visualize related information in a compact way.

State Evolution Flow: First, we want to show how extracted state clusters evolve over time (**R.2.1**). A naive solution is to employ a stacked graph with each layer representing a cluster. This approach can illustrate the linear evolution of state clusters along time, but it fails to well present complicated evolution patterns such as cluster merging and splitting. However, as pointed out by our collaborators, an in-place view to reveal relationships among state clusters pair-wisely is essential for an overall understanding of collective behaviors. Therefore, we adopt an intuitive river flow metaphor which can help a wide audience quickly understand complex collective behaviors and lower the cognitive load.

Fig. 4a shows an example of state evolution flow where the x-axis represents time. The state clusters for each time interval are presented by vertical bars, namely *state bar*, which are aligned vertically at the corresponding time point. The height of a bar indicates the cluster size, and the color stripes between bars present evolution of clusters over time. Like a river flow in the real world, the flow can either split into several branches when corresponding people's states become diverse, or merge with other branches when they get into similar states. Furthermore, to help analysts easily identify and track specific subsets of people chosen in the Multi-facet Filter View, we adopt a color blending strategy. The blending weights are determined by the ratio of people from each subset. However, this color scheme might be misleading as merging of too many subsets likely leads to similar colors. In such cases, the coloring strategy may fail to differentiate people in related subsets, analysts can then track via interactive highlighting.

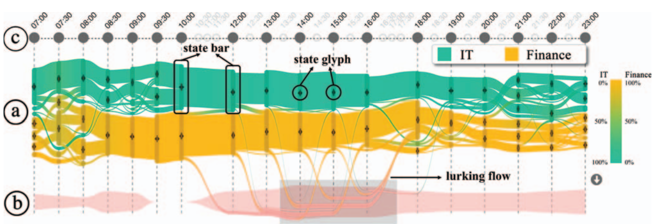


Figure 4: Flow View: (a) State evolution flow provides a visual summary of how state clusters evolve over time. (b) Outlier volume flow conveys the varying number of outliers. (c) Interactive time axis enables a convenient pair-wise analysis.

Outlier Volume Flow: In addition to state clusters, analysts also wish to be aware of outliers (**R.2.1**). To this end, an *outlier volume flow* is appended at the lower part of the view (Fig. 4b), where the varying flow height encodes the number of outliers during each time interval. In some situations, people may split from a cluster and become outliers, stay as outliers for a while, then merge into another cluster, which is important to capture for collective behavior analysis. Thus, we encode this phenomenon with a lurking flow in interior of the outlier volume flow. When analysts brush on the outlier volume flow, all related lurking flows will be generated connecting the splitting and merging clusters in the state evolution flow. To alleviate visual clutter [34], the position of a lurking flow is determined by the timespan remaining as outliers. The longer the timespan, the closer the lurking flow to the bottom.

Interactive Time Axis: Two flows described above allow analysts to see the evolution of clusters and outliers, while the relationship of clusters during non-adjacent time intervals is not deducible, which is a common drawback of flow-based designs. To tackle this issue, we further employ an *interactive time axis*. With this axis, analysts can choose specific time intervals of interest, then corresponding clusters will be directly connected by flows that clearly reveal their relationships (Fig. 4c).

State Glyph: Finally, for a general understanding of each state cluster (**R.2.2**), we overlay a *state glyph* on each state bar in State Evolution Flow to reveal the average CPD. We first considered several design choices, including pie-chart and treemap, but the poor scalability undermines the efficacy of those designs. Ultimately, we come up with the current design whose encoding scheme is illustrated in Fig. 5a. For each state cluster, a bar-chart-like design is used to visually summarize the average CPD, with each horizontal bar representing a cell station and its width encoding the corresponding probability. All bars are sorted in a decreasing order. Due to limited space, we only keep horizontal bars of top k (k=10 in our case) cells in the glyph, and embed a vertical bar with a red part encoding the corresponding proportion of these k cells to provide an awareness of the overall probabilistic distribution. The glyph shape is made symmetric to facilitate comparison. Thus, by observing the glyph shape and the length of red part of the vertical bar, analysts can quickly understand whether a certain state cluster indicates people staying at a fixed place or moving along a similar route. Fig. 5c-d show examples of two typical patterns. Further, analysts can click to highlight a certain state glyph (Fig. 5b), where the background color encodes the cluster density defined by the average distance among contained states.

