International Journal of Research
Available at https://www.internationaljournalofresearch.org/
ISSN: 2236-6124
Volume 09 Issue 01
Jan-June 2019

A Shiny Interface in Exploring the Taxi Trips Data

Ms. BATCHU MOUNIKA **Mr. G.DILIP KUMAR**

B. Tech Student, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem,NH-16, Guntur, Andhra Pradesh, India.

Assistant Professor, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur, Andhra Pradesh, 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 useful 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

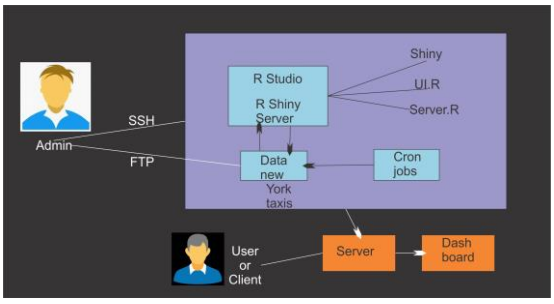
The NYC Taxi & Limousine Commission (NYCTLC) is a governmental agency created in 1971, and is responsible for the licensing and regulating of New York City's yellow taxicabs, for-hire vehicles, para-transit, commuter vans and other luxury limousine services. The NYCTLC licenses and regulates approximately 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 start-ups, 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:



ISSN: 2236-6124
Volume 09 Issue 01
Jan-June 2019

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%.

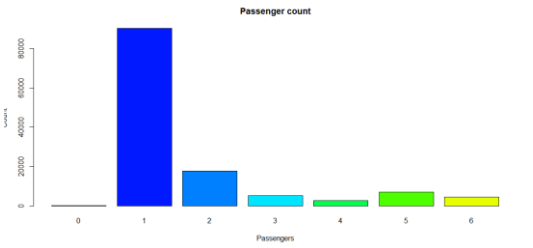
3) Meaningful Fact #3: There is no clear correlation between the negative and/or positive growth of Uber and Green cabs. With the explosive growth of Uber, one could imagine that when

International Journal of Research
Available at
https://www.internationaljournalofresearch.org/
ISSN: 2236-6124
Volume 09 Issue 01
Jan-June 2019

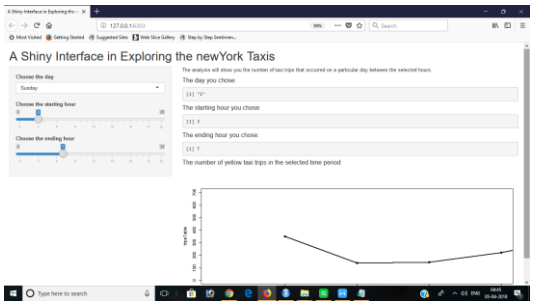
looking at both negative and positive growth, Green cabs would see a negative growth where Uber is experiencing a positive growth, i.e. a 'takeover' growth by Uber. However, as seen in the growth visualization below this is not the case in all areas

	Var1	Freq
1	1	61043
2	2	66367

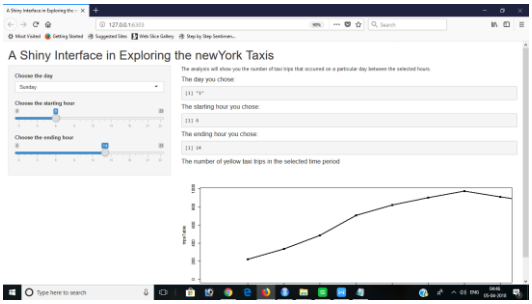
Vendors Green taxis and Yellow Taxis



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 FUTURE ENHANCEMENT

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 R statistics 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 analyzing the data and improved the economic value of the products

