Visual Analysis for Spatio-Temporal Data

Mengying DU

mengying.du@student.ecp.fr

1 Introduction

In this report I introduce the visual analytical approaches for Dino Fun World project. To reveal potential anomalies and unusual patterns in the park, I firstly answered the basic questions like when people come and leave the park or how many attractions people checked in during the day. To build a more meaningful picture of people in the park, I did a classification to help group people together with exhibit similar behavior based on people's check-in patterns. The communication pattern in the park is also important to estimate the time period when crime happened. Based on multiple sources and analytical figures, I could make a hypothesis on the suspicious people. The analysis procedure involved are data preprocessing, data classification and interactive visualization.

2 Data Summary

The data provided are 3 different datasets:

- **park-movement-Sun.csv:** People movement records, contains **10,932,424** rows totally, and **7569** distinct people IDs, data attributes are:

Timestamp	id	type	X	Y
2014-6-08 08:01:04	120617	movement	98	76
2014-6-08 08:01:04	1330021	check-in	63	99

- **comm-data-Sun.csv:** People communication records, contains **1,422,491** rows, with **6116** distinct senders and **5236** receivers, sample data is:

Timestamp	from	to	location
2014-6-08 08:01:05	1401601	839736	Kiddie Land
2014-6-08 08:01:16	1169163	600636	Entry Corridor

- **Attraction-Coordinates.csv:** Index and coordinates of the rides, attractions and park areas according to the given map, with **41** check-in points and four major rides, there are also three different entrances located at **North**, **East** and **West** of the park.

AttractionID	Attraction.x	x	y	ParkArea	CategoryNames
1	Wrightiraptor Mountain	47	11	Coaster Alley	Thrill Rides
N	Park Entrance North	63	99	Entry Corridor	Entry-Exit

3 Method

There are three main steps I followed to investigate the patterns on those large datasets. Beginning with preprocessing like data integration and data filtering aiming to extract useful information to answer basic questions listed in Section 4. For further investigation, I classified and grouped people which similar check-in patterns. Then I imported the enriched dataset into Tableau for visualization, with statistical plots, visually filtering and data animation(paging), I could identify the movement and communication patterns on demand and further detect the outliers and anomalies.

3.1 Data Wrangling and Preprocessing

- Data Integration

To get the spatial movement data regarding to the attractions in the park, I began with integration of two sources from **park-movement-Sun.csv** and **Attraction-Coordinates.csv** based on X and Y coordinates. The joined table then contains:

Timestamp	id	type	X	Y	AttractionID	Attraction	ParkArea
2014-6-08 08:00:19	1130496	check-in	63	99	N	Park Entrance North	Entry Corridor

This data integration approach was performed several times along the whole analysis procedure, each time when there is a new table generated(e.g. classified group data), I joined it with the exist sources to get a more enriched dataset

for further exploration.

- Data Filtering and Extraction

To find the answer of when people arrive and leave at the park during the day, I need to extract the entry and exit timestamps of each people id. By setting **CategoryNames** equal to **"Entry-Exit"**, I got all check-in and movement behaviors of each Id at Park "Entrance-Exit" area. From observation on this data(Figure 1), I noticed the check-in and movement are always paired, and majority people had only two records(started with "check-in" then "movement"), so I could assume that the first check-in time is the arrival time and the latest movement is the time people left park. In an addition, the data indicated that people entered and left at the same entrance point, which also makes people's movement pattern quite trival.

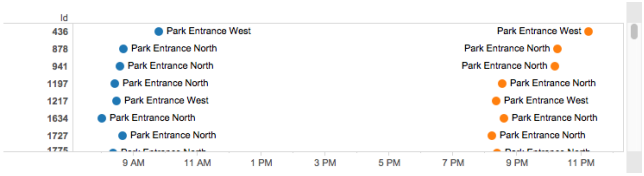

Figure 1: **Time and location people enter and leave the park**

So then I transformed this filtered data in to **enter_exit.csv**(Table 5), which would easily identify when people come and leave, and how long they stay at the park, details are discuss in Section 4 Question 1.

