Crime Rate Inference with Big Data

Qianran Sun

New York University

New York, USA

qs410@nyu.edu

Juanlu Yu

New York University

New York, USA

jy2234@nyu.edu

Yuqian Zhang

New York University

New York, USA

yz2444@nyu.edu

*Abstract*—

(Use a short version of your project proposal here.)

Our project is to predict crime rate using big data. We will use three datasets: the daily crime data in the City of Chicago, Public Health Statistics - Births to mothers aged 15-19 years old in Chicago, and Taxi Trips in the City of Chicago. In other words, we want to study the correlation among these three datasets.

Keywords—analytics, crimes, big data

# Introduction

Crime is one of the most important social problems in the country, affecting society safety, economic status and the development of children. Analyzing what factors cause higher crime is critical for policy makers to reduce crime and increase citizens’ life quality. In our paper, we try to analyze the crime rate inference at Chicago neighborhood level. We collected large-scale taxi flow data and Births to mothers aged 15-19 years old data in the city of Chicago, IL in the USA. By analyzing the relationship between the two datasets and the crime rate data, we could conclude which feather will significantly affect the crime rate in Chicago.

# Motivation

(Write a paragraph describing why you think this application is important.)

# Related Work

(Each team member has read papers related to this project. Please add here the paper summaries and comparisons each team member wrote. Each paper referenced should be added to the References section. When you refer to reference #1 in your paper, for example, use this notation: [1])

We investigated three papers related to our project.

First, in the paper, Fuzzy Association Rule Mining for Community Crime Pattern Discovery[1], authors study the application of fuzzy association rule mining for community crime pattern discovery. This paper summaries some previous work, available data sources, and possible techniques for crime data mining. Fuzzy association rule mining is introduced as a novel means for knowledge discovery in the crime domain, supported by experimental results on the open-source Communities and Crime data set. This paper also concludes with a discussion on directions for further research and proves Fuzzy association rule mining to be useful for crime-related data sets.

Second, abortion may have prevented “unwanted” children from being born. These unwanted children might, if born, have had smaller investments in human capital by their parents and thus been more prone to end up in trouble when they grew older. The absence of unwanted children, following legalization in 1973, led to a reduction in crime 18 years later, starting in 1992 and dropping sharply in 1995. These would have been the peak crime-committing years of the unborn children. Legalized abortion is a primary explanation for the large drops in murder, property crime, and violent crime that our nation has experienced over the last decade. Indeed, legalized abortion may account for as much as one-half of the overall crime reduction.

Third, in this paper, they solve the problem of crime rate inference with new features. They propose to use POI features to assist the demographic features, and to use taxi flow as hyperlinks to supplement the geographical neighbors. Both POI and taxi flow features from a publicly accessible dataset in Chicago are evaluated to be helpful.

Traditionally, researchers have used demographic information to estimate the crime rate in a community. However, such demographic information only contains partial information about the neighborhoods and does not dynamically reflect the changes in the community. Thus, geographical influence is of little help in improving the crime inference on top of demographic feature.

Contrary to the traditional way of using demographic information, newer types of big data: large-scale POI (Point-Of-Interest data) and taxi flow data is introduced in this paper. POI distribution across community areas reflects profiles of the region functionality. Taxi flow models the social interaction among nonadjacent regions, which potentially propagate crime or resources and information used in crime control.

There are four types of features for inference: (The details of features are discussed in section5) Nodal feature: describe the characteristics of the focal region Such features include (1)

demographic information and (2) Point-of-Interest (POI) distribution.

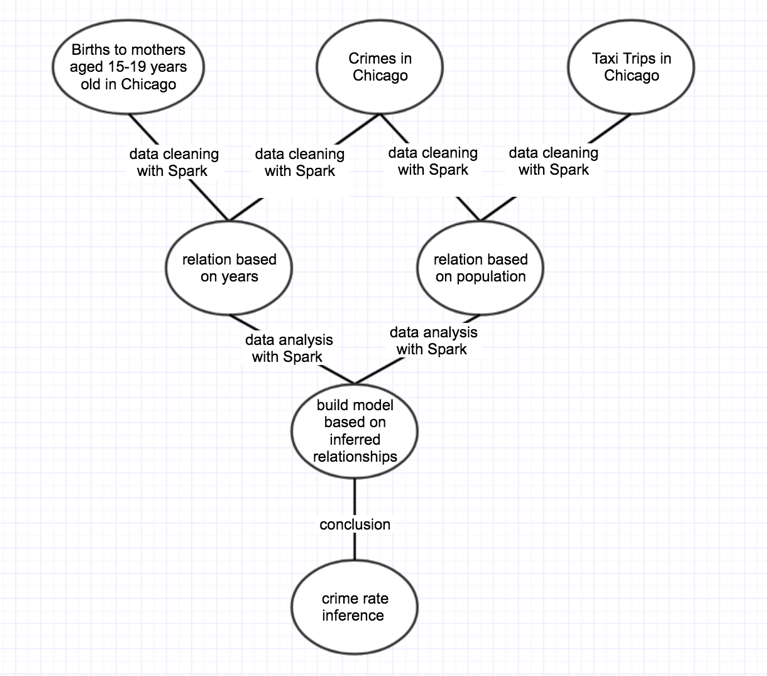
Edge feature: (1) Geographical influence. Geographical influence considers the crime rate of the nearby locations. (2) Hyperlink by taxi flow. Locations are connected through the frequent trips made by humans, which can be considered as the hyperlinks in space. Their hypothesis is that two regions that are more strongly connected through social flow will influence each other’s crime rate.

The inference models (Linear Regression, Negative Binomial Regression) are based on above four types of features.

They adopt leave-one-out evaluation to estimate the crime rate of one geographic region given all the information of all the other regions and evaluate the estimation accuracy under various feature combinations. They run both the linear regression and the negative binomial regression on five consecutive years, 2010 – 2014, with four types of features combinations tested in the various settings.

Finally, they adopt the negative binomial regression modal over the linear regression model, mainly because the count based regression models and guarantees positive prediction, while the linear regression may give negative crime rate as prediction. In the best scenario, the POI distribution and taxi flow reduces the prediction error by 17.6%.

# Design



# Experiments

(In this section, you can describe: Your experimental setup, problems with: data, performance, tools, platforms, etc. Discuss your experiments, describe what you learned. Discuss limitations of the application. Discuss what you would do to expand it given time - how would you improve it, etc.)

# Conclusion

(One paragraph about the value, results, usefulness of your application.)

##### Acknowledgment

(This section is optional. It can be used to thank the people/companies/organizations who have made data available to you, for example. You can list any HPC people who were particularly helpful, if you used the NYU HPC. List Amazon if you used an Amazon voucher.)

##### References

(Add references for all of the papers/texts that you refer to in your paper. You may have websites to reference, the Spark book, the Hadoop book, etc. A reference is added below as an example.)

1. BUCZAK, A. L., AND GIFFORD, C. M. Fuzzy association rule mining for community crime pattern discovery. In ACM SIGKDD Workshop on Intelligence and Security Informatics (2010), ACM, p. 2.
2. JOHN J. DONOHUE III AND STEVEN D. LEVITT THE IMPACT OF LEGALIZED ABORTION ON CRIME. THE QUARTERLY JOURNAL OF ECONOMICS Vol. CXVI May 2001 Issue 2
3. Hongjian Wang, Daniel Kifer, Corina Graif, Zhenhui Li Crime Rate Inference with Big Data