Data-Driven Mobility Models for COVID-19 Simulation

John Pesavento
Thomas Jefferson High School for Science and Technology
Alexandria, Virginia, USA
2021jpesaven@tjhsst.edu

Andy Chen
Thomas Jefferson High School for Science and Technology
Alexandria, Virginia, USA
2021achen1@tjhsst.edu

Rayan Yu
James Madison High School
Vienna, Virginia, USA
rayan.yu@gmail.com

Joon-Seok Kim
George Mason University
Fairfax, Virginia, USA
jkim258@gmu.edu

Hamdi Kavak
George Mason University
Fairfax, Virginia, USA
hkavak@gmu.edu

Taylor Anderson
George Mason University
Fairfax, Virginia, USA
tander6@gmu.edu

Andreas Züfle
George Mason University
Fairfax, Virginia, USA
azufle@gmu.edu

ABSTRACT

Agent-based models (ABM) play a prominent role in guiding critical decision-making and supporting the development of effective policies for better urban resilience and response to the COVID-19 pandemic. However, many ABMs lack realistic representations of human mobility, a key process that leads to physical interaction and subsequent spread of disease. Therefore, we propose the application of Latent Dirichlet Allocation (LDA), a topic modeling technique, to foot-traffic data to develop a realistic model of human mobility in an ABM that simulates the spread of COVID-19. In our novel approach, LDA treats POIs as "words" and agent home census block groups (CBGs) as "documents" to extract "topics" of POIs that frequently appear together in CBG visits. These topics allow us to simulate agent mobility based on the LDA topic distribution of their home CBG. We compare the LDA based mobility model with competitor approaches including a naive mobility model that assumes visits to POIs are random. We find that the naive mobility model is unable to facilitate the spread of COVID-19 at all. Using the LDA informed mobility model, we simulate the spread of COVID-19 and test the effect of changes to the number of topics, various parameters, and public health interventions. By examining the simulated number of cases over time, we find that the number of topics does indeed impact disease spread dynamics, but only in terms of the outbreak's timing. Further analysis of simulation results is needed to better understand the impact of topics on simulated COVID-19 spread. This study contributes to strengthening human mobility representations in ABMs of disease spread.

ARIC'20, November 3–6, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8165-9/20/11…$15.00
https://doi.org/10.1145/3423455.3430305

CCS CONCEPTS

• **Information systems → Geographic information systems**;

KEYWORDS

Latent Dirichlet Allocation Topic Modeling, Mobility Modeling, Agent-Based Modeling, Simulation, COVID-19, Policy Interventions

ACM Reference format:
John Pesavento, Andy Chen, Rayan Yu, Joon-Seok Kim, Hamdi Kavak, Taylor Anderson, and Andreas Züfle. 2020. Data-Driven Mobility Models for COVID-19 Simulation. In *Proceedings of 3rd ACM SIGSPATIAL Workshop on Advances in Resilient and Intelligent Cities, Seattle, WA, USA, November 3–6, 2020 (ARIC'20),* 10 pages.
https://doi.org/10.1145/3423455.3430305

1 INTRODUCTION

SARS-CoV-2 is a highly contagious human respiratory coronavirus resulting in mortality across the United States and worldwide [3]. Decision-makers rely on models to forecast disease dynamics and test the effectiveness of various policy interventions to strengthen preparedness, responsiveness, and urban resilience in the wake of the COVID-19 pandemic. Among the range of models used for this are agent-based models (ABMs), which employ a bottom-up approach to simulate the physical contact and transmission between individuals or "agents" from which COVID-19 spread dynamics emerge (i.e., number of infections and fatalities in a region).

ABMs in epidemiology expand upon traditional assumptions of compartmental SIR models [22] and its variations (e.g., SI, SIS, and SEIR [2]) to capture heterogeneity among the human population, including socio-demographic profiles, the spatial environment, and interaction probabilities [4]. However, many ABMs still lack realistic representations of human mobility, a dynamic process that plays a crucial role in physical interaction and subsequent transmission of disease [14].

Epidemiological ABMs that are spatially-explicit represent agent mobility using activity sequences where agents move between various points of interest (POIs) grouped by type (i.e., home, school,

work) [26, 27]. These activity sequences are informed by time-use and transport surveys or by patterns extracted from phone records [35]. POIs may be randomly generated across the study area [25] or are represented in the model using geospatial data, although the former is more common. An agent's workplace, school, or visit to public POIs such as restaurants or grocery stores are either selected at random [23, 24, 31, 32, 37] or selected by a function of the distance between the agent's home location and the POI [1, 16, 31]. In some cases, visits to public POIs are ignored altogether [42]. Although these activity-based models do well to approximate mobility behavior, it has been shown that human mobility during a pandemic depends on socio-economic factors such as the income level of counties [17]. Thus, a more realistic and data-driven representation of agent mobility and the POIs is needed to simulate disease spread patterns [20, 21].

Therefore, this study leverages fine-scale, real-world foot-traffic data to inform a more realistic human mobility model in an ABM of COVID-19 spread. Foot-traffic data includes check-ins by users at a variety of POI types. We propose using Latent Dirichlet Allocation (LDA) [5], a popular topic modeling technique, to inform agent mobility behavior with such real-data. Here, we group users by their home census block group (CBG) and treat POIs as "words" and agent home CBGs as "documents" to extract "topics" of POIs that frequently appear together in CBG visits. To simulate realistic behavior, we first create a synthetic population of the case study area (Fairfax County, Virginia, USA) and inform a generative model using the learned topic models with realistic human mobility patterns. This approach preserves the statistical relationship between agent home CBGs and POI visits as observed in the real world and facilitates data-driven and realistic representations of agent mobility and interaction, useful for disease spread simulations. The LDA results are used to inform agent mobility in an ABM that simulates COVID-19 dynamics and tests a variety of policy interventions.

The remainder of this study is organized as follows. Section 2 reviews the related work on LDA and epidemiological ABMs. Section 3 outlines the foot-traffic data and census data used in the LDA approach. Section 4 describes the ABM component that is used to simulate the disease spread dynamics. Section 5 details the LDA approach implemented to develop the underlying model of mobility and the experiments. The resulting latent topics are examined in Section 6 and the simulation results are presented in Section 7. Finally, the concluding statements are made in Section 8.

2 RELATED WORK

In this work, we augment Latent Dirichlet Allocation [5] to model topics of POI visits in Fairfax County, Virginia, USA, to inform the mobility of agents in an ABM that simulates the spread of COVID-19. This section provides an overview and related work of these techniques.