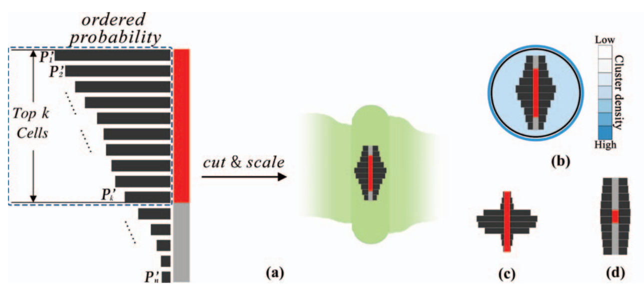


Figure 5: (a) The encoding scheme for *state glyph*. We keep the horizontal bars of top k cells and embed a vertical bar with the grey part encoding the proportion of other related cells. (b) A highlighted state glyph with background color encoding cluster density. (c)-(d) Two typical patterns observed from state glyphs: (c) A fat diamond shape with a long red bar indicates a centralized probabilistic distribution, implying people contained by this state cluster staying together at a fixed place. (d) A flat shape with a short red bar indicates a diverse probabilistic distribution, implying people contained by this state cluster moving along a similar route.

*3) **State View**:* When analysts identify some state clusters of interest in the Flow View, our system supports a detailed state-based exploration (**R.3**) in two aspects via the State View, containing the *radial distribution map* and *radial distance map*.

Radial Distribution Map: To get an intuitive understanding of semantics of a state cluster, we design the radial distribution map to visualize the distribution of people's states in a cluster with the awareness of spatial information (**R.3.1**). Scatterplot is one of the most common ways to reveal distributions for cluster analysis. Meanwhile, to preserve spatial information, we extend the scatterplot by adopting Barycentric coordinates [35]. As shown in Fig. 6a, related cell stations of states in the cluster are regarded as vertices and placed along a circle according to their relative geo-positions. With fixed coordinates of vertices, a person's state can be mapped to a point inside the circle whose coordinates are calculated based on its CPD. Points are colored consistently with the Flow View to differentiate people from different subsets and the color opacity could further encode the certainty of a state. In addition, a color-coded pie chart is embedded in each vertex to visualize average probability on the corresponding cell station.

In addition, to facilitate comparison of different clusters (**R.3.1**), a snapshot function is provided. Analysts can click the snapshot button to store the visualization of current cluster in the clipboard on the right side (Fig. 3c). Moreover, these snapshots of clusters can be selected to compare with the current cluster pair-wisely with shared cell stations highlighted in blue and those non-shared drawn with dotted lines. An example is shown in Fig. 6b.

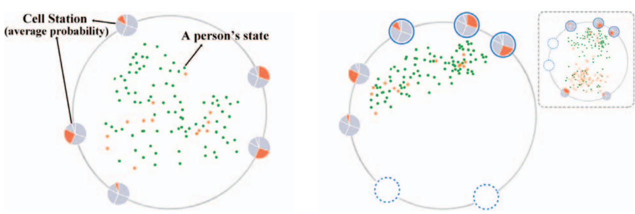


Figure 6: (a) Radial distribution map shows state distribution in a state cluster with barycentric coordinates. (b) Analysts can compare clusters pair-wisely using the snapshot function.