Id	Arrive	Leave	AttractionID
1000279	2014-6-08 08:33:03	2014-6-08 21:39:30	E

3.2 Data Modelling

- Group Classification

In order to get a better understanding of the event occurred at Sunday, I tried to characterize people's movement by determining who came at the park together and visited similar places during the day. Firstly I made a trajectory label on each people based on his check-in attraction ID and check-in time order by a time sequence. For instance, the people who check-in at 08:20:30 a.m. from Park Entrance North, I set his trajectory as **"N-500"** ($500 = 8 * 60 + 20$), accurate to minute), each time he checked in a place, this label should append a piece of chunk with the **"AttractionID-Timestamp"** format. Ideally, people with same trajectory label would be classified in same group, while in reality, the

attraction check-in time for people in same group might be still slightly different, then I change the timestamp to a more rough value as $floor(Timestamp/5)$, then in the previous example it should be **"N-100"** instead, it turns that people check-in the same places within same **5 minutes** period are considered to be same group. To be more reasonable, I also allow **1** check-in differences for people who might skip one attraction or take one more visit compared to his group members during their journey. As a result, there are totally **2639** different groups(**group.csv**, shown in **Table 6**), **21** different group capacities ranging from **1** to **44** people(**Figure 2**), the movement and communication patterns of those groups are demonstrated in result section.

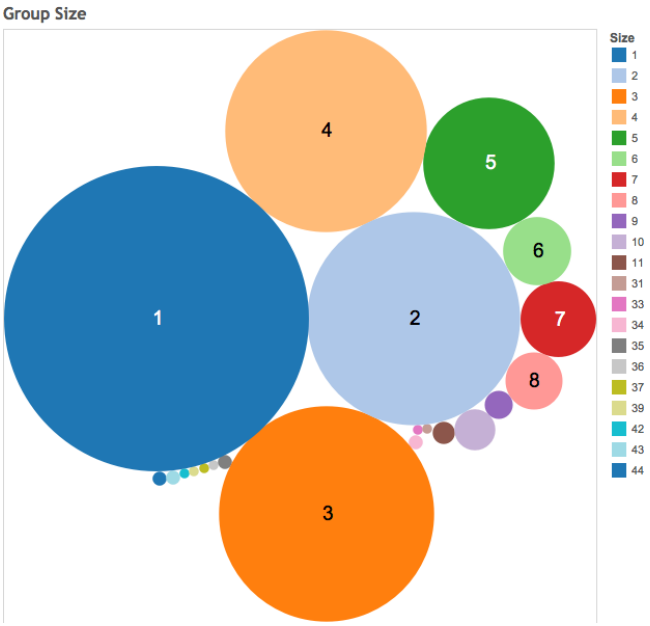

Figure 2: Count of Group by Size

Id	label	GroupId	GroupSize
231413	E-213-30-260-3-265	463	4

3.3 Dynamic Visualization

After data preparation, I moved to data visualization stage to explore the data comprehensively. To do this, I imported all integrated and extracted data (three original datasets with two extracted ones, i.e group.csv and enter_exit.csv)into Tableau. Starting form arranging data to basic worksheets, for example, With

map plot, I could track the movement trajectory of individual or group(Figure 7). Box plots and scatter plots helped me to detect outliers(Figure 6). Then I added related worksheets to a dashboard to make an interactive visual analytics, that helps filter and highlight only the data I interested, also narrow down to the specific group of people. Details for each observation are shown in the result section.

4 Results

4.1 General Pattern Investigation

1. What attractions are the most popular? Temporal patterns when certain places are visited?

Based on number of check-ins, we can see that the **Thrill Rides** are most visits generally during the day, this is because DinoFun Park is an amusement park, that most people came to the park for playing. Whilethere was also a performance given by Scott Jones during the Sunday at **Grinosaurus Stage** (**Attraction 63**) in the morning, so a large amount of people gathered and checked in from **9 a.m.** (**Figure 3(c)**).

