**Identify the Safety Level of Precincts in New York City**

Team 1: Chen Liao, Yu Hong, Tianyu Liu, Xiangxiang He

# Abstract

The New York City is known as Gotham, a city full of opportunities and dangerous. For people living in it or will live in, have a deep understanding of how safe this city is, is somehow necessary. By analyzing the number of complaints in New York City, with complementary information such as offense levels, types of crime and the geological information of the place that each complaint occurred, we will draw a network of the relationship between precincts. Our goal is to identify the safety level of each precinct in New York City, find the typical offense level in high crime rate areas, and representative precincts using network analysis methods.

Key words: crime rate, offense level, network, New York City

# Introduction

Analyze the safety level and crime rate of precincts in New York City is not a new topic. For over a decade, lots of discussions about crimes in New York City. Normally, their goals were focused on how to combat crimes and ensure citizens’ safe. Instead of only fight against crimes, our goals include help New York citizens to avoid crimes as well. In this project, we used the network analysis methods to dig the deep relationship of crimes between precincts to find the most typical precinct in crimes in New York City. Apart from the relationships, the underlying crime features of each precinct, such as the most frequently types, will be revealed as well.

The dataset includes complaints from 2012 to 2015 in New York City in total of 2,714,699 complaints, with 23 variables. We chose 5 of them, which are geological location that crime occurred, the types of crime (attempted/completed), the offense levels(felony/violation/misdemeanor), precincts of occurrence and boroughs of occurrence.

After data preparation, we built networks based on the relationships, crime rate, boroughs, communities and closeness centrality. Moreover, Quadratic Assignment Procedure and Exponential Random Graph Model were applied to analyze the underlying connection between networks.

# Data Understanding and Preparation

## Data Understanding

The original dataset1 has in total of 2,714,699 complaints and 23 columns. In the case of build networks, we chose 6 of them. Crime types include completed and attempted, which shows each crime was completed, or interrupted by police or passengers. Offense levels illustrate three levels of crime base on state law, which are felony (in total of 832,385 complaints), violation (in total of 326,266 complaints) and misdemeanor (in total of 1,556,048 complaints). Felony includes murder, forgery, grand larceny, rape and so on. Violation mainly includes sexual harassment. Misdemeanor includes possess dangerous drugs or dangerous weapons, intoxicated driving, petit larceny and so on. Borough means the districts in New York City such as Manhattan and Brooklyn. Precincts are smaller separation of boroughs, determined by New York Police Department to better organize their work. Latitudes and longitudes show the geological information that the crimes occurred.

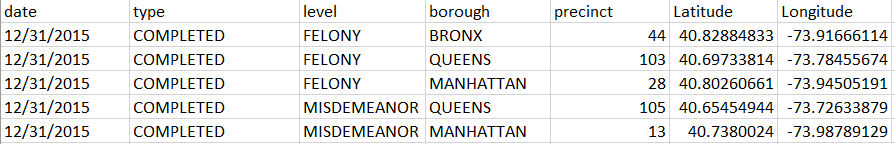


Fig 1. Sample Data

To calculate the crime rate, we also need the population of each precincts. However, we were not able to find the exact population in each precinct, so we used the population in each borough, and divided them by the number of precincts in each specific borough, to get the population in each precinct. Since the original dataset covered a time period from 2012 to 2015, therefore we computed the average population from 2010 to 2017 as the population of each precinct.

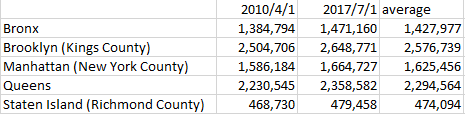
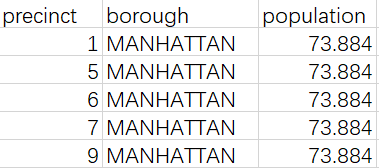
 

Fig 2. Borough Population Fig 3. Sample precincts’ population

The geological information of each precinct (longitudes and latitudes) is calculated by averaging all longitudes and latitudes that described in each complaint. In this situation, the geological information can also illustrate the location that crimes mostly occur.