Radial Distance Map: Another major concern of analysts is to compare state evolutionary trends of different clusters of people, such as how far they come from and will go to (**R.3.2**). In order to facilitate such comparisons, we first implemented a prototype enabling analysts to highlight flows of certain state clusters in the Flow View with varying color intensity encoding distance. But analysts found it difficult to compare different clusters due to frequent flow crossings. Thus, we came up with a stacked radial chart design to visualize the evolutionary trend of each state cluster (Fig. 3d). In this design, we employ a circular time axis for an intuitive visual metaphor. The varying height of each layer indicates the number of people in the cluster with different distance level (e.g., 0-1 km, 1-2 km) from the chosen state cluster at different time intervals. The color encodes distance, where a darker purple means a shorter distance. Note that distance levels can be divided through clustering or specified by analysts. With a snapshot function, comparisons can be conducted conveniently.

*4) **Sequence View**:* By tracking flows in the Flow View, analysts can get an overview of collective behaviors. However, under some circumstances, analysts want to further filter the flows to identify a subset of collective behaviors that supports their tasks. Thus, in this view, we employ a matrix-form design to present collective behaviors in the form of sequences of state clusters, sort them into a linear order, and enable a convenient exploration based on statistics of different attributes (**R.3.3**). As shown in Fig. 3e, the rows of the matrix correspond to sequences representing different collective behaviors, while the columns correspond to state clusters at different time intervals. Each state cluster is presented by a node whose size encodes the cluster size. Intuitively, a larger node represents a larger cluster. And then nodes in each sequence are connected by bands. The color of bands encodes the average distance between two state clusters, where a darker purple means a shorter distance. Furthermore, three bar charts are appended showing statistical information to support joint analysis and filtering of sequences. In particular, the bar chart on the right of the matrix shows the number of people involved in each sequence (i.e., the scale of a collective behavior), the top one shows the number of related sequences during each time interval, and the top-right one summarizes the number of sequences with different length (i.e., how long a collective behavior lasts).

VI. Case Studies

To evaluate the system, we carry out three case studies based on the telco data collected on Aug. 23 - 24, 2015 in Shanghai, China with about 240 GB data size per day. The data contains records of about 2.6 million users on 75039 cell stations mapped to 10118 unique locations covering most of the urban area. The data modeling phase takes about 3.4 hours using a MapReduce system, and then our system can support real time interactions.

A. Analysis of People's Daily Behaviors

The first case study is conducted to show the use of TelcoFlow to help analysts interactively explore collective behaviors in people's daily lives during a work day (i.e., Monday, Aug. 24). Through the histograms in the Multi-facet Filter View, we identify a time period of 6-8 am with a fair amount of people recorded and relatively high activeness. By brushing on the histograms, a multi-layer map is generated showing the distribution of those people with highlighted nodes representing the top 50 cell stations with more active people. We zoom in and brush to choose a region of Lian Yang International Neighborhood (green region in Fig. 3a) which includes several large nodes. Then, we activate the correlation histogram to filter out people who have collective behaviors with the chosen set of people in Lian Yang lasting less than two hours. Based on the updated map, we select another region near Lian Yang, also a well-known international neighborhood called Bi Yun (red region in Fig. 3a). Fig. 3b shows the generated Flow View based on our selections, where we can see that flows from these two chosen regions start to split at 8 am and form five major flows (marked with I - V) at around 9:30 am indicating five major workplaces. The highlighted state glyphs with flat shapes and short red bars at 9 am further confirm that people are likely on their way to work at that time. Through interacting with map, we find that flow I and IV locate near Lian Yang and Bi Yun, indicating people work in nearby office buildings or stay at home, while flow II, III and V locate at Lujiazui, Weifang and Jinqiao which are three well-known official blocks in Shanghai. By observing colors of these flows, we find flow I and IV with purer green and red indicating most people in these two flows come from Lian Yang and Bi Yun respectively, while the other three with light colors indicating a mixed population. In particular, light green of flow II and III shows that Lujiazui and Weifang have more people from Lian Yang, while light red of flow V shows Jinqiao has more people from Bi Yun.