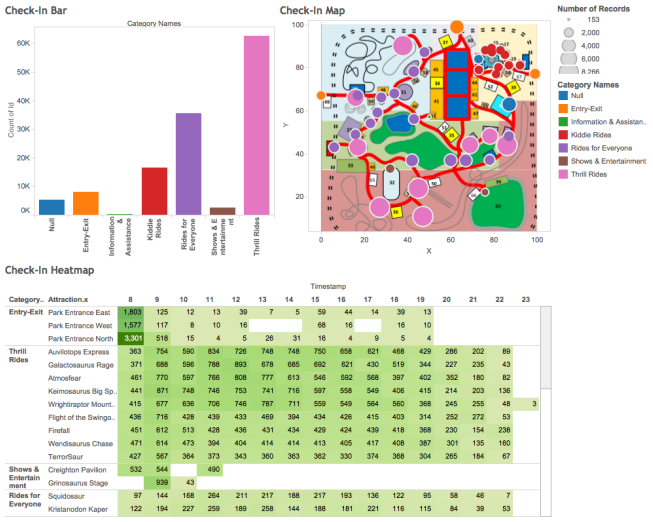

Figure 3: **(a)Bar Chart** - Number of Check-In per Category. **(b)Aggraged Map** - Number of Check-Ins of each Attractions. **(c)Heatmap** - Number of Check-Ins of each Attractions

2. When do people(group) arrive and leave? How long do they stay at the park?

We can see there are three entrances from the park map, large population entered park from **Entrance North** from **Figure 3**, and almost all people

arrived very early in the morning between **8 a.m. and 9 a.m.** and spent whole day in the park. There are also several well separated clusters(**Figure 4(b)**) that some people visited the park in the afternoon or nearly evening who might prefer the night adventure. While the box plot indicates that there is no significant time spending variance between small groups and larger groups, almost were concentrated around **12 hours**.

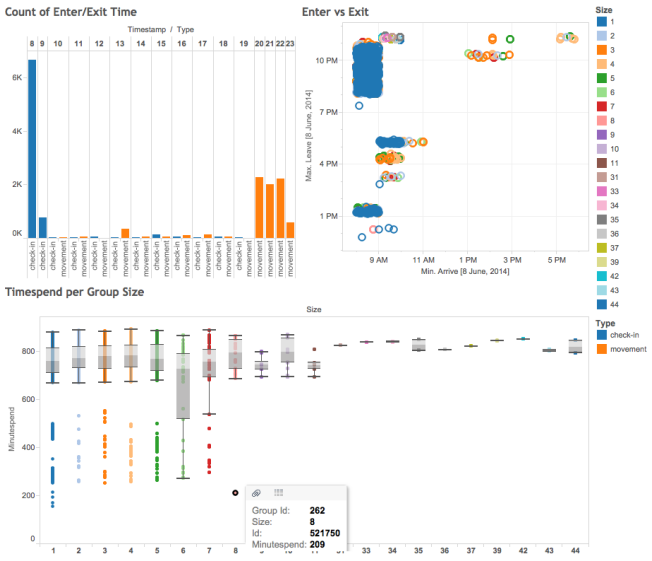

Figure 4: **(a)Bar Chart** - Count of time stamp(in hour) for people enter and leave the park(blue bar stand for enter time, and orange bar for exit time). **(b)Scatter Plot** - Arrival against leaving time per individual (color encoded as size of belonging group). **(c)Box Plot** - Time spend at park for different group size

3. Check-in Distribution and Movement Patterns

Histogram in **Figure 5** represents the distribution of attraction check-in numbers per individual ranging from 1 to 41, which basically follows a normal distributions that majority people checked in **13** to **21** times in the park (**Figure 5(c)**). Then the two scatter plots show the correlation between check-in and movement, check-in against distinct check-in each individual made, which measures people's behavior in the park from the perspective of check-in numbers, it can be told that lots of people visit some attractions more than once, then most people visited 10 to 14 different attractions (**Figure 5(d)**). Finally those scatter plots and box plots helped to clearly identify and detect outliers in Section 4.2.

