Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

SIMULATING SARS-CoV-2 TRANSMISSION IN THE NEW YORK SUBWAY

Alex Washburn

Computer Science
Hunter College CUNY
695 Park Ave, New York, NY 10065

Ye Paing

Computer Science
Hunter College CUNY
695 Park Ave, New York, NY 10065

Pauline Lin

School of Computing and Information Systems (CIS)
The University of Melbourne
Parkville 3010, AUSTRALIA

Felisa Vázquez-Abad

Computer Science, CUNY, New York, and
CIS, The University of Melbourne
695 Park Ave, New York, NY 10065

ABSTRACT

The impact of the spread of the virus on public transport has been a topic of disagreement. Some sources report that travel times are so small that contagion is insignificant, while others claim that the NYC subway was a major contributor to the spread of the virus in 2020. This study addresses this question. While there is an enormous amount of data, it is impossible to know when people got infected. Our approach is to use a model for the virus transmission to simulate contagion during travel, while using data from the turnstiles to make our simulation scenarios realistic. We combine the ghost model for train dynamics and a stopped Continuous Time Markov Chain (CTMC) for the virus transmission to create a hybrid simulation. Preliminary results indicate that our simulation tool may provide accurate answers to the question. In particular, it helps analyze the resulting risk under different transportation policies.

1 INTRODUCTION

The New York City subway is an essential service which transports up to five million people daily (MTA 2021), but during the COVID-19 pandemic it may be dangerous because it may be impossible to keep safe social distancing while on the train. Over 37% of New York's workers use the subway to commute to their workplace (Census 2019). Such a large number of passengers of the city's population may lead to very high infection rates as people are unable to properly socially distance within each subway car, yielding a high infection rate. During the COVID-19 pandemic the ridership of the New York City subway shrunk by more than 70% (MTA 2021) (Silverstone 2020); which improved the chances for each remaining commuter, but it did not completely negate the risk of infection. Under the COVID-19 pandemic, most of the riders of the subway are essential workers who have no alternative transportation to and from their workplace (Silverstone 2020).

The study (Harris 2020) concluded early in 2020 that the subway was a *major disseminator–if not the principal transmission vehicle–* of the coronavirus. That study considers the available data from turnstiles as well as reported data of COVID cases and draws conclusions based on data analytics. The author points out the complexity of the problem, mostly given the strong correlations between consecutive stations, and crossing lines. Some researchers and transit experts have, however, questioned those findings, mostly because the statistical methods used in that paper cannot, by themselves discover causation, only correlations (Meyer 2020).

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

Motivated by the polemic of whether significant contagion happens or not during public transportation, we propose a simulation tool that incorporates explicitly a transmission model for the virus. We show in this paper a case study that uses turnstile data to fit a model for the simulation. We use a discrete representation for the spacial train characteristics that considerably reduces the computational time and combine two simulation methodologies for increased efficiency. We believe that our simulation model provides insight on the impact of the subway on the virus transmission, as a function of passenger arrival rates and train frequencies. Using our tool it will be possible to explore ways to effectively reduce the risk of infection for the riders of the NYC subway.

Most epidemic simulations are animations based on differential equations (SIR-based models) and focus on the spread of virus over days and months. Many of the recent ones related to the COVID-19 pandemic are online publications that have not been peer-reviewed and do not describe the mathematical models behind the simulations (see, for example (Sanderson 2020a; Sanderson 2020b)). Simulation of the virus spread inside Diamond Princess ship (Fang, Zhiming and Huang, Zhongyi and Li, Xiaolian and Zhang, Jun and Lv, Wei and Zhuang, Lei and Xu, Xingpeng and Huang, Nan 2020) produces an animation of differential equations of the SIR-type in days. The similarity to our work is only in the focus of a population closely contained in space. Agent-based simulations of epidemics include mobility (Wilburn and Harris 2021; Chang, Harding, Zachreson, Cliff, and Prokopenko 2020; Ghosh and Bhattacharya 2021). Primarily used for qualitative analysis, those simulations model individual's health state and use the day/month time scales. None of the references that we have found (including posted articles that have not been peer-reviewed) includes a detailed stochastic model of the transmission of the virus, which we do here.

In this work we look at constrained spaces (train cars), for short periods of time (seconds and minutes) and we do not need to keep track of the different stages (incubation, having symptoms, recovering or dying from COVID -19): in our study we only need to consider contagious and non-contagious passengers. We propose a model for the *transmission* of the virus rather than the *spread* of the pandemic. To our knowledge this is the first stochastic model of the way in which the viral infection is transmitted upon close contact, and the time until contagion is specifically modeled, which is not the case with the other simulation models that we have found in the literature.

The ultimate goal is to find a strategy for public transportation when (i) essential workers need to use the system, and (ii) the corresponding demand is higher than the safety level. In particular, the number of passengers that currently ride the subway cannot be placed in the metro cars respecting the social distance recommended for safety. The controls must use a reduction of demand on each car such as increasing the frequency of service and policing to ensure the number of people entering the stations does not exceed a safety bound. In this paper we develop a simulation model that will support the optimization procedures, which are outside the scope of the current paper. However, we illustrate how our procedure may help optimization by comparing various strategies.

