A Visual Analytics Framework for Exploring Theme Park Dynamics

Michael Steptoe, Arizona State University
Robert Krüger, University of Stuttgart
Rolando Garcia, Arizona State University
Xing Liang, Arizona State University
Ross Maciejewski, Arizona State University

In 2015, the top ten largest amusement park corporations saw a combined annual attendance of over 400 million visitors. Daily average attendance in some of the most popular theme parks in the world can average 44,000 visitors per day. These visitors ride attractions, shop for souvenirs and dine at local establishments; however, a critical component of their visit is the overall park experience. This experience depends on the wait time for rides, the crowd flow in the park, and various other factors linked to the crowd dynamics and human behavior. As such, better insight into visitor behavior can help theme parks devise competitive strategies for improved customer experience. Research into the use of attractions, facilities and exhibits can be studied, and as behavior profiles emerge, park operators can also identify anomalous behaviors of visitors which can improve safety and operations. In this paper, we present a visual analytics framework for analyzing crowd dynamics in theme parks. Our proposed framework is designed to support behavioral analysis by summarizing patterns and detecting anomalies. We provide methodologies to link visitor movement data, communication data, and park infrastructure data. This combination of data sources enables a semantic analysis of *who*, *what*, *when* and *where*, enabling analysts to explore visitor-visitor interactions and visitor-infrastructure interactions. Analysts can identify behaviors at the macro level through semantic trajectory clustering views for group behavior dynamics, as well as at the micro level using trajectory traces and a novel visitor network analysis view. We demonstrate the efficacy of our framework through two case studies of simulated theme park visitors.

Additional Key Words and Phrases: Visual Analytics, Trajectory Analysis, Semantic Trajectories, Behavior

This work is supported by the National Science Foundation Grant No. 1350573 and in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

DOI: 0000001.0000001

1. INTRODUCTION

The growth and adoption of technology has increased society's ability to gather data related to human mobility and interactions with their built environment. People leave their homes armed with a variety of sensors, from GPS positioning on mobile phones, to smart watches and wearables tracking heart rate and footsteps, to body cams and life-logging tools that capture images and video of their daily interactions. Such data provides an unprecedented opportunity to study and analyze human behaviors and their relationships with infrastructure. This opportunity has led to the smart city movement where urban planners envision integrating multiple information and communication technologies to help manage city assets and improve the quality of life for their citizens. However, the large scale, heterogeneous nature of the data being captured presents a variety of technical challenges for analyzing human and crowd behavior. As such, there has been an increasing demand for algorithms, methodologies, and frameworks for supporting behavioral analysis. This has led to a variety of recent visual analytics approaches developed to utilize mobility data to reveal behavioral differences between populations. For example, Wood et al. [2011] developed a visual analytics framework to explore bike hire data in London to gain insight into how people utilize the bike rental infrastructure. Other work has developed visual analytics methods for analyzing taxi trajectories [Chu et al. 2014; Ferreira et al. 2013; Zhu et al. 2012] to extract traffic patterns and trends in a city. Such studies exemplify how visual analytics can help support the analysis of data sets that capture information about a person's mobility.

While the previous studies have tended to focus on the large scale dynamics within a city, providing global level insights into behavior in terms of peak hour usage, common commuter routes and so forth, other studies have focused on exploring the use of sensing technology within a well-defined, but densely crowded area [Guest et al. 2013; Kim et al. 2008]. Examples of such locations include university campuses, stadiums, and amusement parks. These locations are increasingly employing sensor technology to gather intelligence about their visitors. For example. universities record keycard access at the fitness center, dorms, and cafeteria, and theme parks provide opt-in wristbands which provide guests with enhanced services (e.g., quicker line queuing) while tracking their information. More recently, Disney has even filed a patent for tracking visitor location by taking images of their shoes [Beardsley and Taneja 2016]. As such, it is clear that understanding crowd behavior is critical for business intelligence (as well as disaster management and emergency response). By understanding where visitors go and what services they utilize, industries can potentially revamp their operations to improve the underlying user experience and increase revenue. Furthermore, such locations are often viewed as potential targets for terrorists. By analyzing crowd behavior in real-time, visitors engaging in anomalous behaviors can also be detected for improved security and response.

In this paper, we present a visual analytics framework for exploring behavioral dynamics for a well-tracked, highly defined area. While the focus of our paper and examples is on theme parks, the proposed framework is applicable to any geographical region where densely sampled trajectory data can be semantically linked to the underlying infrastructure to provide insight into behaviors centered around infrastructure usage. Our goal is to provide analysts with methods to explore the decisions that theme park visitors make. This enables competitive planning for facility and exhibit placement, can help identify the need for general public services (e.g., restrooms), and detect anomalous behavior at the individual level. Our visual analytics framework consists of four parts: 1) constructing trajectories from movement tracks including inferring missing records and aggregating different sampling granularities; 2) enriching trajectories with semantic information from heterogeneous data, including communication data and location types, to help analysts interpret and understand movements; 3) integrating data mining techniques to automate the process of extracting characteristics from large semantic trajectories, especially for crowd patterns and anomaly detection, and; 4) aligning and reporting findings together in a report view.

Our visual analytics framework is not the only one developed for the exploration and analysis of the VAST 2015 Challenge dataset. Other research groups from both university and industry developed their own frameworks consisting of data preprocessing, linked views, and the application of clustering and community detection methods. Ye et al. [2015] developed a collaborative visual exploration system to perform behavior analysis over trajectory records using only the movement data from the first challenge. Their system allows for spatial and temporal exploration with an analysis pipeline adopted from Andrienko et al [2013]. The analysis pipeline consists of: categorizing people who spend most of their time together using DBScan, extracting features from the groups based on each type of point of interest, and calculating the group similarity. Li et al. [2015] designed TMViz and DNViz for collaborative analysis to discover interesting patterns in the communication data from the second challenge. Before visualizing the data, they performed a three-step data preprocessing. Their preprocessing includes community detection, attribute extension, and aggregate calculation. TMViz uses temporal and multifaceted features of the communication data, while DNViz allows for the exploration of communication behaviors in a community or a custom set of individuals from dynamic perspectives. Puri et al. [2015] developed ParkVis, a visual analytics system for tracking unusual patterns of park patrons utilizing both the communication and movement data from the first and second challenges. ParkVis employs two workflows, one focusing on the use of Multidimensional scaling (MDS) with check-in filtering and the second using spatial and temporal filters to identify unusual patterns. Each of these frameworks presented their own methods for exploring and analyzing the dataset: some focusing on one challenge in particular, while others combining both the movement and communication data into a single system. Of the 71 VAST Challenge 2015 submissions, our team was the only one to correctly identify the culprit, and was awarded the Grand Challenge award for Outstanding Comprehensive Submission. The major contributions of our framework include:

— A visual analytics framework to support the discovery of patterns and anomalies in trajectory data.

— A semantic clustering methodology for identifying group behavior by integrating movement traces and location information, and;

— A novel network analysis view for exploring visitor communications with respect to space and time.

In order to demonstrate our framework, we provide examples using two synthetic datasets. Our first dataset uses the theme park visitor modeling work of Solmaz et al. [2015]. A park with an average attendance of 2500 visitors is modeled and analyzed over 30 days. General insights into the model are discussed. Our second dataset is the IEEE Visual Analytics Science and Technology (VAST) 2015 challenge DinoFun World data set consisting of 11,317 park visitors over the course of 3 days and has an embedded robbery event. We report on how the framework can characterize crowd behavior in the park and infrastructure usage.

2. RELATED WORK

Given the abundance of data related to human behaviors coming from fine resolution GPS traces of taxi cabs to more coarse grained social media posts on Twitter or Foursquare, a variety of visual analytics methods have begun emerging. Recently, there has been an increasing amount of literature focusing on helping people define, process, visualize, and analyze movement data. In this section, we will discuss recent work that uses mass mobility data for exploring crowd dynamics and behavior. We summarize techniques for aggregating movement data (Section 2.1) and enhancing such data with semantic information from secondary sources (Section 2.2). Finally, we will cover recent visual analytics methodologies for exploring such data (Section 2.3).

2.1. Trajectory Data Extraction and Aggregation

The most common type of raw movement data is a series of geographical coordinates along with the change of time, often referred to as trajectory data [Andrienko et al. 2011b]. Due to the performance of tracking devices (e.g. GPS locator or RFID), there may exist oversampling or undersampling problems in the geographical coordinates. Andrienko and Andrienko [2010] review the existing approaches to aggregate raw movement data for distilling general features out of fine-detailed particulars, and other recent work by Andrienko and Andrienko [2011] suggests separating the space into several sub-regions and simplifying the trajectory along geographical coordinates to the movement among sub-regions when data is over-sampled. If the movement data is undersampled, which means the position is unknown in some time intervals, a variety of methods to estimate the position using linear or non-linear interpolation have been developed, e.g. [Macedo et al. 2008; Andrienko et al. 2011a]. Our framework builds off of previous work, identifying stops and starts in trajectories. In order to incorporate domain knowledge into the extraction and aggregation phase, a data preprocessing step is incorporated such that the user can choose the temporal level of aggregation from fine-detailed trajectory data as well as define stop and start criteria.

2.2. Semantic Enrichment and Data Alignment

While large amounts of trajectory data are now available for analysis, the data by itself often contains little information regarding behavior. Such data needs to be linked to secondary data sources (such as points of interest for geographic reference) in order to provide semantic meaning to the trajectories, and Laube [2014] lists the semantic gap as one of the critical challenges in movement analysis. Semantic information can be derived from the data itself or complemented by other data sources. From the movement data itself, analysts can link the acceleration, speed, stops, directional change, length, and other physical characteristics in the movement data as features of trajectories [Bogorny et al. 2014]. For example, Wang et al. [2013] extracted traffic jam events from the speed of moving objects in traffic trajectories. Zeng et al. [2013] formulated the interchange pattern from how the moving objects redistribute when entering and passing through a junction in a traffic network, and Wu et al. [2016] focused on the co-occurrence pattern of human mobility data looking at how crowds from different regions visit the same place during the same time span.

While derived semantics (stops, starts, stays, etc.) from the movement data can provide insight, another critical area is to link related datasets to enhance the interpretation and understanding of movement data [Bogorny et al. 2014]. Andrienko et al. [2007] proposed a visual analysis tool that extracted significant places and predefined the types of places in movement data to find common travel routes. Krüger et al. [2015] presented an interactive visual approach to enhance movements with point of interest information from the web and proposed an uncertainty-aware exploration of spatial and temporal patterns. Yan et al. [2009] proposed a multi-layer trajectory model in which raw trajectory data is first reconstructed to structured trajectories and then linked with geographical information and application domain data to fill the semantic gap. Yan et al. [2011] implemented the SeMiTri framework which takes a list of annotations associated with trajectory data as semantic information. Similarly, our framework utilizes derived information from trajectories and links the trajectories to the surrounding geographical information. In this manner, we can gather information about where visitors stopped (and for how long) as well as aggregate this over semantic categories (i.e., were the visitors roller coaster junkies or foodies).