2.1 Latent Dirichlet Allocation

The problem of modeling collections of discrete data can be solved using Latent Dirichlet Allocation (LDA), a generative statistical model of a text corpora [5]. The LDA approach uses words, documents, and corpora to describe data sets. Words represent the basic unit of discrete data. Documents are collections of words, and a corpus is a collection of those documents. Each document contains distributions of latent topics where words characterize topics. LDA can be used for document modeling, collaborative filtering, and dimensionality reduction. In 2003, Blei et al. [5] analyzed a corpus of scientific abstracts of C. Elegans and TREC AP corpus. They showed that LDA consistently performed better than present probabilistic modeling, pLSI, by solving issues like overfitting seen in other techniques. In another study, Văduva et al. [41] applied LDA to a non-textual context of spatial analysis of satellite images. They represented words as spatial signatures, which they defined as the relative positioning of a pair of objects. The document is a tile of the image, and the entire image is the corpus. Unlike orthodox uses of LDA for semantic analyses, Teng et al. [39, 40] utilized LDA as a measure of the diversity of POIs.

2.2 Agent-Based Modeling for Disease Spread

ABMs of disease spread simulate the interaction between individuals or "agents" from which disease spread dynamics emerge. ABMs are touted in epidemiology for their ability to overcome limitations of traditional SIR models and their variations, which treat individuals as homogeneous, interactions as equal and global, and the spatial distribution of individuals as uniform [4, 14]. As a result, ABMs have been developed to simulate seasonal influenza [26, 27], pandemics including H1N1 [8, 15, 29], Ebola [30, 36], and COVID-19 [6, 12, 16, 37], and smaller outbreaks of small-pox [7], anthrax [9], the pneumonic plague [43], and dengue [19]. Agent-based simulations aim to forecast disease spread dynamics, estimate social and economic impacts, develop policy intervention strategies, and better understand the relationship between local processes and disease emergence. These models are often used as a tool by decision-makers to improve the preparedness and responsiveness to outbreaks and pandemics, leading to stronger and more resilient cities and urban areas [10].

3 DATA SETS

This study aims to develop an ABM of COVID-19 spread based on real-world human mobility patterns rather than relying on simplified and often unrealistic assumptions. In this section, we describe the foot-traffic data upon which our mobility model is based.

3.1 SafeGraph Foot-Traffic Data

Data from *SafeGraph Inc.*[1] provides unique and valuable insights into foot-traffic patterns of large-scale businesses and consumer POIs. This work uses *SafeGraph's* "Weekly Patterns" data, which registers GPS-identified visits to POIs (primarily businesses) with an exact location in the United States. For each visit by an individual to a POI, the home census block group (derived from nighttime GPS location) is recorded[2]. Additionally, *SafeGraph* provides a taxonomy of POIs types in a "Core Places" schema, allowing our simulation to test the closure of specific business categories (e.g. restaurants). *SafeGraph* also includes information on the proportion of residents

[1] Attribution: SafeGraph Inc., a data company that aggregates anonymized location data from numerous applications to provide insights about the physical presence in places. To enhance privacy, *SafeGraph* aggregates home locations to census block group level and excludes locations if fewer than five devices visited a POI in a month from a given census block group.

[2] For detailed information, see https://docs.safegraph.com/docs/weekly-patterns.

that either stay home or leave the house on any given day for each CBG in a separate "Social Distancing Metrics" dataset, allowing us to establish CBG-level probabilities of agents leaving their home.

Due to the datasets' sheer size, all of the data we used was filtered to include only POI and CBG data from Fairfax County, Virginia. We chose to use data from the week spanning from October 28 to November 3, 2019, as a representative sample of typical movement patterns before the onset of COVID-19. We filtered POIs only to include those with a large enough sample of aggregate visitors (30 or more) throughout the week-long timeframe. The filtered dataset resulted in 4,130 unique POIs across Fairfax County and 689,731 recorded visits to these POIs.

3.2 United States Census Data

We use United States Census data[3] to map CBGs to their correct geographic locations. This data also facilitates the initialization of agents and agent households.

4 COVID-19 MODIFIED SEIR MODEL

This section describes the model of COVID-19 disease spread used in this study.

4.1 Population Initialization

To initialize our simulation, we first generate households according to CBG-level data provided by the US Census, filling each household with its corresponding number of agents. Between one and seven agents are assigned to each household, with "7-or-more person households" being treated as size seven for simplicity. By default, we simulate approximately 10% of the total Fairfax County population by only generating 10% of the households of each size in each CBG, resulting in a simulation of 106,978 agents. Aside from infection status, agents and households are not assigned any other attributes such as age, income, or race. A small percentage (25%) of agents from a single, randomly selected CBG are initially infected, resulting in a default of 26 initially infected agents. For consistency, this CBG's *SafeGraph* ID is *510594804023* for all of our trials. We used integer *1* as the seed for the pseudo-random number generator in all of our trials for reproducibility.

4.2 Representation of Disease Dynamics

Once the agent population initialization concludes, the simulation begins at midnight and runs until agents are no longer exposed or infected. Each tick in the model represents fifteen minutes, based on the CDC's definition of close contact between individuals [13]. The probability that an agent will leave their home location to visit a POI on any given day is based on *SafeGraph's* "Social Distancing" dataset. We divide the total daily number of people who did not stay home by the total daily number of people in the CBG between October 28 and November 3, 2019. We calculate that the average daily POI visit probability across each CBG in Fairfax County is 74.8%. To roughly approximate the likelihood that an agent would leave the house at each tick, we divide this probability by the number of ticks in a day (96 by default), resulting in a 0.780% average probability. This finding is also consistent with other travel surveys and cellphone based mobility studies [34]. We consider this probability, calculated based on the foot traffic data acquired before the onset of COVID-19 and thus not influenced by the pandemic, as a default parameter of a 100% propensity to leave.

At each 15 minute tick, infectious agents may come into contact with a maximum of five other agents, by default, who are located at the same POI that is not their household. If a susceptible agent comes into contact with an infectious agent, they have a 5% chance of becoming exposed and subsequently infectious by default [28].

Infectious agents also have a chance to spread the virus to susceptible agents in their household. Research indicates that approximately 20.4% of people living in small households (size six or less) will contract the virus if they share a residence with someone infected [18]. This percentage decreases to 9.1% in large households (size seven or greater). Using these numbers and the median infectious period of the virus according to the gamma distributions that we use, as outlined below, we approximated that susceptible agents have a 4.44% and 1.98% chance of contracting the virus from an infected household member each day in small and large households, respectively. For simplicity, household infection occurs at midnight each day, even if a household member is visiting a POI.