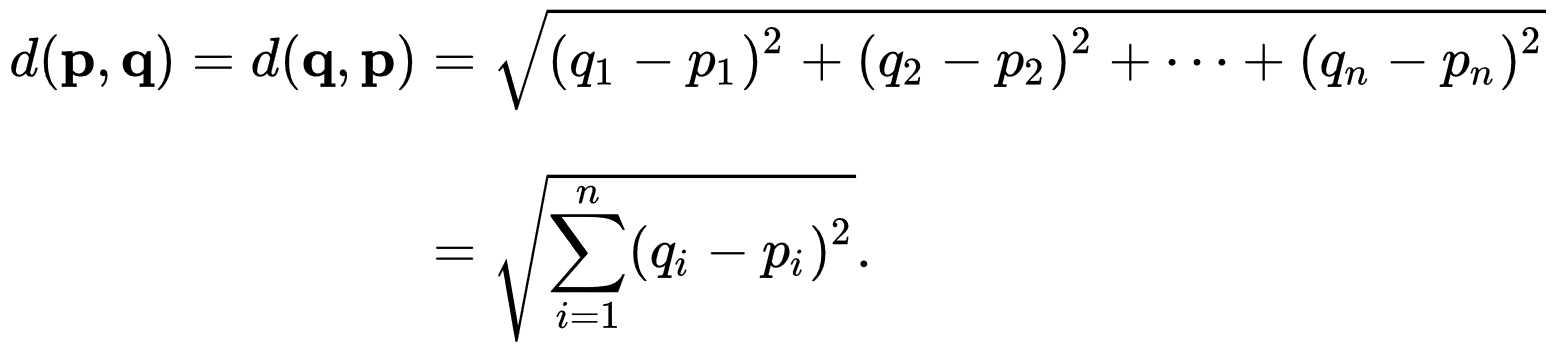
## Data Preparation

In the data preparation procedure, we focused on how to determine the edges in networks. In our first consideration, the edges should illustrate the similarity between precincts, which means we need the similarities among precincts. Therefore, we determined to assign scores to each precinct, based on the crime types and offense levels.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Crime Type | | Offense Level | | |
| Attempted | Completed | Misdemeanor | Violation | Felony |
| Score | 1 | 2 | 10 | 25 | 50 |

Table 1. Base scores

Then, calculate the product of type, level and number of crimes in each precinct, which is the final score. After that, we computed the pair-wise Euclidean distance with scores and geological information (average longitude and average latitude):



Then, we normalized the distance to the range [0,1], and calculate the similarity by:

*Similarity = 1 – Euclidean distance*

Finally, reserve the top 5 similarity for each precinct.

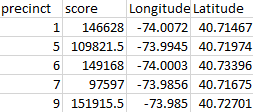
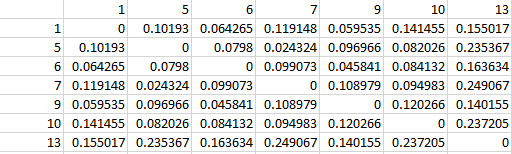
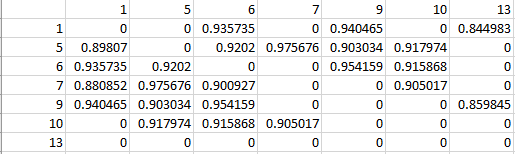
  

Fig 4. Scores to normalized Euclidean Distances to Edges

Apart from edges, we also need the crime rates. Simply divided the scores by population of each precinct, we have the crime rate (per thousand people).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| precinct | borough | population | score | Crime per thousand people |
| 1 | MANHATTAN | 73.884 | 20314 | 274.9445 |
| 5 | MANHATTAN | 73.884 | 15341 | 207.6363 |
| 6 | MANHATTAN | 73.884 | 16192 | 219.1544 |

Table 2. sample data of crime rate

# Networks

## Main Networks

To have an overall view of our network, we plot two types of figures, one with the crime rate per thousand people as node size, and the other with closeness centrality as node size.

