How the Pandemic Influences Travel Mode Choices? Similarity and Network Analysis Leveraging Social Media Data

Xu Chen
Department of Civil Engineering and Engineering Mechanics
Columbia University
500 W. 120th Steet
Newyork, NY 10027
Email: xc2412@columbia.edu

Xuan Di
Department of Civil Engineering and Engineering Mechanics
Data Science Institute
Columbia University
500 W. 120th Steet
Newyork, NY 10027
Email: sharon.di@columbia.edu

Words: 4,810
Figures: 13 (3,250 words)
Algorithms: 0 (0 words)
Total: 8,060

Submitted for the Special Issue of TRR on COVID-19 and Transportation
Submitted: October 31, 2020

Abstract

This paper aims to understand changes in people's travel behavior, mode choice in particular, and assist urban planners and policymakers for greater preparedness and resilience to future pandemics. We explore individual traces through the lens of people's social media activities, including both social media mentions of travel modes and spatial-temporal check-in data. In particular, we leverage two types of social media data, Twitter and Facebook, to extrapolate people's tweet mention of a travel mode along with geo-tagged information, as well as people's geo-locations when they check in using Facebook apps. Building on these data, similarity based network analysis are performed to understand how public opinions and information spread and influence people's opinion of travel mode selection, respectively. This study is mainly focused on New York City across three phases: before the stay-at-home order was enforced, when the stay-at-home order was in place, and after the reopening of NYC. We find that: (1) Users are more active in social media than in the real world after the stay-home order. (2) Compared with other travel modes, users are more easily affected by tweets about subway after the stay-home order due to the fear of travel modes which cannot prevent users from being infected with coronavirus. (3) In the reopening phase, topics about travel modes are hot debated among users who get back to work while the number of subway and taxi trips remains at a low level due to the fear of coronavirus.

Key-words: Mode Choice, Social Media, Similarity, Network Analysis

1 Introduction

The COVID-19 pandemic has affected over 30 million people across the globe, among them 7 millions are Americans with death toll surpassing 200,000 (1). It is estimated one in five residents in New York City (NYC) might have been infected by COVID-19 (2). Fearing that the subway system accelerates virus spreading and infection, a majority of transit commuters have shifted to buses (3), individual cars or bikes (4), leading to adversarial impact including more speeding tickets (5), surging bike traffic (6), and more crashes with cyclist injuries (7). As the lifeline of NYC, mass transit including subways and buses used to carry over 6 million person trips daily and 1.5 billion annually (4), but this ridership has dropped by approximately 70% (for subways) and 50% (for buses) compared to 2019 equivalent day (8). Unfortunately, such a substantial drop in public transportation ridership has also been seen across the globe (9, 10). On the other hand, people from low-income communities (mostly Hispanic and black minorities) have been hit the hardest by the coronavirus (11), and are surely going to be in a disadvantaged position due to lack of accessibility to safer travel modes, given that 75% of essential works are people of color and 60% of whom are renters who spend on average 1.5 hours commuting on public transportation (12).

The **goal** of this paper is to understand changes in people's travel behavior, mode choice in particular, and assist urban planners and policymakers for greater preparedness and resilience to future pandemics. While (13–16) provided a comprehensive picture of how aggregate traffic patterns evolve during the pandemic, this paper aims to delve into pattern changes in individuals travel mode choices by inspecting inter-personal similarity in travel modes across time.

Individual mobile traces are challenging to obtain due to privacy issues. Instead, we resort to a proxy of individual traces through the lens of their social media activities, including both social media contextual mentions of transportation topics and spatio-temporal check-ins. In particular, we leverage two types of social media data, Twitter and Facebook, to extrapolate one's travel mode topics in the social media world, or spatial-temporal mobile traces in the real-world whenever one checks in using social media. Recent years have seen a growing trend of using social media data in travel behavior studies, including crowd sensing and routing (17), activity pattern classification (18), activity location and pattern inference (19), travel activity estimation (20), and longitudinal travel behavior inference (21). Interest readers can refer to (22, 23) for a comprehensive review of the value of social media for travel behavioral studies. In spite of limitations of social media data such as sparse time and geolocation information, it is safer, quicker, and more reliable to collect in the face of emergency, compared to traditional methods via travel diaries or deployment of tracing devices. (24) explored the impact of social media in information sharing and spreading during Hurricane Sandy. (25) leveraged geo-tagged tweets to understand people's travel mode changes before and after Uber and Lyft withdrew their service from Austin, Texas. (26) analyzed the number of views of coronavirus related contexts and found that social media has become an ubiquitous platform for health information sharing and education.

This paper aims to understand pattern changes in people's travel mode choices prior to and post the pandemic leveraging social media data. We will analyze people's mode choices in the real-world through their activities in the social media world using both similarity analysis and network effect analysis. We will also establish a relationship between these two types of behaviors by comparing both social media activities and aggregate traffic data from the NYC Open dataset (27). This study is mainly focused on NYC for several reasons. First, NYC was the epicenter of the pandemic, which caused a substantial disruption to people's travel activities. Social media will be an insightful source to understand the "new normal" once the stay-at-home order was lifted. Second, NYC's multi-modal transportation infrastructure system, comprised

of subways, buses, bicycles, taxis, pedestrians, provides an idea platform to understand how travel mode choices are shifted among various travel modes and to what extent transit is influenced due to people's fear of coronavirus spreading in closed environment. Third, NYC Open dataset contains aggregate traffic counts of each travel mode, which is a valuable source for comparison with social media data.

The contributions of this paper include:

1. Two types of data are used for analysis, including contextual information (i.e., topic mentions on travel modes) and spatial information (i.e., geo-locations from tweets and check-in locations from Facebook). The former manifests people's opinions about travel modes, while the latter reveals individual spatial-temporal mobility traces.

2. Three phases are analyzed, including before the stay-at-home order was in place, during, and after the stay-at-home order was lifted. We find that each phase exhibits quite different patterns, which can provide insights into people's attitudinal evolution towards different travel modes.

3. Both similarity analysis and network regression analysis are implemented to understand how inter-personal similarity evolves across three phases and how social media quantitatively influence people's opinion of travel mode selection.

The remainder of the paper is organized as follows. Section 2 describes the data format we will use for analysis. Section 3 introduces two similarity measures: contextual similarity and spatial-temporal similarity, which are used for phase and topic analysis, respectively. Section 4 performs network analysis and network regression to analyze how information spreads in the social network and influences social media users' travel mode mentions. Section 5 concludes.

2 Social Media Data

In this section, we introduce data sample collected from social media: Twitter and Facebook. The data sample consists of two parts: textual data and spatial-temporal data.

