YORK CITY TAXIES

BANDARI SRINATH, Dr.T.PRABAKARAN
M. Tech Student, Department of CSE, malla reddy Engineering College, Village maisammaguda, Mandal medchal, District RangaReddy, Telangana, India.
Professor, Department of CSE, malla reddy Engineering College, Village maisammaguda, Mandal medchal, District RangaReddy, Telangana, India.

Abstract *- The widespread use of location-based services has led to an increasing availability of trajectory data from urban environments. These data carry rich information that are us eful for improving cities through traffic management and city planning. Yet, it also contains information about individuals which can jeopardize their privacy. In this study, we work with the New York City (NYC) taxi trips data set publicly released by the Taxi and Limousine Commission (TLC). This data set contains information about every taxi cab ride that happened in NYC. A bad hashing of the medallion numbers (the ID corresponding to a taxi) allowed the recovery of all the medallion numbers and led to a privacy breach for the drivers, whose income could be easily extracted. In this work, we initiate a study to evaluate whether "perfect" anonymity is possible and if such an identity disclosure can be avoided given the availability of diverse sets of external data sets through which the hidden information can be recovered. This is accomplished through a spatio-temporal join based attack which matches the taxi data with an external medallion data that can be easily gathered by an adversary. Using a simulation of the medallion data, we show that our attack can re-identify over 91% of the taxis that ply in NYC even when using a perfect pseudonymization of medallion numbers. We also explore the effectiveness of trajectory anonymization strategies and demonstrate that our attack can still identify a significant fraction of the taxis in NYC. Given the restrictions in publishing the taxi data by TLC, our results indicate that unless the utility of the data set is significantly compromised, it will not be possible to maintain the privacy of taxi medallion owners and drivers.*

Keywords-Big social data, Social set analysis, Social business, Visual analytics, geo-spatial, GIS, Taxi, Green cabs, Uber..

I INTRODUCTION

The NYC Taxi & Limousine Commission (NYCTLC) is a governmental agency created in 1971, and is responsible for the licensing and regulating of New York City's yellow taxicabs, for-hire vehicles, para-transit, commuter vans and other luxury limousine services. The NYCTLC licenses and regulates approximately 50,000 vehicles and counts 100,000 drivers. The paper presented here will focus on the Green cabs, that were introduced by the Five-Boro Taxi Plan, a NYCTLC initiative that aims to meet the demand surplus for taxi rides in the outskirts of New York City. In August 2013, the NYCTLC introduced a fleet of Green cabs to the city of New York. These Green cabs were introduced with the goal of providing the residents of Brooklyn, Queens, the Bronx, and Upper Manhattan more access to metered taxis. Considering that the Yellow cabs prefer to operate in the areas of NYC that are most dense in pick-ups (Manhattan and the airports), the availability of Yellow cabs tends to be low in the outer boroughs of NYC. Hence, Green cabs are not allowed to pick up street hails from the largest part of Manhattan (below 110th St. on the West Side, and below 96th St. on the East Side), or either of JFK or LaGuardia airports.

II RELATED WORK

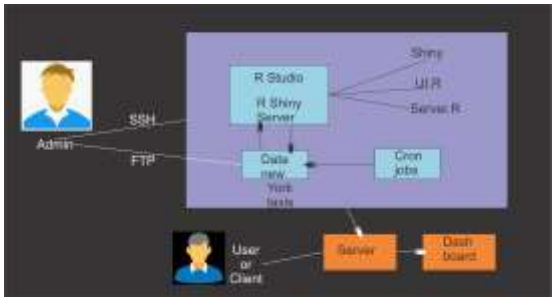
The NYC Taxi & Limousine Commission (NYCTLC) is a governmental agency created in 1971, and is responsible for the licensing and regulating of New York City's yellow taxicabs, for-hire vehicles, para-transit, commuter vans and other luxury limousine services. The NYCTLC licenses and regulates approximately 50,000 vehicles and counts 100,000 drivers. The paper presented here will focus on the Green cabs, that were introduced by the Five-Boro Taxi Plan, a NYCTLC initiative that aims to meet the demand surplus for taxi rides in the outskirts of New York City

The "NYC Taxi Data Set," a historical repository of 750 million rides of taxi medallions over a period of four years (2010-2013). This data set provides rich (batch) information on the movements in an urban network as its citizens go about their daily life. We present a spectral analysis of taxi movement based on the graph Fourier transform, which necessitates the spectral decomposition of a large directed, sparse matrix. Important considerations toward handling this matrix are discussed. Preliminary results show that our method allows us to pinpoint locations of co-behavior for traffic in the Manhattan road network.

Today, there are about 13,000 taxis in use in New York City every day—but by design they usually pick up and drop off a single passenger or group. Some popular transportation startups, such as Uber and Lyft, offer ride-sharing options, but vehicles typically have space for only two passengers at most.

Research published in Proceedings of the National Academy of Sciences in 2014 found that 80 percent of Manhattan taxi trips could be shared by two riders, but the work didn't take into account new riders joining after a trip has already begun. In addition, the 2014 work and other studies of ride sharing either limit the number of riders or they don't study the effects of letting customers choose different pick-up and drop-off locations from each other, Alonso-Mora says. So the real benefits for large-capacity vehicles haven't been determined before.