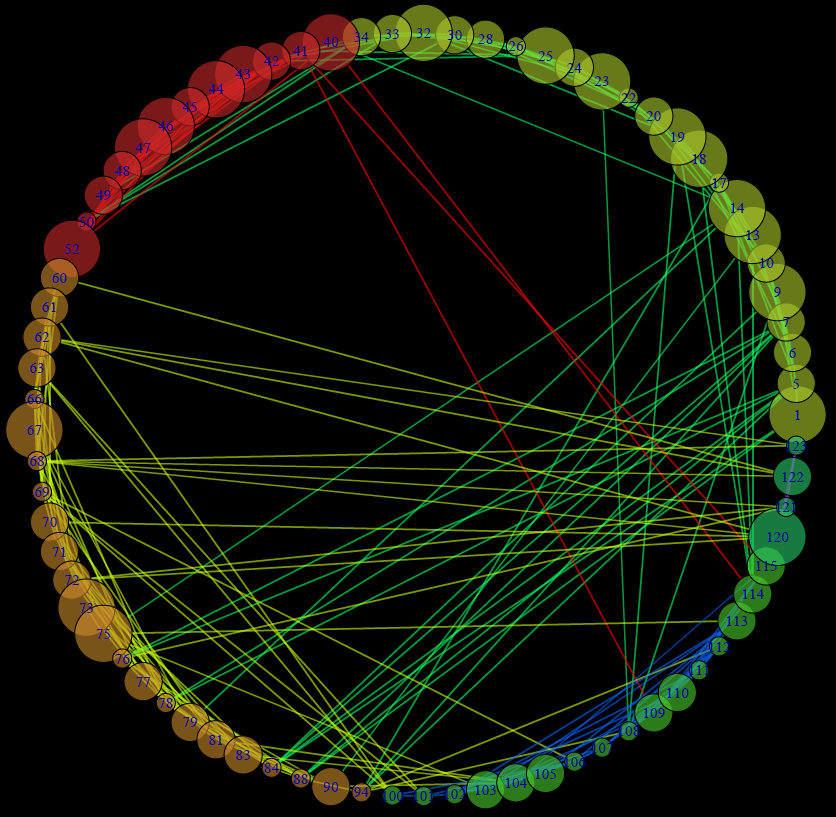


Fig 5. Main Network with Crime rate as node size

The number at the center of each node indicates the precinct that it represents.

According to figure 5, we can easily find out that the crime pattern of Lower town Manhattan (precinct 1 to 10) is similar with some places in Brooklyn (precinct 84, 88 and 94). It is also obvious that most precincts in Brooklyn have lower crime rate per thousand people than Manhattan, however Brooklyn is always considered as more dangerous than Manhattan.

It is also surprising that Queens (from precinct 100 to 115) has even lower crime rate than Manhattan, and the crimes are more concentrated in several precincts, which may easier for citizens to avoid being attacked in Queens by just don’t show up in those areas.

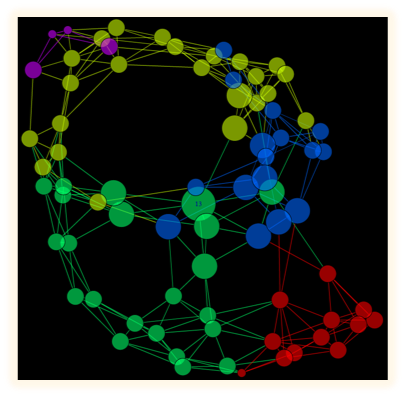


Fig 6. Main Network with Closeness Centrality as node size

Closeness centrality is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes2.

Because our network was constructed base on ‘similarity’, therefore by calculating the closeness centrality, may we find the most central node, which is also the most representative one. In our case, precinct 13 in Manhattan is the most the representative one, as it may has averaged number of crimes of every offense level.

## Communities:

Community refers to a group of nodes that are similar with each other or say, have the same module. is Community is formed by individuals such that those within a group interact with each other more frequently than with those outside the group. We are grouping all the closely connected precincts in the New York City. Now colors also indicate different communities. We can also find seven colors from this graph. In this case, we can see the relationship between each precinct more clearly. The figure connects the nodes with the closest relationship. What’s different is that the Manhattan is divided into 3 communities because Manhattan includes the lower, middle and upper downtowns so it has more different features and is very complicated.

