Behavior Analysis of Tourists on Citi Bike Network

December 12, 2015

Linfeng Zhou, Tianyi Gu, Xiaoge Wu and Yi Zhang

Contents

Introduction

With incomparable museums, attractions, theaters, entertainment and shopping, New York City has always been one of the greatest city that attracted millions of tourists from all over the world. According to the statistic results from The NYC Official Guide, the total number of visitors to New York City in 2014 reached 56.5 million, with 44.5 million domestically and 12 million internationally. The total direct spending of visitors reaches $41 billion[1]. Tourism has brought significant economic and social impacts on the entire city.

As New York's bike sharing system, Citi Bike has over 12,000 of bikes at hundreds of stations all around the city. In order to create a green city and healthy lifestyles, Citi Bike has been highly encouraged for use by both commuters and tourists. Being available for use 24 hours a day and all year round, Citi Bike has become a great transportation alternative for NYC tourists.

With a huge customer base, Citi Bike has a great potential on profit maximization by utilizing its large exposures to customers on advertisements. Our team is very interested in figure out how to make appropriate advertisements on Citi Bike stations, aiming at not only maximizing the utility from the perspective of Citi Bike operator but also improving the enjoyment of tourists. Namely, we expect to locate stations with relatively high frequencies of visiting by tourists and find out communities in the Citi Bike network. In this case, we are able to give recommendations to Citi Bike operator on how to allocate advertisements based on the flows of tourists and the pattern of communities in New York City.

---

[1] NYC. (n.d.). *NYC Statistics*. Retrieved from NYC The Official Guide nycgo.com: http://www.nycgo.com/articles/nyc-statistics-page

Data Description

We are able to get the historical data of Citi Bike trips from the official website of NYC Citi Bike. The data includes trip duration, start/stop time and date, start/end station name and ID and station Lat/Long. The data also includes information about users, including user type, gender and year of birth. Since we are focusing on NYC tourists, we choose the trip data of July 2015 as our dataset because it is the peak season for tourism. For the category "User Type", "customer" is defined as 24-hour pass or 7-day pass user and "subscriber" is defined as annual member. In our study, we make an assumption that tourists are "customer" because of their short-term use of Citi Bike. Sorting the count of renting and returning of each station, we find that the stations that have top 5 count of renting or returning, are all near to tourist attraction (Central Park, City Hall, World Trade Center, Columbus Circle and High Line), which positively demonstrate the reasonability of our assumption.

Methodology and Data Analysis

Betweenness centrality

Our analysis is based on the Citi Bike network, which consists of thousands of station nodes that are connected with each other by links. In order to find the stations with relatively high frequencies of visiting by tourists, we use betweenness centrality to allocate nodes with large influences through the network. Betweenness centrality equals to the number of shortest paths from all nodes to all others that pass through that node. Each node in a weighted network should be weighted in proportion to its capacity, influence and frequency. In our study, we weighted each node based on its count of visiting by customers.

The top 5 nodes with the highest betweenness centrality measurements are displayed in Table 1 and visualized in Figure 1.

Table 1 Location of Top 5 Nodes in Terms of Betweenness Centrality

Node number	Location
217	Old Fulton Street
387	Centre Street & Chambers Street
412	Forsyth Street & Canal Street
532	S 5 Pl & S 4 Street
539	Metropolitan Avenue & Bedford Avenue