2.1 Textual Data

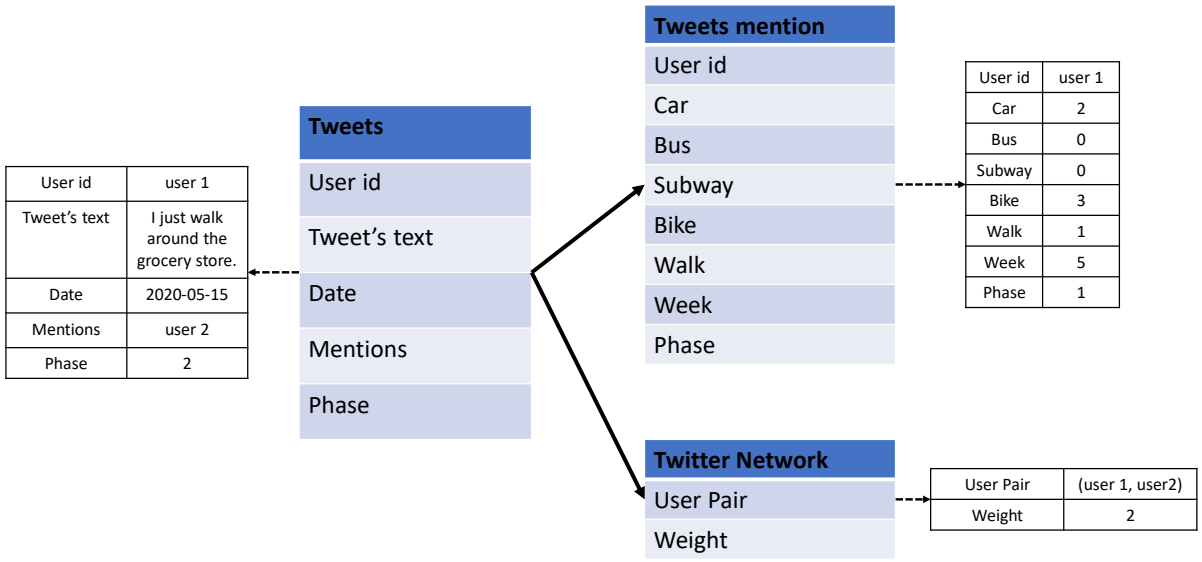


Figure 1: The diagram of textual data

Social media provide users' textual data like tweets. We use Twitter official API to fetch tweets related to travel modes based on regions, time intervals and keywords. As aforementioned, we focus on geo-locations that fall within NYC. The temporal range is from January 1, 2020 to July 20,2020, covering the entirety of the lockdown period (March 22-June 8) in NYC. We divide the temporal range into three phases:

1. Phase 1: 01/01-03/22 (before the stay-home order in NYC began).
2. Phase 2: 03/22-06/08 (when the stay-home order in NYC was in place).
3. Phase 3: 06/09-07/20 (when the 4th stage of reopening in NYC began).

Basic keywords used to fetch tweets include transportation types and transportation services. We fetch around 32425 tweets from 5694 users.

To better understand the textual data, we use a diagram in Fig. 1 to illustrate three types of data used for analysis. The table "Tweets" is about tweets from all users. Each tweet has the following fields: "Tweet's text" (what content it included), " Mentions" (users mentioned in it), "Date" (when it was posted), "User id" (by whom it was tweeted) and "Phase" (which phase it belongs to). The table on the side of "Tweets" shows one tweet posted by user 1 in Phase 2.

In the table "Tweets mention", each record represents a user's tweets mention, i.e., the number of tweets regarding some topic posted by the user. Topics are identified by five main modes of transportation: car, bus, subway, bike and walk. To decide whether a user mentions some of the five topics, we use keyword lists to filter the user's content. The keyword list of topic "Car" is: automobile, cab, drive, Lyft, taxi and Uber. The keyword list of topic "Bus" is: bus. The keyword list of topic "Subway" is: metro and subway. The keyword list of topic "Bike" is: bike, Citibike, cycling. The keyword list of topic "Walk" is: walk. The table on the side of "Tweets mention" shows that user 1 posted two tweets about car, three tweets about bike and

one tweet about walk in the 5th week of Phase 1. Fig. 2 plots tweets mention of an individual user from Phase 1 to 3.

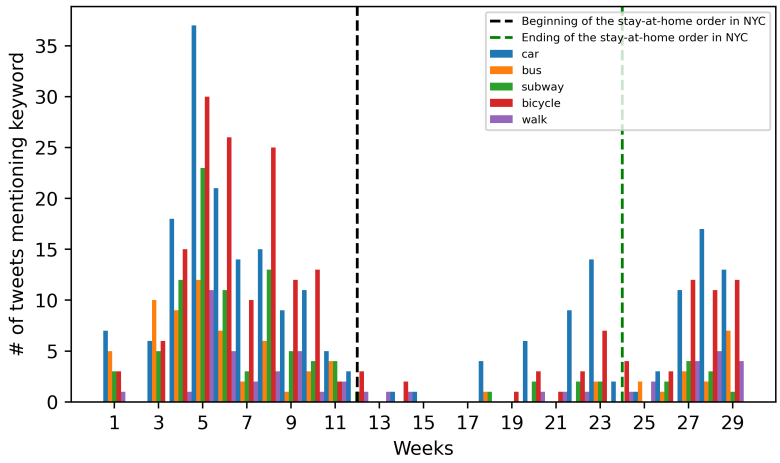


Figure 2: Tweets mentions of five travel modes from a specific Twitter user

In the table "Twitter Network", each record represents a pair of users. The field "Weight" represents the strength of users' connection, which is determined by the field "Mentions" in the table "Tweets". The table on the side of "Twitter Network" illustrates one example of user pairs. If user 1 mentions user 2 (@ user 2) in one tweet and user 2 mentions user 1 in another tweet, then the weight is 2. If users do not mention each other in any tweet, the weight is 0, i.e., they are not connected with each other.

2.2 Spatial-Temporal Data

In social media, geo-tagged posts provide users' spatial-temporal (S-T) data by multi-day check-in records. In our spatial-temporal data, users' check-ins are collected from Twitter and Facebook based on regions, time intervals and keywords, which are consistent with those in the textual data. We fetch 28053 check-in records from 4732 users. For each check-in record, it is known when, where (latitude and longitude") and by whom the record was provided. Fig. 3 plots part of S-T records of an individual user.

Figure 3: S-T records of a specific Twitter user

2.3 Pattern Analysis

To give readers a preliminary idea of how Tweets mention and S-T records evolve across three phases, in this subsection, we will present their aggregate patterns and how real-world traffic patterns are correlated with Tweets mention.