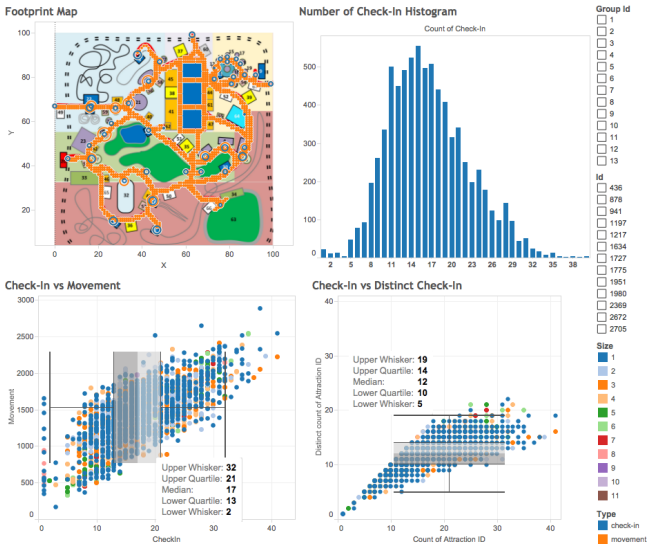

Figure 5: **(a)Footprint Map** - Records for people's movement **(b)Number of Check-In Histogram** - Distribution of count of check-ins people mode. **(c)Check-In vs Movement Scatter Plot** - Count of check-In against movement records per ID(colored coded by group size). **(d)Check-In vs Distinct Check-In Scatter Plot** - Count of check-In against distinct check-In records per ID(colored coded by group size)

4.2 Outliers and Anomalies Detection

4. Distinct people who work at and who visit the park? Id of Scott Jones and the people working with him.

Continue with the people check-in and movement scatter plot, I found there are bunch of people who did not check-in any attraction but park entrances only(**Figure 6**), so I speculated those people are working at the park like security guards. Futhermore, I found there are 8 people who are in a group that traveling together, the group ID is **262**, which is also highlighted as an outlier who check-in only once and stay very short in the park (**Figure 4(c)** and **Figure b(a)**). When visualize their movements(**Figure 7**), I noticed they entered from Entrance East and directly to the stage then leave right after the performance, then I can assumed those people are Scott Jones and his team.

5. Communication Pattern in the Park

Another observation on people's behavior in the park is to look the communication data. To identify the top communicators, I firstly made a plot to show the number of messages sent by individuals**(Figure 8)**. And I found two obvious outliers, ID **1278894** and **839736**, who sent tremendous amount of messages**(Figure 8(a))** and also to large amount of distinct receivers**(Figure 8(b))**. To figure out where and when these two ID sent communication, I plotted

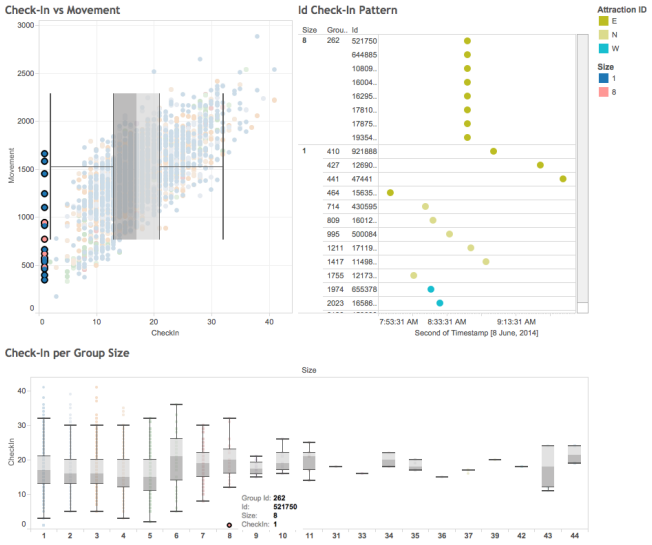

Figure 6: **(a)** Highlighted IDs Check-In Only Once. **(b)** Listed of Highligted ID's with the Group ID. **(c)** Box plot of number of Check-Ins per Group Size.