2.3. Visual Analysis of Trajectory Data

Cleaning, aggregating, and augmenting trajectory data are critical steps in the analysis process; however, once these steps are completed, analysts still need to explore the data. As such, a variety of visual analytics methods for the analysis of movement data have been developed. When visualizing movement data over static maps, the most common challenges are visual clutter and occlusion [Andrienko and Andrienko 2013]. Multiple solutions to alleviate these problems have been proposed. For example, Andrienko and Andrienko [2011] proposed a spatial generalization and aggregation strategy in which the space is divided into compartments and trajectories are transformed to the moves between compartments. Other work by Andrienko and Andrienko [2013] focused on using animated trajectories within a selected time interval from interactive time filters. Bouvier and Oates [2008] presented a technique named “staining” in which the moving objects will be stained with colors when passing by marked items as a means of identifying common intersections. Furthermore, in Bouvier et al.’s system, users can see when and where the moving objects were stained by reversing the animations.

Along with animation, other researchers have focused on three-dimensional visualizations. For example, Tominski et al. [2012] visualized trajectory attributes using three-dimensional stacked trajectory bands and encoded attribute values by color. Bach et al. [2016] proposed a series of space-time cube operations to turn a part of the three-dimensional visualization into an easily-readable two-dimensional visualization for analyses. Other work has focused on using small multiples and coordinated views. Boyandin et al. [2012] conducted a user study over animations and small multiples. Results indicated that animation was good for exploring geographically local events and changes between subsequent time frame while small-multiples helped identify long-term trends that can be missed during animation. Zhao et al. [2014] designed a coordinated multiple view system that visualized both network traffic data and IDS alert data together. Chen et al. [2016] explored geo-tagged Weibo posts and visualized movement patterns using a matrix view, timeline view, map view, bar chart and others.

One critical challenge these visualization face is the sheer volume of trajectory data. GPS traces from taxis or wearable devices can provide information at sub-second resolution. This means that for a single individual, the trajectory trace can be extremely large. Once these are collected for an entire city, the sheer volume of data can often overwhelm the visualization and reduce interactivity. As such, a variety of machine learning techniques are often employed as part of the visual analytics pipeline, specifically, clustering and classification are frequently adopted, and several recent works have surveyed the application of clustering techniques in spatio-temporal data [Kisilevich et al. 2009; Diggle 2013; Bermingham and Lee 2015; Zheng 2015].

Clustering and classification techniques have been used to identify behaviors from individual trajectories or groups of trajectories. For example, Andrienko et al. [2009] proposed four types data mining techniques for identifying behaviors: 1) clustering trajectories sharing similar features (e.g. shape, direction, and speed); 2) classifying trajectories into predefined classes (e.g. types of transportation in traffic data); 3) uncovering shared paths of trajectories, and; 4) identifying similar trajectories (e.g. bird flock). However, work by Andrienko et al. [2009] also notes that while data objects are often represented by feature vectors, moving objects in trajectories cannot be properly represented in this way. Each trajectory may have different lengths and their sequential and spatial properties are difficult to encode in the feature vector. Therefore, when using traditional clustering algorithms, such as K-Means and KD-Tree, on trajectory data, it is important to utilize proper distance metrics for clustering. Zheng [2015] reviews a variety of distance metrics for trajectory analysis, and Sun and Wang [2015] detail a comparative study over different features and distance metrics for measuring similarities of trajectories.

Critical to clustering and classifying trajectories is deciding on the length and features to use. Zheng et al. [2008] propose partitioning a trajectory into several segments and extracting a set of features (such as the heading change rate, stop rate, and velocity change rate) for use in a Decision Tree Classifier to classify the transportation types rather than simply relying on simple rules such as velocity-based approach. Pelekis et al. [2012] cluster the entire trajectory using several distance measures based on different derived parameters of trajectories (speed, acceleration, and direction). Rather than clustering over the full-length trajectory, da Silva et al. [2016] utilize an incremental algorithm to continuously capture the evolution of trajectory data streams and performed clustering over newly captured trajectory patterns. Other work first classify trajectories into regions-of-interest segments before performing clustering. For example, Giannotti et al. [2007] identified regions-of-interest using the density of trajectory data, and then proposed an algorithm to detect the sequential patterns of the region of interest.

While clustering and classification focus on finding commonalities and groups of trajectories, another critical task is outlier detection. There are two primary methods of labeling a trajectory as an outlier. First, an outlying trajectory cannot fit into any cluster or does not show up frequently [Zheng 2015]. For example, Pan et al. [2013] identified anomalous taxi routes by checking if their routes are different from their usual routing behaviors, and Liu et al. [2011] measured the extremeness of data points both from their temporal and spatial neighbors using Mahalanobis distance. Second, an outlying trajectory could be anomalous not due to the trajectory, but due to other attribute features. For example, in a traffic jam, even though all taxis follow the same route, an outlier can be recognized as a low-velocity trajectory [Wang et al. 2013].

3. ANALYTIC TASKS AND DESIGN REQUIREMENTS

Although the visual analytics solution presented in the following (Section 4) is applicable to a range of datasets and tasks (e.g., emergency response, smart cities, etc.), it was originally motivated by the 2015 VAST Challenge that dealt with amusement park visitor behavior of a fictitious theme park (see Section 5.3 for a more detailed description). The grand challenge required combining information gained from two mini-challenges to solve a crime and identify the culprit(s). Movement and communication data of the park visitors were given. From this data, a variety of analytic tasks were outlined. The analytic tasks focused on exploring the visitor's behaviors to understand how park infrastructure was used. Analytic requests from the contest included:

(1) Characterize park attendance including group behavior.
(2) Characterize and identify any abnormal behavior (either by groups or individuals).
(3) Characterize communications patterns within the park.
(4) Characterize operational behavior of attractions in the park.

Based on these analytic questions, we developed the following design requirements:

(1) The framework will enable the user to seamlessly switch focus from aggregate trajectories to group trajectories to individual trajectories in a nonlinear and iterative fashion, and in a manner that supports effective hypothesis formation and refinement.
(2) The framework will enable the user to obtain a pre-attentive grasp of the scale and order of time and space, in a manner that supports fast judgment about which groups or patrons restricted their activities to an extraordinarily small amount of space, visited some attraction for an extraordinarily long time, and visited which attraction before/after which other individuals.
(3) Immediate access to information about crowd and communication density at each attraction is necessary at regular and reasonable time intervals. The information shall be provided to the user free of clutter and in a manner that is unobtrusive so the user can access the information on-demand.
(4) The framework will enable the user to make fast judgments about which attractions have extraordinarily high or low levels of attendees, and be able to quickly compare crowd densities across different attractions. Crowd density information shall be linked to attraction geo-location information on-demand.
(5) The system should support fast judgments about the attraction-visit history of each individual of interest, as well as the personality profile of each visitor, in addition to being able to compare individuals easily to determine who moves with a group and who moves alone.
(6) It should be possible to determine whether any coordination occurred between visitors, despite those operating at a distance and moving independently, by analyzing and superimposing communication data over movement data.
(7) The user shall be able to generate reports containing movement, communication, and network information on identified aggregates, groups, and individuals.

Because analysis tasks, data structures, and analytical means to solve the tasks can be complex, we consider a team of experts, similar to [Arias-Hernandez et al. 2011], for an optimal usage of the visual analytics framework.

— *The Visual Analytics Expert* is able to work with the interactive visual interface, to filter and select data, steer algorithms and choose visualizations well-suited for the analysis task. The visual analytics expert is able to communicate with domain experts that define the tasks and analysis goals.

- *The Domain (Subject Matter) Expert* may either have knowledge about the amusement parks infrastructure and security guidelines, or has a background in behavior psychology, crowd dynamics, and has access to other relevant information.
- *The Decision Maker* acts on the information presented by the domain expert (or may be the same person), has significant authority, and bears liability for mistakes. Hence, the decision maker will require the domain expert to make a strong case regarding whether to pursue some course of action.

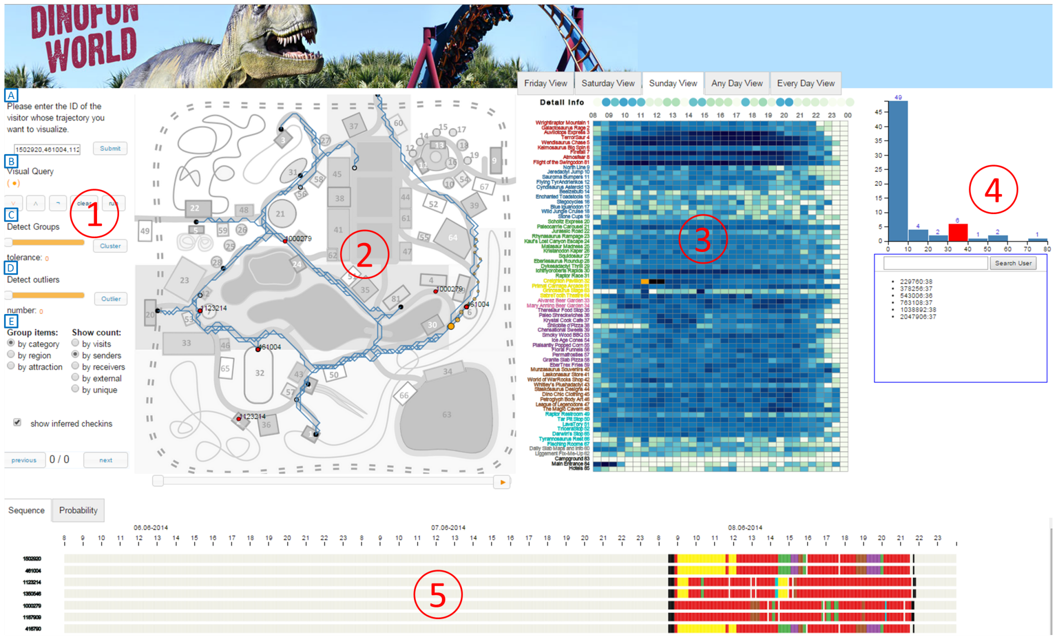


Fig. 1. System overview. View (1) shows the analytics interface which allows the user to adjust the systems parameters. View (2) shows a map of DinoFun World and the trajectories of selected visitors. The blue line represents the patron's traversed space in the park, and the red dots represent other park patrons, at their actual locations at that time, who were contacted by this patron. View (3) shows a calendar which represents each attraction in the park and displays the count of visitors in 30 minute intervals. View (4) shows a histogram which displays the number of calls made during a time period. View (5) shows a pixel based representation of a visitors trajectory in 5 minute intervals.

4. VISUAL ANALYTICS FRAMEWORK

Based on the design requirements, we developed a visual analytics framework to support group dynamics analysis and implemented the framework in a web-based analytics system (Figure 1). The system provides multiple coordinated views that are enhanced by automated methods such as clustering, community, and outlier detection for coping with heterogeneous and complex datasets. Different views allow for an analysis of spatio-temporal as well as communication related data aspects. A calendar view [Van Wijk and Van Selow 1999], where columns are time slices and rows represent locations, gives an overview over the monitored days and serves as a starting point for the analysis. For further spatio-temporal analysis, individual traces are displayed on a map view. Group dynamics are explored through a pixel based trajectory cluster view. Communications between patrons are also captured, and details of who-when-where are tracked in the communication view. Finally, a summary view is also linked to explore semantic details about visitor behavior (e.g., groups that ride roller coasters, groups that strongly communicate with each other, stay at the park hotel, etc.) and present findings to a wider audience. As such, our system harnesses data mining techniques for data exploration, enabling powerful visualization techniques, comprehensive analyses, and fast response times.