2.3.1 Tweets mentions

To study how users' tweets mentions change across three phases, we use Fig. 4 to illustrate the bar chart for aggregated tweets mentions regarding different topics of all 5694 users in three phases. The x-axis represents the week-base timeline and the y-axis represents the number of tweets regarding some topic. The black dashed line stands for the time when the stay-home order starts and the green dashed line represents the time when the reopening starts.

It is shown that the number of tweets about car, walk and bike (represented by blue, purple and red bars) keeps increasing while the number of tweets about subway and bus (represented by green and orange bars) decreases in Phase 2. It means that users talk more about traffic modes which can prevent people from being infected with coronavirus after the stay-home order. In addition, the number of tweets about traffic modes has a significant increase in Phase 3. This is probably because in Phase 3, many people who get back to work are worried about transportation services and they ask for opinions about which traffic modes are safer.

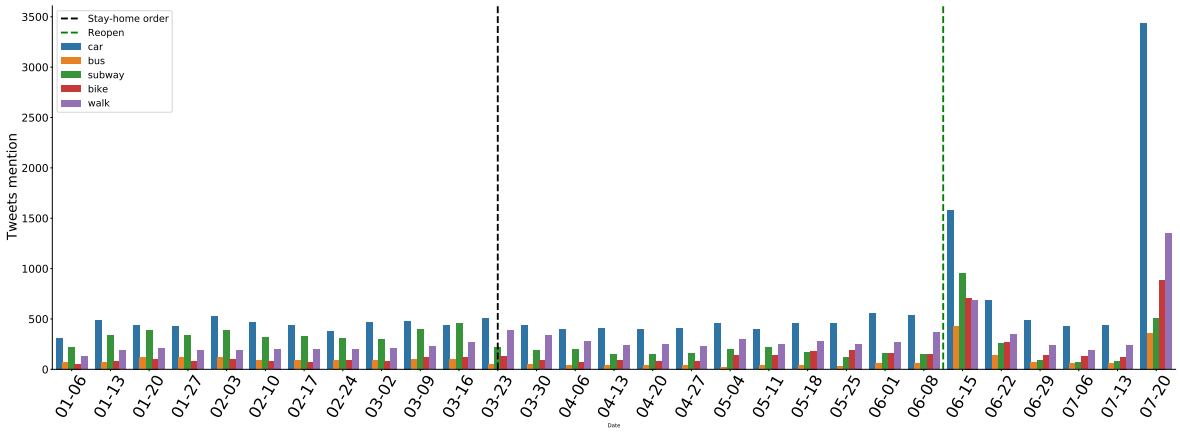


Figure 4: Tweets mentions of five travel modes across three phases

2.3.2 S-T records

We filter S-T data by considering users who have at least one S-T record in a phase. There are 6788 S-T records among 552 users, 9239 S-T records among 549 users, and 4973 S-T records among 598 users in Phase 1, Phase 2 and Phase 3, respectively.

Based on S-T records, we investigate users' gyrations. Gyration is a measurement of travel distance, which is calculated as

$$g_i = \sqrt{\frac{\sum_{l=1}^{k}(r_i^l - \bar{r_i})^2}{k}}, \bar{r_i} = \frac{\sum_{l=1}^{k} r_i^l}{k}, i = 1, ..., N \tag{1}$$

where g_i is the gyration of user i, $r_i^l, l = 1, ..., k$ is the lth coordinates in user i's S-T records, k is the number of user i's S-T records and $\bar{r_i}$ is the centroid of user i's coordinates.

To study how users' gyrations change across three phases, we use Fig. 5 to illustrate the gyration performance. The left y-axis in Fig. 5 represents users' gyration in average. It is shown that the gyrations in Phase 2 and 3 are smaller than that in Phase 1. The right y-axis represents the proportion of users who has only one coordinate in S-T records, i.e., gyration is 0. The blue bars show that the proportion of users who do not travel around increases in Phase 2. The explanation is that after the stay-home order, people are not willing to travel outside due to the fear of being infected with coronavirus.

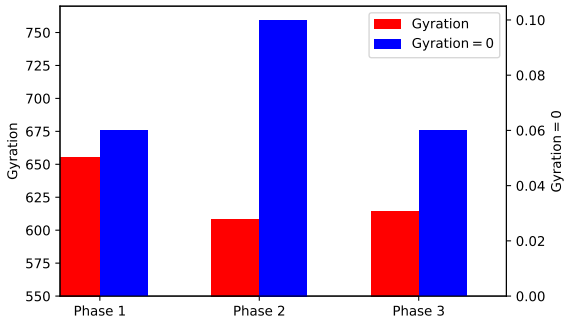
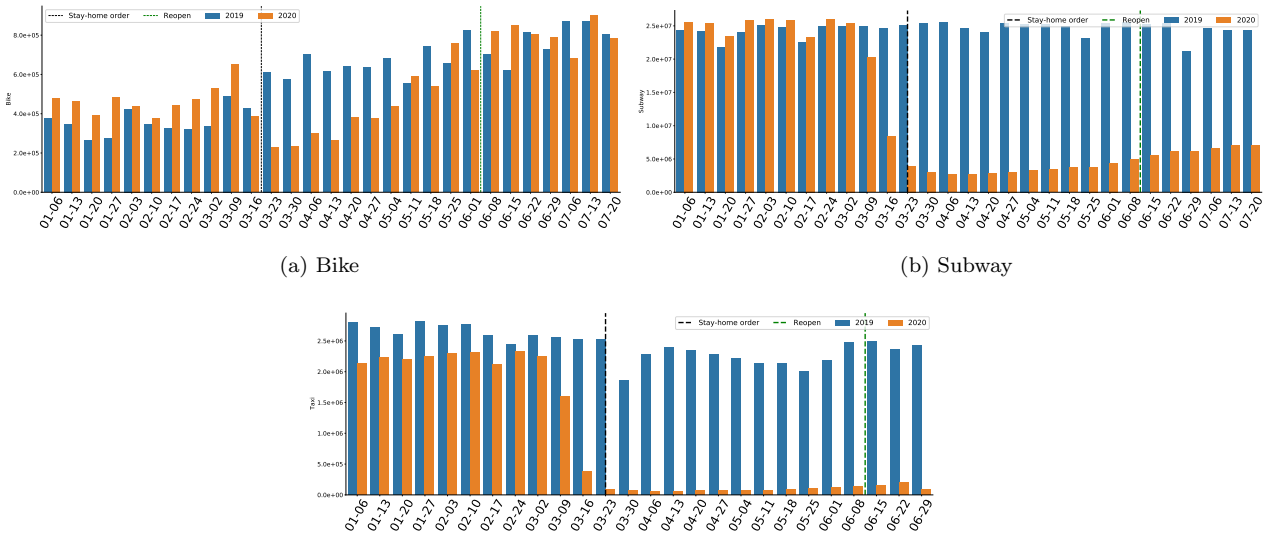


Figure 5: Gyration performance

2.3.3 From social media to the real-world

Here we would like to demonstrate how the distribution of Tweets mentions of five travel modes is related to traffic patterns of these travel modes in the real-world.

