# Analysis of Time-Varying Network

Xiaoyi Wen

May 7, 2023

## 1. Introduction

Networks are among the most widely-applied tools in machine learning and statistics with a large number of successful applications, including biology, climatology, sociology and epidemiology. The fundamental property that networks capture is the interaction (edges) of different entities (nodes), a description well-suited to numerous domains. There exist various generative models to capture the key characteristics of networks arising in practice. One of the simplest random graph models. Random graphs form an important building block in the correlation between graphs or graphs and their nodes' attributes, a task which arises in many applications.

In this project, we want to explore how COVID-19 cases will affect the New York yellow taxi traffic flow network. Specifically, we mainly focus on the following goals:

- Establish a traffic flow network of yellow taxis in New York and analyze the structure of the network, including node and path analysis;
- Construct an effective model to analyze the time-varying network and the correlation with COVID-19 cases and determine the region most affected by COVID-19.

## 2. Data Description

The yellow and green taxi trip records on pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types and driver-reported passenger counts, collected by New York City Taxi and Limousine Commission(NYC TLC), are publicly available at `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. Additionally, NYC Coronavirus Disease 2019 (COVID-19) data are available at `https://github.com/nychealth/`

`coronavirus-data`, where one can find citywide and borough-specific daily counts of probable and confirmed COVID-19 cases in New York City since February 29, 2020. This is the date at which according to the Health Department the COVID-19 outbreak in NYC began.

The data we want to analyze include:

- The yellow taxi trip records data collected by New York City Taxi and Limousine Commission(NYC TLC), including April to August 2020;
- NYC Coronavirus Disease 2019 (COVID-19) data from April to August in 2020.

The boroughs in NYC are divided into 6 regions including the islands, and the zones in NYC are divided into 259 regions. The pickup location and dropout location has an edge in a random graph. The yellow taxi zones (excluding islands) as delimited by New York City Taxi and Limousine Commission (NYC TLC) is shown in Figure 1.

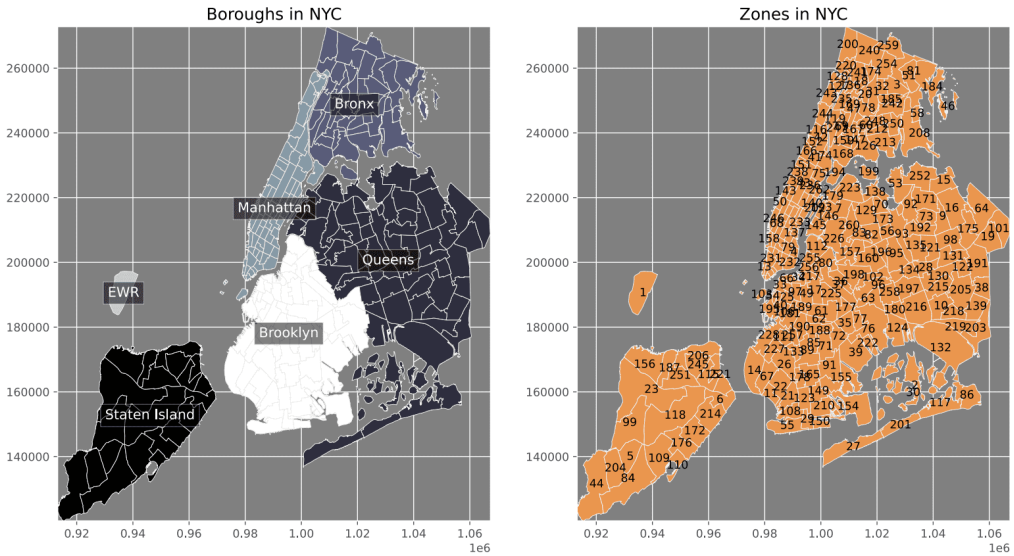


Figure 1: NYC Zone.

## 3. Time-Varying Network Representation

Dubey and Müller (2022) proposed a method based on the Fréchet Mean to find a Fréchet variance trajectories representation of the time-varying network data. Aftering transforming the time-varying

network into the normal vector, we can do further analysis using the classical statistical method. First, we want to introduce the Fréchet Mean and Fréchet variance.