VII REFERENCES

[1] C. Bialik, A. Flowers, R. Fischer-baum, and D. Mehta, "Uber is serving new yorks outer boroughs more than taxis are. but most of its rides, like those of taxis, still start in manhattan." http://fivethirtyeight.com/features/ uber-is-serving-new-yorks-outer-boroughs-more -than-taxis-are/, 08 2015. 1, 3, 6

[2] S. Silverstein, "These animated charts tell you everything about uber prices in 21 cities," http://www.businessinsider. Com/uber-vs-taxi-pricing-by-city-2014-10?IR=T , 10 2014. 1

[3] A. Tangel, "Green taxis gaining ground in new York city," http://www.wsj.com/articles/ green-taxis-gaining-ground-in-new-york-city-14 03145481, 06 2014. 1

[4] B. P. Loo, B. S. Leung, S. Wong, and H. Yang, "Taxi license premiums in Hong Kong: Can their fluctuations be explained by taxi as a mode of public transport?" International Journal of Sustainable Transportation, vol. 1, no. 4, pp. 249–266, 2007. 2

[5] J. M. Cooper and W. Faber, "Taxi demand modelling to ensure appropriate taxi supply. do both open access and model based restrictions fail?" Traffic and Transportation Studies 2010, p. 60, 2010. 2

[6] T. J. Kim, "Metadata for geo-spatial data sharing: A comparative analysis," The Annals of Regional Science, vol. 33, no. 2, pp. 171–181, 1999. 2

[7] M. A. Yazici, C. Kamga, and A. Singhal, "A big data driven model for taxi drivers' airport pick-up decisions in new York city," in Big Data, 2013 IEEE International Conference on. IEEE, 2013, pp. 37–44. 2

[8] T. H. Savage and H. T. Vo, "Yellow cabs as red corpuscles," in Big Data, 2013 IEEE International Conference on. IEEE, 2013, pp. 22–28. 2

[9] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urbandata:Astudy of new York city taxi trips," Visualization and Computer Graphics, IEEE Transactions on, vol. 19, no. 12, pp. 2149– 2158, 2013. 2

[10] L. Zhu, D. M. Gorman, and S. Horel, "Alcohol outlet density and violence: a geospatial analysis," Alcohol and alcoholism, vol. 39, no. 4, pp. 369–375, 2004. 2

[11] S. Funk and P. Piot, "Mapping ebola in wild animals for better disease control," Elife, vol. 3, p. e04565, 2014. 2

[12] NYC Taxi and Limousine Commission, "2015 hail market analysis," http://www.nyc.gov/html/tlc/downloads/pdf/hail market analysis 2015.pdf, 2015. 2, 3

[13] "TLC trip record data," http://www.nyc.gov/html/tlc/ html/about/trip record data.shtml, 2016. 3

[14] Amazon Web Services, "Amazon relational database service (amazon rds)," https://aws.amazon.com/rds/, 2016.

[15] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo." Commun. Acm, vol. 53, no. 6, pp. 59–67, 2010. 4

[16] N. Gale and W. C. Halperin, "A case for better graphics: The unclassed choropleth map," The American Statistician, vol. 36, no. 4, pp. 330–336, 1982. 4

[17] V. Mahajan, E. Muller, and R. K. Srivastava, "Determination of adopter categories by using innovation diffusion models," Journal of Marketing Research, pp. 37–50, 1990.

[18] A. Hern, "New York taxi details can be extracted from anonymised data, researchers say,"

http://www.theguardian.com/technology/2014/jun/27/

new-york-taxi-details-anonymised-data-researchers-warn, 06 2014.

AUTHORS :

MARRI MANEESHA : B. Tech Student, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem NH-16, Guntur Andhra Pradesh, India.

GANNAVARAPU JEEVAN JYOTHI:

B. Tech Student, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur Andhra Pradesh, India.

ALA DURGA TIRUMALA RAYUDU:

B. Tech Student, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur Andhra Pradesh, India.

BATTU VIJAY SAI : B. Tech Student, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur Andhra Pradesh, India.