Before we present the correlation between Tweets mentions and traffic counts computed from NYC Open data, we will first give readers a comparison of counts of three travel modes in 2020 and 2019 equivalent week. Fig. 6 plots the bar charts for NYC open data of bike, subway and taxi in 2019 and 2020. The blue and orange bars represent 2019 and 2020, respectively. Fig. 6(a) plots bike trips from Citibike and bike count data of DOT. The y-axis represents the number of bike trips. Fig. 6(b) visualizes subway turnstile data from MTA. The y-axis represents the number of exits in subway stations. Fig. 6(c) is about NYC green and yellow taxi data. It is shown that compared with the number of trips in 2019, the number of trips of all three travel modes have a significant decrease after the stay-home order in 2020. The explanation is that people have a fear of being infected with coronavirus when travelling around.

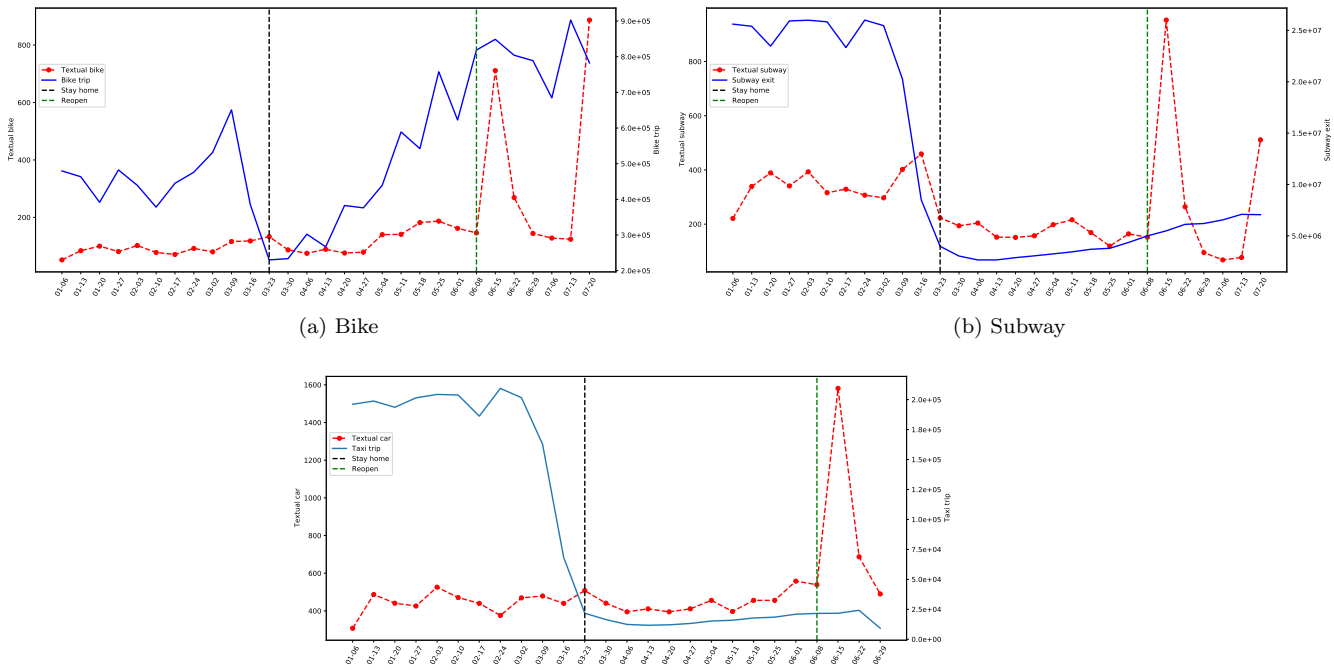


(a) Bike

(b) Subway

(c) Taxi

Figure 6: NYC open data: 2019 v.s. 2020



(a) Bike

(b) Subway

(c) Car

Figure 7: Tweets mentions v.s. NYC open data

We then make a comparison of tweets mentions and NYC open data in Fig. 7. The red dashed lines represent tweets mentions regarding topics “Bike”, “Subway” and “Car”. The blue lines represent the

number of trips related with bike, subway and taxi in NYC open data. Fig. 7(a) shows that the number of trips related with bike first decreases at the beginning of Phase 2 and then has an increasing trend while tweets mention about bike has a significant increase at the beginning of Phase 3. The explanation is that at the beginning of Phase 2, people are afraid of travelling outside, leading to the decrease of bike usage. At the beginning of Phase 3, many people who need to get back to work discuss a lot about which travel modes are better in the pandemic, making the topic "Bike" hot debated. Similarly, the significant increase of tweets mentions about subway and car at the beginning of Phase 3 shows that topics "Subway" and "Car" are hot debated. However, the number of trips related with subway and taxi remains at a low level after the stay-home order. This is because people have a fear of travel modes which cannot prevent them from being infected with coronavirus. The more they discuss about subway and taxi in tweets, the more they are worried about these transportation services.

3 Similarity Analysis

In this section, we use similarity analysis based on social media data to investigate the extent to which users are alike (similarity score), and how users' similarity scores change from Phase 1 to Phase 3.

Textual and S-T similarity analysis are conducted on textual and S-T data in Section 2, respectively.

3.1 Textual Similarity

Textual similarity measures the extent to which users are alike based on their multi-day tweets. In this section, we propose a textual similarity metric and use the metric to analyze how users' content mentions change regarding to travel modes in the pandemic.