2 FORMULATION OF THE PROBLEM

2.1 SARS-CoV2

The coronavirus SARS-CoV2 that is currently affecting the world causes a disease called COVID-19. Following the onset of the virus in the Chinese province of Wuhan in November 2019, many scholars, epidemiologists and medical experts have studied and described the manner in which this virus affects the human respiratory system. For our purposes we mention here a few characteristics that will be part of our model. A person can be infected with the virus days before showing any symptoms, and during this incubation period the person can transmit the virus to others (Patrozou and Mermel 2009). In addition, this virus has an extremely long lifespan outside of a body, being able to remain active and infectious for up to 72 hours on plastic and steel surfaces (Doremalen et al. 2020). To account for these issues, we label each passenger as *contagious* or *non-contagious* rather than describing whether or not the person is infected, symptomatic or recovered. Simply touching another person may lead to infection from a person who is not infected, but contagious. Contagious people are very likely to become infected themselves, unless they follow very strict hygiene measures (Santarpia et al. 2020). Further differences from previous SARS viruses include an adhesion rate to cell membranes that is ten times higher, thus far fewer virus particles need to enter an organism in order to begin multiplying (Hoffman et al. 2020), so the infection happens as soon as viral particles enter an organism, which in turn becomes contagious. In this study we assume that people who show symptoms are not allowed to ride the subway, as it is now common practice to monitor access to closed spaces.

According to the CDC guidelines, “close contact” is defined as spending 15 minutes or more within 6 feet (≈ 1.8 m) of a contagious person (Chu et al. 2020). Viruses can “jump” from one person to another within this range (the actual distancing rules vary in different countries, ranging from 1.2 to 2 meters). The rate of transmission between humans has been estimated, yielding the 15 minute “rule” as a guide. Based on this we create a stochastic model for the time until contagion, as described later.

2.2 Risk Measure

Definition 1 The risk measure $R(y)$ is defined as the expected fraction of novel transmissions in one day as a function of the initial number of contagious passengers. Let N denote the number of passengers initially free of the virus, and Y the initial number of contagious passengers that board the subway on a given day. Call X the number of novel transmissions happening during the subway ride in that day. Then

$$R(y) = \mathbb{E}\left[\left(\frac{X}{N}\right) | Y = y\right]. \tag{1}$$

To illustrate the impact of contagion during the subway train rides, we now present a brief and simplified argument that motivates the use of the risk measure defined above.

Lemma 1 Suppose that the number of riders per day is constant. Let Y_n be the number of contagious passengers that board trains on day n. Then, assuming that individuals always come back on the following day, and that none of them stops using the subway, Y_n follows a (stochastic) geometric progression.

Proof. Denote by $r_n(y)$ the random fraction of novel transmissions on day n when there are y contagious passengers boarding that day. Specifically $r_n(y) = X_n/N_n$, where the index n denotes the day. Then we obtain the following difference equations:

$$\begin{aligned} X_n &= r_n(Y_n)N_n \\ N_{n+1} &= N_n - X_n = N_n(1 - r_n(Y_n)) \\ Y_{n+1} &= Y_n + X_n. \end{aligned}$$

Expressing Y_n as a function of the initial number N_0 of non-contagious passengers, we obtain

$$Y_{n+1} = Y_0 + \left(\sum_{k=0}^{n} r_k(Y_k) \prod_{\ell=0}^{k-1} (1 - r_\ell(Y_\ell))\right) N_0.$$

From the assumptions, it follows that $\{Y_n\}$ is non decreasing w.p.1. If the rates of novel contagion were equal to a constant α, then this expression would correspond to the usual geometric progression and $Y_n = Y_0 + N_0(1-(1-\alpha)^{n-1})$, which naturally converges to $Y_0 + N_0$, which is the total number of passengers (everybody becomes infected). □

The above simplified model does not consider the fact that when a person is contagious, they may or may not develop the disease, and when they do then they stop using the subway. Under this more realistic scenario we would add a stochastic element that makes Y_n possibly decrease. However, we remark that symptoms are usually not shown within a fortnight, during which the contagious passengers keep riding the subway. Therefore it is important to assess the value of the risk measure $R(y)$ in order to further study the impact of the subway transmissions in the spread of the virus.

There are many reasons why simulation is justified as the best method for this study. Mathematical models (SIR epidemiological models, Markov chain models, data analysis, etc) usually require to make simplifying assumptions (as we have done for the Lemma 1), or do not explicitly model causality. Our approach will consider specifically the geometry of the subway cars and proximity of passengers. In addition, looking at a single day we will account for the fact that most riders that go from an origin to a destination in the morning, will then ride the subway the reverse way in the end of the day. That is, if a passenger becomes one of the novel infected passengers in the morning, then he/she will add to the initial number of contagious passengers later in the day. We propose to use only one day to better extract the effect of the risk $R(y)$. Our simulation tool can then be used to explore various policies and scenarios.

3 DISCRETE GEOMETRY MODEL FOR CALCULATION EFFICIENCY

When calculating distances between humans, one must first determine the placement of the people and their diameter. Using cartesian coordinates for calculation of all the pair-wise distances entails evaluation of costly operations (such as the square root) in the order of $\mathscr{O}(P^2)$ where P is the number of passengers in each train car. Instead, we performed a reduction in computation time by discretizing the car geometry.

The average American body's cross-section was calculated from a 17 year study (Fryar et al. 2018). The average radius of New York City adults is 0.15 meters. Even on a packed subway car, passengers require a minimal amount of space surrounding them to allow themselves and others to maneuver in the subway car. Therefore we define a "human bubble" adding an extra 0.05 meters all around, as being the space effectively occupied by a person. The bubbles help us discretize the space in the subway cars in terms of "places" for passengers. To do so, we considered the dimensions of a subway car. Given the manufacturer's specifications of the R143 New Technology Train cars (Kawasaki Rail Car 2017), each subway car has a length of 18.35 meters and a width of 2.98 meters. To model the interior of the car with hexagons of 0.2 meter radius, we construct a 46×8 hexagonal grid, see Figure 1. With this construction, the hexagonal grid has a length of 18.4 meters and a width of 3 meters, which very closely models the actual train car dimension.