We represent COVID-19 dynamics use a generalized SEIR (Susceptible, Exposed, Infectious, Recovered) model [2] that is modified to include subclinical, preclinical, and clinical subclasses of the Infectious stage. Agents undergo the following stages:

(1) Susceptible: An agent who has never been infected or exposed to the virus, but has the potential to become exposed.
(2) Exposed: An agent who has caught the virus and will become contagious (infectious) after an incubation period.
(3) Infectious: An agent who can infect others and is contagious. We define three subclasses of infectious agents:
    - Subclinical: An asymptomatic infected agent. It is estimated that 40% of infections are subclinical. As these agents will never show symptoms, they are estimated have a 75% relative infectiousness compared to clinical agents.
    - Preclinical: An infected agent who is pre-symptomatic (not currently symptomatic) but will enter the clinical stage and become symptomatic in the future. All agents that enter the clinical stage first pass through the preclinical stage. As preclinical agents do not exhibit symptoms, they are also estimated to have a 75% relative infectiousness compared to clinical agents.
    - Clinical: An infected agent who shows symptoms of the virus and is fully infectious. It is estimated that the remaining 60% of infections progress to the clinical stage.
(4) Recovered: A previously infected agent that is noncontagious and immune to the virus. An agent is classified as recovered as long as they cannot actively spread the virus, even if they have lasting complications or symptoms.

The duration of each stage in the SEIR model in days is determined by drawing from the following gamma distributions [11]:

- Exposed stage duration: $gamma(\mu = 3.0, k = 4)$
- Subclinical stage duration: $gamma(\mu = 5, k = 4)$
- Preclinical stage duration: $gamma(\mu = 2.1, k = 4)$
- Clinical stage duration: $gamma(\mu = 2.9, k = 4)$

[3] https://www.census.gov

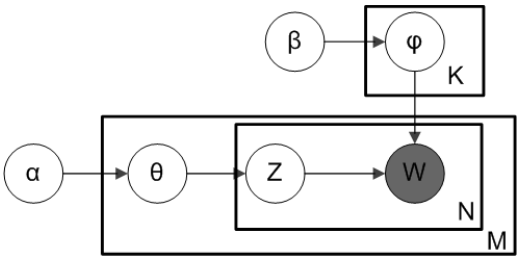


Figure 1: Graphical Model in *plate notation* of LDA-based topic modeling. Boxes represent entities (M visitor home CBGs, N POI visits per visitor home CBG, K latent topics). Nodes correspond to random variables, shaded nodes are observable random variables, and arrows indicate stochastic dependencies.

5 LDA TOPIC MOBILITY MODEL

This section describes our approach for a data-driven mobility model that applies Latent Dirichlet Allocation (LDA) to SafeGraph foot traffic data for topic extraction.

5.1 Foot Traffic Topic Extraction

Latent Topic Modeling. Given the CBG foot traffic data for each POI as obtained from the SafeGraph data, we apply topic modeling using LDA [5] – a generative probabilistic model described in Section 2.1. While traditionally used to find K latent topics among a corpus of M text documents containing N words per document, we employ LDA in our simulation to find K latent topics among a subset of M CBGs each containing N distinct POI visits. Specifically, we define the following analogous terms as follows for the purposes of our study:

- A word, representing the basic unit of data, is defined in our context as a single POI visit designated by a unique POI ID assigned by *SafeGraph*.
- A document, representing a single visitor home CBG, is defined as a collection of words (POI visits). Like POIs, *SafeGraph* provides unique CBG IDs derived from US Census data.
- The corpus is defined as the complete collection of documents (Fairfax County visitor home CBGs).

A graphical representation of our LDA model is shown in Figure 1. We used a uniformly distributed vector α of length K to parameterize the apriori distribution of topics. The parameter K corresponds to the number of latent topics we want to find. When a CBG is created in our simulation, we assume that its topics are chosen following a *Dirichlet distribution* having distribution parameter α which we use to obtain a topic distribution θ for each of our M CBGs. Thus, the large plate in Fig. 1 corresponds to a set of all M CBGs, each having a topic distribution θ drawn randomly (and Dirichlet distributed) from α. To infer the topics of CBGs, LDA uses a generative process similar to Monte-Carlo sampling with an iterative refinement of the distributions α and θ. More details on LDA can be found in [5].

Transposing Foot-Traffic Data into Text for LDA Modeling. We generate documents as we process each POI's data; the frequency of a specific POI's ID in a document generated for LDA corresponds to the number of people from the document's analogous visitor home CBG that visited the POI within the timeframe. The order of words is not considered while generating the LDA model. It should be noted that *SafeGraph* does not provide visitor home CBGs for a POI if only 1 visitor from the given home CBG visited the POI of focus throughout the entire week. Additionally, *SafeGraph* provides "4" as the number of visitors from the given home CBG to the POI if the real number is 2 or 3 to maximize privacy. To account for this, we replaced "4" with a randomly generated number in the range [2, 4] whenever "4" was provided.

We now have many documents filled with words, and LDA can now extract topics from the corpus. LDA generates these topics without the usage of background knowledge. LDA treats each word in the document as a sample for a mixture model, where mixture components are viewed as representations of underlying latent topics. This allows LDA to assign words to a "representing" topic. Thus, LDA provides a word probability distribution for each topic upon completion. Besides, LDA also provides a topic probability distribution for each document. These two probability distributions enable us to create realistic mobility behavior in the ABM based on real-world data.

5.2 LDA Distributions for Agent Generation

In using the LDA approach, each CBG's POI visits are a mixture of underlying latent topics and each topic has a latent distribution of more and less likely POIs. In that respect, LDA provides two distributions: 1) a topic probability distribution for each visitor home CBG and 2) a POI probability distribution for each topic. These two distributions allow an ABM to be constructed such that 1) agents are generated and assigned a specific LDA topic according to the topic probability distribution of their home CBG and that 2) agents visit POIs based on the POI probability distribution of their assigned topic.

We are now able to generate agents with specific attributes based on their home CBG, providing the groundwork for our ABM. First, for each home CBG in Fairfax County that *SafeGraph* provides data for, agents are generated according to the real population of the CBG, with each being assigned a topic according to the first distribution. Agents' topics are static and cannot change throughout the simulation.

After agents are assigned topics, we may randomly sample from the second distribution to determine which POI agents will visit if they decide to leave their house. Due to LDA's nature, it is unlikely that two unrelated POIs, such as a nightclub and a library, will have relatively equal weights in this probability distribution, resulting in a vast improvement from the uniform probability distribution used to select POIs in traditional ABMs. This probability distribution, however, does not provide information on whether or not an agent will decide to visit a POI in the first place.