3.1.1 Similarity Metric

We first introduce a matrix representation of users' textual data. One user i, where $i \in \{1, ..., N\}$ is the index of the user in all N users, has a textual record whose elements are $w^i(m,t)$ where $m \in \{$bus, bike, car, subway, walk$\}$ and $t \in T$. $w^i(m,t)$ represents the "topic mention" about travel mode m at time t. We divide the temporal range T into time windows $T_1, T_2, ..., T_l$ based on some timestamp (simplified to weekly basis or other levels). We can reformulate user i's textual data in the table (Table 1):

	car	bus	subway	bike	walk
T_1	$\sum_{t \in T_1} w^i(car, t)$	...			
T_2	...	...			
...					
T_l					$\sum_{t \in T_l} w^i(walk, t)$

Table 1: Textual data reformulation

Based on Table 1, we can calculate a density distribution of topic mention with respect to topics and

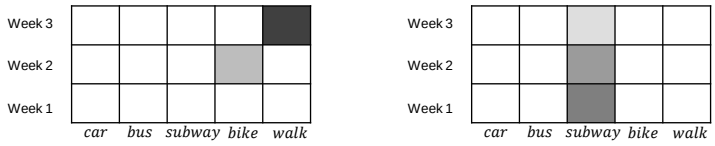
time zones and formulate the matrix representation of user i's textual data as follows:

$$H^i = \begin{bmatrix} q^i_{11} & q^i_{12} & q^i_{13} & q^i_{14} & q^i_{15} \\ q^i_{21} & q^i_{22} & q^i_{23} & q^i_{24} & q^i_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q^i_{l1} & q^i_{l2} & q^i_{l3} & q^i_{l4} & q^i_{l5} \end{bmatrix}$$

with each entry as:

$$q^i_{km} = \frac{\sum_{t \in T_k} w^i(m,t)}{\sum_m \sum_k \sum_{t \in T_k} w^i(m,t)},\ k = 1, ..., l, m = 1, ..., 5,$$

where $\sum_m \sum_k \sum_{t \in T_k} w^i(m,t)$ denotes the total number of tweets mention of the user. Fig. 8 illustrates part of two users' textual matrix. Nonzero entries are colored, and a darker color indicates a higher q^i_{km} value.



(a) Part of user i's textual matrix

(b) Part of user j's textual matrix

Figure 8: Part of two users' textual matrix

Characterizing a user's textual data by a matrix representation, which preserves both the textual and temporal information, we aim to define an appropriate similarity measure which is specially designed for the matrix representation. The textual similarity (similarity score) between user i and j is formulated as:

$$S^{(C)}_{ij} = \sum_{k=1}^{l} \sum_m \sqrt{q^i_{km} \cdot q^j_{km}},\ i, j = 1, 2, ..., N, m \in \{\text{car, bus, subway, bike, walk}\} \tag{2}$$

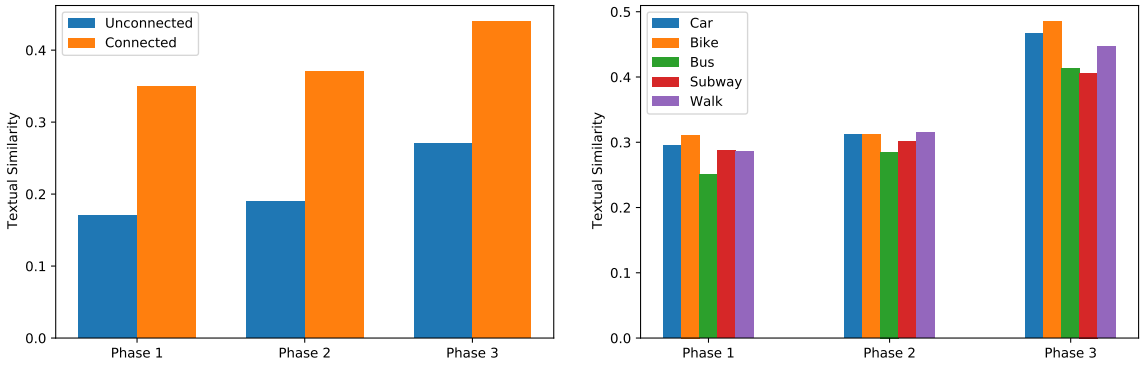
The textual similarity metric $S^{(C)}_{ij}$ satisfies the following properties: (1) $0 \leqslant S^{(C)}_{ij} \leqslant 1$. (2) $S^{(C)}_{ij} = 1$ if and only if user i and j have the same textual matrix. (3) $S^{(C)}_{ij} = 0$ if and only if the Hadamard product of user i's and j's textual matrices is 0. (4) $S^{(C)}_{ij}$ is symmetric, i.e., $S^{(C)}_{ij} = S^{(C)}_{ji}$.

3.1.2 Phase-level Analysis

In this subsection, we investigate the performance of textual similarity in three phases. Note that there are many users who may tweet nothing in some phase, which should not be considered in textual similarity analysis. Therefore, we filter the data sample by considering users who have more than 3 tweets in a phase. There are 407, 323 and 317 users whose tweets mentions are larger than 3 in Phase 1, Phase 2 and Phase 3, respectively.

Fig. 9(a) visualizes the textual similarity for connected and unconnected users in three phases. User connections are determined by the field "Mentions" (@ someone) in Section 2. Two users are connected if at least one of them mentions the other one in tweets, i.e., the weight in Fig. 1 is larger than 0. The blue and orange bars in Fig. 9(a) represents the textual similarity in average of unconnected and connected

user pairs, respectively. From Phase 1 to 3, the textual similarities of unconnected and connected user pairs both increases. It is shown that the textual similarity of connected user pairs is always larger than that of unconnected user pairs, meaning that users are more likely to share the same topics with each other if they are in connection. Fig. 9(b) plots the textual similarity regarding different topics in three phases. It is shown that from Phase 1 to 3, the textual similarities regarding five topics all increases. The intuitive explanation is that after the stay-home order, topics related with different travel modes are hot debated.



(a) Textual similarity for connected and unconnected users

(b) Textual similarity for different topics

Figure 9: Textual similarity performance

3.2 Spatial-Temporal Similarity

Spatial-temporal similarity measures the extent to which users are alike based on their multi-day check-ins. In this section, we use spatial-temporal similarity to analyze how users' travel pattern changes in the pandemic.

3.2.1 Similarity Metric

We first introduce S-T similarity defined in (25) based on S-T data. User i ($i \in \{1, ..., N\}$) has an S-T record set whose elements are coordinates (r_l^i, t_l^i) where $r_l^i \in \mathcal{Z}$ and $t_l^i \in \mathcal{T}$. r_l^i represents the latitude and longitude of check-ins. t_l^i represents the corresponding timestamp and l means that (r_l^i, t_l^i) is the l^{th} S-T record of user i. Zone Z is divided into a zone set: $\mathcal{Z} = \{z_1, z_2, ..., z_Z\}$. The temporal range is divided into a time interval set: $\mathcal{T} = \{\tau_1, \tau_2, ..., \tau_T\}$. Therefore, the density distribution of user i's check-ins can be formulated as an S-T matrix:

$$ST^i = \begin{bmatrix} p_{11}^i & p_{12}^i & \cdots & p_{1Z}^i \\ p_{21}^i & p_{22}^i & \cdots & p_{2Z}^i \\ \vdots & \vdots & \vdots & \vdots \\ p_{T1}^i & p_{T2}^i & \cdots & p_{TZ}^i \end{bmatrix}$$