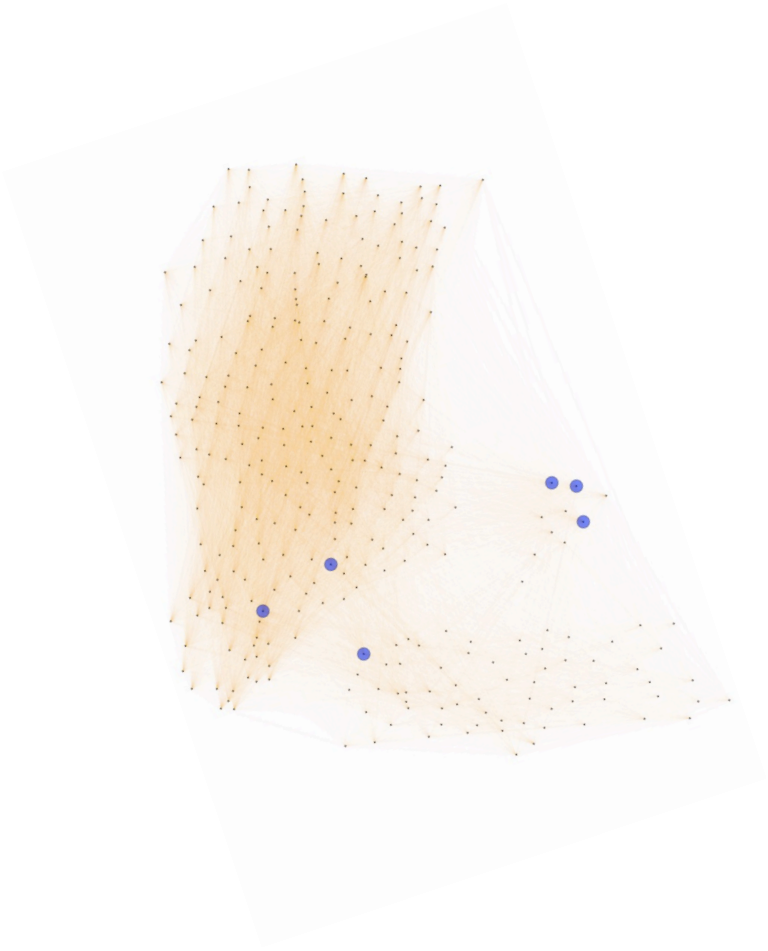

Figure 1 Top 5 Nodes in NYC in Terms of Betweeness Centrality

From the graph above we could observe that Nodes 217 and 387, which correspond to Old Fulton Street and Centre Street & Chambers Street separately, are the two Citi Bike

stations that locate near the two ends of Brooklyn Bridge; Node 412, corresponding to Forsyth Street & Canal Street, is located on the Manhattan side of Manhattan Bridge; Nodes 532 and 539, corresponding to S 5 Pl & S 4 Street and Metropolitan Avenue & Bedford Avenue, are the stations near the Brooklyn side of Williamsburg Bridge.

According to the definition, the betweenness centrality of a node is a measure of how often it connects any other nodes in the graph in the sense of being in between other nodes[2]. Therefore, we conclude that the stations that are located around the bridges have the most critical influence on the current network in terms of visitor flow rate. The high betweenness centrality of these five Citi Bike stations make them perfect locations for advertisements.

Community Detection

In order to investigate whether there is any internal community structure present in the costumer network, the community detection is accomplished and visualized (see Figure 2 below). We also output a visualization through CartoDB for a better view of geographical position of different communities (Figure 3).

Based on the behavior of customers, we noticed that three distinct communities are distinguished: Downtown Manhattan, Midtown Manhattan, and Brooklyn.

First of all, we found that the three stations 224, 387 and 412, which are located inside downtown Manhattan, actually belong to the community of Brooklyn. By looking at the

[2] Russell, M. (2013). Mining GitHub: Inspecting Software Collaboration Habits, Building Interest Graphs, and More. In *Mining the Social Web* (2nd ed., p. 323). Sebastopol, CA: O'Reilly Media.

address of these stations, we discover that all three stations are located around the connection bridges between Brooklyn and Manhattan.

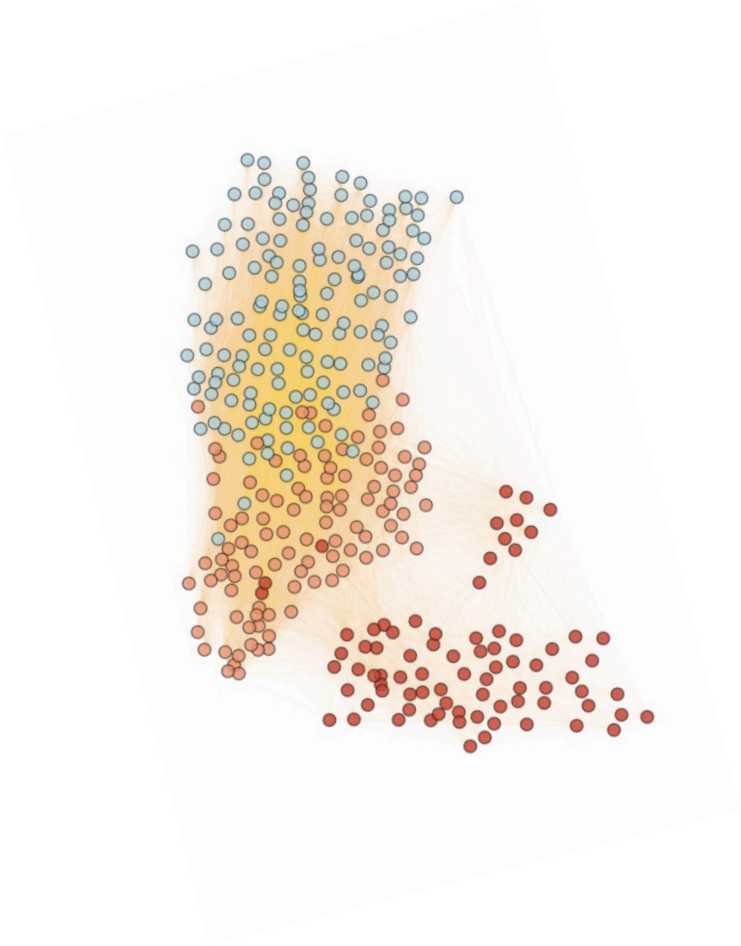

Figure 2 Community Detection

Specifically, station 224 and 387, which are located at Spruce Street & Nassau Street and Centre Street & Chambers Street separately, resident near the Manhattan side of the Brooklyn Bridge; while station 412, corresponding to Forsyth Street & Canal Street, is located on the Manhattan side of Manhattan Bridge. There is a large possibility that customers use bikes only for the visit of the bridges and return them after they go across the bridge. Thus, it is reasonable for the three stations located near the bridges to be classified to the community of Brooklyn.