### *3.1. Fréchet variance trajectories*

Consider a random object $X \sim F_X$ taking values in a metric space $(\Omega, d)$. The Fréchet mean and Fréchet variance of random objects in metric spaces are defined as:

$$\omega_{\oplus} = \arg\min_{\omega \in \Omega} E\left[d^2(X, \omega)\right], \quad V_{\oplus} = E\left[d^2(X, \omega_{\oplus})\right].$$

For the $\Omega$-valued stochastic process $\{X(t)\}_{t\in[0,1]}$ and sample of random object trajectories $X_1, \cdots, X_n$. Given $t \in [0,1]$, the population and sample Fréchet mean trajectories at $t$ are defined as

$$\mu(t) = \arg\min_{\omega \in \Omega} E\left(d^2(X(t), \omega)\right), \quad \hat{\mu}(t) = \arg\min_{\omega \in \Omega} \frac{1}{n} \sum_{i=1}^{n} d^2(X_i(t), \omega).$$

The target functions for our analysis would be the functions:

$$V_i^* = d^2(X_i(t), \mu(t)),\ t \in [0,1],$$

which correspond to the pointwise squared distance functions of the subject trajectories $X_i$ from the population Fréchet mean function $\mu = \mu(t)$ for the subject trajectories $X_i$. Population Fréchet variance trajectories:

$$V_i = d^2(X_i(t), \hat{\mu}(t)),\ t \in [0,1],$$

since $\mu$ is unknown and needs to be estimated from the data.

### *3.2. Network representation*

Let $G = (V, E)$ be a network with a set of nodes $V = \{v_1, \cdots, v_m\}$ and a set of edge weights $E = \{w_{ij} : w_{ij} \geq 0, i, j = 1, \cdots, m\}$, where $w_{ij} = 0$ indicates $v_i$ and $v_j$ are unconnected. Some basic and mild restrictions on the networks $G$ we consider here are as follows.

(C1) $G$ is simple, i.e., no self-loops or multi-edges exist.

(C2) $G$ is weighted, undirected, and labeled.

(C3) The edge weights $w_{ij}$ are bounded above by $W \geq 0$, i.e., $0 \leq w_{ij} \leq W$.

The graph Laplacian is defined as $L = D - W$, where $D$ is the degree matrix. The corresponding space of graph Laplacians given by:

$$\mathcal{L}_m = \left\{ L = (l_{ij}) : L = L^{\mathrm{T}}; L1_m = 0_m; -W \leq l_{ij} \leq 0 \ \text{ for } \ i \neq j \right\},$$

We usually analyse the random graph by the Laplacian matrix. Next, we define a metric to measure the distance between graphs. A common choice of metrics for the space of graph Laplacians $\mathcal{L}_m$ is the Frobenius metric, defined as

$$d_F(L_1, L_2) = \|L_1 - L_2\|_F = \left\{ \mathrm{tr}\left[ (L_1 - L_2)^{\mathrm{T}} (L_1 - L_2) \right] \right\}^{1/2},$$

Denote the space of real symmetric positive semi-definite $m \times m$ matrices by $\mathcal{S}_m^+$. Defining matrix power maps:

$$F_\alpha(S) = S^\alpha = U \Lambda^\alpha U^{\mathrm{T}} : \mathcal{S}_m^+ \to \mathcal{S}_m^+,$$

where the power $\alpha > 0$ is a constant. The power metric between graph Laplacians is

$$d_{F,\alpha}(L_1, L_2) = d_F \left[ F_\alpha(L_1), F_\alpha(L_2) \right].$$

## 4. Data Applications

### *4.1. Exploratory data analysis*

We plot the 'CASE COUNT' and 'DEATH COUNT' of New York as shown in Figure 2. We can see