4.1. Data Preprocessing

As previously mentioned, a major challenge in analyzing trajectory data is the scale and the need to link to secondary data. For a theme park, data may be captured

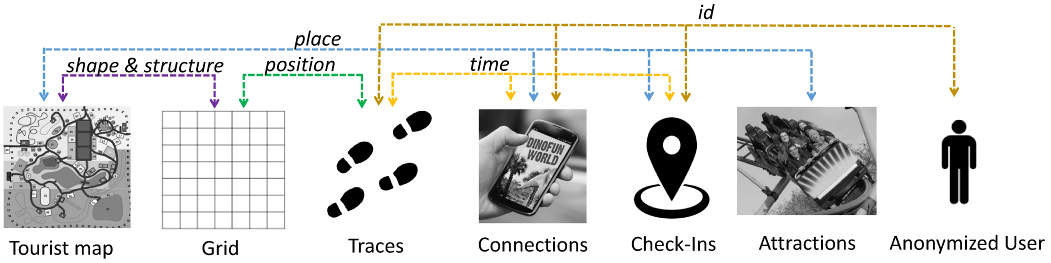


Fig. 2. The synthetic datasets comprise multiple heterogeneous data sources that have to be fused in order to enable behavior analysis. From left to right: tourist map, spatial grid, visitor traces, connections per visitor, attraction check-ins, attraction information (region, category, etc.), anonymized user information.

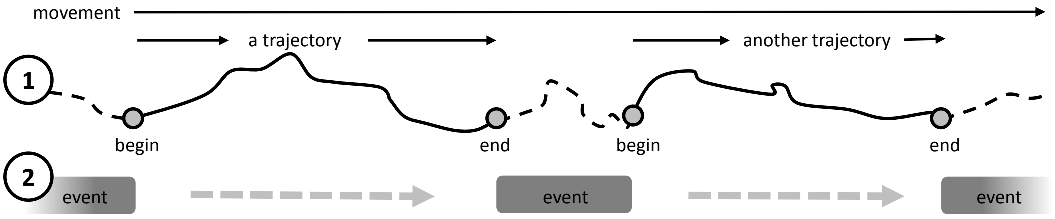


Fig. 3. Based on an illustration by [Spaccapietra et al. 2008] 1) Visitor movements are described by a set of trajectories (focus on the movement itself). 2) In between movements a person follows different activities taking place at a specific time at a specific place (event). Depending on the application of the movement itself (1) or the sequence of events taking place (2), an event of interest can be identified.

about the park infrastructure, movement, communication between visitors, etc. Figure 2 sketches a simplified version of the dependencies noted within the theme park data.

We first set up a relational database model and subsequently create the links between the individual data sources based on common timestamps, IDs, and attractions visited. Specifically, given a map of the park, a user defined temporal threshold, τ, and a spatial threshold, γ, an event is defined within each trajectory such that if the trajectory of visitor i is located within some γ + $epsilon_{\gamma}$ of some known park infrastructure for a time span $[t, t + \tau]$ then that trajectory segment can be semantically linked to the location as an *event*. Otherwise, the trajectory is considered to be in a *move* state.

After the trajectories are enriched with contextual data, the visitor behavior can be described as a sequence of events $s := e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow ... \rightarrow e_n$. Here, an event is a visitor using infrastructure, e.g., a park visitor that is checked in to a certain attraction. When information about visitor-visitor communication is available, events can also be defined as instances when one visitor contacts another visitor. While communication events could also be formulated as the spatial links, the density of park visitors would add too much uncertainty in those instances, and future work will explore methods of identifying relevant visitor-visitor interactions at locations. Figure 3 illustrates the semantic trajectory extraction. This process is general for any spatiotemporal trajectory data where point-of-interest data can be utilized to semantically enhance trajectories.

4.2. Spatiotemporal Analysis

Once the data is pre-processed, the primary view of our system focuses on spatiotemporal analysis for identifying group behaviors and individual anomalies. Figure 1 provides an overview of the spatiotemporal views as applied to the *DinoFun World* dataset. The main window consists of five coordinated views (the analytics interface, the map view, the calendar view, the distribution view, and the event sequence view) and is complemented by communication views (Section 4.3) and a reporting view (Section 4.4).

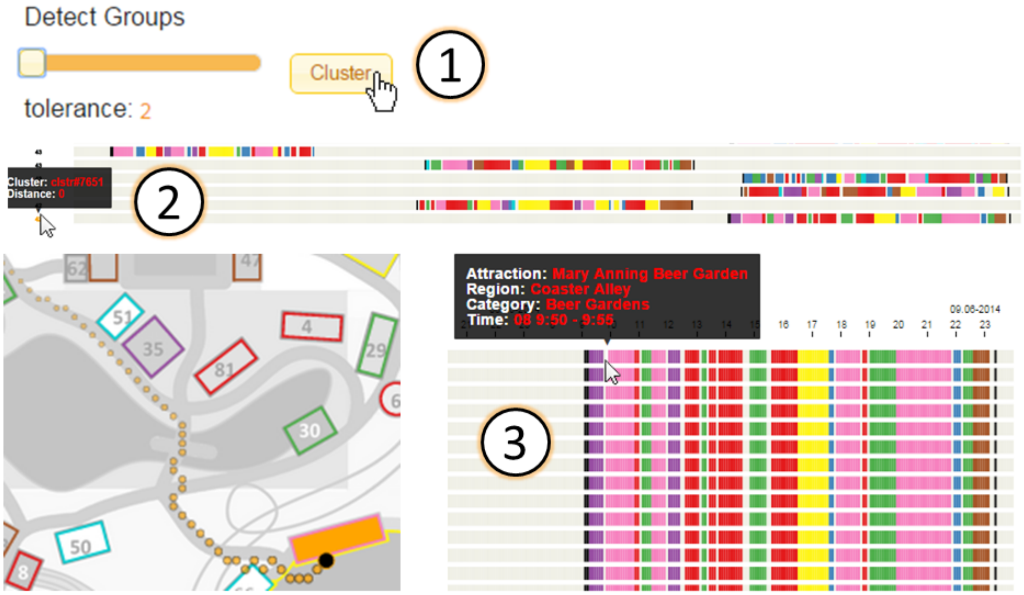
The Analytics Interface A main component to steer the spatiotemporal analysis is the analytics interface. It allows the analyst to adjust the system's parameters and filter the visualizations depending on the objective of their exploration. The development of this interface was driven by design requirement (1), which allows the user to effectively control and refine the information being displayed from the highest level of abstraction, the aggregate, down to the individual park visitors. The panel is made up of five sub-controls referenced with letters A-E in Figure 1-1. In control A), the analyst can input a list of visitor IDs and view their trajectories on the map view, Figure 1-2. Pressing play will animate the trajectories. The sequence of check-ins for each visitor is displayed in the event sequence view, Figure 1-5. The communications data is also visualized as points on the map that appear at the time of the call. Using control B), a user can select time and location intervals on the calendar view (Figure 1-3) and create a visual query with logic operators (AND/OR/NOT). This query will return, for example, the IDs of all park visitors that were at attraction 38 at 4PM and at attraction 45 at 9PM on Friday. This is the primary feature for exploring behavior hypotheses (e.g., do people visit a certain souvenir shop prior to leaving) as well as finding users that were at locations of interest at particular times. IDs and trajectories that are returned from the query are plotted in the event sequence view.

Along with filtering individual trajectories, our visual analytics system is also designed to cluster visitors as a means of identifying group behavior within the park. All visitor trajectories are clustered using an agglomerative hierarchical clustering with a Levenshtein distance (for details, see Section 3.5). Control C) is designed to allow users to select a clustering tolerance. If a tolerance of 0 is selected, the resultant clusters consist of the IDs with identical trajectories (in terms of locations visited at the same time). Reducing the tolerance provides fuzzier clusters (i.e., they have visited mostly the same locations at the same time during their stay). The groups found are plotted in the event sequence view where the trajectory of each cluster is represented by a single row of pixels. Since there are multiple trajectories within each cluster, the result shown is the most representative trajectory of the group.

In addition to exploring group dynamics, our system also contains methods for outlier detection to detect unusual individual behavior, control D). Using the Levenshtein distances calculated during clustering, trajectories that are very far from any other trajectory can be considered as a unique event sequence. This slider returns the top n-IDs with the largest distance (see Section 3.5 for a technical description of the approach). Results are plotted in the event sequence view.

Finally, the calendar aggregation options of control E) define how the rows in the calendar view are sorted (by region, attraction, or ride type), which event-granularity is used for the clustering, and how the data is plotted in the cells of the calendar view (data can be the number of visitors at a ride or the number of sent/received/external/unique calls sent from a ride at time t if communication data is also captured, Figure 1-4).

Map View Next to the analytics interface is the map view, (Figure 1.2) which provides a spatial representation of the trajectory and communication data. This view was developed in alignment with our design requirements (2-4) enabling users to grasp locality of the attractions, groups, and visitors within the park and supports coordinated filtering for cross-view analysis. The map view can be utilized to



tolerance: 2

Cluster: clstr#7651
Distance: 0

Attraction: Mary Anning Beer Garden
Region: Coaster Alley
Category: Beer Gardens
Time: 08 9:50 - 9:55

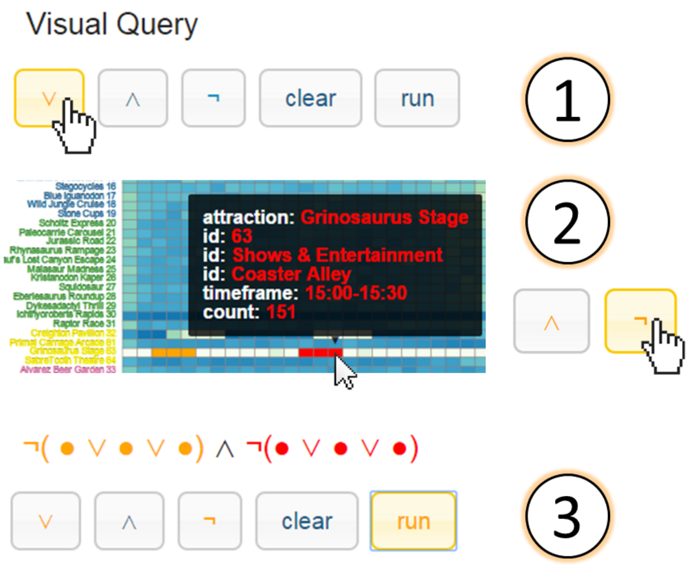
Fig. 4. Three steps to run and explore the clustering. 1) Define the tolerance level (hierarchy cut). 2) Explore the cluster representatives. 3) Unfold and explore the individual cluster members.

show either individual movements or aggregated movements of clusters of visitors. For individuals, the visitor's traces and communications can be animated over time and are linked to the event sequence view. In addition, a so called footprint visualization is also shown, when the user brushes over time slices in the event sequence view (Figure 4 lower left). For aggregated trajectories, a bi-variate heatmap view colors each pixel on the map by the number of times a visitor passed the area. The color-coding ranges from white to blue, indicating the number of passing visitors and to green, indicating the standard deviation of the distribution, accordingly.

Calendar View The calendar view (Figure 1-3) provides a temporal overview of attraction check-in volumes for each day. In order to support our design requirements (3-4), the calendar view was developed to give users immediate access to information about crowd density at regular and reasonable intervals allowing the user to quickly compare densities across different attractions. Each row represents an attraction in the park and each cell is colored based on the calendar aggregation control as previously defined. Data can be viewed for each day, or as an aggregate.