with each entry as

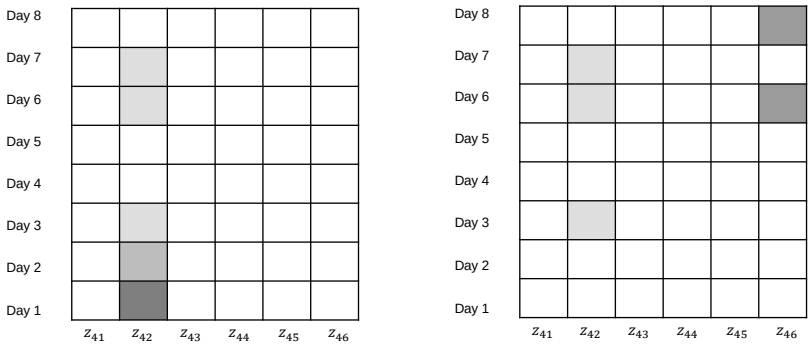
$$p_{km}^i = \frac{\sum_{l=1}^{R_i} \mathbf{1}_{r_l^i \in z_k} \times \mathbf{1}_{t_l^i \in \tau_m}}{R_i} \tag{3}$$

where R_i denotes the total number of user i's S-T records. p^i_{km} is numerically equivalent to the proportion of user i's S-T records falling into zone z_k during time interval τ_m.

S-T similarity (similarity score) between user i and j is formulated as:

$$S^{(G)}_{ij} = \sum_{k=1}^{T} \sum_{m=1}^{Z} \sqrt{p^i_{km} \cdot p^j_{km}},\ i, j = 1, 2, ..., N. \tag{4}$$

To encode spatial-temporal data into S-T matrix, we first divide NYC into 100 zones. The latitude range is $[39.95, 41, 50]$ and the longitude range is $[-74.86, -73.07]$. We evenly split the latitude and longitude range into 10 intervals, formulating 100 zones, which are encodes as: $z_{11}, z_{12}, ...z_{nn}$ where $n = 10$. Time intervals are arranged in a sequence and each interval represents one day. Fig. 10 illustrates part of two users' S-T matrix after encoding spatial-temporal data. Nonzero entries are colored, and a darker color indicates a higher p^i_{km} value.



(a) Part of user i's S-T matrix (b) Part of user j's S-T matrix

Figure 10: Part of two users' S-T matrix

3.2.2 Phase-level Analysis

In this subsection, we study the performance of S-T similarity and make a comparison of S-T and textual similarity. Fig. 11 shows phase-level analysis on S-T data. S-T similarities in three phases are represented by the blue bars. From Phase 1 to 3, S-T similarity of users keeps decreasing. This is probably because many users cancel their travelling plans with other people after the stay-home order. In addition, the decreasing trend of S-T similarity is opposite to the increasing trend of textual similarity in Fig. 9. The intuitive explanation is that after the stay-home order, people are more active in social media than in the real world.

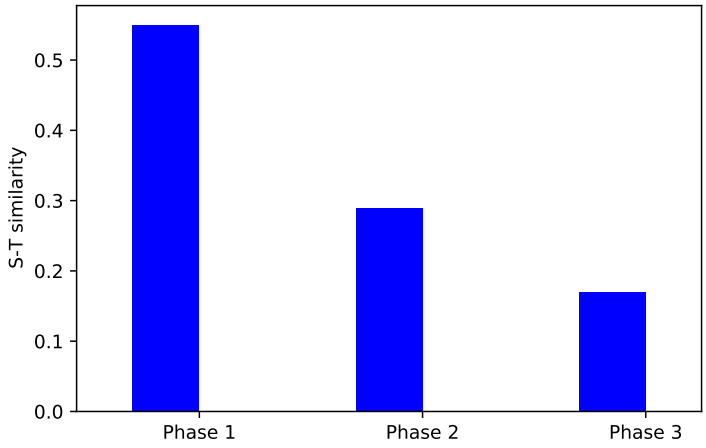


Figure 11: S-T similarity

4 Network Analysis

In this section, we first use Twitter network data in the table "Twitter Network" (Fig. 1) to construct a user network. Based on the user network, we apply a network effect model to investigate how users' tweets mentions of travel modes influence one another. The goals of the network effect model are to understand:

1. In which phase tweets mentions of individual users are more easily affected by other users?
2. Under which topic of travel modes individual users are more easily affected by other users?

4.1 User Network

A user network is constructed in the following way: the vertices of the network represent all users in Twitter network. Two users are connected by an edge if at least one of them mentions the other one in tweets, i.e., the weight in the table "Twitter Network" (Fig. 1) is larger than 0. The weight of the edge is the number of times that they are mentioned in tweets (Fig. 1). Fig. 12 plots the cumulative degree distribution of the user network on a log–log scale. The orange dots represent the cumulative degree distribution of the user network, which coincides with the blue dashed line, representing a power-law distribution.

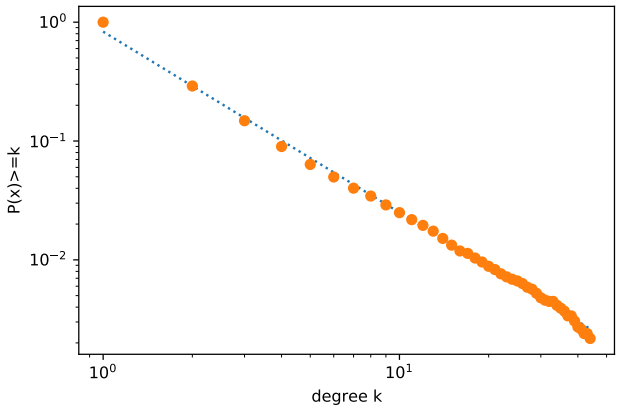


Figure 12: Cumulative degree distribution of the user network (log-log scale)

4.2 Network Effect Model

The network effect model is a linear network regression model (28), capturing the relationship between individual users and the whole user network. In the user network constructed in Section 4.1, tweets mentions of individual users are denoted as $Y = [y_1, y_2, ..., y_N]^T$ where $y_i,\ i = 1, ..., N$ is user i's tweets mention, i.e., the number of tweets user i has posted. The network regression model is defined as:

$$Y = \rho(A_{N\times N} \cdot Y) + \epsilon \tag{5}$$

where,

ρ: network effect parameter, which represents the influence of the user network.

ϵ: error term $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_N]$ where ϵ_i $(i = 1, ..., N)$ are i.i.d. and normally distributed.