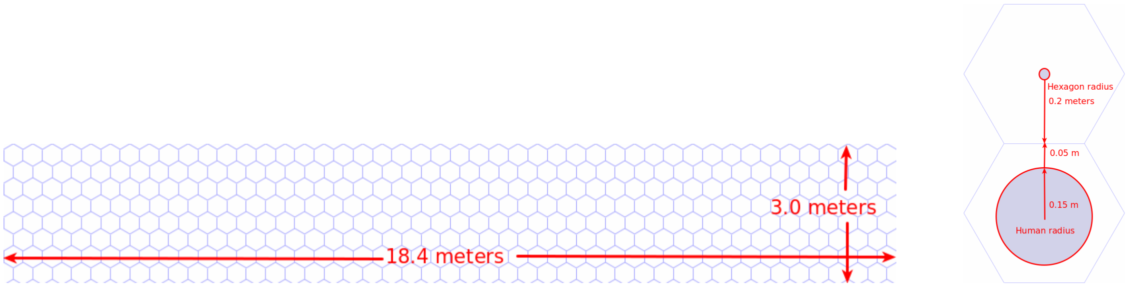


Figure 1: Subway car modeled as a honeycomb grid. Hexagons and "human bubbles" shown to the right.

With this geometry, our system of coordinates labels rows and columns as shown in Figure 2. With the grid system in place, we can now easily calculate if two passengers are socially distanced by simply counting the hexes in between the two passengers. We calculated that two passengers being 4 hexes apart (that is, 3 hexes between them) represents 1.3 meters of separation, which is not considered safe social distancing. However two passengers being 5 hexes apart, which represents about 1.7 meters of separation, is the minimum required amount of hexes needed between passengers to be considered safely socially distanced for our simulation.

By using a discrete space of the hexagonal grid, the pairwise distance computations are much faster than using real values and euclidean distance. The finite and small space of the hexagonal grid allowed for distances between all possible coordinates to be pre-computed, something which cannot be done when using real values. The real-valued euclidean distance computation has mean and median computation times of 3.29ns and 3.30ns, respectively. The hexagonal grid lookup table has mean and median computation times of 1.71ns and 1.70ns, respectively. On the 3Ghz processor on which the benchmarks were performed, this micro-optimization of using a hexagonal coordinate system represents an improvement from 10 clock cycles to 5 clock cycles per distance calculation.

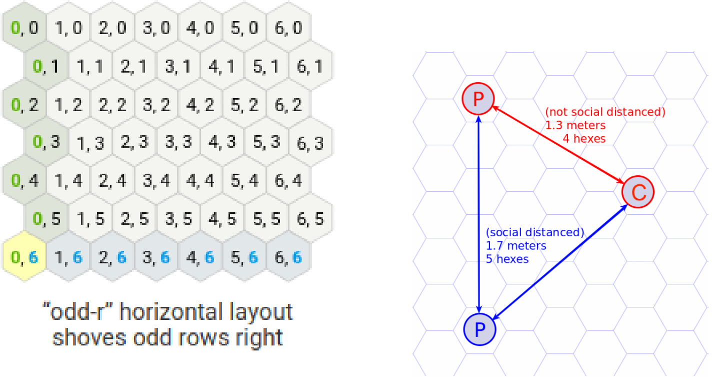


Figure 2: Honeycomb coordinates for human places in train cars and calculation of distances.

Under this model a train car can hold a maximum of 24 passengers while being safely socially distanced, see Figure 3. In this work we assume that the “safe” places are clearly indicated by a cross or other visible mark, so passengers know which are the recommended seats.

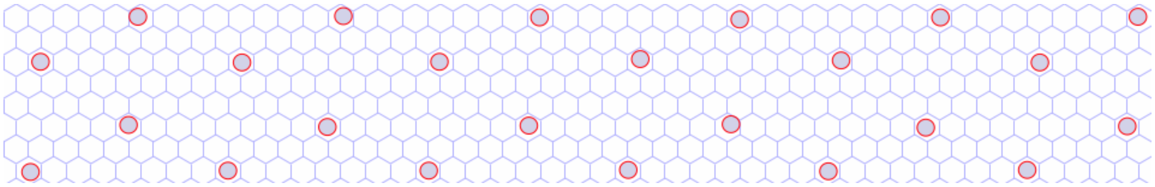


Figure 3: Social distance calculation and safe passenger layout

4 SIMULATION MODEL

4.1 Case Study Dataset

We use New York City’s L-train subway line as a “proof of concept” for the model. The L line was chosen for possessing a number of desirable characteristics. First, the L-line service is 24-hours. Second, the L-line is linear, meaning it does not bifurcate at either end to sending half the trains down one line, and half down another. The L-line does not “run express, ” which allows some trains to skip stations. Compared to subway lines of similar volume, the L-line does not have as many stations at which transfers can occur. These characteristics allow us to ignore different line configurations dependent on the time of day, which are present in other subway lines (for example some lines have different stops during rush hour) and minimize the impact of our simplifying assumptions described below for the estimation of demand rates. This choice is justified in order to focus on the development of our simulation tool and to better assess the impact of the subway travel in virus transmission, by isolating the virus spread effects in our simulation.

