VI. Passenger privacy

New York City Taxi and Limousine Commission provided taxi data including exact pick up and drop off longitude and latitude from January 2009 to June 2016, but they changed their policy to only provide taxi zone after July 2016. It was because passenger privacy issue was raised, mainly from the 2nd Annual NYC Taxi and Limousine Commission Hackathon on October 2016. Therefore, we were curious how much de-anonymization can be done on anonymized public big dataset. We thought it could be possible to de-anonymize certain passenger if they have unusual tipping behavior or take taxi at unpopular places.

Firstly, we gathered all the taxi trips that have tip greater than $15. Since the algorithm was basic going over the trips and filtering tip amount, which has run time of O(n), it was faster to use run MapReduce in multiple instances rather than using Google Cloud Dataproc. Variable name, data type, and order changed couple times (2010/01, 2010/08, 2014/01, 2015/07), but it was not a big problem with MapReduce.

After combining all the trips with $15 plus tips, we added another filter of tip rate greater than 50% to find unusual trips. Tip rate was calculated by dividing tip amount with fare amount, and tip amount are provided only when passengers had paid by credit card. Then we ran a second MapReduce to find top 150 pick up and drop off locations. After manually searching coordinates in the Google Maps, we found top places were John F. Kennedy International Airport, LaGuardia Airport, Pennsylvania Station, and the Grand Central Terminal. However, there were also some interesting places such as near famous nightclubs, company headquarters, and luxurious apartments. Figure 1 shows pick up and drop off location of trips that have tip amount greater than $ 15.

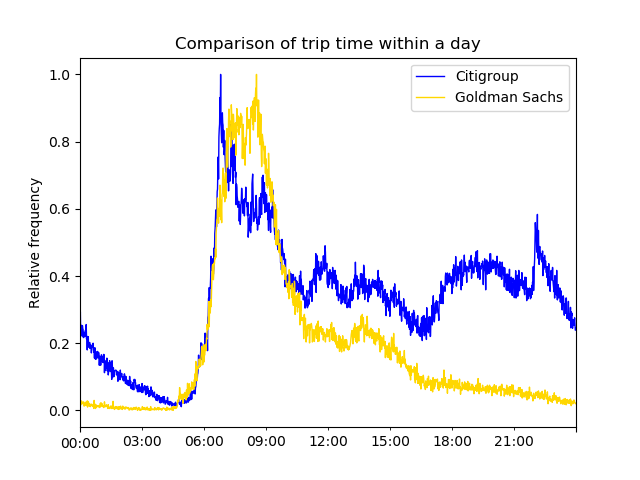
**Figure 1. Pick up and drop off location of trips with tip amount $15 or more**



Then we found lists of nightclubs, company headquarters, and luxurious apartments in New York City and filtered location that are not in the center of Manhattan as it is hard to tell where passenger is coming from. Then we ran third MapReduce to label trips that have pick up or drop off location neighboring known places. However, we only analyzed trips from January 2015 to June 2016 because businesses could not have been in the same places as now if we go back too far. Also, this MapReduce also had run time linear to the number of trips, so it would be simply spending more time to run with more month.

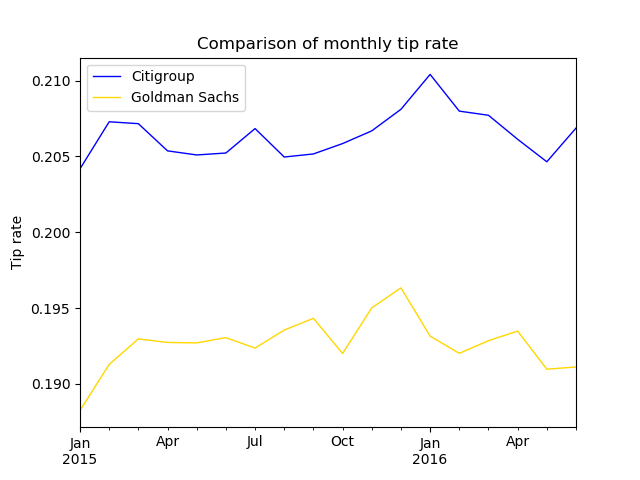
After labeling each trip, we were able to compare businesses. Figure 2 shows number of taxi trips near two investment banks in New York City (Citigroup and Goldman Sachs).

**Figure 2.**



Pick of trips surge earlier for Citigroup than Goldman Sachs in the morning, suggesting workers or visitors of Citigroup taking Yellow taxi to these places come earlier compared to Goldman Sachs. Figure 3 shows monthly tip rate for trips near these companies.