The every day view shows the counts of IDs that were at the same place at the same time every day and provides functionality for exploring repeat visitor behavior. Each cell represents 30 minutes of time. By activating the visual query mechanism in the analytics interface, the analyst can click on time slices of interest, setting up different conditions for the query. Figure 5 shows an example of this interaction.

Event Sequence View The event sequence view (Figure 1-5) allows analysts to explore temporal semantic patterns of either individuals or groups of visitors. Using this view, analysts can quickly determine how long visitors and groups are spending at attractions and which attractions their activities are restricted to. Analysts can also use this view to make fast judgments about attraction-visit history and easily

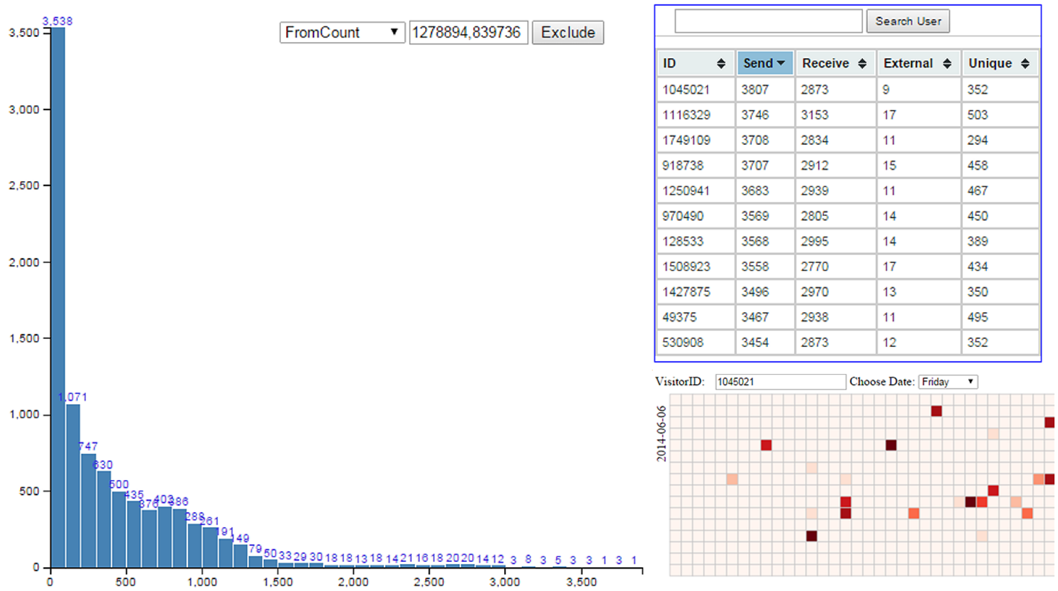


attraction: Grinosaurus Stage
id: 63
id: Shows & Entertainment
id: Coaster Alley
timeframe: 15:00-15:30
count: 151

Fig. 5. The visual querying is defined directly in the calendar view. 1) The analyst sets the mode (here OR) in the analytics interface. 2) The analyst marks areas that have match trajectory sequences. 3) The analyst runs the query. Results are displayed in both the event sequence view and map view.

compare individuals to determine who moves with a group or alone, supporting design requirements 2 & 5. At the individual visitor level, each row is a pixel based representation of a visitors event sequence, where events were calculated as defined in Section 4.1. Each cell represents a 5 minute time interval (which is adjustable in the pre-processing stage but not interactively due to the large data size) that is colored based on the location where a user is in the park. By hovering over a cell, the attraction or path associated (depending if the cell is an *event* or *move* segment) with a visitor is highlighted on the map. When clusters of visitors are explored, each row becomes a representative trajectory from the cluster. By clicking on the representative trajectory, the cluster unfolds and each cluster member can be explored individually. The view provides pagination to browse through larger amounts of clusters and individuals. Figure 4 shows three interactive steps carried out by the analyst to run and explore the clustering results.

Distribution View The distribution view provides users with immediate access to information about communication density at each attraction fulfilling the remaining goals of our design motivations (3.3). It thereby links the communication data to attraction and movement information. A rough overview of communication volumes is given in the histogram view (Figure 1-4). The histogram view shows the number of sent/received/external/unique calls made during a time period. The y-axis is the number of IDs, and the x-axis is the number of communications made. Users can click a bin to see all the IDs in a bin. In this way, we can find those IDs with extraordinarily large amounts of communications. Using this visualization as a starting point, communication patterns can be explored more in-depth in the communications view (Section 4.3). Note that all other views still function if such data is unavailable in order to allow for generalizability of the framework.



ID ⬍	Send ▾	Receive ⬍	External ⬍	Unique ⬍
1045021	3807	2873	9	352
1116329	3746	3153	17	503
1749109	3708	2834	11	294
918738	3707	2912	15	458
1250941	3683	2939	11	467
970490	3569	2805	14	450
128533	3568	2995	14	389
1508923	3558	2770	17	434
1427875	3496	2970	13	350
49375	3467	2938	11	495
530908	3454	2873	12	352

VisitorID: 1045021 Choose Date: Friday ▾

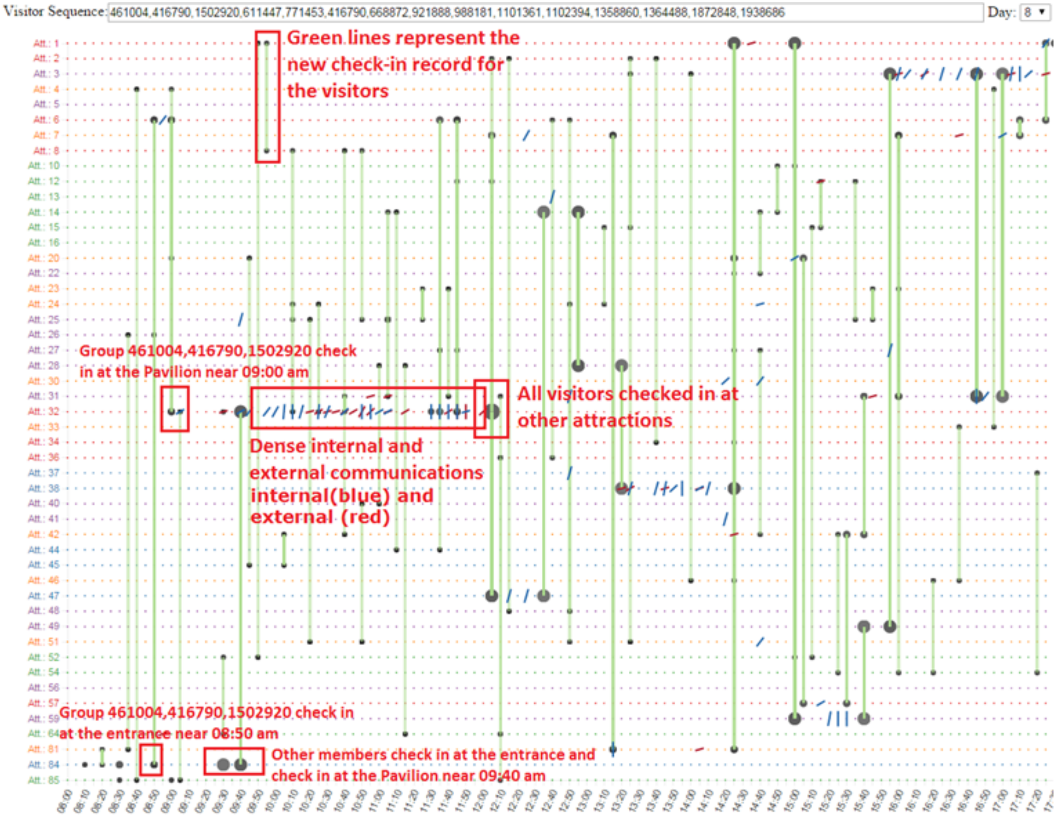
Fig. 6. The communication view shows a histogram with the distribution of sent messages. The x-axis represents the sent message volume and the y-axis represents the number of visitors. The top-right section of the view is a table showing the specific amount of the four types of messages for all visitors within a user selected bin. The lower-right section contains a matrix view showing the distribution of sent messages for visitor 1045021, which is the visitor sending the largest amount of messages in the left histogram view.

4.3. Communication/Network Analysis

Our primary analysis view focused on spatiotemporal relationships of individuals and their built environment (visitor-attraction interactions); however, visitor-visitor interaction can also be captured (e.g., mobile phone records). While such data is relatively uncommon, in the VAST Challenge 2015 dataset, messaging data between visitors (which contains a timestamp, the originators ID, the recipients ID, and the park area from which the message was sent) is captured. These connections between individuals allow for even finer grained behavioral analysis (e.g., identifying parents and children who visit different attractions and then call to arrange meeting back up).

Compared to trajectory data, messaging data is composed of communication traffic throughout the network and communication networks between message originators and recipients. To extract communication patterns from the message data, researchers have analyzed the network structure and traffic using measures of network symmetry, centrality, traffic volume and other factors [Onnela et al. 2007; Eagle et al. 2009; Ying et al. 2010]. The system's communication view focuses on identifying influential nodes in the network structure and critical stages in communication time series and consists of two unique views (Figure 6 and Figure 7). We derived four features, sent/received/external/unique message amount, and combined a bar chart and matrix view together to display and characterize visitors' messaging behaviors from which we can find meaningful communication patterns and infer their roles and locate the anomalies. Our two views allow analysts to determine coordination between visitors by analyzing and superimposing communication with movement data, meeting design motivation 3.6.

Detecting Influential Nodes While a simple message is built on the communication between two or more nodes, the direction and volume of the message can help identify



Visitor Sequence: 461004,416790,1502920,611447,771453,416790,668872,921888,988181,1101361,1102394,1358860,1364488,1872848,1938686

Day: 8

Fig. 7. The network analysis view, the x-axis represents the time from 08:00 to 24:00 and the y-axis represents the attractions that the visitors are near. Circles represent the movement of the visitors from one attraction to another, and the circle size encodes the number of visitors. Green lines connecting attractions represent a new check-in record for visitors. Blue lines on attractions represent internal communications and red lines represent external communications.

behaviors. We categorized the communication data into four categories: sent messages, received messages, external messages and unique messages. The sent and received messages are to/from communications between visitors located in the park. External messages are messages whose recipients are outside the park. Unique messages is a count for messages sent out simultaneously. For example, if a visitor sends out multiple messages at the same time, all of them will only be counted once in the unique volume. These four features can be explored in the distribution view defined earlier. Clicking in the distribution view of Figure 1 will lead to an expanded matrix view to further explore the messaging behaviors of outstanding nodes, Figure 6.

In the specialized communication view, Figure 6, the left part is a histogram showing the distribution of sent messages. Users can change the type of the message being explored (sent, received, external, unique) through the dropdown in the top middle of the view. The x-axis represents the sent message volume and the y-axis represents the number of visitors. In this view, users can identify the amount of sent and received messages. Clicking on a bin creates a list view of the IDs (and their communication meta-data), and clicking on an individual ID will populate another calendar matrix view. The matrix view in Figure 6 shows that the selected ID's communications are not smoothly distributed over time but burst out at certain time intervals. Users can also identify that this ID is present in the park for all of the data and can begin developing hypotheses (for example, one might hypothesize that such an ID is likely to be park staff).