$A_{N\times N}$: weighted adjacency matrix of the user network, which is illustrated in Fig. 13. According to the user network in Section 4.1, the entry w_{ij} in the adjacency matrix $A_{N\times N}$ is the weight of the edge between user i and j. We assume user i is not connected with herself, i.e., $w_{ii} = 0$. For user i, we then have

$$y_i = \rho(\sum_{j=1}^{N} w_{ij} y_j) + \epsilon_i \tag{6}$$

where $\sum_{j=1}^{N} w_{ij} y_j$ is the weighted sum of tweets mentions of users who are connected with user i.

Remark. 1. Based on Equation 6, $A_{N\times N} \cdot Y$ in Equation (5) can be reformulated as:

$$A_{N\times N} \cdot Y = [\sum_{j=1}^{N} w_{1j} y_j,\ ...,\ \sum_{j=1}^{N} w_{Nj} y_j]^T. \tag{7}$$

Accordingly, the linear network regression model can be reformulated as an ordinary linear regression model:

$$Y = \rho X + \epsilon \tag{8}$$

where $X = [x_1, x_2, ..., x_N]^T$ and $x_i = \sum_{j=1}^{N} w_{ij} y_j,\ i = 1, ..., N$. In other words, the explanatory variable in the regression model is the weighted aggregation of tweets mentions of users who are connected with each individual user in the network.

2. The term $\rho(A_{N\times N} \cdot Y)$ in Equation (5) is also called "contagion effect" in economy, which can be interpreted as a phenomenon that tweets spread out and affect users in the network through their connections with each other.

3. When $w_{ik} = 0$ for some k in the adjacency matrix, we have $w_{ik}y_k = 0$ in the term $\sum_{j=1}^{N} w_{ij}y_j$, implying that the tweets mention of user k cannot affect user i if they are not connected with each other. In other words, users can influence each other only when they are connected in the network.

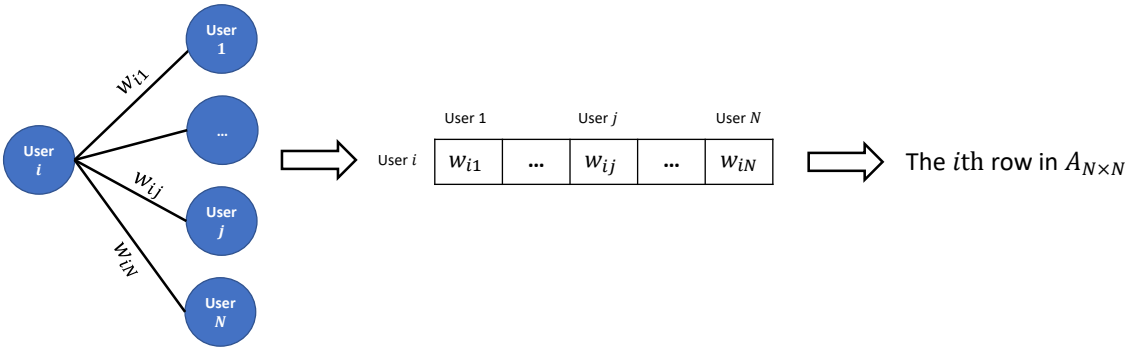


Figure 13: The ith row in $A_{N\times N}$

4.3 Network Effect Results

In this subsection, we apply the network effect model to analyze how users influence each other when phases and topics vary. The data to fit the network effect model comes from the user network and users' tweets mentions of three phases.

The network effect results are displayed in Table 2, where ρ, R^2 and p-value of the network regression model are shown. The coefficient ρ represents the strength of the impact of the network on individual users' tweets mentions. R^2 measures the goodness-of-fit of the regression model. The p-value measures the statistical significance of the model. Each column in Table 2 illustrates the network effect results regarding different topics of users' tweets mentions in three phases.

For a phase-level analysis, we study the network effect results in three phases under the same topic. From Phase 1 to 3, the column "Bike" shows that the network effect ρ first decreases in Phase 2 and then increases in Phase 3. It means that users in the network are less easily affected by tweets regarding the topic "Bike" in Phase 2. Similarly, the column "Bus" illustrates that compared with Phase 1 and 3, users are less easily affected by tweets regarding the topic "Bus" in Phase 2. The column "Car" shows that the network effect ρ keeps decreasing. It means that after the stay-home order, users are less easily affected by tweets about car. Similarly, users are less easily affected by tweets about walk after the stay-home order. However, the column "Subway" shows that from Phase 1 to 3, the network effect ρ keeps increasing, meaning that users are more easily affected by tweets about subway after the stay-home order. In summary, users are less easily affected by tweets about bike, car, walk and bus while they are more easily affected by tweets about subway in Phase 2. The intuitive explanation is that after the stay-home order, the fear of subway that cannot prevent people from being infected with coronavirus spreads out among users.

For a topic-level analysis, we study the network effect results regarding different topics within the same phase. In Phase 1 and 2, the network effects ρ regarding "Bike", "Car" are larger than those regarding

"Walk", "Bus" and "Subway" (around 0.2). It means that compared with topics "Walk", "Bus" and "Subway", users are more easily affected by tweets mentioning "Bike", "Car". In Phase 3, users are more easily affected by other users mentioning "Bike", "Bus" and "Subway" compared with "Car" and "Walk". In summary, users are more easily affected by tweets about bike compared with other travel modes no matter in which phase. This is probably because users tend to choose individual transportation like bike in the pandemic.

ρ_c (R^2)	Bike	Car	Walk	Bus	Subway	Aggregated
Phase 1	0.57 (0.32)***	0.40 (0.21)***	0.21 (0.06)***	0.25 (0.07)***	0.21 (0.06)***	0.43 (0.24)***
Phase 2	0.44 (0.22)***	0.33 (0.13)***	0.19 (0.04)***	0.17 (0.03)	0.25 (0.06)***	0.37 (0.17)***
Phase 3	0.49 (0.31)***	0.25 (0.10)***	0.14 (0.03)***	0.38 (0.19)***	0.30 (0.11)***	0.41 (0.22)***
$***: p < 0.001$	$**: p < 0.01$	$*: p < 0.05$				

Table 2: Network effect

5 Conclusions and Future Work

This paper aims to understand changes in people's travel behavior, mode choice in particular, through the lens of people's social media activities. To this end, we leverage two types of social media data, Twitter and Facebook, to extrapolate people's tweet mention of a travel mode along with geo-tagged information, as well as people's geo-locations when they check in using Facebook apps. Building on these data, similarity and network analysis are performed to understand how public opinions and information spread and influence people's opinion of travel mode selection, respectively. We have several major findings. (1) Users are more active in social media than in the real world after the stay-home order. (2) Compared with other travel modes, users are more easily affected by tweets about subway after the stay-home order due to the fear of travel modes which cannot prevent users from being infected with coronavirus. (3) In the reopening phase, topics about travel modes are hot debated among users who get back to work while the number of subway and taxi trips remains at a low level due to the fear of coronavirus.