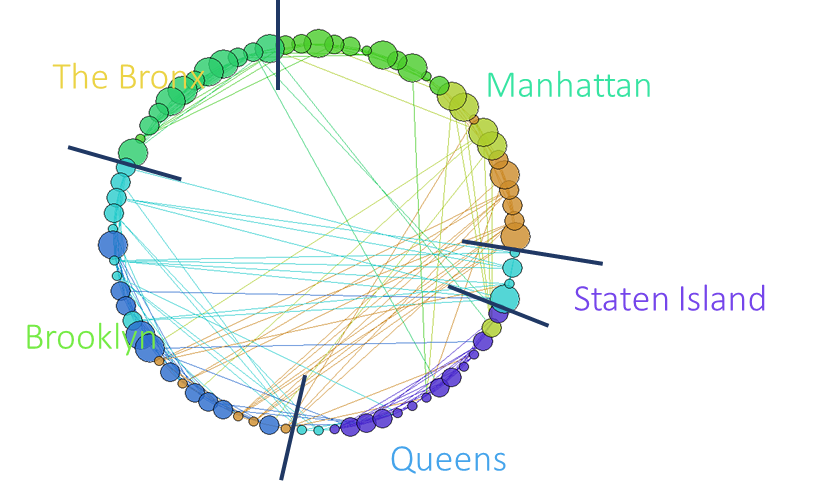
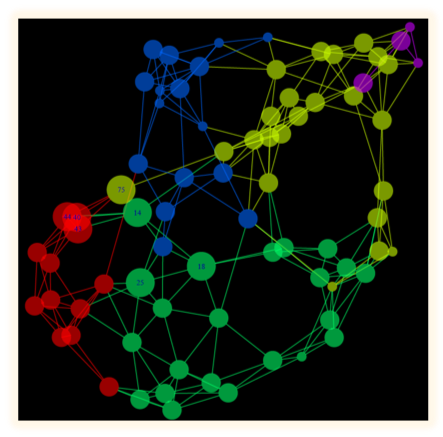
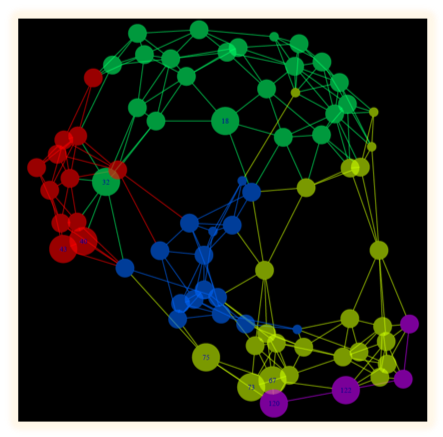
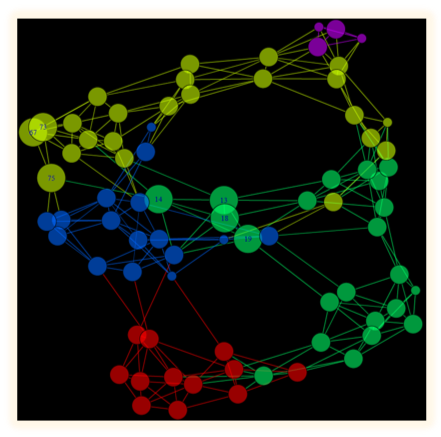


Fig 7. Communities

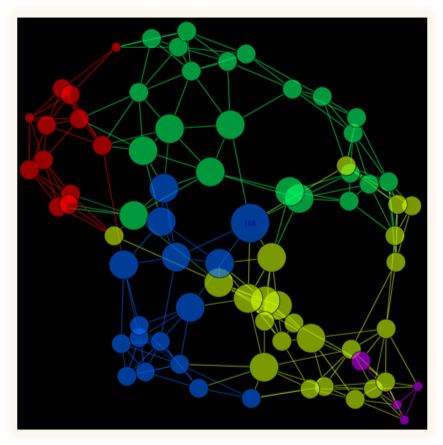
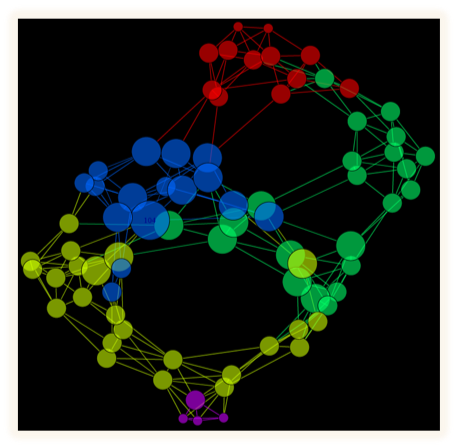
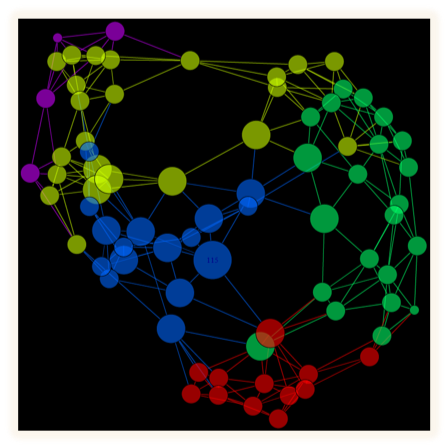
Networks based on different offense level:

In this part, we construct three different networks base on different offense levels. From the three figures, we can easily determine which precinct has more type of crimes. These three figures can help us analyze the crime situation in different parts of New York more deeply than the previously overall figure. For example, for some precincts in Queens, there are most violation but a little other two types of crimes.