Network View Analysis of the spatiotemporal communication data also requires the examination of visitor communications with respect to location, time, and frequency. We have designed a specialized visualization component, Figure 7, to enable the analyst to explore when and where visitors communicate, assemble/disband, and the frequency of user interactions. Input to the view is a list of visitor IDs and the day to examine. The x-axis represents the time from 08:00 to 24:00 and the y-axis represents the attractions that the visitors are near. Circles represent the movement of the visitors from one attraction to another and the circle size encodes the number of visitors. A circle consists only of visitors that have a link in the communication data. Looking at Figure 7, a small black circle can be seen in the lower left at attraction 84. The small black circle represents multiple people simultaneously checking-in at that time. Later, part of the group at attraction 84 moved to attraction 1. A green line is then drawn to emphasize this movement. By hovering over the line we show if visitors joined or left this group. If the circle changes to red, this means that a person left the group at this attraction. However, if the circle changes to blue a person joined the group.

Blue lines represent the internal communications among the visitors and red lines represent their external communications. While communication information can be animated on the map view, the challenges of remembering when communications occurred (or overlap if data is plotted as a single image) remain as challenges. As such, this view provides an alternative to animation, plotting movement and communication events simultaneously. Future extensions will explore methods of reordering spatial locations on the x-axis for reducing visual clutter.

4.4. Reporting

Our final visualization component is the reporting view which is designed to allow analysts to create summarized views of group behaviors for reporting and analysis. The reporting view was designed in conjunction with design requirement (6). This view contains components (referred to as ”cards”) for movement and communication data by visitor, group, or group type. Each component is placed in its own small multiple view to allow for a quick overview and analysis. This view is designed primarily to enable the analyst to save and compare findings of interest in a sandbox like setting, Figure 8.

Movement View Card In this small multiple selection (Figure 8 - Movement View), a user can select multiple visitors, groups, or group types to examine their spatiotemporal movement data. This view provides users with a space to compare the movements of visitors, groups, and group types. For each selected item a bivariate heatmap is shown along with the trajectory traveled. The mean and standard deviation of visitor movements are calculated for each point on the grid and displayed using the bivariate heatmap. The trajectory can be displayed for any day of data or an aggregate of all days and can be colored by category, region, or attraction. When viewing individual visitor movements the user can click on a visitor card and view the movements of all of the visitors that they communicated with. When viewing a group or group type the user can click on either of those cards and view the movements of the visitors that make up the entity.

Feature View Card In this small multiple selection (Figure 8 - Feature view), a user can examine the types of attraction visitors frequent in the park through a histogram displaying attractions visited by category. The x-axis represents the category of attraction visited and the y-axis represents the number or average number of attractions visited within that category. The icon in the upper-left hand corner of each card shows how the visitor entered the park, in the case of the VAST Challenge they could enter from the main entrance, the hotels, or the campground. The icon in the center of the card above the histogram shows how many days the visitor, group, or group type spent in the park. Similar to the functionality within the movement view card, the user can click on a visitor card to view the features of the visitors that they communicated with or they can click on a group or group type and decompose it into the individual visitors.

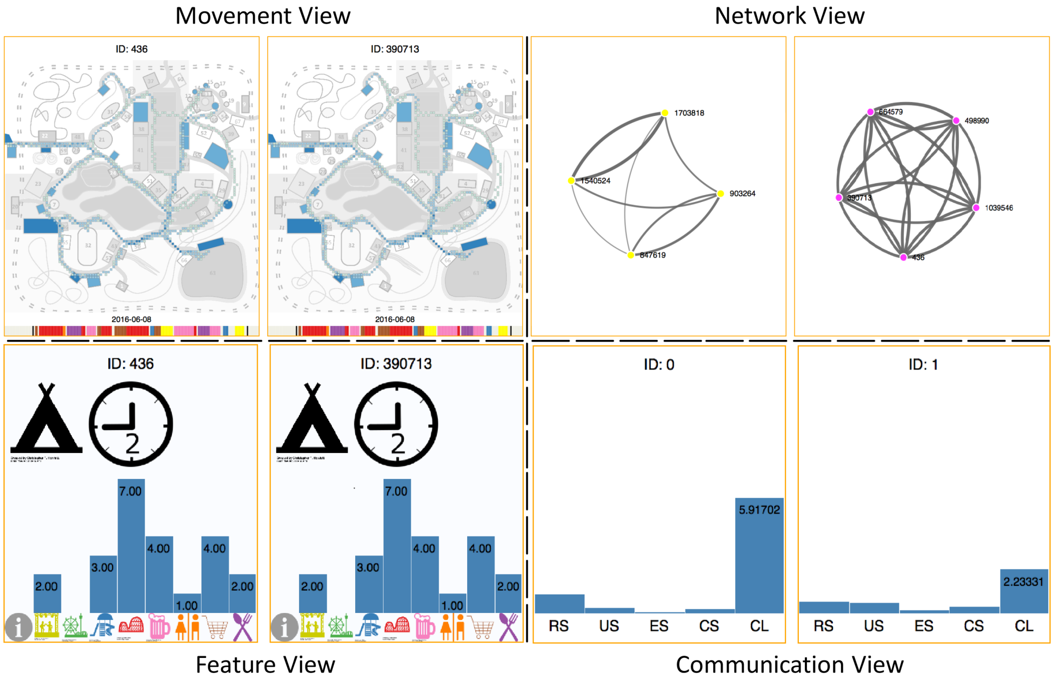


Fig. 8. The reporting view enables users to examine each component of the dataset for visitors, groups, or group types. The movement view displays a bivariate heatmap along with the trajectory traveled. The feature view shows a histogram that displays the attractions visited by category. The network view displays a network graph visualization of the communication connections between visitors. The clustered communication view displays a histogram of the different types of communication among the selected subgroups.

Network View Card In this small multiple selection (Figure 8 - Network View), analysts can inspect the communication networks of visitors, groups, and group types all in the same space. A network graph visualization is used to show the communication connections between visitors. Coloring is used to denote a difference in groups allowing users to quickly see if a visitor is communicating with visitors outside of their group. This color also helps the user determine if communication is happening within the group.

Clustered Communication View Card Finally, in this small multiple selection (Figure 8 - Clustered Communication View), users can explore clusters of groups by their communication types. This enables the analyst to quickly identify which types of communication were prominent and the users that fit into those subgroups. Clustering is performed using k-means on the user selected features.

4.5. Analytical Methods

As hinted at in the previous subsections, underlying each of the visualization components are a variety of analytical methods. The goal of the proposed visual analytics framework is to enable the exploration of visitor's semantically enriched traces to give insights into the behavior of an individual. Understanding the visitor's habits allows park managers to improve their infrastructure accordingly and to stay competitive with other amusement parks. While individual experiences are important, much of the behavioral identification is identifying what do various subgroups of individuals tend to do at the park. Furthermore, identifying outlying behavior of visitors is of interest to ensure safety in the park. Individuals that behave very differently to the rest of the visitors in terms of check-in and communication patterns may be subject for a closer inspection in case of any criminal activity. Our system employs two analytical approaches to identify groups of common behavior patterns as well as outliers: temporal clustering and feature based clustering.

Temporal Clustering and Outlier Detection Two of the major analysis tasks our system supports is identifying groups that travel the park together, thus share a temporal behavior of attraction visits and communication, and identifying behavior that differs from the expected. For the detection of groups that travel the park together, or at least have certain overlapping activities, we apply an agglomerative clustering approach using the Levenshtein distance [Levenshtein 1966] metric on the event sequences extracted in the pre-processing phase. This metric defines the distance between two sequences as the number of insert, edit, and delete operations necessary to transform one sequence into the other. The Levenshtein distance between two strings a, b is given by $\mathrm{l}_{a,b}(|a|, |b|)$ (see Equation 1).

$$\mathrm{l}_{a,b}(i,j) = \begin{cases} \text{if } (\min(i,j) == 0): \ \max(i,j), \\ \text{else: } \min \begin{cases} \mathrm{l}_{a,b}(i-1,j)+1, \\ \mathrm{l}_{a,b}(i,j-1)+1, \\ \mathrm{l}_{a,b}(i-1,j-1) \\ \qquad +[a_i \neq b_j] \end{cases} \end{cases} \tag{1}$$

where the indicator function $[a_i \neq b_j]$ equals zero if $a_i == b_j$ and 1 otherwise.

The Levenshtein distance between each event sequence is then compiled into a matrix containing similarities for each pair of sequences. Because sequences can be of different lengths (visitors may arrive later or leave earlier) we additionally normalize the number of editing operation with respect to the length of the longer sequence. We apply an agglomerative hierarchical clustering with average linkage (see Equation 2).

$$D_{\text{avgLink}}(A,B) := \tfrac{1}{|A||B|} \sum_{a \in A, b \in B} d(a,b) \tag{2}$$

The reason for this decision is that we do not know the sizes of groups traveling the park in advance. The hierarchical methods allows to investigate different levels of detail. While the bottom of the cluster hierarchy shows each visitor individually, we expect smaller clusters to contain smaller groups, while larger groups may be found at a higher level of the hierarchy. A drawback of hierarchical clustering, however, is that it comes with high computational costs. Especially, the Levenshtein distance has a high complexity of $O(n * m)$. To maintain computational and interaction scalability we precompute the distance matrices based on attractions, regions, and attraction categories. To shorten the preprocessing time, the preprocessing was carried out in parallel for each row of the matrix and run on a 40 kernel processor. The clustering is precomputed as well to allow for interactive browsing.

Similarly, outliers can be easily detected. Having the Levenshtein-based distance matrices already computed, a lookup on a visitor basis allows users to quickly calculate visitors that show very different temporal behavior compared to others via a slider component (Figure 1-1). Therefore, we simply count all distances per row and sort the results in descending order.

Feature-based Clustering Along with identifying group trajectory behavior, we also focus on identifying visitors that share similar interests within the park. With these profiles, the park owners can adjust their offers accordingly to attract different groups or increase capacity in the most popular areas. To create visitor group profiles, we employ k-means clustering [Lloyd 1982]. While this clustering method has some drawbacks, it has a low complexity and can be computed at runtime utilizing a back-end server.The complexity of the algorithm used is $O(n^(dk + 1))$, where n is the number of entities and d is the number of dimensions. At first, a feature vector is set up. The analyst can interactively define which features (e.g., restaurant and attraction visits, overnight stays) are to be considered for the clustering. Subsequently, a feature vector for each visitor is built, describing the visitor's interest profile. The vectors $(x_1, x_2, ..., x_n)$ are then used as input for the k-means clustering with the objective to find a grouping that minimizes the distances as in Equation 3).The number k of clusters C and features used in clustering can be interactively controlled in the Reporting View, Figure 8.

$$\arg\min \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 \tag{3}$$

5. CASE STUDIES

In order to demonstrate the effectiveness of our proposed visual analytics framework, we explore two case studies on simulated theme park data.