Moreover, after 5 pm, these major flows start splitting and merging again, which implies people leaving their offices. Surprisingly, many flows merge into flow I and form a large light-green flow (marked by a black rectangle in Fig. 3b). We click to highlight several state glyphs on that flow and observe light background colors, indicating a low cluster density of these state clusters worth of further exploration. Thus, we further click on a state glyph (marked by a black arrow in Fig. 3b) and compare the corresponding state cluster with another cluster in flow I during working hours (i.e., 3 pm) through the radial distribution maps shown in Fig. 3c and Fig. 6b respectively. We find that the CPD of these two state clusters share three common cell stations highlighted in blue on the top which cover some residential areas, while the other two unshared cell stations at 3 pm (Fig. 6b) and 7 pm (Fig. 3c) cover some office buildings and a local shopping mall respectively. In addition, as shown in Fig. 3c, there are obviously two sub-clusters at 7 pm, one of which mainly consists of green points representing people from Lian Yang, and the other contains many people from Bi Yun shown by red points. Thus, we infer that such a flow merge after work is due to people dining at restaurants in the local shopping mall near Lian Yang. Note that we conduct an informal field study and talk to some shop owners in that shopping mall to verify our hypothesis.

Furthermore, as required by our collaborators, we consider the following scenario of using TelcoFlow for investigating people's collective behaviors to choose proper places for a car company to *set up billboard advertising* targeting at people from these two international neighborhoods. The preferred places should locate on people's way home after work with many people passing by. In order to facilitate information spreading, people with collective behaviors lasting for a longer time have a higher priority in our consideration. To find places satisfying these requirements, the Sequence View could be helpful. First, by brushing on the top and top-right bar charts, we extract sequences containing state clusters during 5-7 pm and with a length of more than three hours. Then we sort these sequences based on the number of people involved through the bar chart on the right. As shown in Fig. 3e, we focus on the top 5 sequences. Among them, we are more interested in sequences with bands in light purple, implying people commute for a longer distance for their work who are potential customers to buy cars. Besides, we also want sequences containing relatively larger nodes during 5-7 pm which indicate larger state clusters, implying places with more people passing by. In this way, three sequences are chosen and highlighted in the Flow View. We can easily identify three state clusters (Fig. 3f) where the chosen flows converge, implying three candidate places. By further interacting with map, we find two of them are metro stations near Lujiazui and Weifang, and another one is a busy intersection near Lian Yang.

B. Comparison of Finance and IT Professionals

The second case demonstrates the use of TelcoFlow in comparative analysis on collective behaviors of people working in finance and IT industry. Through the Multi-facet Filter View, we filter people who appear at Lujiazui and Zhangjiang, two well-known official blocks for finance and IT industry in Shanghai, during 10-11 am on Monday (Aug. 24). The flows generated are shown in Fig. 4, where we can see that people in these two fields have a quite different time schedule. In particular, the green flow merges at 10 am and splits after 8 pm, implying people in IT industry work from 10 am till late at night, contrasting sharply to those people in finance industry (yellow flow) who start work much earlier (8:30 am) and leave offices at 6 pm. In addition, two flows have few crossings, indicating relative independency of finance and IT industry. Furthermore, by observing the outlier volume flow, we identify an increase of outliers during 2-4 pm, implying more people have unique states during that time period. By brushing on the outlier volume flow, the lurking flows are generated, showing that most outliers come from the yellow flow representing people in finance. Thus, we infer that people in finance industry tend to go out to meet clients in the afternoon, which might be a working habit shared in the financial field.

C. Exploration of Customer Behaviors

In this case, our collaborators are keenly interested in collective behaviors of customers at different shopping malls. As suggested, we extract data of customers at three shopping malls during 1-2 pm on Sunday (Aug. 23) through the Multi-facet Filter View, including a large shopping center in CBD (i.e., IFC mall), a popular department store (i.e., Next Age store) and a local shopping mall in a neighborhood (i.e., Thumb Plaza). Fig. 7 shows the generated Flow View. We find that although IFC mall and Next Age store have more customers than Thumb Plaza during 1-4 pm as indicated by the larger height of blue and red flows, Next Age store loses customers significantly around dinner time (the red flow splits at 5 pm). Those customers are attracted by Thumb Plaza (part of the red flow merges with the green flow, and generates a yellow flow). Moreover, we can see that after 8 pm there are still several yellow branches, which implies similar customer sources and the potential competitive relationship of Next Age store and Thumb Plaza.