There are three approaches for analysis when using historical data. The first is to apply statistics and data analytics, which usually describe static phenomena. For our project, the available data is limited because it does not describe the transmission of the virus: it is not known when riders actually got infected, or how many passengers board the trains when they are in a contagious stage. Simulation models provide insight into the dynamics of the contagion, and there are two ways for using real data when simulating. The first is to implement a data-driven simulation, where the data is read from the database and directly fed into the simulation (train and passengers arrival and exit data). We use here the second way, which is to create a simulation model for passenger behavior and fit the hyper parameters of the model using the historical data. The main reason for this approach is that we wish to estimate the risk of contagion over one "random" (or typical) day. Use of a model allows us to generate identically distributed instances of such passenger processes in order to better estimate the risk in (1).

For our simulation the train dynamics is taken directly from the Metropolitan Transit Authority (MTA) timetable.While it is true that the actual travel time may be subject to randomness, this is usually due to either delays in passenger boarding and disembarking (which happens usually only when the system is over crowded) or due to shocks such as malfunctioning of trains or accidents. For our study we assume that these situations are not present, mainly because we wish to isolate the question of the transmission of the virus during "typical" train rides, and adding more randomness will not add much more insight. If anything, delays due to accidents or malfunctions will cause more novel transmissions because the exposure times are larger.

4.2 Poisson Processes

For our simulation we need to be able to generate passenger arrivals for each direction of Line L and, for each passenger, we need to generate his/her corresponding destination. The turnstile data available from the MTA website (NYC-MTA 2020) provides the actual counts of entries and exits to each station in the New York subway, where each day is divided into six, four-hour time slots beginning at midnight. While the turnstile data provides the counts of people entering and exiting the *stations*, there is no way that the system may know which lines people took at stations serving multiple lines. To account for this, we make a simplifying assumption that the number of passengers entering a station will be evenly divided between the number of accessible subway lines at that station and passengers don't transfer from other stations. We assume that the arrival processes are homogeneous Poisson processes on each time slot. We now explain how we estimate the origin-destination demand rate per time slots, from the available data.

Notation: In New York, some stations have separate entrances for each direction, so the turnstiles only serve a single platform. To account for this, we label the stations using the following convention. Stations are labeled in order from 8th Ave. Manhattan (station 0) to Canarsie-Rockaway Pkwy (last station). When a station has dedicated entrances per direction, we distinguish the entrances by giving them adjacent labels $k, k+1$, where the label k refers to the Manhattan-bound entrance and the label $k+1$ to the Brooklyn-bound entrance. Thus the total number of labels n is larger than the actual number of stations. Call $\mathscr{L} = \{0, 1, \ldots, n-1\}$ the total number of labels. Define the subsets $M, B \subset \mathscr{L}$ as the labels corresponding to exclusive entrances towards Manhattan (or Brooklyn, respectively).

For our estimates we downloaded the turnstile data from from NYC MTA's website for weekdays from Jan 2020 to Dec 2020 and made the average over approximately 260 days to obtain the average count per time slot. This number is further divided by 240 (because the time slots are 4 hours = 240 minutes long) to estimate $\lambda_i(t), t = 0, 1, 2, 3, 4, 5$ as the arrival rate per minute at station i on Line L during the four-hour time slot t. Similarly we calculate the corresponding exit counts per time slot $\mu_j(t), t = 0, \ldots, 5$.

Because we have only limited information, we assume that for a given origin station for Line L the distribution of the corresponding destinations is consistent with the exit counts of all other stations along the line L. This will include both directions on line L, so we produce a model for the probability of passengers' destinations, as follows. Fix the time slot t. For each station i on the line, call $d(i) \in \mathscr{L} \setminus i$ the possible destinations. For each station $i \in M$ the possible destinations are upstream, that is, $d(i) < i$ and we set

$$\mathbb{P}(d(i) = j) = P_{ij}(t) = \frac{\mu_j(t)}{\sum_{\ell=0}^{i-1} \mu_\ell(t)\, \mathbf{1}_{\{\ell \notin B\}}}; \quad j < i, j \notin B \tag{2}$$

The formula for stations $i \in B$ is obtained in a similar way (omitted here for space limitations). For all other labels $i \in \mathscr{L} \setminus (M \cup B)$ we define the destination distribution as:

$$\mathbb{P}(d(i) = j) = P_{ij}(t) = \begin{cases} \frac{\mu_j(t)}{\sum_{\ell \neq i} \mu_\ell(t)\, \mathbf{1}_{\{\ell \notin B\}}} & \text{if } j < i, j \notin B \\ \frac{\mu_j(t)}{\sum_{\ell \neq i} \mu_\ell(t)\, \mathbf{1}_{\{\ell \notin M\}}} & \text{if } j > i, j \notin M \end{cases} \tag{3}$$

The above uses the fact that for arrivals at station i that go towards Manhattan, their possible destinations satisfy $d(i) < i$ and $d(i)$ cannot be one of the Brooklyn-bound exits (and similarly for the opposite direction). For these stations, where the direction of travel is not specified from the data, (3) is consistent with the observed "popularity" of the exit stations and correctly assigns $\mathbb{P}(d(i) = i) = 0$, as we assume that all passengers travel to another station different from their origin.