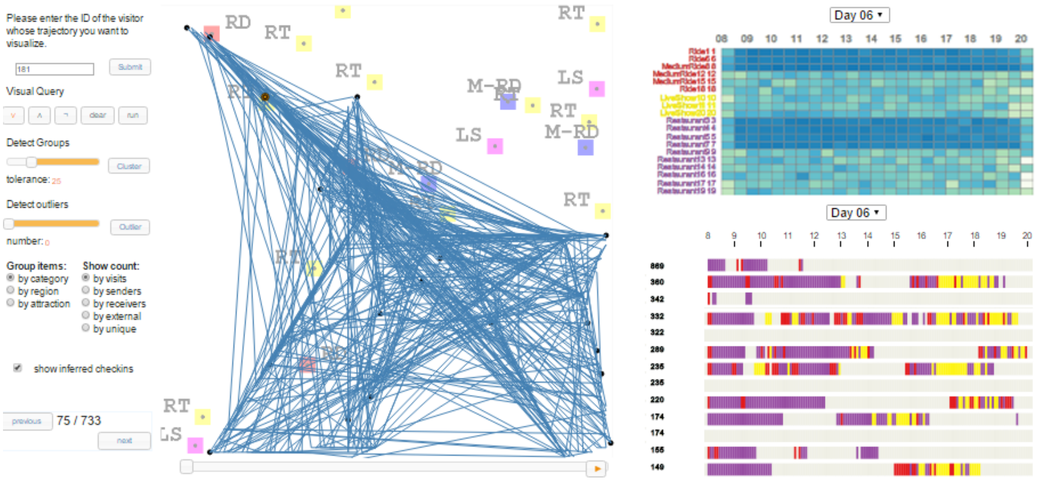
5.1. Simulated Theme Park Data

The first simulated theme park scenario is an amusement park generated in an area of 1,000 x 1,000 meters using the software developed in Solmaz et al. [2015]. The theme park contains 20 attractions in total that consist of main rides, medium-size rides, restaurants, and live shows. The simulation is run for thirty continuous days with a simulated time of twelve hours per day, starting at 8am and ending at 8pm. For each of these days 2500 visitors trajectories are recorded and other semantic information, such as average waiting times at attractions, and proportion of time spent waiting are captured. The raw trajectory data is a list of log events which contains a unique visitor ID, park location (x,y coordinates), a time stamp and a tag showing whether the visitor is moving or checking-in at an attraction.

5.2. Simulation Analysis and Exploration

To begin the exploration of the simulated amusement park, we first load the dataset into the main view (Figure 9). Turning to the analytics interface, we apply several clustering procedures and examine the various clusters that are generated. Starting with a high tolerance level, groups of visitors that spend a single day at the park are found across each of the thirty days of simulation. As the tolerance is lowered in the clustering, groups that span multiple days begin to appear (such as groups of visitors with similar park usage patterns but visit on different days).

Sticking to a higher tolerance with well-defined groups, we examine visitors' sequences in the sequence view. In this dataset, the groups that tend to spend two to four hours at the park are larger in size and only visit a few restaurant before leaving. Other groups found are those that come for approximately the same time and ride one to two rides in addition to dining.



Please enter the ID of the visitor whose trajectory you want to visualize.

181

Fig. 9. Overview of the system with the first simulated them park scenario. In this exploration, a group's trajectory is plotted showing the paths taken from one attraction to another. The calendar view has the sixth day of simulation selected showing heavy traffic for several rides and restaurants. The sequence view also has the sixth day of simulation selected showing the detected clusters and their sequences.

Decreasing the tolerance in clustering, we can explore the larger groups of visitors that spend a long period of time at a particular restaurant, stop at a live show or two, and visit many different rides. To further examine these groups, we can expand the cluster in the sequence view to investigate the individual sequences. Then, we can view their trajectories in the map view and animate their paths traveled. In the animation, we can see visitors clustering at different attractions and separating to visit other attractions. As the animation moves along, and some visitors show little signs of movement, we hypothesize that these attractions are experiencing heavy traffic times. We switch to the calendar view for clarification. Using the calendar view, we can quickly detect the peak times for the various attractions and query for visitors who get stuck in them. From this, we can see there are visitor profiles more likely to get hit with longer wait times, and that there are visitors that are likely to leave after burning up their visit time in these pools.

Such analyses can help identify park bottlenecks, key attractions, and under-utilized space. Note that this example illustrates limitations in the system when trajectories are more noisy (as in this simulation). In those instances, the map view is less useful; however, the event view serves as a critical component for exploration in such instances.

5.3. VAST Challenge 2015 - DinoFun World Data

Our second simulated theme park data set comes from the IEEE Visual Analytics Science and Technology (VAST) Challenge 2015. These challenges have been held annually since 2006 and aim to advance visual analytics through a series of competitions [Cook et al. 2014; Whiting et al. 2015]. The presented visual anlaytics system (Section 4) served as a basis for solving the VAST Challenge 2015, and we won the grand challenge award (only one grand challenge award was awarded) for the Outstanding Comprehensive Submission from a pool of 71 submissions.

In the VAST Challenge 2015, the story line is set in an amusement park which covers a space of 500x500 m^2 and is comprised of ride attractions, restaurants, food stops, souvenir shops, game stores, and other attractions [Whiting et al. 2015]. Data is provided for three days: Friday, Saturday, and Sunday. The movements and text-message meta-data of each of 11,317 visitors were recorded, averaging nearly 10 million records per day. The raw trajectory data is a list of log events of the same format as the simulated theme park data (Section 5.1). The raw communications data is also a list of messaging events which contains the originators ID, recipients ID, a time stamp and the park area from which the message was sent. The challenge was subdivided into two mini-challenges (MC) and a grand challenge (GC) that required the contestant to use information from both mini-challenges.

MC1 contained structured location and time data for park visitors, in addition to check-in information. Contestants were prompted to characterize the park attendance for the weekend. This challenge asked contestants to describe different types of groups at the park on the weekend and to notate their characteristics. Details such as how big the group types are, the individuals point of entry, the commonality of the group, inferences about the group type, and whether to include improvements to the park to better meet the groups needs. Next, contestants were asked to determine if there were notable differences in park patron activity over the course of the weekend. Finally, contestants were asked to report on any anomalies or unusual patterns they discovered in the data. The tasks related to this challenge were focused on analyzing groups, overall park movement, and anomalous behavior within the trajectory data set. These factors lead us to devise the tools and views from our system that centered around clustering, group detection, and outlier detection utilizing the trajectory data set.

MC2 contained structured communication meta-data. this challenge focused on the characterization of communication patterns over time, and discovery of visitor communities. Contestants were asked contestants to identify IDs that had large volumes of communication. After identifying those IDs contestants were asked to characterize the visible communication patterns. From the patterns they were then asked to hypothesize about those IDs. The second task asked contestants to describe communication patterns in the data and characterize who is communicating, with whom they were communicating, as well as when and where. Lastly, contestants were asked to hypothesize about when the vandalism was discovered from the communication data. Turning our focus towards the communication data, these tasks required identification of specific individuals, analysis of overall communication patterns, and the detection of anomalous communication behavior. Our visual analytics system design was driven by the need to not only have a high level view of the communication within the park, but to also provide interactions to drill down to specific times and individuals within the data providing context for the activities within the park that were observed.

The **GC** required contestants to blend knowledge gained from both mini-challenges to solve a crime that occurred during the weekend for which data was provided. The crime focused on soccer star Scott Jones from a town near DinoFun World. Jones rose to international prominence by winning an Olympic gold medal and the World Cup. The 2015 VAST Challenge stated that during a weekend tribute to Scott Jones at DinoFun World, in which Jones was scheduled to appear on stage twice a day, the pavilion displaying Jones' memorabilia was vandalized, and his Olympic medal was stolen. Contestants were asked to determine Scotts arrival and departure times at the park, who he was spending his time with, and the route that he followed. Next contestants were asked to identify issues with park operations during the weekend of inspection. The last task was related to the crime that occurred over the weekend. Contestants were asked to identify when and where the crime took place as well as identifying suspects. In previous challenges, only one of the two data sets was necessary to complete the assigned tasks. However, this challenge required the analysis of both data sets. In particular, these tasks required the unity of group and anomaly detection across both data sets. This lead to the develop of views coordinating across previous challenges and providing interactions to drill down and connect information across both data sets.

5.4. Dinofunworld Analysis

Due to the massive scale of datasets collected from sensor and telecommunications networks, the visual exploration of forensic data from these sources poses special challenges. In the following case study, we apply our visual analytics framework to analyze movement and communications data in *DinoFun World*, a fictitious theme park created for the 2015 Visual Analytics Science and Technology (VAST) Challenge.

Analysis of Visitor Profiles To improve the park's attractions based on visitors' interests we want explore the behaviors of visitors that travel the park together, this includes exploring the average group sizes, as well as visitor profiles, i.e., groups that share similar interests. We start our analysis in the main view (Figure 1) and utilize the clustering capability in the analytics panel. Initially, we define a high tolerance level. This reveals groups of visitors that spend a single day at the park only as well as groups of visitors that spend the full three days. To get more details, we reduce the tolerance level stepwise until we are satisfied with the cluster results. Next, we switch to the event sequence view and use the pagination to browse through the found temporal clusters. The representatives give a first indication of the clusters' members. By hovering over individual representative ids, a heatmap in the map view reveals the spatial profiles of the clusters. This way, we can identify groups that, e.g., went to thrill rides frequently, but barely visited the kids area and vice versa. By clicking on a representative in the event sequence view, the cluster unfolds and we can inspect the individual visitor sequences. We find out that most clusters have a size between 2 and up to 20 visitors that mostly have exactly the same visitor behavior. We hypothesize that these might be friends, families or larger groups such as school classes.

Subsequently, we are not only interested in groups that share temporal behavior but also in visitors that have a similar profile of interest. By identifying such groups the park owners can create special offers for them and stay competitive to other amusement parks. We therefore switch to the reporting view where we can inspect the found groups per id. We then apply the k-means clustering using the reporting view to cluster these groups based on common interest profiles. We then explore selected interest profiles and give an example of how theme parks could use this information to increase revenue. For example, we identify groups who only come to the park to visit beer gardens, and special microbrew events could be offered as an attraction to this crowd. Specific profiles extracted include:

— **Tourists** Visitors identified with this profile are likely to arrive via bus, show little interest in the shows and pavilions, explore most of the park, and visit the beer gardens repeatedly. To attract more *tourists*, we recommend a discount on high-volume ticket purchases.

— **Thrill Seekers** Visitors identified with this profile spend most of their time in thrill rides, and are unlikely to use overnight accommodations. To attract more *thrill seekers* and incentivize spending on lodging, we recommend offering fast passes (a means to save time in line) to thrill rides for patrons staying in a theme park hotel.

— **Foodies** Visitors identified with this profile are likely to spend two to five hours at restaurants and are predominantly active in the mornings. To attract more *foodies*, we recommend offering an eating pass that includes samples at various restaurants, and doing a more thorough consumer analysis to offer popular brunch options.

— **Shoppers** This is the most common interest profile. Visitors identified with this profile visit shops later in the day or before leaving the park, and spend several hours at the shops. *Shoppers* always buy lunch at the park and are big spenders. To attract more *shoppers*, we recommend a promotional campaign featuring the park's theme.

These identified profiles were comparable to the ground truth data embedded in the VAST challenge.

Analysis of Communication Patterns While clustering trajectories can give insight into consumer types, MC2 focused on communications between visitors. Communication information enables the inference of social networks and provides further indications about who visits the park together, how the staff reaches the visitors when broadcasting, and how information diffuses following a shocking event.

We used a calendar view and histogram (Figure 6) to identify visitors, staff, and bots with high volumes of communications. We categorize the volume of communications into four types: sent, received, external, and unique (i.e., if a visitor sends multiple communications at the same time, it will only count once in the unique volume). We can explore the sent message histogram and interactively select bins that have a high volume of messages to extract the identification numbers. As seen in the distribution view of Figure 1, there are several histogram bins indicating a small number of IDs having an extraordinarily high volumes of sent messages.