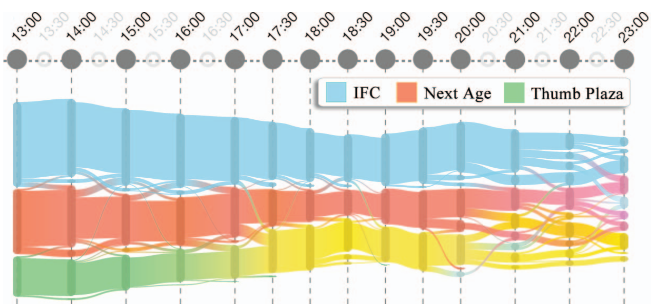
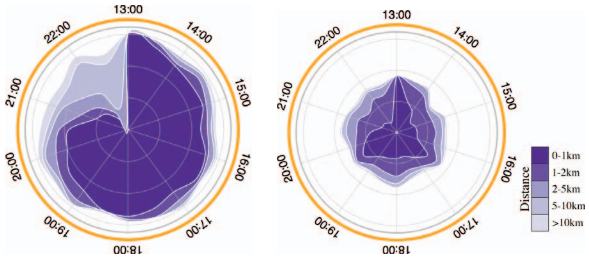


Figure 7: Flow View for customer behavior exploration.

To better understand customer sources of these shopping malls, our collaborators want a more detailed comparison about how long customers will stay and how far they will go after shopping. Thus, three radial distance maps are generated for three state clusters at 1 pm corresponding to three shopping malls respectively. We can see that most part of the radial distance map for Thumb Plaza (Fig. 8b) is in dark purple, implying that the major customer sources of this local shopping center is residents living in the nearby neighborhood. In contrast, there are obviously thicker layers in light purple in the left part of the radial distance maps for IFC mall (Fig. 8a) and Next Age store (Fig. 3d), implying some customers of these two shopping malls live in distant places. Moreover, the radial distance map for IFC mall has a relatively smaller part in light purple showing that customers leave later at night. Thus, we infer that there are more entertaining events in IFC mall, which is also verified by local people. In this way, the radial distance map can be viewed as a visual signature characterizing customers of a shopping mall, which helps analysts get an idea about its customer constitution and business mode.



(a) (b)

Figure 8: Radial distance maps for (a) IFC and (b) Thumb Plaza.

VII. Expert Interview

We demonstrated our system, presented the use cases to four domain experts invited by our collaborators, and conducted one-on-one interviews to collect their feedback. Two of them are project managers (Expert A & B) from a telecom corporation, another is an analyst (Expert C) from an urban planning bureau, and the last one is a professor (Expert D) who has studied trajectory data for more than ten years. Their feedback is summarized as follows:

Interactive Visual Design: All experts are impressed by the design of TelcoFlow. Expert A commented "*It's an excellent idea to adopt the state-based model which is an interesting and expressive description of human behaviors*." Expert B added "*Data uncertainty is always a challenge in my work. This method can alleviate this problem and help to extract hidden patterns*." They also considered to extend this method to other studies on telco data. In addition, Expert C had a particular preference to the Flow View with its glyph design and interactive features, commenting that "*The Flow View provides a vivid presentation about people [state] distributions and the overall trends, which is helpful to identify collective behavior patterns*." He also thought that such a composite design is "*great*" and "*allows me to easily connect the abstract model with underlying behaviors*". Furthermore, the flexible explorations supported by Multi-facet Filter View are also highly appreciated by the experts, and Expert D further highlighted that "*This kind of integrated methodology, visualization coupled with flexible navigation, is essential to address complex analytical tasks nowadays*." Lastly, Expert A and B considered detail analysis valuable for them to deeply inspect the data and model. They mentioned that "*The rich interactions, including snapshot, highlighting and interactive time axis, can greatly facilitate in-depth analysis*."