4.3 Hybrid simulation model

4.3.1 Train Dynamics

Our simulation considers each direction of line L separately. We use a ghost model for simulation as explained in (Vázquez-Abad 2013) as follows. The code has two nested loops: the outer loop goes from one train to the next, the inner loop makes the current train consecutively visit each of the stations in order. At each station, passengers with this destination disembark and waiting passengers in the platform embark the train (see below). Once train n has reached the final station, the code must turn back the clock because train $n+1$ may have started its route before the previous train arrived at the last station. Timing coordination is done by keeping a clock at each station that records the time when the most recent train departed. Because trains do not overpass each other, then train $n+1$ always visits stations right after train n. Thus the train dynamics follow a tick-based simulation model with simple `For` loops.

On this part of the simulation we generate the passenger arrivals at the platforms. Rather than generating the Poisson processes of arrivals at each station, in order to increase the efficiency of the simulation we use *retrospective simulation* conditioning on the time elapsed since the last train departed, as was done in (Vázquez-Abad and Zubieta 2005) and (Vázquez-Abad 2013), generating a Poisson random variable to represent all the passengers that are waiting at the platforms. Specifically, if the time since the last train departure is denoted by ΔT then the number of passengers waiting at station i with destination j is a Poisson random variable with mean $\lambda_i(t) P_{ij}(t)\, \Delta T$, when the current time is in the time slot t, using (2) or (3), To increase efficiency, we used the binary search for the inverse function method for generating a Poisson random variable, as explained in Chapter 4.2 of (Ross 2006). In addition to their destinations, each passenger is represented in our code as having a "state" (1: contagious, or 0: non-contagious).

Embarking and disembarking passengers are coded following these rules:

- Disembarking passengers leave their places and the remaining passengers seek safe places.
- If the number of passengers waiting at the station is smaller than the available safe seats then all passengers use a safe seat (refer to Figure 3).

- The excess passengers are assigned random hexes in the train. Distance calculations determine which passengers are too close.

For each train car, once the passengers are seated we store a set of contagious passengers' coordinates and a set of non-contagious passengers' coordinates. We loop over the set of contagious passengers, and generate the set of coordinates 3 hexes from the contagious passenger. Note that this coordinate set is always the same size, so generating the coordinate set is $\mathscr{O}(36)$. We then perform a set intersection between the coordinate-range set and the non-contagious passenger set. If there are C contagious passengers and W non-contagious passengers, the runtime is $\mathscr{O}(C\min(36,W) * \log(\max(36,W)/\min(36,W)))$. In the worst case when there are more than 36 non-contagious passengers, this is of the order of $C * \log(W/36)$, which is asymptotically quite good for collecting all the exposed passengers. This operation allows us to keep another set V of "vulnerable" passengers per car: all those that are virus-free (with state = 0), but in close proximity to contagious passengers (with state = 1). Elements in this set are a pair of values: the honeycomb coordinates j and the number of contagious people that are within the 5 hexes, denoted $k(j)$.

4.3.2 Virus dynamics

The above description helps to code the train dynamics at consecutive stations using the ghost simulation model. During the travel time we simulate the viral activity and transmission inside the train.

We model the process as follows. Viral particles can jump from a contagious person to another human (via droplets, saliva sweat, etc) within one unit of time (say, a minute) with probability p. Therefore, the vulnerable person will be afflicted with the virus at a random time T that has the distribution of the first "success" in a Bernoulli trial, namely a geometric distribution with parameter p on $\{1,2,\ldots\}$. If time is discretized with finer grids, a similar argument leads to an approximation of the geometric random variable by an exponential with mean $m = (1-p)/p$ minutes. To our knowledge there has not been an empirical study to estimate m. People who have been in contact for 15 minutes or more and that have been tested do not always result positive, so the guidelines consider a conservative bound. In this study we propose to model the exposure time to contagion as an exponential random variable with mean $m = 15$. Suppose now that the vulnerable person is in close proximity of k contagious people (less than the recommended 6 feet). Then each of the sources of contagion will generate its viral jump process and the time to contagion will now be the minimum between the k independent exponentially distributed times, which has an exponential distribution with mean m/k.

The simulation of viral transmission during travel is done for each car independently and can be implemented using parallel cores for speed up. The model for the viral transmission corresponds to a stopped CTMC (continuous time Markov chain), as follows. Let T denote the time of arrival of the train at the following station. For each of the coordinates $j \in V$ we define the contagion rate as $k(j)/m$, where $k(j)$ is the number of viral sources in close range to j. Because the viral processes are assumed independent, then the next event that happens is either a contagion or arrival at a station. Call $K = \sum_{j\in V} k(j)$. At time t in the simulation we set $\delta t \sim \text{Exp}(v)$, where $v = K/m$. If $t + \delta t > T$ then the train arrives at the station at time T and the sets are not changed. Otherwise the coordinate i (passenger) that changes state is chosen with probability $k(i)/K$. Then the sets are updated: coordinate i is removed from V, and it is moved from the set of non-contagious to the set of contagious passengers, and a new evaluation of V is made. At this moment other passengers may be in close proximity (less than 5 hexes) of the newly infected person, so their coordinates must be moved to the set V. An illustration of transitive spread is provided in Figure 4.

Due to the memoryless property of Markov chains, it is not necessary to keep residual timers or to do a search over the list of possible events, as is the case with a discrete event model for simulation, so the stopped CTMC is more efficient. Updating the sets is also done using only the differences between the old sets and the new ones, in order to calculate the new set of vulnerable passengers.