The first of the two IDs has more than 189,000 sent and received messages; however, the unique volume for this ID is only 180 and it did not have any external communications. By looking at the calendar view of this ID, we find that it has a high volume of communications every five minutes, starting at 12:00 PM and ending at 9:00 PM every day. In this case, we hypothesize that this ID belongs to a park auto message bot that broadcasts announcements.

The second of the two IDs has almost identical sent and received message counts (over 60000) and has 0 external communications, and also has a large amount of unique communications. Its calendar view does not indicate any obvious temporal patterns, so we hypothesize that this ID is a park ID that replies to visitors like an information service center.

We now present selected communication patterns that were identified using our system, give an interpretation of whom the ID could belong to, and hypothesize about their reason for having the observed communication pattern.

— **Park broadcasters** We observe that some IDs broadcast large amounts of messages over a short period of time while others send messages intermittently. We hypothesize that the park uses a broadcast system to send messages to many patrons simultaneously. We assume that all messages sent by one ID within one minute are the same message and we provide a count of all unique messages in the data. We can view the distribution of unique messages at each location using the calendar view and can observe the high volume of sent/received messages as well as the large amount of unique messages being sent.

— **Crime Reporters** Because we suspect the crime described in the GC occurred at the Pavilion between 9:45 and 11:30 AM, we explore the communication patterns of people located near the pavilion during this time. First, we use the histogram to visualize those people who have external communications over that time. Then, we use our communication explorer tool (figure 7) to explore those IDs and found that three IDs have abnormal behaviors: Unlike most IDs in this pattern whose behavior could be characterized as gossip-like, three IDs came into the park together and then went to the Pavilion around 9:00am. Over the next 30 minutes, they stayed there with few communications. However, they had a lot of communications after 09:40AM, which was very different from the previous 30 minutes until 12 pm. We hypothesize that these IDs belong to janitorial staff who began communicating after discovering evidence of vandalism.

5.5. VAST Grand Challenge Analysis

Along with the analysis of frequent visitor patterns for marketing and sales, security is an important aspect in large facilities such as DinoFun World. In this section, we document the analysis procedure performed using our visual analytics system to correctly identify the criminal embedded in the VAST Challenge. As a starting point for the analysis, the 2015 VAST Challenge stated that during a weekend tribute to Scott Jones at DinoFun World, in which Jones was scheduled to appear on stage twice a day, the pavilion displaying Jones' memorabilia was vandalized, and his Olympic medal was stolen. All location names (e.g., pavilion, stage) were available from the park data and are semantically linked to trajectories in the pre-processing stage.

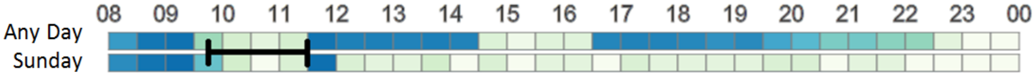


Fig. 10. A heatmap displaying attendance density at the Creighton Pavilion. The Pavilion is closed after 11:30 AM on Sunday, because the crime was discovered and police were called.

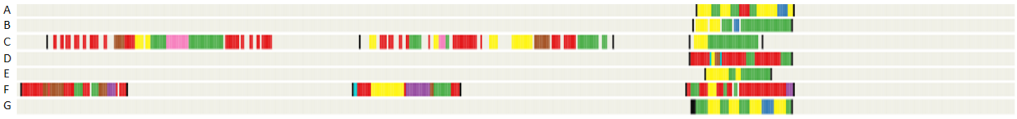
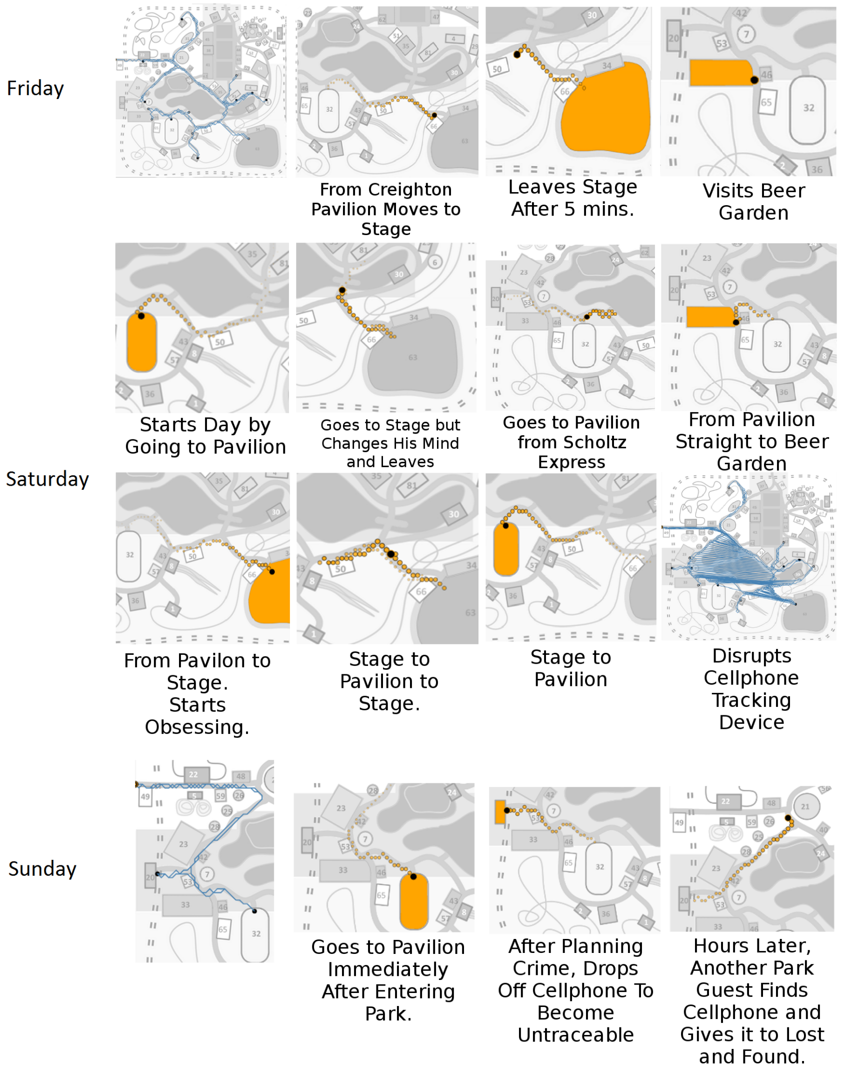


Fig. 11. A pixel-based visualization of suspicious patrons' trajectories. Trajectory C was the actual culprit.

Hypothesis Generation and Visual Querying First, we begin by exploring the calendar view of Figure 1. A close examination of the row related to the pavilion reveals that every day, the pavilion showcasing Jones' memorabilia receives zero visitors from 9:30 AM to 11:30 AM (indicating the pavilion is closed), and the security personnel are moved from the pavilion to the stage to ensure higher security for the guests. Whenever the pavilion is open, it is densely popular. On Sunday, the Pavilion is closed and re-opened according to schedule, but then is closed again almost immediately (Figure 10). We hypothesize that the crime occurred when security was low and few witnesses were on the scene, Sunday between 9:45 AM and 11:30 AM. Additionally, we assume that the culprit flees the scene of the crime and does not return. Next, we filter suspects via a visual query incorporating these hypotheses and assumptions.

Outlier Detection 39 visitors pass the filtering criteria. Of the 39, only seven visitors moved alone (Figure 11). We used the system described in the previous sections to explore the trajectories of each of these 7 visitors, and selected trajectory C from Figure 11 as the culprit because of its behavior.

Story Formation Based on the contest background, hypotheses and assumptions, we looked for envious, resentful, obsessive, and depressed behavior. We found the trajectory of an obsessed and conflicted person who could not bear to witness the performance of Scott Jones, but would compulsively return to the stage time and time again. After this person, who we conjecture was one of Jones' childhood friends, showed evidence of resentful and obsessive behavior, he would visit the beer gardens, as though to drown his sorrows (Figure 12). We correctly named this visitor as the culprit in our VAST Challenge submission. Along with these observations, there were several critical highlights observed in the suspects daily trajectories. For example, on Saturday, we can observe the GPS traces of the device going haywire as if a person was tampering with the device (Figure 12 - Disrupts cellphone hacking device). Then, on Sunday, we observe the suspect board the train ride, at approximately 9:15AM. However, the trajectory information indicates that the suspect stays on the train 3 times longer than any other person who boarded the ride near the same time. This lead us to believe that the suspect abandoned his tracking device on the train, robbed the pavilion and left the park. The device was later found by another visitor and returned to the main entrance. A storyboard detailing key events is provided in Figure 12, and the evidence chain presented here corresponds to the known ground truth embedded in the Grand Challenge.A video detailing the analysis procedure and



Friday

From Creighton Pavilion Moves to Stage

Leaves Stage After 5 mins.

Visits Beer Garden

Saturday

Starts Day by Going to Pavilion

Goes to Stage but Changes His Mind and Leaves

Goes to Pavilion from Scholtz Express

From Pavilion Straight to Beer Garden

From Pavilon to Stage. Starts Obsessing.

Stage to Pavilion to Stage.

Stage to Pavilion

Disrupts Cellphone Tracking Device

Sunday

Goes to Pavilion Immediately After Entering Park.

After Planning Crime, Drops Off Cellphone To Become Untraceable

Hours Later, Another Park Guest Finds Cellphone and Gives it to Lost and Found.

Fig. 12. A story-based exploration of culprit's behavior. The first row corresponds to the culprit's actions on Friday, the next two rows correspond to the actions on Saturday, and the last row corresponds to the actions on Sunday.

tool usage can be found at https://www.youtube.com/watch?v=LUZr3qEt7Qo

6. CONCLUSIONS

This article introduces a visual analytics framework for exploring theme park dynamics. While the focus of our paper and examples are on theme parks, such work is applicable for exploring behavioral dynamics within any well-tracked, highly defined area. While initially designed for the VAST Challenge 2015, we demonstrate the efficacy of our tool through a second case study and provide detailed results of applying our system to identify group behaviors and detect anomalies. The benefit of using synthetic data is that there is known ground-truth embedded, and we evaluate our results within the context of the known results. This provides confidence that our visual analytics framework and developed system would be effective for real-world scenarios.

Underscoring the evaluation of our framework are three main contributions. Our first contribution consists of a mechanism for semantic trajectory clustering. As part of the pre-processing step, we automatically extract events from trajectories and link these to underlying infrastructure and visitor interactions. We then consider these event and movement segments like letters within a string, and apply a Levenshtein distance metric to extract group dynamics from within our semantic trajectory data, as discussed in § 4.5. While different distance metrics between trajectories can be explored, our results of the visitor groups in both simulations matched well with the known groups within the data. Furthermore, the use of an aggregated event-movement trajectory string with enriched semantic information provides a more compact representation of the trajectory and serves as a convenient representation for visualization as well as for accelerating the proposed clustering algorithms.

Our second contribution is the novel spatiotemporal network view that was employed to track group communication patterns over the park as describe in §3.3. Here, we abstract the spatiotemporal data into a 2D view in order to track communication and group movement over time. Users are able to select visitors of interest and identify locations where communications took place (both in person and via mobile devices). While there are still obvious visual issues with such a view, we believe that exploring new representations of complex spatiotemporal patterns is necessary. Plotting such data in a space-time cube results in overplotting and occlusion, and animations can increase the cognitive load on analysts. While our proposed view is complicated, we were able to demonstrate how to utilize such a view to track complex dynamics within a group, identify group splits and merges within the park, and find time periods and locations of interest.