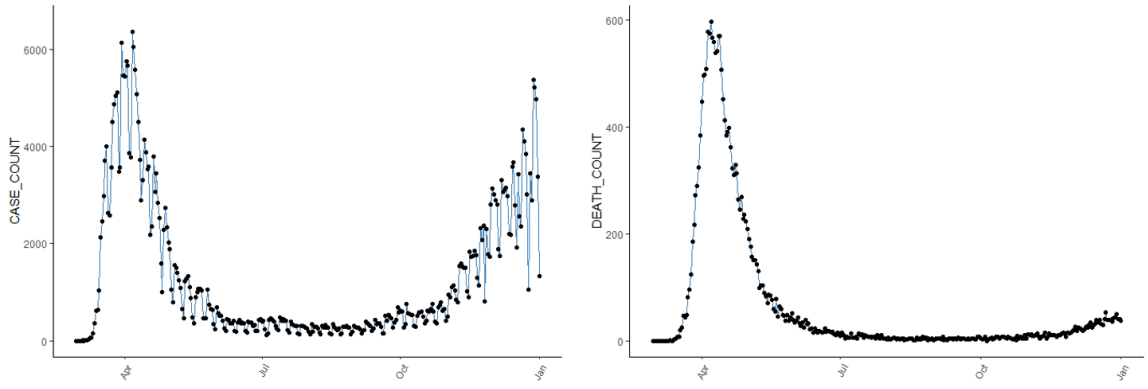


Figure 2: 'CASE COUNT' and 'DEATH COUNT' of COVID-19 cases in NYC.

that the case count of COVID-19 cases peaked in April and then decreased. During the analysis of the COVID-19 cases, we found that in April 2020, the number of COVID-19 cases was the highest. Then we choose the traffic flow in April 2020 as an example and plot in Figure 3. Figure 3 shows

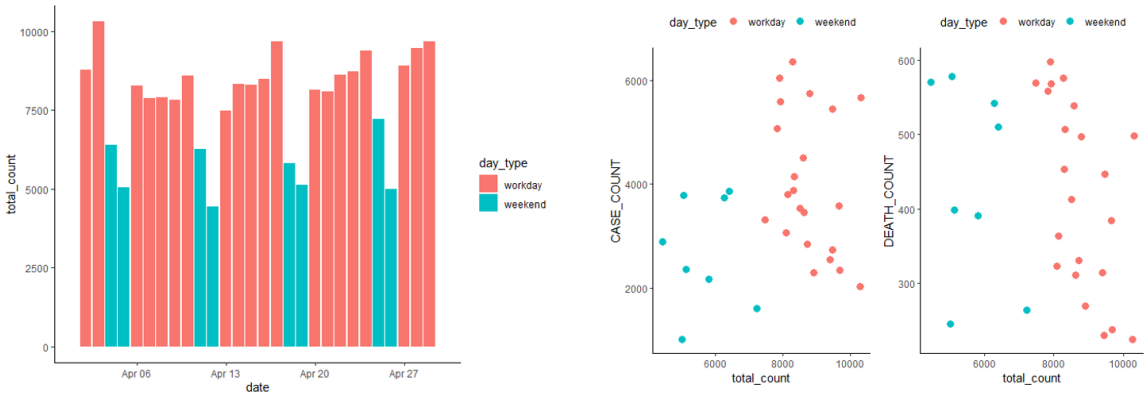


Figure 3: Traffic Flow in April 2020.

that the traffic count differs between weekdays and weekends. Moreover, we can plot the density heatmap of the pick-up and drop-off locations. We select two days, one is a workday, and the other is a weekend. We can see a slight difference in the density of pick-up and drop-off locations on