Modeling Hourly Visit Patterns. So far, LDA generates distributions from data that contains the number of visitors from each CBG to each POI across an entire day. However, this approach is slightly flawed because in reality, the number of visits is dependent on the

time of day. For example, a restaurant would likely have higher concentrations of visits at noon and in the evening and lower concentrations during the mid-afternoon. Similarly, visits to a school POI during the evening would be less likely. To remedy this issue, we use additional *SafeGraph* data that provides individual POI visits for each hour over the entire week-long timeframe and create 24 distinct POI probability distributions for every topic, one for each hour of the day. We do this by re-weighting the topic's base POI probability distribution 24 times according to each POI's proportion of visits during the given hour. For each topic, a weighted distribution of visits that take place each hour is constructed according to the topic's POI distribution and the number of visits to those POIs that take place during the given hour. This distribution provides the weighted percentage of visits on the topic that takes place in the given hour compared to the entire day. For example, the midnight hour may have a probability of 0.01, while the noon hour may have a probability of 0.07. By modifying the likelihood, we allow simulating agent POI visits more accurately. Given our simulation's default parameters for the hour of noon example, an agent's chance to visit a POI between 12:15 and 12:30 would be approximately $0.748 * 0.07/4$, or 1.31%, markedly higher than the generic average probability of leaving each tick of 0.780%.

5.3 Dwell Time Distributions for POIs

The *SafeGraph* data allows us to use a data-driven approach to modeling dwell time by fitting a suitable probability distribution to a POI's bucketed dwell time data (provided by *SafeGraph*) and random sampling the said distribution for every agent that "visits" the POI.

SafeGraph Bucketed Dwell Time Data. SafeGraph provides a "bucketed" version of each POI's dwell times, where only the number of visits within a range of dwell time is quantified, i.e., *"<5 minutes": 266, "5-20 minutes": 4184, "21-60 minutes": 3597,"61-240 minutes": 2492, ">240 minutes": 892.* While providing an initial perspective on the potential probability of dwell time's for an individual agent's visit to a POI, the raw *SafeGraph* data is not adequate for direct usage due to its non-specificity.

Fitting Probability Distributions. To compensate for the bucketed format of the data ("5-20 minutes":4184), we first impute the bucket ranges of each POI's dwell data by random uniform sampling: for a range of 5-20 minutes with 266 visits, we fill the bucket with 266 random uniform samples ranging from 5-20. With a full range of dwell data for each bucket, we move to methods of sampling. We determine that employing probability distributions allows for the most optimal method of a random sample due to their "smoothing" of minor irregularities that may occur from the random uniform imputation of each POI's dwell time buckets. From those, we approximate using the parametric function with the best fit for each POI dwell time distribution. For example, restaurants might have a more normal distribution around a mean stay time of 1 hour. In comparison with malls, where a large proportion might drop off or drive by with a large proportion of visits under 5 minutes so might be better represented using an exponential curve. For each POI, we test the fit of 10 of SciPy's most common probability distributions for a continuous random variable—normal, generalized extreme value, exponential, gamma, Pareto, lognormal, double Weibull, beta, Student's t, uniform—and select the most optimal based on the goodness of fit test. See [38] for more information on the probability distributions used and the package employed. We initialize and cache the fitted distribution for the POI in question; the distribution is randomly sampled five times for each agent visit to said POI. The median is returned, representing an estimated *SafeGraph* data-based dwell time. If the median dwell time is less than one tick, then one tick is returned. Alternatively, if the median dwell time is greater than 16 hours, then 16 hours is returned. The median dwell time is rounded to the nearest tick in all other cases.

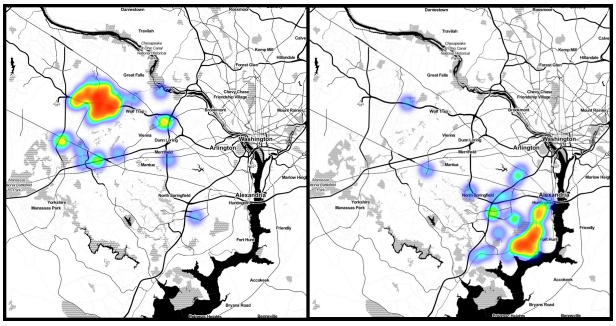


Figure 2: Heatmaps of the top 100 POIs for two selected topics with topic count $k = 50$ (left: topic 37, right: topic 43)

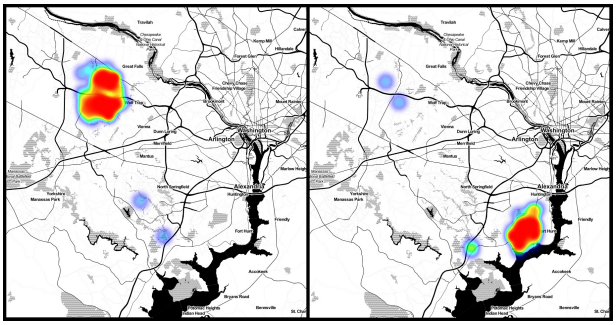


Figure 3: Heatmaps of the home CBGs of agents assigned to two selected topics with topic count $k = 50$ (left: topic 37, right: topic 43)

6 QUALITATIVE ANALYSIS OF LATENT TOPICS

In this section, we qualitatively evaluate the spatial distribution of the topics generated, as described in Section 2.1. Recall that in the LDA approach, each CBG is described by a topic distribution that maps each topic to the probability that the CBG is an instance of this latent topic. It seems intuitive that different CBGs have different preferences, both spatially and semantically. For example, more wealthy CBGs may more frequently choose to visit high-class restaurants, whereas less wealthy CBGs may more frequently choose more affordable eateries. It is also intuitive that topics should capture spatial preferences, as people are more likely to choose nearby POIs than more distant ones. Figure 2 shows the

Table 1: POI Distributions of Latent Mobility Topics

Topic ID	POIs (Probabilities in %)
#16	[('Holiday Inn', 3.47), ('Tysons Corner Center', 2.51), ('Dulles Corner Park', 1.90), ('Westin Hotels & Resorts', 1.63), ('Village Center At Dulles', 1.51), ('Westfields Marriott Washington Dulles', 1.28), ('Campagna Kids Wm Ramsay Edc', 1.26), ('Tysons Cafe', 1.18), ('Reston Town Center', 1.13), ('Sully Plaza', 1.10), ...]
#25	[('Hilton International', 6.11), ('CoCell Used Phones', 2.80), ('Village Center At Dulles', 2.64), ('Atlantic Union Bank', 1.50), ('Imagination Learning Center 2', 1.49), ('Neiman Marcus', 1.30),('Mclearen Square', 1.28), ('sweetgreen', 1.17), ('Coppermine Run', 1.17), ('Spring Hill Elem', 1.10), ...]
#28	[('Sully Plaza', 4.65), ('Chantilly Crossing', 3.32), ('Franklin Farm Village Center', 2.74), ('Chantilly High School Academy', 1.88), ('Chantilly High', 1.81), ('Greenbriar Town Center', 1.75), ('Costco Wholesale Corp.', 1.73), ('Mclearen Square', 1.63), ('Village Center At Dulles', 1.34), ('Chantilly Governor's Stem Academy', 1.34), ...]