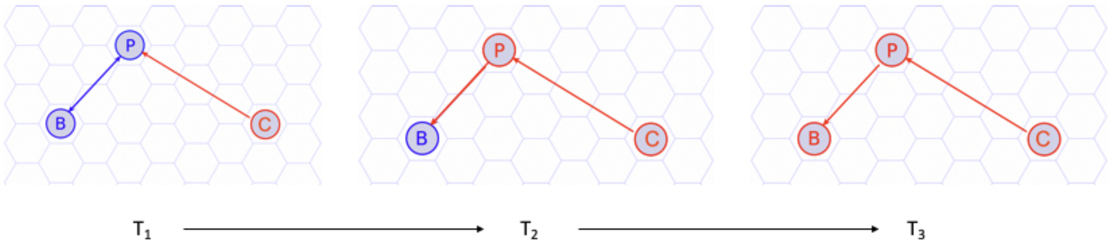


Figure 4: P is vulnerable (in the set V), while C is contagious. At time T_1 of boarding B is not close to a contagious person. At time T_2, P becomes contagious and B becomes vulnerable. B is infected at time T_3.

5 SIMULATION RESULTS AND DISCUSSION

The purpose of this section is to illustrate the potential uses of our simulation engine for decision making. This work has focused on the development of the simulation code and corresponding mathematical models. The results below underestimate the risk due to the following. (a) As explained before, this work focuses on only one line of the MTA subway network as a case study, (b) the estimation of arrival rates in this work has used simplifying assumptions that significantly underestimate the arrival rates at platforms on the Brooklyn-bound direction. Looking at Figure 5 it is evident that the Manhattan part of the line contains a significant number of crossing lines, particularly at Union Station. For this study we divided entry counts equally among lines, but we did not add any extra sources of arriving passengers from those other lines. We are currently studying methods for data handling and analysis that will aim at better estimation of the arrival rates. Most of the stations in Brooklyn are single line stations, and there are few crossing lines. Errors in estimation of platform arrival rates for the Manhattan bound are expected to be few and not too significant because few passengers arriving at Manhattan stations will be boarding on this direction. (c) Our model does not account specifically for indirect sources such as contact with surfaces or airborne transmission, which may be relevant inside the cars (Doremalen et al. 2020). Finally, (d) the effect of returning passengers is not included in our simulations (for lack of time and space), which is an additional source of underestimation of the effective platform arrival rates for the evening. The treatment of data and corresponding modeling is left for a separate contribution. We emphasize that the simulation engine that we have developed will require minor changes in parameters once the data analysis is completed.

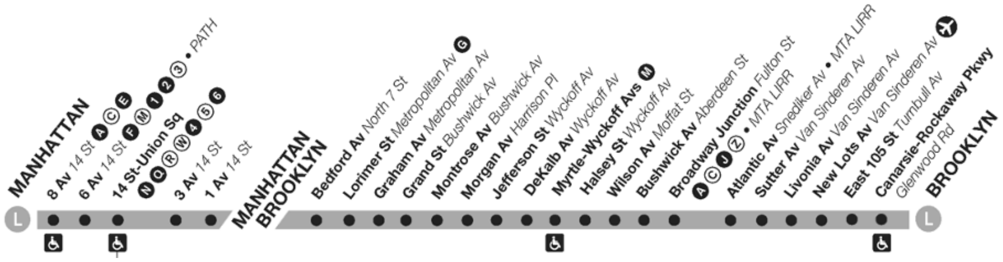
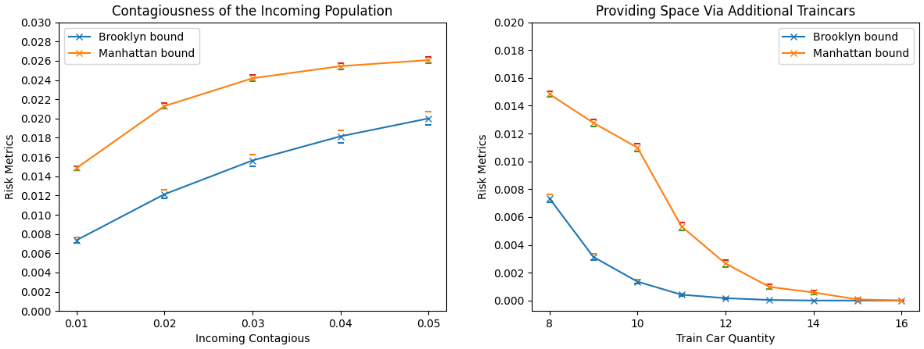


Figure 5: Line L.

Figure 6a shows the plots of the estimated risk measure as a function of the initial number of infected passengers. As expected, the risk is reported much higher in the M-bound direction. This plot shows the trend of the risk. Figure 6b shows the results of the estimation when we increase the number of cars on each train. Currently trains in the MTA subway have 8 cars. We have considered “what-if” scenarios where we assume that there are more cars on each train. It is apparent that these plots can provide a helpful visual aid to decision makers. For example, from our simulations it follows that the train capacity should be increased about 50% in order to significantly decrease the risk (naturally the plots will show different results for different lines and under different parameters). Notice that this increase in capacity may be infeasible. On-going research will result in future contributions where we will specifically focus on the use of our simulator for constrained optimization.