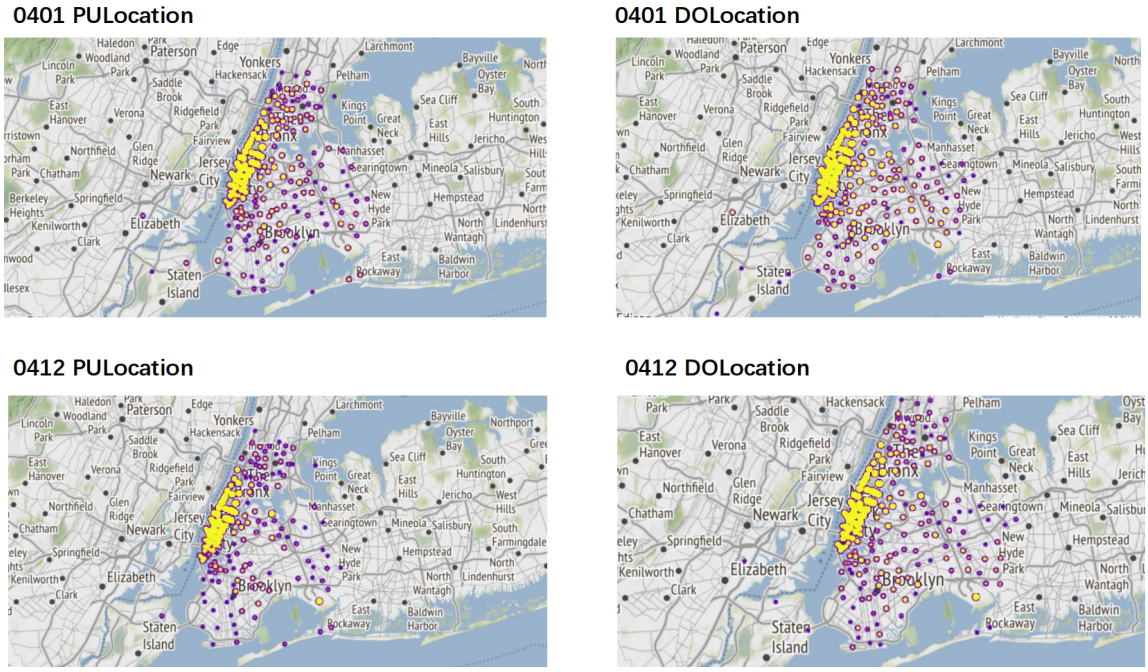


Figure 4: Pick-up and Drop-off density for the selected date.

weekdays and Sundays. It may not be noticeable in Manhattan, where there is a lot of ridership, but there are significantly more weekdays than Sundays for ridership points around Manhattan. Next, we also drew the traffic network map for April 1st and 12th, as shown in the figure below.

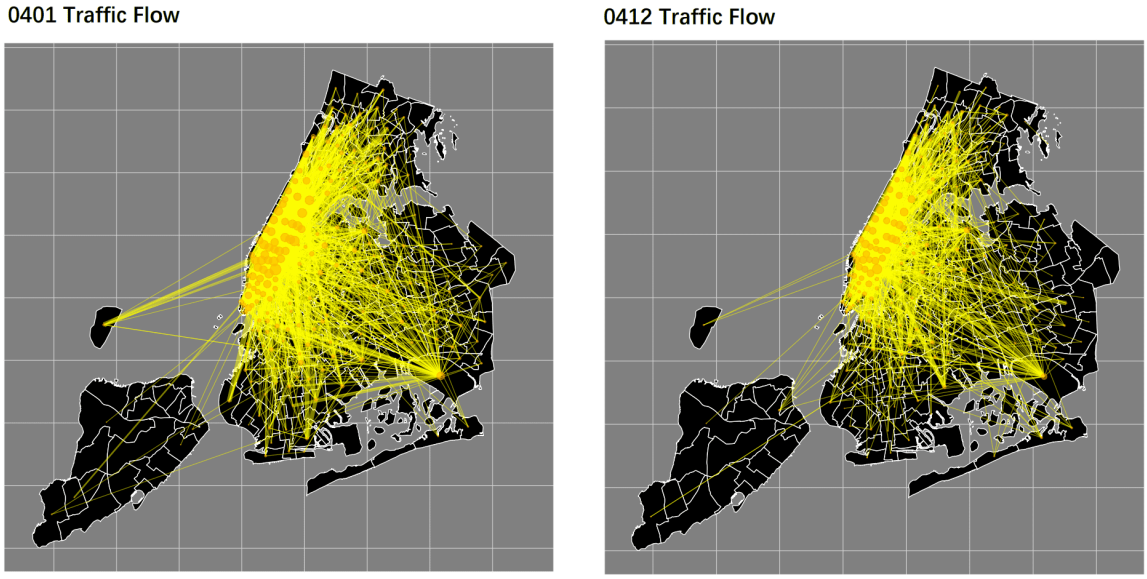


Figure 5: Traffic flow network for the selected date.