POI distribution of two topics, 37 and 43, resulting from our topic modeling approach using $k = 50$ latent topics. The heatmap in Figure 3 shows the geographic distribution of agents assigned to the two topics by their home CBGs. First, we observe a clear spatial pattern, as the two topics correspond to different areas.

Beyond spatial topics, we observe that different topics capture different semantic types of POIs. Table 1 presents the Top 10 POIs, and their respective probability within a topic for topics #16, #25, #28 for our latent topic model using 50 topics. These three topics exhibit mutually similar spatial distribution, which can be seen using our web demonstrator [33] . While spatially similar, these topics differ semantically. Topic 16 includes numerous hotels, attractions, and cafes, hinting that this topic may capture visitors or tourists. Topic 25 also includes hotels such as Hilton International, but also includes Atlantic Union Bank and Neiman Marcus, which targets professionals. In contrast, Topic 28 captures POIs such as Chantilly High School, Greenbriar Town Center (next to the school), and Carson Middle School, which targets students. For additional details to explore the spatial distribution of topics, an interactive map that allows users to select combinations of topics and explore the resulting topic density map can be found at https://jpes707.github.io/safegraph-simulation.

7 EXPERIMENTAL EVALUATION

This section presents the results for the LDA approach and the simulation results obtained from the disease spread model with LDA informed mobility behavior. We compare our LDA topic informed mobility model and the resulting disease spread model to two models with competitor approaches for representing mobility, described in Section 7.1. Next, in Section 7.2, we evaluate how changes to the different model parameters affect the number of agents entering the "Exposed" stage on each simulation day also known as an epidemic curve. The resulting area under the curve can be interpreted as the total number of agents that have become infected (at any time). Finally, in Section 7.3, we present a proof of concept of how our simulation model can be used to evaluate different policy interventions such as agent quarantine or POI closure. All experiments are outlined in Table 2.

For all of the experiments we compare between the epidemic curves produced using mobility behavior that is informed by the LDA approach where $k = 1$ topics, $k = 10$ topics, and $k = 50$ topics. We note that in all experiments, it appears that the number of topics may have an effect on the timing of the outbreak where in some cases the peak of the outbreak is accelerated or delayed. This is interesting, as in the case of $k = 1$, all agents share the same global POI-visit distribution, thus allowing to spread diseases globally. In contrast, in the case of $k = 50$, many topics are very localized, thus making it more likely to meet agents from nearby CBGs. While it seems intuitive that a more local mobility pattern would mitigate a disease outbreak, we see in all experiments that this is not the case. This is even more surprising since the initial infected agents are from a single CBG only and even still, it appears that localized mobility behavior (such as restaurants and grocery stores) does not significantly affect the epidemic curve.

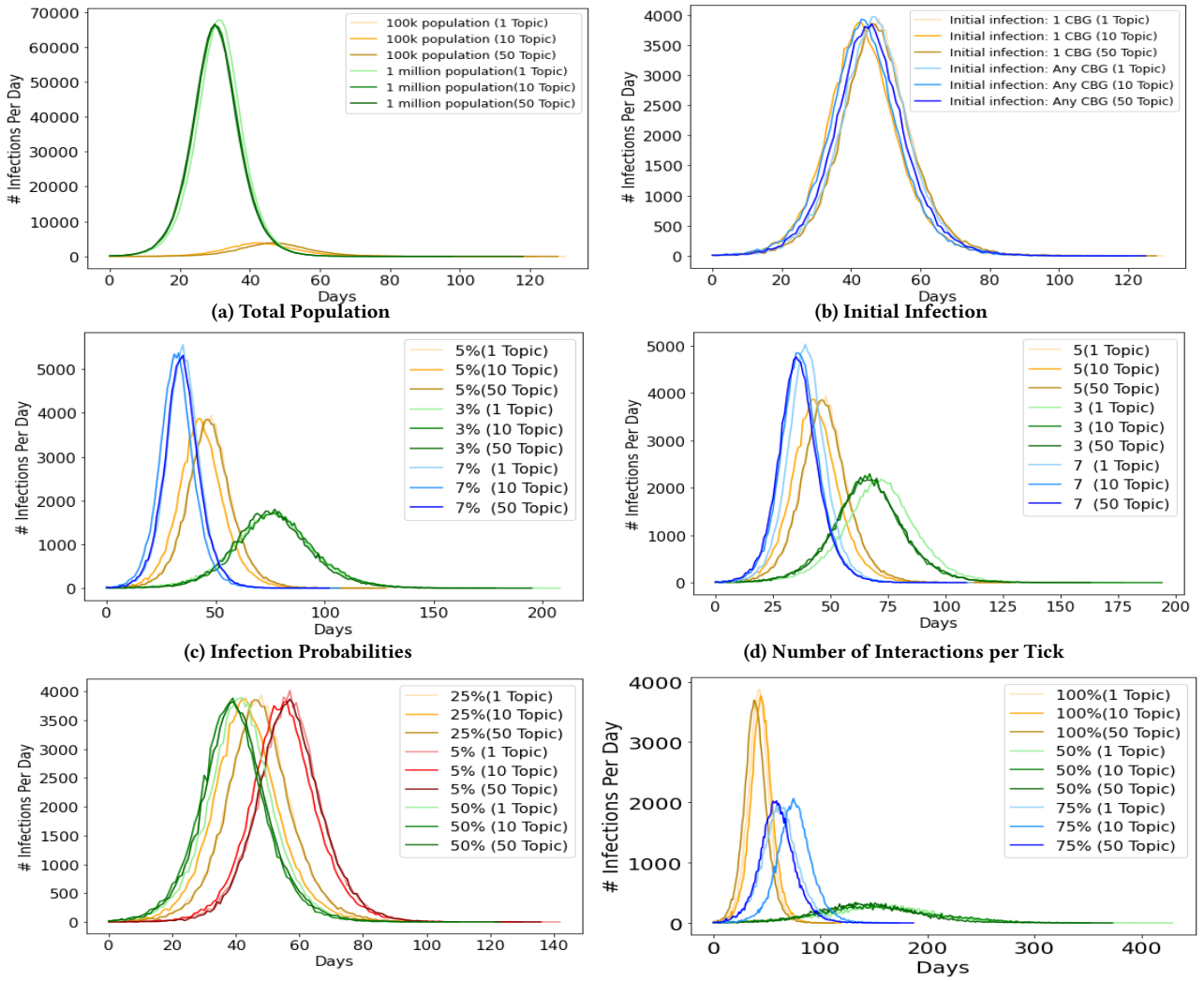
This may be due to our relatively small study area, in which many POIs have visitors across all CBGs and thus allow a disease to quickly disseminate across space. What is particularly interesting about this result, is that SafeGraph uses the average distance travelled per day as an indicator of social distancing. Our results suggest that a better metric of social distancing would be based on proximity between users, such as measuring the frequency of users staying within a distance of less than two meters for more than fifteen minutes.