Our final contribution is the overarching visual analytics framework that enables pattern directed exploration of movement data within a theme park setting. This tool enables the directed exploration of group behaviors and dynamics within the park and provides insight into how multiple data sources can be fused for further knowledge extraction. A detailed description of the tool and its functionality are provided, and use case examples and results are demonstrated. By allowing an analyst to filter and tune clustering parameters, we are able to enhance the overall system support, and we believe that the described visual analytics framework and the developed system would be applicable for a wide variety of behavioral analysis for mobility data. Furthermore, the linked clustering and anomaly detection methods provide unique opportunities for both macro and micro level behavior analysis. By using clustering, group dynamics can be explored as demonstrated in Section 4.4, and by utilizing anomaly detection, unique behaviors between groups can further be explored.

More generally, our proposed framework is suitable for behavior analysis over well-defined geographic regions where the underlying built environment is being heavily used. In locations where structures of use are sparsely located, the string comparison approach may prove computationally unfeasible, as few localized patterns may exist;

instead, broader geographic patterns may be of interest (such as general commuter routes between suburbs and the city). Furthermore, such work is directly applicable to emergency response situations where large scale crowd dynamics need to be evaluated. In these cases, such a framework could be used in a pre-planning stage to help train workers and identify potential bottlenecks and problem areas. Combining such work with planning and optimization constraints should be considered for the future both for disaster planning and for improving and optimizing infrastructure use and flow.

Our framework also has several known limitations. First, the pre-processing step removes the ability of the analyst to interactively inject domain knowledge into the semantic matching and aggregation process. This choice was made due to the computational demands of the large data; however, future versions of developed system should be modified to allow many of the pre-processing threshold choices to be made interactively. Second, the outlier detection focuses only on trajectories that deviate from other clusters. What is needed is to also link semantic information about the underlying infrastructure within the anomaly detection methodology. For example, in the VAST Challenge, the criminal was identified by the length of time spent on a particular ride; however, this was done through a manual inspection. Future work should better incorporate the semantic information into the anomaly detection procedures. Third, the framework was developed and used by visual/data analysts without user evaluation from field experts. In future work with the input of domain experts, the framework will provide greater usability for general theme park analysis and allow for better integration of domain expertise.

REFERENCES

Natalia Adrienko and Gennady Adrienko. 2011. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics* 17, 2 (2011), 205–219.

Gennady Andrienko and Natalia Andrienko. 2010. A general framework for using aggregation in visual exploration of movement data. *The Cartographic Journal* 47, 1 (2010), 22–40.

Gennady Andrienko, Nathaliya Andrienko, Peter Bak, Daniel Keim, Slava Kisilevich, and Stefan Wrobel. 2011a. A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages & Computing* 22, 3 (2011), 213–232.

Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, and Stefan Wrobel. 2013. *Visual Analytics of Movement*. Springer Publishing Company, Incorporated.

Gennady Andrienko, Natalia Andrienko, Christophe Hurter, Salvatore Rinzivillo, and Stefan Wrobel. 2011b. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 161–170.

Gennady Andrienko, Natalia Andrienko, Salvatore Rinzivillo, Mirco Nanni, Dino Pedreschi, and Fosca Giannotti. 2009. Interactive visual clustering of large collections of trajectories. In *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 3–10.

Gennady Andrienko, Natalia Andrienko, and Stefan Wrobel. 2007. Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter* 9, 2 (2007), 38–46.

Natalia Andrienko and Gennady Andrienko. 2013. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization* 12, 1 (Jan. 2013), 3–24.

Richard Arias-Hernandez, Linda T Kaastra, Tera Marie Green, and Brian Fisher. 2011. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 1–10.

Benjamin Bach, Pierre Dragicevic, Daniel Archambault, Christophe Hurter, and Sheelagh Carpendale. 2016. A Descriptive Framework for Temporal Data Visualizations Based on Generalized Space-Time Cubes. *Computer Graphics Forum* (2016).

Paul Beardsley and Aparna Taneja. 2016. System and method using foot recognition to create a customized guest experience. (19 07 2016).

Luke Bermingham and Ickjai Lee. 2015. A general methodology for n-dimensional trajectory clustering. *Expert Systems with Applications* 42, 21 (2015), 7573–7581.

Vania Bogorny, Chiara Renso, Artur Ribeiro Aquino, Fernando Lucca Siqueira, and Luis Otavio Alvares. 2014. CONSTAnT–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS* 18, 1 (2014), 66–88.

Dennis J Bouvier and Britian Oates. 2008. Evacuation Traces Mini Challenge award: Innovative trace visualization staining for information discovery. In *IEEE Symposium on Visual Analytics Science and Technology*. 219–220.

Ilya Boyandin, Enrico Bertini, and Denis Lalanne. 2012. A Qualitative Study on the Exploration of Temporal Changes in Flow Maps with Animation and Small-Multiples. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 1005–1014.

Siming Chen, Xiaoru Yuan, Zhenhuang Wang, Cong Guo, Jie Liang, Zuchao Wang, Xiaolong Luke Zhang, and Jiawan Zhang. 2016. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 270–279.

Ding Chu, David A. Sheets, Ye Zhao, Yingyu Wu, Jing Yang, Maogong Zheng, and George Chen. 2014. Visualizing Hidden Themes of Taxi Movement with Semantic Transformation. In *IEEE Pacific Visualization Symposium*. 137–144.

Kristin Cook, Georges Grinstein, and Mark Whiting. 2014. The VAST Challenge: history, scope, and outcomes: An introduction to the Special Issue. *Information Visualization* 13, 4 (2014), 301–312.

Ticiana L Coelho da Silva, Karine Zeitouni, and Josỳ AF de Macỳdo. 2016. Online Clustering of Trajectory Data Stream. In *IEEE International Conference on Mobile Data Management (MDM)*, Vol. 1. 112–121.

Peter J Diggle. 2013. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press.

Nathan Eagle, Alex Sandy Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* 106, 36 (2009), 15274–15278.

Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva. 2013. Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 2149–2158.

Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 330–339.

Jack Guest, Todd Eaglin, Kalpathi Subramanian, and William Ribarsky. 2013. Visual analysis of situationally aware building evacuations. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 86540G–86540G.

SungYe Kim, Ross Maciejewski, Karl Ostmo, Edward J Delp, Timothy F Collins, and David S Ebert. 2008. Mobile analytics for emergency response and training. *Information Visualization* 7, 1 (2008), 77–88.

Slava Kisilevich, Florian Mansmann, Mirco Nanni, and Salvatore Rinzivillo. 2009. *Spatio-temporal clustering*. Springer.

Robert Krüger, Dennis Thom, and Thomas Ertl. 2015. Semantic Enrichment of Movement Behavior with Foursquare–A Visual Analytics Approach. *IEEE Transactions on Visualization and Computer Graphics* 21, 8 (2015), 903–915.

Patrick Laube. 2014. *Computational movement analysis*. Springer.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, Vol. 10. 707.

Juncai Li, Quan Wang, Pin Luo, Yuan Zeng, Ying Zhao, and Fangfang Zhou. 2015. Applying advanced analytic techniques to visually explore communication patterns in mobile data. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 135–136.

Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1010–1018.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137.

J Macedo, Christelle Vangenot, Walied Othman, Nikos Pelekis, Elias Frentzos, Bart Kuijpers, Irene Ntoutsi, Stefano Spaccapietra, and Yannis Theodoridis. 2008. Trajectory data models. In *Mobility, Data Mining and Privacy*. Springer, 123–150.

J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104, 18 (2007), 7332–7336.

Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. 2013. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 344–353.

Nikos Pelekis, Gennady Andrienko, Natalia Andrienko, Ioannis Kopanakis, Gerasimos Marketos, and Yannis Theodoridis. 2012. Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems* 38, 2 (2012), 343–391.

Abishek Puri, Dongyu Liu, Shaoyu Chen, Siwei Fu, Tianyu Wang, Yeukyin Chan, and Huamin Qu. 2015. ParkVis: A visual analytic system for anomaly detection in DinoFun World. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 123–124.

Gürkan Solmaz, Mustafa İlhan Akbaş, and Damla Turgut. 2015. A Mobility Model of Theme Park Visitors. *IEEE Transactions on Mobile Computing* 14, 12 (Dec 2015), 2406–2418.

Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. 2008. A conceptual view on trajectories. *Data & knowledge engineering* 65, 1 (2008), 126–146.

Zhanhu Sun and Feng Wang. 2015. A Comparative Study of Features and Distance Metrics for Trajectory Clustering in Open Video Domains. In *New Research in Multimedia and Internet Systems*. Springer, 47–56.

Christian Tominski, Heidrun Schumann, Gennady Andrienko, and Natalia Andrienko. 2012. Stacking-based visualization of trajectory attribute data. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2565–2574.

Jarke J. Van Wijk and Edward R. Van Selow. 1999. Cluster and Calendar Based Visualization of Time Series Data. In *Proceedings of the IEEE Symposium on Information Visualization*. IEEE Computer Society, Washington, DC, USA, 4–.

Zuchao Wang, Min Lu, Xiaoru Yuan, Junping Zhang, and Huub Van De Wetering. 2013. Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2159–2168.

Mark Whiting, Kristin Cook, Georges Grinstein, John Fallon, Kristen Liggett, Diane Staheli, and Jordan Crouser. 2015. VAST Challenge 2015: Mayhem at dinofun world. In *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 113–118.

Jo Wood, Aidan Slingsby, and Jason Dykes. 2011. Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica* 46, 4 (2011), 239–251.

Wenchao Wu, Jiayi Xu, Haipeng Zeng, Yixian Zheng, Huamin Qu, Bing Ni, Mingxuan Yuan, and Lionel M Ni. 2016. TelCoVis: Visual Exploration of Co-occurrence in Urban Human Mobility Based on Telco Data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 935–944.

Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. 2011. SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th international conference on extending database technology*. ACM, 259–270.

Zhixian Yan, Stefano Spaccapietra, and others. 2009. Towards Semantic Trajectory Data Analysis: A Conceptual and Computational Approach. In *VLDB PhD Workshop*.

Tangzhi Ye, Youfeng Hao, Zhenhuang Wang, Chufan Lai, Siming Chen, Zongru Li, Jie Liang, and Xiaoru Yuan. 2015. Behavior analysis through collaborative visual exploration on trajectory data. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. 131–132.

Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S Tseng. 2010. Mining user similarity from semantic trajectories. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*. ACM, 19–26.

Wei Zeng, Chi-Wing Fu, Stefan Müller Arisona, and Huamin Qu. 2013. Visualizing interchange patterns in massive movement data. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 271–280.

Ying Zhao, Xing Liang, Xiaoping Fan, Yiwen Wang, Mengjie Yang, and Fangfang Zhou. 2014. MVSec: multi-perspective and deductive visual analytics on heterogeneous network security data. *Journal of Visualization* 17, 3 (2014), 181–196.

Yu Zheng. 2015. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 29.

Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 247–256.

Yin Zhu, Yu Zheng, Liuhang Zhang, Darshan Santani, Xing Xie, and Qiang Yang. 2012. Inferring taxi status using GPS trajectories. *arXiv preprint arXiv:1205.4378* (2012).