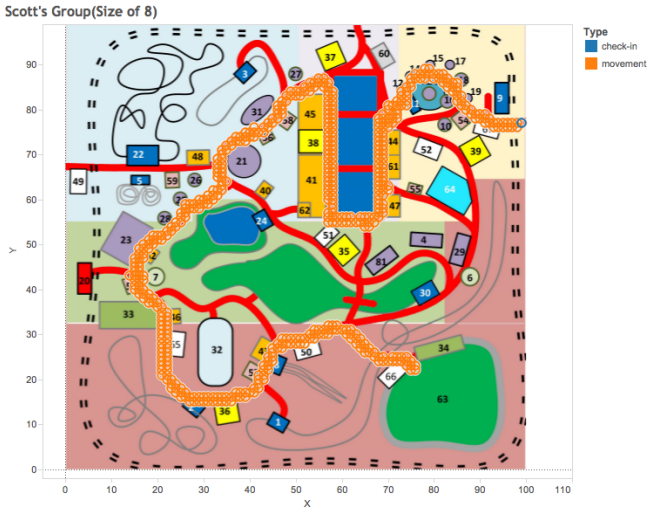

Figure 7: Trajectory of Scott Jones and His Group

another view based on location and time**(Figure 8(c))**, and it turned out that they were just sending messages from **Entry Corridor** and only to the internal IDs(**Figure 9,10**), so I hypothesized these two IDs are park communicators who sent official and security announcements throughout the day. However,

theses two communicators have a quite different communication patterns, that **1278894** sent equally same number of messages every two hours only in the afternoon(**Figure 9**), so I guessed this ID was sending public information to visitors. While ID **839736** started to send bursts messages at **12 p.m.**(**Figure 10**), and it can be interpreted that large number of security announcements need to be sent after the crime was discovered. It can be also noticed from **Figure 8(c)** and **8(d)**, there is a communication peak both internal and external around **11:45 a.m.** to **12:00 p.m.** inside Wet Land and Coaster Alley, which might be the time people discovered the vandalism.

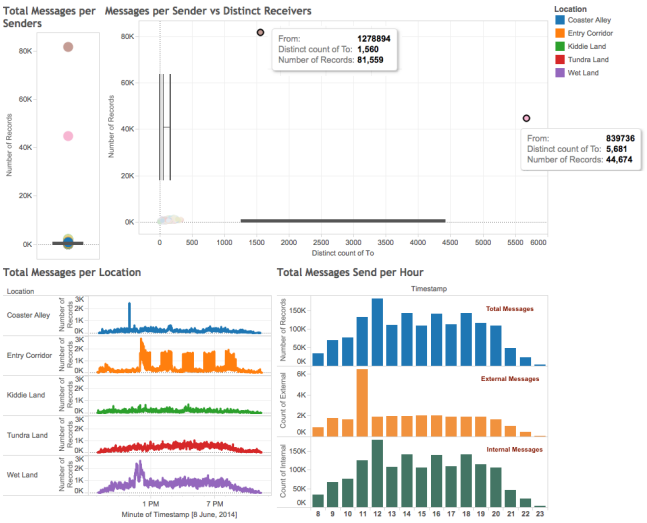

Figure 8: **(a)Box Plot** - Number of Messages Send per Senders. **(b)Scatter Plot** Messages Send per Senders Against Distinct Receivers. **(c)** Messages Send per Location, per Hour. **(d)** Internal and External Messages Send