This work can be extended in several ways. First, sentimental analysis of textual data will further help us identify the correlation between social media mentions and real-world traffic patterns. Second, we can identify official accounts that may be "influencers" on public opinions of travel mode selection.

Acknowledgements

This work is partially sponsored by Columbia's "Innovations for Urban Living in the Face of COVID". The second author would also like to thank Columbia's School of Engineering and Applied Science for organizing the Student Transit Design Challenge in Summer 2020, in which this research idea was nurtured.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Xuan Di, Xu Chen; analysis and interpretation of results: Xu Chen, Xuan Di; draft manuscript preparation: Xu Chen, Xuan Di. All authors reviewed the results and approved the final version of the manuscript.

References

[1] Johns Hopikins Unviersity of Medicine. Coronanirus resource center. https://coronavirus.jhu.edu/map.html, 2020. [Online; accessed 11.01.2020].

[2] David Goodman and Michael Rothfeld. 1 in 5 new yorkers may have had covid-19, antibody tests suggest. https://www.nytimes.com/2020/04/23/nyregion/coronavirus-antibodies-test-ny.html, 2020. [Online; accessed 11.01.2020].

[3] Christina Goldbaum and Winnie Hu. Why new york buses are on the rise in a subway city. https://www.nytimes.com/2020/07/06/nyregion/mta-buses-nyc-coronavirus.html, July 6, 2020. [Online; accessed 11.01.2020].

[4] MTA. Subway and bus ridership for 2019. https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2019, 2020. [Online; accessed 11.01.2020].

[5] Fox News. Coronavirus outbreak sees surge in reckless driving on america's empty roadways. https://www.fox29.com/news/coronavirus-outbreak-sees-surge-in-reckless-driving-on-americas-empty-roadways, Apr 19, 2020. [Online; accessed 11.01.2020].

[6] Winnie Hu. A surge in biking to avoid crowded trains in N.Y.C. https://www.nytimes.com/2020/03/14/nyregion/coronavirus-nyc-bike-commute.html, March 14, 2020. [Online; accessed 11.01.2020].

[7] Julianne Cuba. Nypd: Bike injuries are up 43 percent during coronavirus crisis. https://nyc.streetsblog.org/2020/03/25/experts-senate-bills-25b-for-transit-wont-be-enough/, Mar 19, 2020. [Online; accessed 11.01.2020].

[8] MTA. Day-by-day ridership numbers. https://new.mta.info/coronavirus/ridership, 2020. [Online; accessed 11.01.2020].

[9] Christina Goldbaum. Is the subway risky? it may be safer than you think. https://www.nytimes.com/2020/08/02/nyregion/nyc-subway-coronavirus-safety.html?referringSource=articleShare, Aug 2, 2020. [Online; accessed 08.03.2020].

[10] Janette Sadik-Khan and Seth Solomonow. Fear of public transit got ahead of the evidence. https://www.theatlantic.com/ideas/archive/2020/06/fear-transit-bad-cities/612979/, 2020. [Online; accessed 08.03.2020].

[11] CDC. Health equity considerations and racial and ethnic minority groups. https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Fracial-ethnic-minorities.html, July 24, 2020. [Online; accessed 11.01.2020].

[12] New York City Comptroller Scott M. Stringer. New york city's frontline workers. https://comptroller.nyc.gov/reports/new-york-citys-frontline-workers/, March 26, 2020. [Online; accessed 11.01.2020].

[13] Camille Kamga, Bahman Moghimi, Patricio Vicuna, Sandeep Mudigonda, and Rodrigue Tchamna. Mobility trends in new york city during covid-19 pandemic: Analyses of transportation modes throughout may 2020. *University Transportation Research Center*, 2020.

[14] Fan Zuo, Jingxing Wang, Jingqin Gao, Kaan Ozbay, Xuegang Jeff Ban, Yubin Shen, Hong Yang, and Shri Iyer. An interactive data visualization and analytics tool to evaluate mobility and sociability trends during covid-19. *arXiv preprint arXiv:2006.14882*, 2020.

[15] Suzana Duran Bernardes, Zilin Bian, Siva Sooryaa Muruga Thambiran, Jingqin Gao, Chaekuk Na, Fan Zuo, Nick Hudanich, Abhinav Bhattacharyya, Kaan Ozbay, Shri Iyer, et al. Nyc recovery at a glance: The rise of buses and micromobility. *arXiv preprint arXiv:2009.14019*, 2020.

[16] Ding Wang, Brian Yueshuai He, Jingqin Gao, Joseph YJ Chow, Kaan Ozbay, and Shri Iyer. Impact of covid-19 behavioral inertia on reopening strategies for new york city transit. *arXiv preprint arXiv:2006.13368*, 2020.

[17] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 344–353, 2013.

[18] Samiul Hasan and Satish V Ukkusuri. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 44:363–381, 2014.

[19] Xinhu Zheng, Wei Chen, Pu Wang, Dayong Shen, Songhang Chen, Xiao Wang, Qingpeng Zhang, and Liuqing Yang. Big data for social transportation. *IEEE Transactions on Intelligent Transportation Systems*, 17(3):620–630, 2015.

[20] Jae Hyun Lee, Adam W Davis, Seo Youn Yoon, and Konstadinos G Goulias. Activity space estimation with longitudinal observations of social media data. *Transportation*, 43(6):955–977, 2016.

[21] Zhenhua Zhang, Qing He, and Shanjiang Zhu. Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. *Transportation Research Part C: Emerging Technologies*, 85:396–414, 2017.

[22] Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, 68:285–299, 2016.

[23] Taha H Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S Waller. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 75:197–211, 2017.

[24] Arif Mohaimin Sadri, Samiul Hasan, Satish V Ukkusuri, and Manuel Cebrian. Exploring network properties of social media interactions and activities during hurricane sandy. *Transportation Research Interdisciplinary Perspectives*, 6:100143, 2020.

[25] Zhenyu Shou, Zhenhao Cao, and Xuan Di. Similarity analysis of spatial-temporal travel patterns for travel mode prediction using twitter data. *the 23rd IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2020.

[26] Adrian Wong, Serene Ho, Olusegun Olusanya, Marta Velia Antonini, and David Lyness. The use of social media and online communications in times of pandemic covid-19. *Journal of the Intensive Care Society*, page 1751143720966280, 2020.

[27] NYC. NYC OpenData. https://opendata.cityofnewyork.us/. [Online; accessed 11.01.2020].

[28] J. Neidhardt, Nataliia Rümmele, and H. Werthner. Predicting happiness: user interactions and sentiment analysis in an online travel forum. *Information Technology Tourism*, 17:101–119, 2017.