We conducted our experiments in a Windows environment with 6 CPUs at 4.0 GHz and 64 GB of RAM. For reproducibility, the interested reader is referred to our implementation at [33].

7.1 Competitor Approaches

We use two competitor models to compare our proposed solutions: 1) a naive mobility model which assumes that agents choose POIs uniformly at random, and 2) a baseline mobility model in which agents choose POIs at random, weighted by empirical foot-traffic visit frequencies. Real-world POI visit frequencies are preserved in the latter, but all agents have the same POI visit distribution regardless of their location and latent topics. These approaches are described in detail in the following.

Naive Mobility Model. To create a naive human mobility model, we assume that agents visit POIs uniformly at random and that the dwell time is regularized to one hour. Additionally, the hourly POI visit probability weighting described in Section 5.2 is not implemented. Thus, each POI has the same probability of being visited by an agent at any time of day. Agents choose randomly from the list of POIs provided by *SafeGraph*, but no visitor count or dwell time data is used except for filtering out POIs that see less than 30 visitors per week. Using this naive model, we find that the virus



(a) Total Population **(b) Initial Infection**

(c) Infection Probabilities **(d) Number of Interactions per Tick**

(e) Fraction of Initially Infected Agents **(f) Propensity to Leave Home**

Figure 4: Disease Spread Simulation Results with Various Model Parameterization

only infects 80 agents total and stops spreading within 50 days with no government interventions. This is due to a large number of POIs (4130), allowing agents to spread out and rarely meet other agents. We conclude that due to the unrealistic, uniform distribution of agents across POIs, agents' concentration in a single POI is not high enough for the virus to spread. This result shows that a realistic disease simulation should consider realistic human mobility, which follows a long-tail distribution with a few highly frequented POIs and has many POIs with very few or no visits.

Baseline Mobility Model (1 Topic). A straightforward way of informing a simulation with foot-traffic data is to have agents randomly choose POIs to visit, but weighted by the marginal visit probability across all POIs. The resulting simulation model yields visit frequencies similar to observed frequencies. However, in this model, any agent uses the same POI visit distribution, regardless of location and preferences. We note that this baseline is equivalent to using our LDA-based approach by setting the number of latent topics k to one. In this case, there is only one global topic which is chosen by all agents.

7.2 Effect of Model Parameters

This section presents disease simulation results using the SEIR disease model described in Section 4. A variety of experiments are implemented to test the effect of the number of topics and the model parameterization on the epidemic curve.

Total Population. Our study area (see Section 5.1 for details) has a population of ~1.1M. By default, our experiments use a population of 100K to allow fast simulation results and multiple replications. Figure 4a compares the epidemic curves resulting from ~100K versus ~1M agents. With 1M agents, we observe a much larger magnitude of the disease peak. Furthermore, the disease spreads much faster, such that the academic peak is reached after only ~30 days vs. ~45

Table 2: Experiment settings (Default parameters in bold)

Experiment Type	Parameter	Values
Effect of Model Parameters	Latent topics	**1** / **10** / **50**
	Total population	**100K** / 1M
	Initial infection	Random / **One CBG**
	Fraction of initially infected agents in the CBG (% CBG population)	5% / **25%** / 50%
	Infection probabilities	3% / **5%** / 7%
	Number of interactions per tick	3 / **5** / 7
	Propensity to leave home	50% / 75% / **100%**
Effect of Policy Interventions	Generic quarantine (only infected agents)	**0** / 4 / 6 / 10 days
	Household quarantine (all household individuals)	**0** / 4 / 6 / 10 days
	Closure of POIs	**None** / Schools closed / Restaurants closed / Nonessentials closed

days with 100K agents. Since there are a greater number of agents and the number of POIs do not change, it becomes more likely for an infectious agent to find other agents in any POIs. In contrast, in the case of 100K, agents that visit POIs with a low visit frequency may interact with less than five agents, the maximum number of interactions by default.

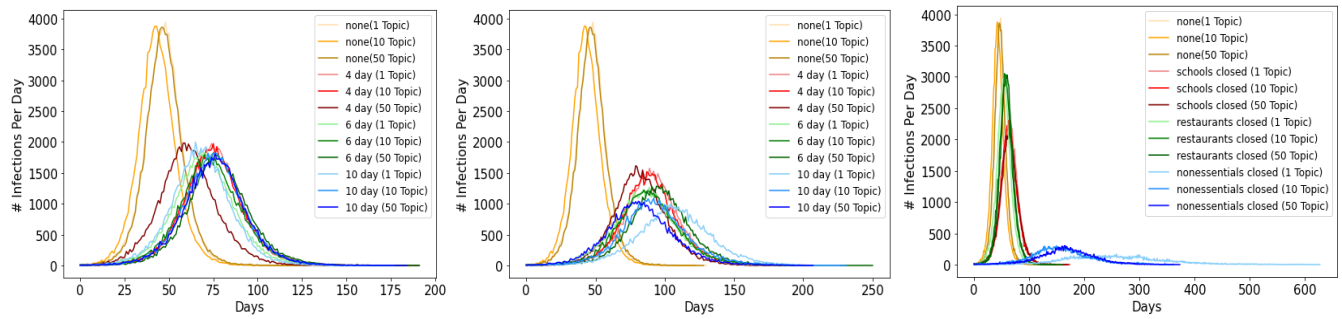
Initial Infection. Figures 4a and 4b compare the epidemic curve resulting from a random selection of initially infected agents across all CBGs and one CBG only. For both experiments, we compare between $k = 1$ topics, $k = 10$ topics, and $k = 50$ topics. In the case where the initially infected agents are distributed randomly across the map, the disease progression is very similar across all topics. This is due to the maximum number of infections made by each agent per fifteen minutes. Initially, as POIs are crowded, infected agents can always find the maximum number (per default five) agents to infect. While using different topics, the location where these interactions and infections may change, but the total number of infections remains the same. This somewhat changes once we set the location of initially infected agents to a single CBG. Intuitively, in the baseline modeling where $k = 1$ topic, the infected agents in this CBG will randomly disperse across the simulation. In the case of increasing k, these infected agents are more likely to visit POIs among their topic, thus spatially and semantically relevant to their CBG. In this case, infected agents are more likely to remain local, and visit fewer POIs, but the total number of infections remain the same.