Closeness centrality of three new networks

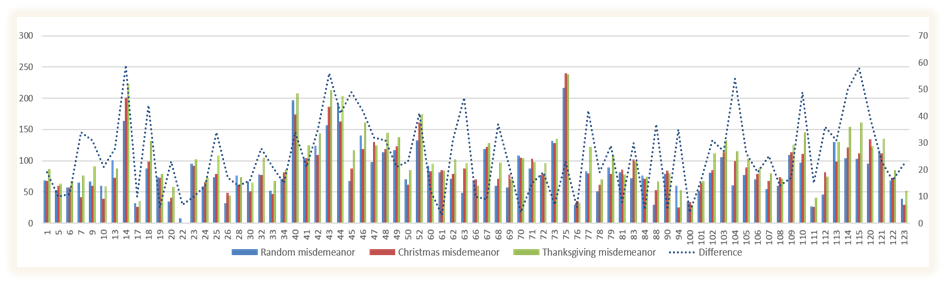
These three graphs are all the closeness centrality of the last three graphs. In this case, we can find the most representative nodes in three crime types. For example, in the misdemeanor figure, the largest circle is the precinct numbered 108. What’s interesting is that for all three types of crimes, the most representative nodes in three different graphs are all precincts in Queens. After that, we can conclude that most typical districts are all in Queens. Compared with other graphs, they are more specific.

# Analysis of Networks

## Analysis of the Special Days

To have a more specific view of the crime, we analyzed the number of crimes in different time periods. We chose the Thanksgiving holiday, the Christmas holiday and ordinary times. In order to balance the time period length, we chose 4 days for each three categories since every year Thanksgiving have 4-day-long holiday.



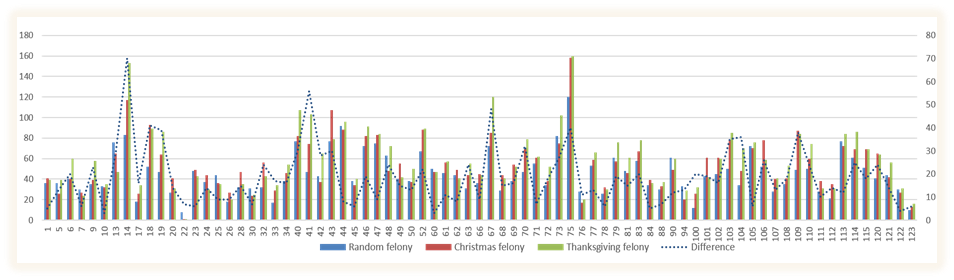
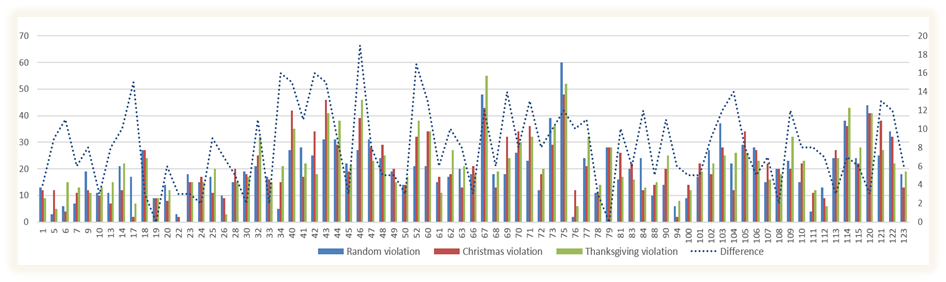


Fig 5. From top to bottom, Misdemeanor, Violation, Felony

According to the figures above, during the Thanksgiving and Christmas, the misdemeanor and felony offences happen more than ordinary times, but the violation attack rate decreases. However, in overall view, NYC is more dangerous in special days than ordinary times. To be more specific, the NYC police should suggest citizens to stay at home or don’t visit specific precincts in these days to avoid being hurt.

## Crime Rate Visualization

Then we plot a heat map for crime rate of each precinct to have a more directly view.