III DESIGN OF THE WORKFLOW:



Here first collect real dataset from DATA.GOV. Now divide real data into different chunks. To perform this task we applied fixed size chunking algorithm. In fixed chunking algorithm initialize the number of chunks and size of chunks is to be generated for example size of 64 MB. It indicates file is divided into various chunks of size 64MB.

subset is the process of determining which reducer instance will receive which intermediate keys and values. Each mapper must determine for all of its output (key, value) pairs which reducer will receive them. It is necessary that for any key, regardless of which mapper instance generated it, the destination partition

IV EXPERIMENT RESULTS

Using trip data records, how does NYCTLC's share of rides per zip code compare to Uber's in the outer neighbourhoods of New York?.

1) Meaningful Fact #1: Green cabs are just as popular as Uber on the weekend. The distribution of rides according to weekends versus weekdays comparison is very similar in regards to Green cabs and Uber as shown in Fig. 4. Also, the distribution is close to equal in both cases with approximately 40% of the rides occurring during the weekends. It should be noted, though, that the distribution is not really equal in terms of days as the weekends constitute 2.5 day and the weekdays 4.5 days. This means that even though the visualization deceives the interpreter to think of the distribution as a close to 50/50, one should realise that there are more rides taken place during one weekend day than one week day.

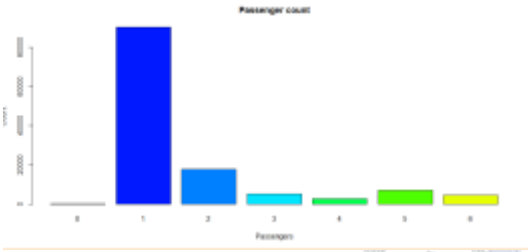
2) Meaningful Fact #2: Weekdays versus weekend rides per hours. To make up for this difference in days with 4.5 weekday days and 2.5 weekend days, we took the total of number of rides occurring during weekdays and divided them by the total number of hours in 4.5 weekdays. Similarly we took the total number of rides occurring in weekends and divided that by the total number of hours in those 2.5 weekend days. The bar charts in Fig. 5 show the difference between the average rides per hour in weekends and weekdays for Green cabs and Uber respectively in total numbers, on the left, and in percentage increase, on the right. The difference is clearer in the right bar chart, as it shows a 3% higher increase of Uber rides per hour in the weekends i.e. compared on average hours, Uber increases during

weekends by 48% while Green cabs increase by 45%.

3) Meaningful Fact #3: There is no clear correlation between the negative and/or positive growth of Uber and Green cabs. With the explosive growth of Uber, one could imagine that when looking at both negative and positive growth, Green cabs would see a negative growth where Uber is experiencing a positive growth, i.e. a 'takeover' growth by Uber. However, as seen in the growth visualization below this is not the case in all areas

	Var1	Freq
1	1	61043
2	2	66367

Vendors Green taxis and Yellow Taxis



maximum passenger count

Peak hour trips made with in the time lap

maximum rides @ time intervals

V CONCLUSION

In this paper, we conducted classification model for analyzing the data of taxis which found to be more effective than the statistical models, creating the subsets for the required data based on the probability models. The results obtained are pretty much helpful for the organization in arranging the cabs in peak hours at different location which in return provides enough profits for the company and the analysis helps in identifying the better locations for the cabs to be maintained

.

VI REFERENCES

The analysis is done for only the data available online. The same can be implemented to the data of national and local taxis like OLA Cabs, Radio Cabs etc.. This will results in economic growth of the company and also improve the financial value of the cabs. The application seems to works with almost 200000 lines of the data, The same can be implemented with hadoop and R. This can be applied using the Rstatistics which works well in the hadoop and R framework. This produces a huge amount of data analytical platform with the best use of all available resources. Finally implementing the analysis with shiny makes ease of anlysing the data and improved the economic value of the products

VI REFERENCES

[1] Qinlu He, Zhanhuai Li and Xiao Zhang, "Data Deduplication Techniques", 2010 International Conference on Future formation Technology and Management Engineering, IEEE 2010, pp. 430-433.

[2] Won, Lim and Min, "MUCH: Multithreaded Content-Based File Chunking". IEEE Transactions on Computers, IEEE 2015, pp. 1-6.

[3] Wen Xia, Hong Jiang, Dan Feng and Lei Tian, "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low

Overheads", IEEE Transactions on Computers, IEEE 2015, pp.1-14.

[4] Yukun Zhou, Dan Feng, Wen Xia, Min Fu, Fangting Huang, Yucheng Zhang and Chunguang Li, "SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management", IEEE 2015, pp. 1-4.

[5] Zhi Tang and Youjip Won, "Multithread Content Based File Chunking System in CPU-GPGPU Heterogeneous Architecture", 2011 First International Conference on Data Compression, Communications and Processing, IEEE 2011, pp. 58-64.

[6] E. Manogar and S. Abirami, "A Study on Data Deduplication Techniques for Optimized Storage", 2014 Sixth International Conference on Advanced Computing(lCoAC), IEEE 2014, pp.161-166.

[7] Bin Lin, Shanshan Li, Xiangke Liao and Jing Zhang, "ReDedup: Data Reallocation for Reading Performance Optimization in Deduplication System", 2013 International Conference on Advanced Cloud and Big Data, IEEE, pp.117-124.

[8] Apache-Hadoop, http://www.hadoop.apache.org/wiki/apachehive

[9] http://hadoop.apache.org