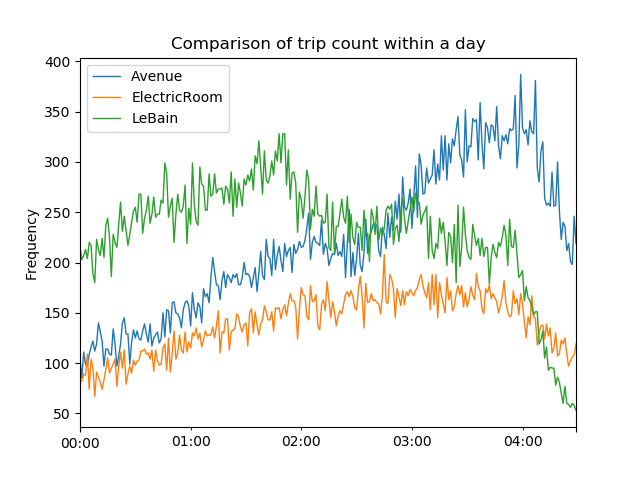
**Figure 3.**



Tip rates for trips near Citigroup building were consistently higher than the trips near Goldman Sachs building over the period January 2015 to June 2016. Also, two tip rates had similar trend suggesting people in each group are affected by similar macroeconomic factors when making decision how much to tip.

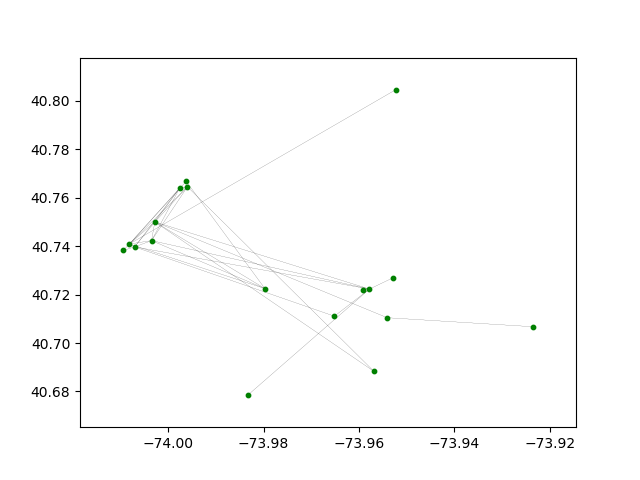
Figure 4 shows number of weekend (Saturday, Sunday, Monday), late night (12 a.m. to 4:30 a.m.) taxi trips near three nightclubs (Avenue, Electric Room, Le Bain).

**Figure 4.**



Three nightclubs had similar average tip rate (Avenue: 0.2052, Electric Room: 0.2064, Le Bain: 0.2058), but the pick times were different (Avenue: 3:59 a.m., Electric Room: 2:45 a.m., Le Bain: 1:50 a.m.). Then we did network analysis between nightclubs to find which network is in the center of the network. Figure 5 shows the network graph between nightclubs.

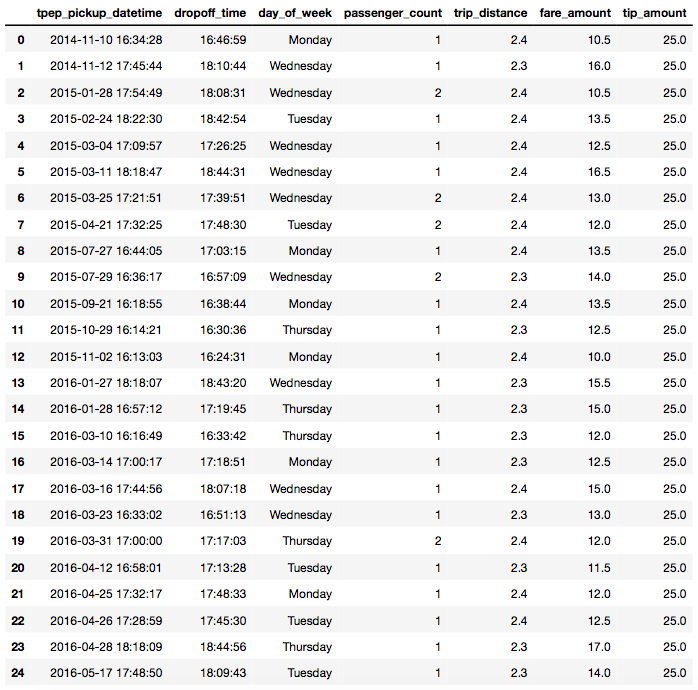
**Figure 5. Network graph of night clubs (weekend, late night trips)**



It is not clear which nightclub is most popular, but it suggests there are more trips between Manhattan clubs than Brooklyn clubs.

Then we ran MapReduce to find top pick up, drop off, and tip amount pair with tip rate greater than 50% to find frequent user of Yellow taxi. Although it was hard to tell if the trips were from one person, there was one interesting trip happening on weekdays afternoon in New York City. Table 1 shows the trip records of this individual.

**Table 1. Trip records of individual with $25 tip**



Locations were intentionally left out to protect the innocent, but it is clear that these trip records are coming from single person (sometimes with a partner).

Our last step was trying to find a pair of trips within a day that had same tip amount but with a pick up and drop off location switched after rounding to the third digit. However, there was not a single trip within our labeled sample trip from January 2015 to June 2016.

In conclusion, we were not able to de-anonymize as much as we wanted at first due to the crowd nature of New York City, but were able to do some analysis of businesses. Privacy concerns could be larger for when taxi trips are published by less populated city. However, we were able to identify an individual, though this person had highly unusual tipping behavior. Therefore, we argue that privacy should always be considered when publishing a big dataset.