From the traffic flow network, we can also see the difference between weekdays and weekends. Therefore, weekdays and weekends should be analyzed separately when analyzing whether COVID-19 cases will impact the taxi traffic network.

### *4.2. Time-Varying Network Representation*

We choose four main regions, namely Bronx, Brooklyn, Manhattan and Queens, and divide the traffic network from April to September into weekdays and weekends, using the representation based on Fréchet Mean we mentioned earlier method, get the Fréchet Variance Trajectory of weekdays and weekends in the four regions, as shown in the Figure 6.

From the obtained Fréchet Variance Trajectory, we can see that there are indeed significant differences between the trajectories of weekdays and weekends, especially in Manhattan, where the traffic volume is the largest. The Fréchet Variance Trajectory in Manhattan and Queens both had a trough in July, which means that traffic volatility is minimal between June and August. We can combine the left graph of Figure 2 to analyze that in 'case count', the number of COVID-19 cases gradually

decreased in May and maintained a minimum value during June and August but began to steepen in October. The volatility of the transportation network is probably related to this phenomenon.

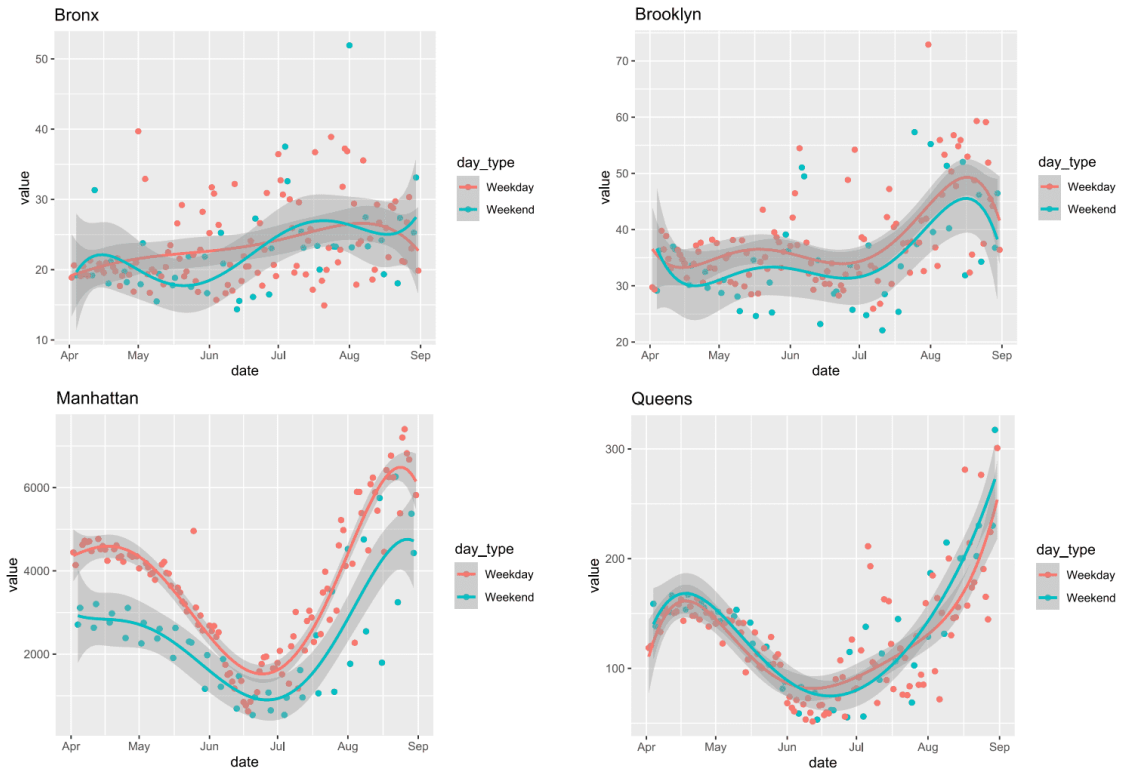


Figure 6: Traffic flow network for the selected date.