Fraction of Initially Infected Agents. In the next experiment shown in Figure 4e, we scale the fraction of agents initially infected in the CBG where the disease originates from. What is interesting is that that the magnitude of the disease curve seems unaffected by the number of initially infected disease. In all cases of having 5%, 25%, and 50% of initially infected agents, we observe that the disease curve hits about 4000 new infections per day. The only difference is the time until this peak is reached. We conclude that having an initially higher number starts the simulation at a later stage of the disease progression.

Infection Probabilities. Next, we scale the infection probability, as shown in Figure 4c. This parameter defines the probability of an interaction between an infectious and a susceptible agent to expose the susceptible agent. As expected, we observe that when we increase the infection rate to 7%, we observe a steep increase in infections. On the flip-side, decreasing the infection probability yields slower disease progression and a flatter curve. While the areas under the curve in the three cases are similar (98%, 95%, and 87%, for infection rates of 7%, 5%, and 3%, respectively), we observe that the number of infections is not affected much by the infection rate. Still, the magnitude of the peak is strongly affected. This experiment shows the importance of using means to reduce infection probabilities (such as wearing masks). It drastically flattens the curve, thus reducing the number of people infected at a time.

Number of Interactions per Tick. While the probability of infection in a fifteen minutes meeting is well studied [28], we assumed the number of such meetings that can happen concurrently, and we assumed that any type of POIs allows the same number of interactions. Figure 4d shows the resulting epidemic curves when changing this parameter. We observe results similar to changing the infection probabilities. This is expected, as having more interactions, each resulting in an infection chances, increases the expected number of infections.

Propensity to Leave Home. A final disease spread parameter that we evaluate is the propensity to leave home, which defines the hesitation of agents to leave their home. When this parameter is set to 100% (the default value), then agents leave their home in a "normal" frequency as observed in the foot-traffic data. Setting this value to lower values allows to simulate the effect of agents deliberately staying at home to avoid infection. Figure 4f shows that a lower propensity to leave home most drastically reduces the spread of the disease. By setting this parameter to 50% and thus, reducing the number of trips made by agents by 50%, we see that the academic peak drops from 4000 infections per day down to 500 infections per day. This superlinear decrease in infections comes from the fact that fewer agents are leaving home, but when they do, they also find fewer other agents to infect.



(a) Intervention: Generic Quarantine **(b) Intervention: Household Quarantine** **(c) Intervention: Closure of POIs**

Figure 5: Disease Spread Simulation Results after Prescribing Interventions

7.3 Effect of Policy Interventions

A variety of experiments are implemented to test the impact of various public health interventions on the spread of COVID-19 and the subsequent effect on the epidemic curve.

Generic Quarantine. This intervention requires any agent that is infectious and aware of symptoms (in subclinical stage or beyond) to stay at home for a specific number of days. Figure 5a shows the infection curves for zero, four, six, and ten days of quarantine. We observe that a quarantine of four days is sufficient to drastically flatten the number of infections per day from to 2000. However, as 40% of agents are subclinical and not aware of their infection, these agents are not quarantined and continue to spread the disease.

Household Quarantine. This intervention additionally requires that all agents that share the same household as the quarantined agent remain at home. This intervention is significantly more effective, flattening the curve further to about 1500 infections per day. This result is intuitive as agents living in the same household are at highest risk of becoming infected. We also observe that a longer quarantine duration is significantly more effective. Specifically, agents are quarantined before they show symptoms, thus dropping the infections per day to 1000 using a ten day quarantine.

Closure of POIs. We test three interventions related to the closure of POIs where each intervention targets the closure of a specific POI type including schools, restaurants, and non-essential places. Figure 5c presents the effect of POI closure on the epidemic curve. Instead of visiting the closed POI, the agent will instead decide to stay at home. We observe that closure of restaurants yields a significant reduction of disease spread from a disease peak of 4000 agents per day down to 3000 agents per day. Interestingly, the closure of schools is far more important, reducing the peak to about 2000. This is likely due to the difference in dwell time between schools and restaurants. As agents typically dwell at school POIs for up to eight hours a day, it becomes extremely likely that during any 15-minute tick they are successfully exposed to the virus by an infectious agent. In contrast, dwell times at restaurants are usually less than one hour (this also includes fast-food restaurants), drastically reducing the probability of becoming exposed by a collocated infectious agent. We also see that if we choose to close all POIs that are classified as non-essentials (using the POI classification provided by SafeGraph), we observe that the disease is nearly eradicated.

8 DISCUSSION, CONCLUSIONS, AND FUTURE WORK

This study implements LDA using *SafeGraph* foot traffic data to extract realistic mobility patterns for agents belonging to different CBGs in Fairfax County, Virginia. The LDA results are used to inform a mobility model to simulate the spread of COVID-19 and test the effect of changes to various parameters (e.g. public health interventions, number of topics) on disease dynamics. We examine the simulated number of cases over time and find that the number of topics does indeed impact the epidemic curve, but only concerning the outbreak's timing. Further investigation is needed to understand better the impact of topics on simulated disease spread dynamics. Likely, measures that capture the spatial distribution and variation of infected individuals and that examine the relationship between POIs and different CBGs may better highlight the effect of topics on disease spread dynamics.

As a proof of concept and to allow us to run many experiments efficiently, the LDA approach informs an ABM of COVID-19 spread across only 10% of the population of Fairfax County, Virginia. Furthermore, Fairfax County is a well-connected region bordered by Washington, D.C., Maryland, and the rest of Northern Virginia. Thus, we cannot account for interactions with agents that do not belong to Fairfax County but travel to Fairfax County POIs and vice versa. Future work may focus on scaling up the simulation to a more extensive and inclusive study area and population, such as the entire Washington D.C. metropolitan area.

Existing parameter settings are based on available research; however, finding a consensus on empirical observations to inform these parameters is difficult due to the novelty of COVID-19. Additional empirical research is needed to refine model parameterization and to validate the model. Further experimentation and model testing may reveal the best number of topics empirically to represent mobility behavior for ABMs of disease spread. Furthermore, we may consider increasing the model's complexity to account for hospitalization, deaths, and various other COVID-19 spread factors.

Given the different public health interventions implemented in our simulations, we may also consider conducting prescriptive analytics to determine the relationship between an intervention's benefit and the resulting losses. Research to find an optimal solution to mitigate disease spread while also decreasing community impact can benefit government officials in writing intervention policies.

In this case, the novel LDA approach is used to develop data-driven mobility behavior to better inform an ABM of COVID-19 spread. However, there is a potential to use this approach to model mobility in various ABMs, not just in application to COVID-19. In future work, we aim to explore the LDA approach's implementation on public and openly available location-based social network data, such as Twitter data.