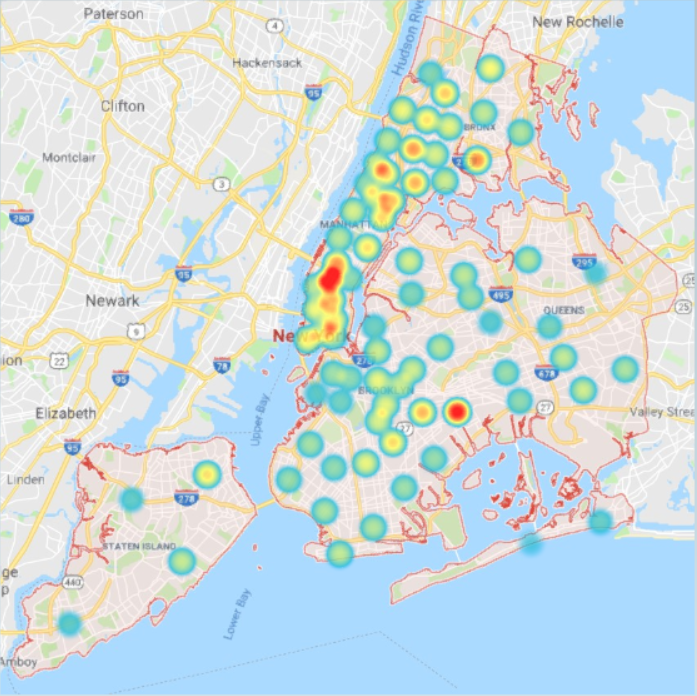


Fig 8. Crime rate visualization

As we can see from this heat map, the heat of the points means the crime rate in NYC. The crime rate with high value will have red color, and low crime rate place will have blue or no special colors.

The result is very impressive, to our surprise, people tend to believe the stereotype that Manhattan is the safest place in NYC, however, it’s incorrect. Manhattan is the most dangerous place in NYC which is with the highest crime rate. And Brooklyn and Queens are not as dangerous as people think. The NYC police should notice people to watch out in Manhattan, by this means they can control the crimes in NYC.

## QAP Tests

To understand the relationship between networks better, we applied the QAP Test to each pair of networks.

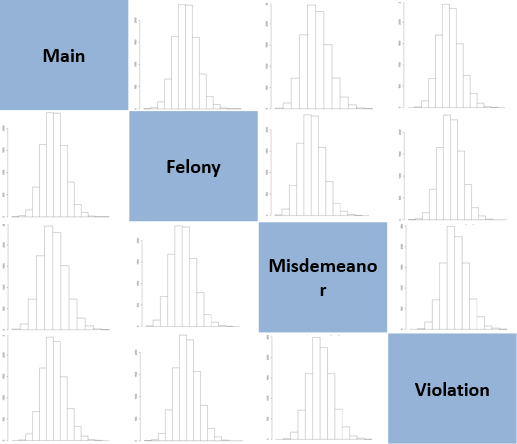


Fig 10. QAP Test result

QAP (Quadratic Assignment Procedure) test is a test that we can compare the interdependency between different networks. The structural relationship between different networks can be described by the correlation coefficients. The network will be changed randomly for many times to test if its structural relationship with other networks is special.

In our case, the plot above shows that nearly all the correlations of networks are normally distributed in different situations, which means that the relationship between each network is based on their structure rather than other features, as the QAP majorly tests the structure relativity.

Also, we can figure out how much they related with each other by calculating the correlation coefficients. The correlation coefficients’ matrix as shown below:

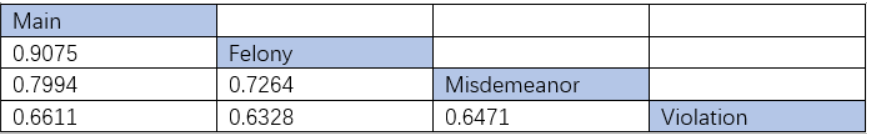
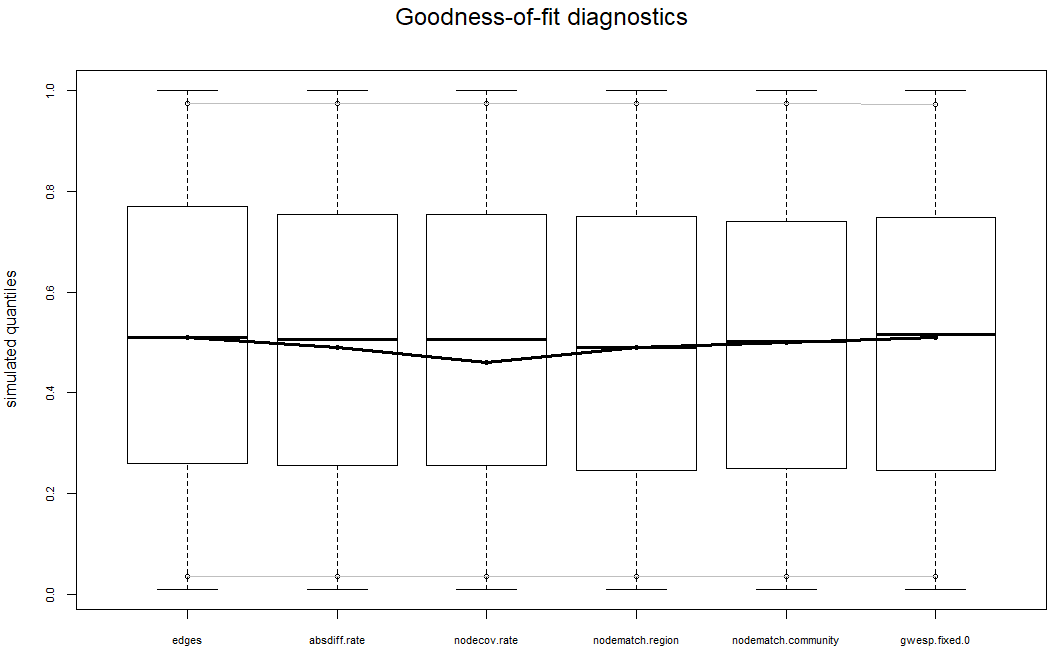


Fig 10. Correlation Coefficients

From this table, we can understand that the correlation among felony, misdemeanor and violation is high. To be more specific, the overall crime rate is highly correlated with the felony offense. So, the NYC police can take the crime rate of felony as an evaluation of the overall crime rate of NYC.

## ERGM Analysis

In order to find out the attributes that influenced the structure of networks, we need the help from ERGM.

 A close up of a map

Description automatically generated

Fig 10. ERGM results

ERGM (exponent random graph modeling) is a method which we can assign different attributes to the network and form a new randomly created network. By comparing the structure of the new network and the original network can we understand the most important attributes that influence the format of the original network.

In this plot, we found that the most important attributes that trigger the special structure of the NYC crime rate network is the absolutely distance of the crime rate in different precincts, the community of the precincts and the transitivity of the network.

Community is the cluster of the nodes based on their own attributes, in this project, the attributes of the node is the crime rate, the region of the precincts, and the degrees (that is, how many connections exist in a single node) of them.

The transitivity is that if whenever an element a is related to an element b and b is related to an element c then a is also related to c. Transitivity (or transitiveness) is a key property of both partial order relations and equivalence relations. (quoted by Wikipedia)

# Conclusion

With the results and analysis, we may give some conclusions and further, provide an inspiring perspective for the police department to understand crime situations in New York City. For instance, compared to the stereotype that Manhattan is safer than other districts, actually it is the most dangerous place.

Overall, both precinct 14 in Manhattan and precinct 75 in Brooklyn have the largest number of complaints, and correspondingly the highest crime rate.

Moreover, if we take dimension of time into consideration, however, the frequency of crimes varies by date and criminal categories. During Thanksgiving and Christmas, more misdemeanor and felony type of crimes could happen more frequently in most precincts, while there seems to be less violation type of crimes in these special dates than in normal days.

# Outlook and discussion

Beyond this project, there are abundant potential applications expected. For example, with the most representative precinct found (precinct 13 in this model), the police department could monitor criminal situation of the whole city by learning cases in this relatively smaller area.

For further study, it may be helpful to take other attributes like race, gender and time of criminal act into account. Based on them, we may predict more accurate “dangerous rate” for particular persons being at particular places and push individual alerts.

# Reference

1: Dataset from New York Police Department Public Data: <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>

2: Definition of Closeness Centrality, Wikipedia: <https://en.wikipedia.org/wiki/Closeness_centrality>

3. Explanation of ERGM, Wikipedia:

<https://en.m.wikipedia.org/wiki/Exponential_random_graph_models>