Besides, the separation line of the other two communities, Manhattan downtown and midtown, approximately follows the path of 14th street, which happens to be the boundary line of the downtown district and midtown district.

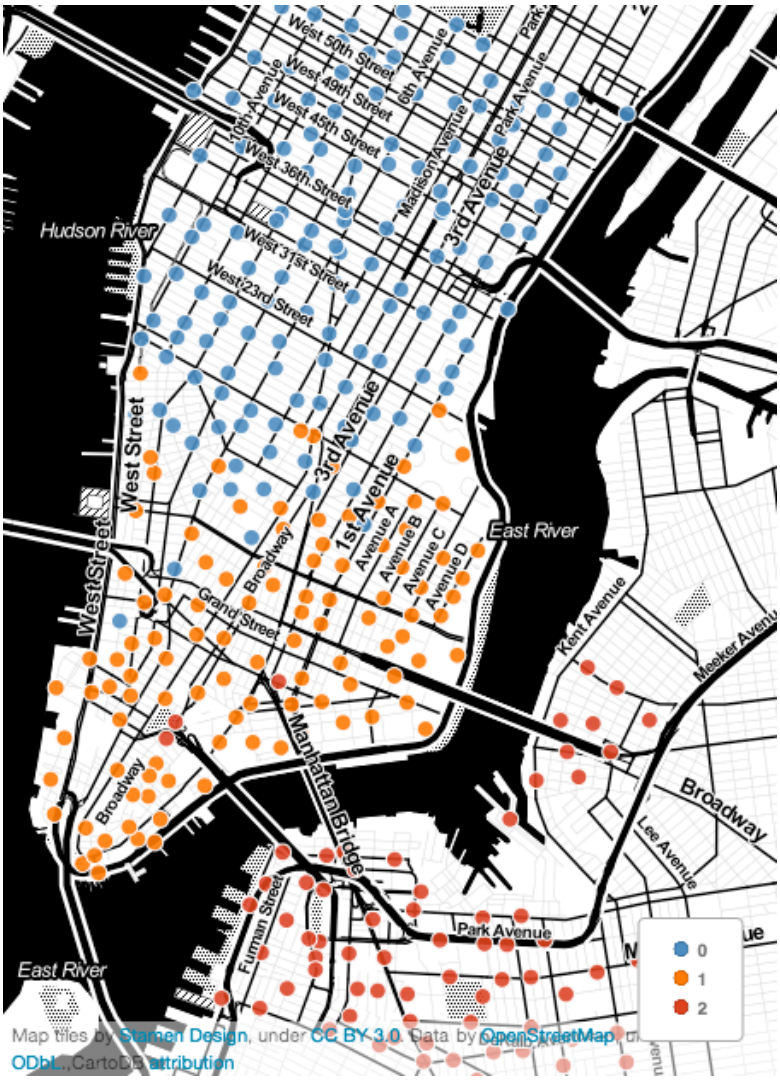

Figure 3 Community Detection with CartoDB

On account of the existence of the distinct community structure, we could draw the conclusion that it might be more efficient and profitable for the operator of Citi Bike to make different advertisements in stations of different communities.

Discussion and Conclusion

According to our betweenness centrality analysis, we find that the top 5 highest centrality stations are all near to the entrance or exit of the Manhattan Bridges, Brooklyn Bridge or Williamsburg Bride, which connect Manhattan and Brooklyn. It is understandable that if citibike users plan to commute between Manhattan and Brooklyn, they must ride through one of the bridges, thus, centrality of the nodes that connected by those three bridges are inevitably high. In consequence, the business value of those nodes are higher than the other nodes.

From our community analysis, we can observe that some Citi Bike stations are located inside one community, but belong to another community. For example, the stations 224, 387 and 412, which are located inside the community of downtown Manhattan, however belong to the community of Brooklyn. Since those three stations are all near the bridges that connected Manhattan to Brooklyn, we can draw conclusion that the customers who pick up citibike from those stations, prefer to ride through the bridges rather than ride to other stations inside Manhattan. We can also make an interference that there is a large number of customers ride from Brooklyn, would return the bike as soon as they get through the bridges.

Additionally, we find that the separation line of the two communities in Manhattan approximately follows the separation line of Midtown District and Downtown District. Considering a certain group of customers might feel more likely to stay around a certain type of district, and people are less likely to make long distance commute by bike, we should do further research on different functionality of Midtown and Downtown district

and the difference between people in those two district, thus propose more appropriate advertisement for each community.

Contribution

- Introduction and Data Description (Xiaoge Wu, Tianyi Gu)
- Data Manipulation (Linfeng Zhou)
- Map Visualization (Linfeng Zhou, Tianyi Gu)
- Centrality Analysis (Yi Zhang, Xiaoge Wu)
- Community Analysis (Yi Zhang, Tianyi Gu)
- Discussion and Conclusion