Usability and Improvements: All experts confirmed the usefulness of the system and expressed interests in applying it to their work. Expert A highlighted "*By using TelcoFlow, analysts can get an overview of the data and focus on specific collective behavior analysis conveniently without programming requirements or complex settings*." Expert C also commented "*TelcoFlow is not only useful for data analysis but also helpful to easily communicate my findings with colleagues or to a wide audience*." Furthermore, valuable suggestions are also provided to improve the system. Expert B suggested that we should support more data filtering operations, such as filtering based on different state cluster size. Expert D also suggested that TelcoFlow had a great potential to further explore people's behaviors for more advanced applications by combining the analysis of social media and even supporting a real-time analysis.

VIII. Discussion and Future Work

In this paper, we present TelcoFlow, a novel visual analytics system for the interactive exploration of collective behaviors based on massive telco data. Considering data characteristics and important analytical tasks raised by the domain experts, TelcoFlow incorporates state-based behavior modeling and a set of novel visualization designs, enabling analysts to integrate information for analysis from different aspects and at different scales. Case studies and expert interviews demonstrate the effectiveness and usefulness of our system.

Nevertheless, there exist some limitations of the current prototype that we would like to address in the future. From the research perspective, our system focuses on visualization and lacks sufficient support for automatic analysis, and current exploration is performed based on statistical analysis and suggestions by domain experts. To support a more efficient analysis, TelcoFlow can be easily extended to incorporate existing automatic algorithms to search for patterns systematically. Besides, in the state-based behavior model, the probabilistic distribution is derived based on one-dimensional Gaussian distribution. In the future, we can try more advanced models in civil and transportation research area, which may improve the descriptive capacity under more complex situations. Moreover, in the case study, we visually analyze three subsets of customers' behaviors using TelcoFlow. It can naturally support analysis of more subsets by adding more flows in the Flow View. But due to the limited screen space and capability of users for comparison tasks [36], the analysis of more than four subsets is not recommended. Even so, our system will still be faced with scalability issues when the size of the chosen subset is considerably large. To cope with this problem, we can extend filtering strategies in the Multi-facet Filter View to the whole system. More levels of detail and abstraction as well as advanced clutter reduction methods can also be introduced to mitigate this problem. From the application perspective, we acknowledge that information derived from the system may not be precise, as we are currently unable to obtain detailed profiles of mobile phone users due to privacy issues and can only explore based on spatial and temporal features. Nonetheless, we consider this issue is common to all existing work on studying human behaviors and plan to actively work with our collaborators for further exploration.

In addition, we'd like to extend TelcoFlow with real-time monitoring abilities with live telco data streams to support time-critical applications. It's also our plan to test our system on a larger dataset (e.g., covering longer time), and conduct controlled experiments to get more feedback from end users for further improvements.

Acknowledgment

The authors thank the anonymous reviewers for their valuable comments. This research was supported in part by HK RGC GRF 16241916, the University of Macau Grant SRG2015-00050-FST, and the National Key Basic Research and Development Program of China (973) Grant 2014CB340303. Wenchao Wu is the corresponding author.

REFERENCES

[1] H. Blumer, "Collective behavior," *New outline of the principles of sociology*, pp. 166–222, 1951.

[2] G. Naldi, L. Pareschi, and G. Toscani, *Mathematical modeling of collective behavior in socio-economic and life sciences*. Springer Science & Business Media, 2010.

[3] X. Hu, M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan, and J. Zeng, "Differential privacy in telco big data platform," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1692–1703, 2015.

[4] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni, "Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 935–944, 2016.

[5] B. P. Clarkson, "Life patterns: structure from wearable sensors," Ph.D. dissertation, Citeseer, 2002.