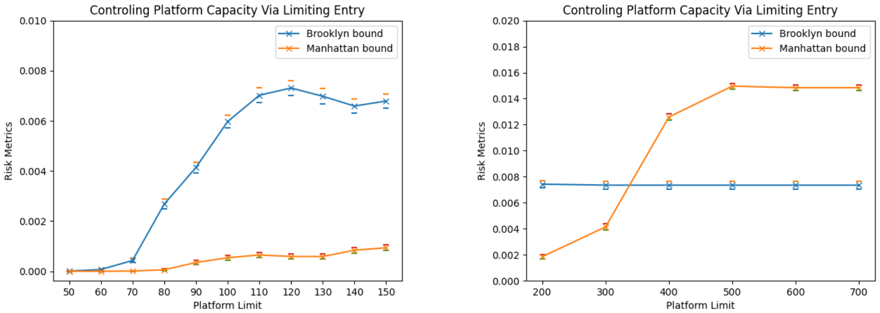


(a) Varying initial fraction of contagious passengers.

(b) Varying number of cars on each train.

Figure 6: Simulation results using 1000 replications of a day to estimate the risk measure for each scenario.

In order to explore the consequences of demand reduction, Figure 7 shows the “what-if” scenarios if we implement the following restriction: each platform will accept up to L passengers between trains. The plots clearly show the trends as functions of the number of passenger arrivals. Arrival control for a public transportation system entails complicated analysis, because decision makers must handle the actual demand using alternative ways. As mentioned before, optimization is the subject of a future contribution.



(a) Manhattan-bound direction.

(b) Brooklyn-bound direction.

Figure 7: Simulation results using 1000 replications of a day to estimate the risk measure limiting the number of arrivals at each platform.

6 CONCLUSIONS AND ON-GOING WORK

We have presented a simulation model that combines real data with a model for virus transmission (for which there is no available data) in order to estimate a risk measure that reflects the impact of virus spread on the subway. Our model can be adapted to other urban train networks. In order to create an efficient model we used a discrete geometry together with efficient data structures for placing and labeling passengers. The actual simulator is a hybrid simulator that uses the ghost model for the train dynamics and a stopped CTMC model for the virus spread. We are currently focusing on data processing, in order to better estimate the origin-destination rates of arrivals. For this, more complicated algorithms will be implemented that consider all stations and turnstile statistics. An important part of our current extension of the model is the "recursive" aspect of virus spread. Specifically, if we estimate that there is an increase of 6% passengers that are infected on line L, then this will have to be reflected on all other lines where these passengers may transfer to. Because this is information that cannot be measured in real life, we propose a bootstrapping method to re-simulate the various interacting lines. Our preliminary study shows that the virus spread inside the subway is a real concern and we believe that our simulator can be an important tool for analyzing scenarios. In a next stage we will look into the optimization aspect of the problem. For example, while our results indicate that limiting the number of passengers that enter the stations does have a significant decrease in transmission, we have not addressed the important question of what will the rejected passengers do? How will they get to their destinations? These are important questions that we will study using our simulator as a basic model for future research.

ACKNOWLEDGMENTS

This research was partially supported by the CUNY Institute for Computer Simulation, Stochastic Models and Optimization. Some of the original code was prepared by Hae Yang Lee.

REFERENCES

United States Census 2019. "U.S. Census Sex Of Workers By Means Of Transportation To Work For Workplace Geography". https://data.census.gov/cedsci/table?q=ACSDT1Y2017.B08406&g=1600000US3651000&tid=ACSDT1Y2017.B08406, accessed 25th June 2021.

Chang, S. L., N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko. 2020. "Modelling transmission and control of the COVID-19 pandemic in Australia". *Nature communications* 11(1):1–13.

Chu, D. K., E. A. Akl, S. Duda, K. Solo, S. Yaacoub, and H. J. Schnemann. 2020. "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis". *The Lancet* 395:1–15.

Doremalen, N. V., T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, and S. I. Gerber. 2020. "Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1". *New England Journal of Medicine* 382(16):1564–1567.

Fang, Zhiming and Huang, Zhongyi and Li, Xiaolian and Zhang, Jun and Lv, Wei and Zhuang, Lei and Xu, Xingpeng and Huang, Nan 2020. "How many infections of COVID-19 there will be in the "Diamond Princess"-Predicted by a virus transmission model based on the simulation of crowd flow". https://arxiv.org/abs/2002.10616 accessed 25th June 2021.

Fryar, C. D., D. Kruszon-Moran, Q. Gu, and C. L. Ogden. 2018. "Mean Body Weight, Height, Waist Circumference, and Body Mass Index Among Adults: United States, 1999-2000 Through 2015-2016". *PubMedl* 122:1–16.

Sayantari Ghosh and Saumik Bhattacharya 2021. "Computational Model on COVID-19 Pandemic Using Probabilistic Cellular Automata". Journal of Nature Public Health Emergency Collection https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8061453/ accessed 25th June 2021.

Jeffrey E. Harris April 19, 2020. "The Subways Seeded the Massive Coronavirus Epidemic in New York City, National Bureau of Economic Research Working Paper No. 27021". https://www.nber.org/papers/w27021 accessed 25th June 2021.

Hoffman, M., H. Kleine-Weber, S. Schroeder, N. Krager, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, and A. Nitsche. 2020. "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor". *Cell* 181(2):271–280.

Kawasaki Rail Car, Inc 2017. "R143 Train Car". https://www.kawasakirailcar.com/R143 accessed 25th June 2021.

David Meyer 2020. "MIT study: Subways a major disseminator of coronavirus in NYC". https://nypost.com/2020/04/15/mit-study-subways-a-major-disseminator-of-coronavirus-in-nyc/ accessed 25th June 2021.