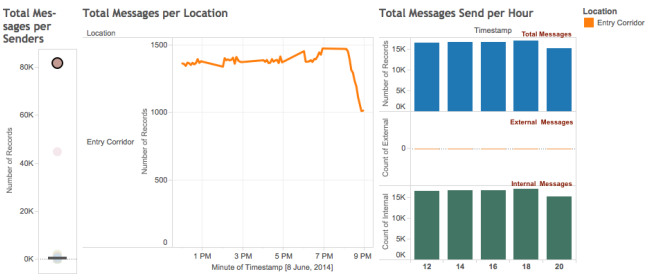

Figure 9: Messages Send by ID **1278894**

6. Suspicious Person and Group

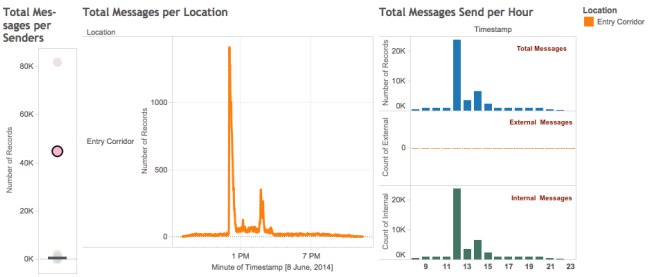

Figure 10: Messages Send by ID **839736**

As the news reported, **Creighton Pavilion (Attraction 32)** is the place vandalism happened. It is also mentioned that, "Creighton Pavilion was closed and locked up tight before each show", so there should be no people inside the pavilion during show. Based on the time that the crime was discovered, I assumed the criminal stole the medal between Scott's performance from **9:30 a.m.** to **11:30 a.m.** Then I tried to find out who checked in the pavilion but not checked out between this time period, the result was narrowed down to the ID **1520920** who checked in at the last minute and did not move to anywhere else until 11:30 a.m(**Figure 11**). He also had 2 other companions (**Group 966**) who did not move out pavilion as well. This group is the most suspicious one might commit the crime. For further exploration, I also found two related groups with 4 more people had a frequently communication with group **966**, which are group **920** and **957**(**Figure 12**). Finally, I put all these three groups onto a map to check if they had any common trajectories, using the Paging View in tableau I found they had several connections in the park(**Figure 13**), it might be speculated as exchanges of the stolen goods.

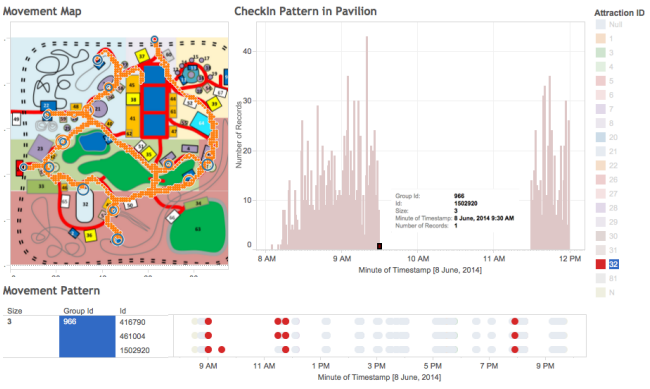

Figure 11: Suspicious People and Group stay in Pavilion Between **9:30AM** and **11:30AM**

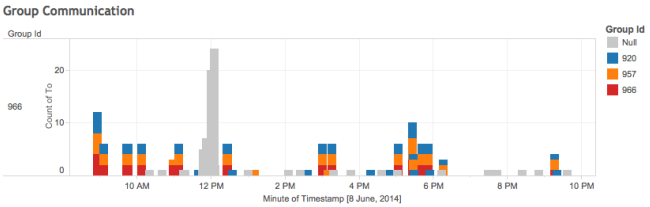

Figure 12: Two Groups Communicated with Group **966**

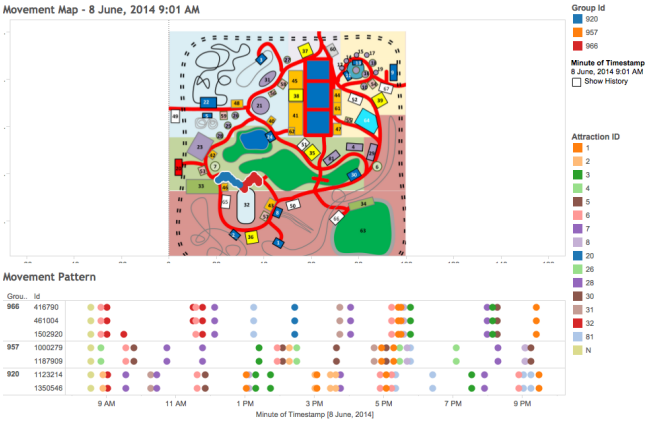

Figure 13: Movement Pattern of Suspicious Groups

5 Conclusion

Exploring park movement and communication data can be complicated and misleading if not well dealing with hidden relationship. The data are embedded with spatial and temporal stamps which hold crucial information that can help analyze the anomalies. My approaches firstly emphasize on data integration and modeling. Then with the prepared data and the help of tableau, I successfully identified the suspicious groups involved in the crime. However, there is still some unexplored areas such as finding the relationship of different groups according to communication patterns, which is better visualized with a network representation and that can be a part of future work.