[6] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[7] M. Kim and D. Kotz, "Periodic properties of user mobility and access-point popularity," *Personal and Ubiquitous Computing*, vol. 11, no. 6, pp. 465–479, 2007.

[8] N. Eagle and A. S. Pentland, "Eigenbehaviors: Identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[9] K. Farrahi and D. Gatica-Perez, "What did you do today?: discovering daily routines from large-scale mobile data," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 849–852.

[10] J. Zheng and L. M. Ni, "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 153–162.

[11] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: a survey of research," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 25, 2015.

[12] G. Di Lorenzo, M. Sbodio, F. Calabrese, M. Berlingerio, F. Pinelli, and R. Nair, "Allaboard: visual exploration of cellphone mobility data to optimise public transport," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 2, pp. 1036–1050, 2016.

[13] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Politz, "Discovering bits of place histories from people's activity traces," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE, 2010, pp. 59–66.

[14] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15 888–15 893, 2014.

[15] Y. Ma, T. Lin, Z. Cao, C. Li, F. Wang, and W. Chen, "Mobility viewer: An eulerian approach for studying urban crowd flow," *Intelligent Transportation Systems, IEEE Transactions on*, pp. 1–10, 2015.

[16] Y. Zheng, W. Wu, Y. Chen, H. Qu, and L. M. Ni, "Visual analytics in urban computing: An overview," *Big Data, IEEE Transactions on*, vol. 2, no. 3, pp. 276–296, 2016.

[17] 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*, vol. 7, no. 3-4, p. 173, 2008.

[18] P. Lundblad, O. Eurenius, and T. Heldring, "Interactive visualization of weather and ship data," in *Information Visualisation, 2009 13th International Conference*. IEEE, 2009, pp. 379–386.

[19] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "Stacking-based visualization of trajectory attribute data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2565–2574, 2012.

[20] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering, "Visual traffic jam analysis based on trajectory data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2159–2168, 2013.

[21] D. Guo and X. Zhu, "Origin-destination flow data smoothing and mapping," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 2043–2052, 2014.

[22] Y. Zheng, W. Wu, H. Qu, C. Ma, and L. M. Ni, "Visual analysis of bi-directional movement behavior," in *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 2015, pp. 581–590.

[23] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual analytics of movement*. Springer Science & Business Media, 2013.

[24] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.

[25] X. Huang, Y. Zhao, J. Yang, C. Zhang, C. Ma, and X. Ye, "Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 160–169, 2016.

[26] M. Aron, N. Bhouri, and Y. Guessous, "Estimating travel time distribution for reliability analysis," *Transportation Research Arena, TRA2014, paper*, vol. 19638, 2014.

[27] J. R. Benjamin and C. A. Cornell, *Probability, statistics, and decision for civil engineers*. Courier Corporation, 2014.

[28] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. Zhang, and J. Zhang, "Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 270–279, 2016.

[29] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: some insights from statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 251–256.

[30] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2011.

[31] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007, pp. 593–604.

[32] M. Sun, C. North, and N. Ramakrishnan, "A five-level design framework for bicluster visualizations," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 1713–1722, 2014.

[33] H. Liu, Y. Gao, L. Lu, S. Liu, H. Qu, and L. M. Ni, "Visual analysis of route diversity," in *Visual Analytics Science and Technology, IEEE Conference on*. IEEE, 2011, pp. 171–180.

[34] Y. Wu, N. Pitipornvivat, J. Zhao, S. Yang, G. Huang, and H. Qu, "egoslider: Visual analysis of egocentric network evolution," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 260–269, 2016.

[35] M. S. Floater, "Generalized barycentric coordinates and applications," *Acta Numerica*, vol. 24, pp. 161–214, 2015.

[36] S. Yantis, "Multielement visual tracking: Attention and perceptual organization," *Cognitive psychology*, vol. 24, no. 3, pp. 295–340, 1992.