After getting the Fréchet Variance Trajectory of the transportation network of the four regional weekends and weekdays from April to August, we can draw a scatter plot of the Fréchet Variance Trajectory and the number of cases, and annotate the correlation coefficients between the Fréchet Variance Trajectory and the number of cases on weekdays and weekends in the upper left corner of the figure. We can see that the correlation between weekday traffic flow and the increase in COVID-19 cases is more significant. In contrast, on weekends, perhaps because of the small amount of data, the correlation is not significant.

## 5. Discussion

Considering the framework for analyzing time-varying object data, where the random objects can take values in a general metric space, by defining a generalized notion of mean function in the object space. The time-varying networks can be transferred to a time series using the Freéchet variance trajectories.

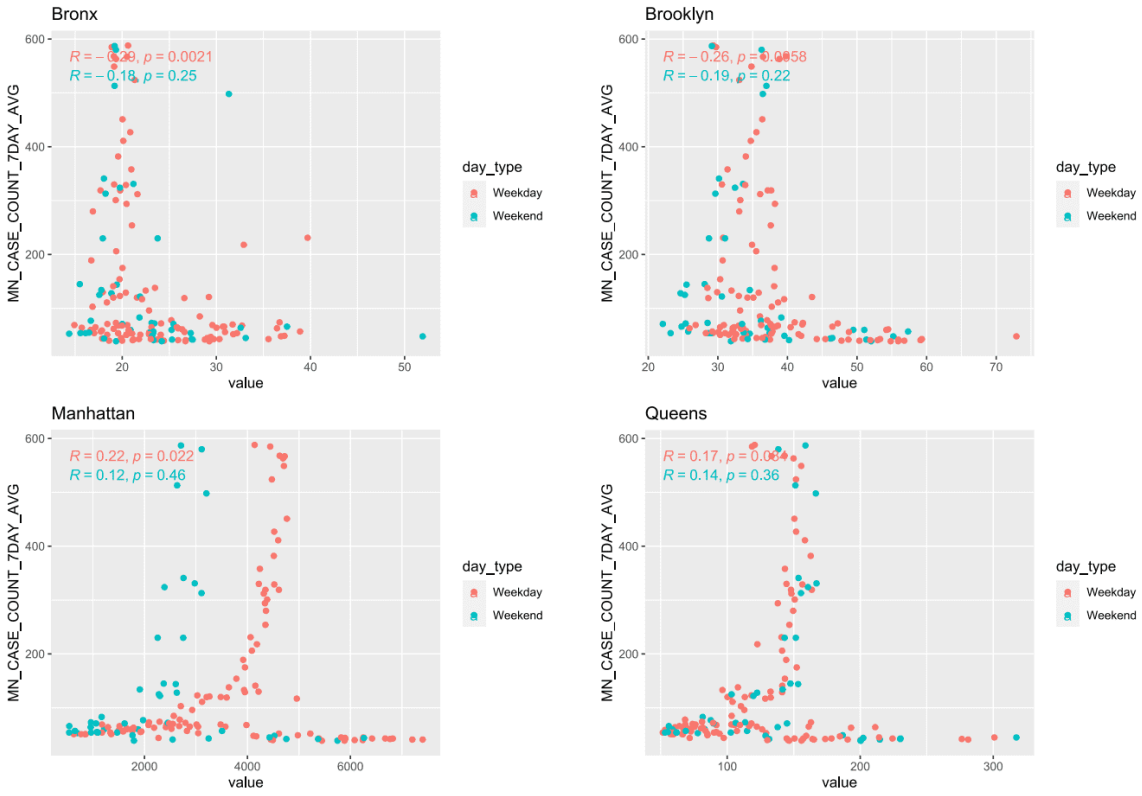


Figure 7: Traffic flow network for the selected date.

In the real data analysis, we can know that COVID-19 cases are indeed affecting the taxi transportation network in New York. The increase in the number of COVID-19 cases has a greater impact on traffic flow on weekdays but less on weekends. As the area with the largest traffic volume, Manhattan is also the most affected by the new crown.

Since we use the variance trajectory of the time-varying network mainly to describe the volatility of the network structure, the negative correlation between the variance trajectory and COVID-19 becomes difficult to interpret.

## References

Dubey, P. and Müller, H.G. (2022). "Modeling time-varying random objects and dynamic networks." *Journal of the American Statistical Association*, **117(540)**, 2252–2267.