MTA 2021. "MTA service during the coronavirus pandemic". https://new.mta.info/coronavirus/ridership accessed 25th June 2021.

NYC-MTA 2020. "Turnstile Data". http://web.mta.info/developers/turnstile.html accessed 25th June 2021.

Patrozou, E., and L. A. Mermel. 2009. "Does influenza transmission occur from asymptomatic infection or prior to symptom onset?". *Public Health Rep* 124(2):193–196.

Ross, S. M. 2006. *Simulation, Fourth Edition*. USA: Academic Press, Inc.

Sanderson, G. 2020a. "Exponential growth and epidemics". https://www.youtube.com/watch?v=Kas0tIxDvrg accessed 25th June 2021.

Sanderson, G. 2020b. "Simulating an epidemic". https://www.youtube.com/watch?v=gxAaO2rsdIs accessed 25th June 2021.

Santarpia, J. L., D. N. Rivera, V. L. Herrera, M. J. Morwitzer, H. M. Creager, G. W. Santarpia, K. K. Crown, D. M. Brett-Major, E. R. Schnaubelt, M. J. Broadhurst et al. 2020. “Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care”. *Scientific reports* 10(1):1–8.

Tom Silverstone April 10, 2020. “Risking coronavirus on the New York City subway: ’I feel guilty but have no choice”’. https://www.theguardian.com/world/video/2020/apr/10/new-york-city-subway-coronavirus-video accessed 25th June 2021.

Vázquez-Abad, F. 2013. “Ghost Simulation Model for Discrete Event Systems, an Application to a Local Bus Service”. In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Hill, S. H. Kim, M. Kuhl, R. Pasupathy, and A. Tolk, WSC ’13, 655–666. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Vázquez-Abad, F., and L. Zubieta. 2005. “Ghost simulation model for the optimisation of an urban subway system”. *DEDS Journal* 15:207–235.

Thomas Wilburn and Richard Harris 2021. “How Herd Immunity Worksâ And What Stands In Its Way”. https://www.npr.org/sections/health-shots/2021/02/18/967462483/how-herd-immunity-works-and-what-stands-in-its-way accessed 25th June 2021.

AUTHOR BIOGRAPHIES

ALEXANDER J. WASHBURN is a Masters student at Hunter College of the City University New York. He is interested in cryptography, high performance computing, and high assurance software design. He developed novel heuristics for both phyologentic network optimization and multiple string alignment as a Technical Lead at the American Museum of Natural History from 2015 to 2020, and mentored post-baccalaureate Helen Gurley Brown fellows. He is also an Adjunct Lecturer in the Computer Science department at Hunter College. He earned his B.S. with a double major in Computer Science and Mathematics from the University of Wisconsin – Milwaukee in 2014. His email address is academia@recursion.ninja.

YE W. PAING is a Masters student at Hunter College of the City University New York, currently finishing up his final project for the program, researching in mining software artifacts/repository comments and discussions. He is also working full-time as a Software Engineer at American Express, working primarily on consumer facing products. His interests in the Computer Science field includes working with big data, stream processing and data visualization techniques. Outside of his academic studies and his professional career, he is also currently doing volunteer work through a non-profit organization, teaching computer programming to high school students. He earned his B.A in Computer Science, with a minor in Mathematics, from Hunter College. His email addresses are ye.paing89@myhunter.cuny.edu, ye@y3p.io.

PAULINE LIN is a Lecturer in the School of Computing and Information Systems. Her interests include data linkage, data mining, graph analysis, anomaly detection and uncertainty modeling. Dr. Lin specialized in data structures and algorithms for pattern search during her PhD in Computer Science at RMIT University. She has over 10 years industry experience in financial intelligence, network analysis, anomaly detection, data cleaning and data linkage. She competed her PhD in Computer Science at RMIT University. Her email address is pauline.lin@unimelb.edu.au.

FELISA J. VÁZQUEZ-ABAD is Professor of Computer Science (CUNY) and Principal Investigator at the School of Computing and Information Systems (University of Melbourne). She is interested in the optimization and computer simulation of complex systems under uncertainty, primarily to build efficient self-regulated learning systems. She has applied novel techniques for simulation and optimization in telecommunications, transportation, medical ad biological models, finance and insurance and she is interested by real life problems. She co-authored a US patent for an optical network switch and was research consultant to the Melbourne Airport. She obtained a B.Sc. in Physics (1983) and a M.Sc. in Statistics and Operations Research (1984) from the Universidad Nacional Autónoma de México. In 1989 she obtained a Ph.D. in Applied Mathematics from Brown University. She was postdoctoral researcher at the INRS-Telecommunications in Montreal, Canada from 1990 to 1993. She was a professor at the University of Montreal (1993-2004), and then a professor at the University of Melbourne, until 2009 that she moved to New York to join the CUNY Faculty. In 2000, she was a recipient of the Jacob Wolfowitz award for advances in the mathematical and management sciences. She was the team leader for the finalist in the MEDSTART competition (2014). In 2017 she mentored the award-winning Hunter Hawks team in the CUNY-IBM Watson Competition. She has participated in Grant Selection Committees and has been Associate Editor for IEEE Transactions on Automatic Control, Management Science, and Operations Research Letters, Area Editor of the ACM Transactions on Computer Modeling and Simulation, and web editor of the INFORMS College on Simulation. She actively participates in events and programs to encourage women and minority students to succeed in Computer Science. Her email addresses are felisav@hunter.cuny.edu, felisav@unimelb.edu.au.