ACKNOWLEDGEMENTS

This research is supported by National Science Foundation "RAPID: An Ensemble Approach to Combine Predictions from COVID-19 Simulations" grant DEB-2030685 and by the Aspiring Scientists Summer Internship Program (ASSIP) at George Mason University.

REFERENCES

[1] T. Anderson and S. Dragićević. Neat approach for testing and validation of geospatial network agent-based model processes: case study of influenza spread. *International Journal of Geographical Information Science*, pages 1–30, 2020.

[2] J. L. Aron and I. B. Schwartz. Seasonality and period-doubling bifurcations in an epidemic model. *Journal of theoretical biology*, 110(4):665–679, 1984.

[3] P. Auwaerter. Johns Hopkins' ABX Guide. https://www.hopkinsguides.com/hopkins/view/Johns_Hopkins_ABX_Guide/540747/all/Coronavirus_COVID_19__SARS_CoV_2_. Accessed: 2020-04-11.

[4] L. Bian. A conceptual framework for an individual-based spatially explicit epidemiological model. *Environment and Planning B: Planning and Design*, 31(3):381–395, 2004.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] A. Bossert, M. Kersting, M. Timme, M. Schröder, A. Feki, J. Coetzee, and J. Schlüter. Limited containment options of covid-19 outbreak revealed by regional agent-based simulations for south africa. *arXiv preprint arXiv:2004.05513*, 2020.

[7] D. S. Burke, J. M. Epstein, D. A. Cummings, J. I. Parker, K. C. Cline, R. M. Singa, and S. Chakravarty. Individual-based computational modeling of smallpox epidemic control strategies. *Academic Emergency Medicine*, 13(11):1142–1149, 2006.

[8] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr. Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput Biol*, 6(1):e1000656, 2010.

[9] L.-C. Chen, K. M. Carley, D. Fridsma, B. Kaminsky, and A. Yahja. Model alignment of anthrax attack simulations. *Decision Support Systems*, 41(3):654–668, 2006.

[10] A. Cheshmehzangi. Preparedness through urban resilience. In *The City in Need*, pages 41–103. Springer, 2020.

[11] N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, R. M. Eggo, C. C.-. working group, et al. Age-dependent effects in the transmission and control of covid-19 epidemics. *MedRxiv*, 2020.

[12] F. Dignum, V. Dignum, P. Davidsson, A. Ghorbani, M. van der Hurk, M. Jensen, C. Kammler, F. Lorig, L. G. Ludescher, A. Melchior, et al. Analysing the combined health, social and economic impacts of the corovanvirus pandemic using agent-based social simulation. *arXiv preprint arXiv:2004.12809*, 2020.

[13] C. for Disease Control and Prevention. Contact tracing for covid-19. *Centers for Disease Control and Prevention*, Aug 2020.

[14] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 57–64. IEEE, 2011.

[15] N. Halder, J. K. Kelso, and G. J. Milne. Developing guidelines for school closure interventions to be used during a future influenza pandemic. *BMC infectious diseases*, 10(1):221, 2010.

[16] N. Hoertel, M. Blachier, C. Blanco, M. Olfson, M. Massetti, M. S. Rico, F. Limosin, and H. Leleu. A stochastic agent-based model of the sars-cov-2 epidemic in france. *Nature Medicine*, pages 1–5, 2020.

[17] X. Huang, Z. Li, Y. Jiang, X. Ye, C. Deng, J. Zhang, and X. Li. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the us during the covid-19 pandemic. *medRxiv*, 2020.

[18] Q.-L. Jing, M.-J. Liu, Z.-B. Zhang, L.-Q. Fang, J. Yuan, A.-R. Zhang, N. E. Dean, L. Luo, M.-M. Ma, I. Longini, et al. Household secondary attack rate of covid-19 and associated determinants in guangzhou, china: a retrospective cohort study. *The Lancet Infectious Diseases*, 2020.

[19] S. Karl, N. Halder, J. K. Kelso, S. A. Ritchie, and G. J. Milne. A spatial simulation model for dengue virus infection in urban areas. *BMC infectious diseases*, 14(1):1–17, 2014.

[20] H. Kavak. *A Data-Driven Approach for Modeling Agents*. PhD thesis, 2019.

[21] H. Kavak, J. J. Padilla, C. J. Lynch, and S. Y. Diallo. Big data, agents, and machine learning: towards a data-driven agent-based modeling approach. In *Proceedings of the Annual Simulation Symposium*, pages 1–12, 2018.

[22] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. ii.—the problem of endemicity. *Proceedings of the Royal Society of London. Series A, containing papers of a mathematical and physical character*, 138(834):55–83, 1932.

[23] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, G. Hart, K. Rosenfeld, P. Selvaraj, R. C. Nunez, B. Hagedorn, L. George, et al. Covasim: an agent-based model of covid-19 dynamics and interventions. *medRxiv*, 2020.

[24] J. S. Kim, H. Jin, H. Kavak, O. C. Rouly, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Location-based social network data generation based on patterns of life. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 158–167, 2020.

[25] J.-S. Kim, H. Kavak, and A. Crooks. Procedural city generation beyond game development. *SIGSPATIAL Special*, 10(2):34–41, 2018.

[26] J.-S. Kim, H. Kavak, U. Manzoor, and A. Züfle. Advancing simulation experimentation capabilities with runtime interventions. In *SpringSim 2019*, pages 1–11. IEEE, 2019.

[27] J.-S. Kim, H. Kavak, C. O. Rouly, H. Jin, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Location-based social simulation for prescriptive analytics of disease spread. *SIGSPATIAL Special*, 12(1):53–61, 2020.

[28] M. Klompas, M. A. Baker, and C. Rhee. Airborne transmission of sars-cov-2: theoretical considerations and available evidence. *JAMA*, 2020.

[29] B. Y. Lee, S. T. Brown, G. W. Korch, P. C. Cooley, R. K. Zimmerman, W. D. Wheaton, S. M. Zimmer, J. J. Grefenstette, R. R. Bailey, T.-M. Assi, et al. A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 h1n1 influenza pandemic. *Vaccine*, 28(31):4875–4879, 2010.

[30] S. Merler, M. Ajelli, L. Fumanelli, M. F. Gomes, A. P. y Piontti, L. Rossi, D. L. Chao, I. M. Longini Jr, M. E. Halloran, and A. Vespignani. Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, 15(2):204–211, 2015.

[31] G. J. Milne, J. K. Kelso, H. A. Kelly, S. T. Huband, and J. McVernon. A small community model for the transmission of infectious diseases: comparison of school closure as an intervention in individual-based models of an influenza pandemic. *PloS one*, 3(12):e4005, 2008.

[32] J. Parker and J. M. Epstein. A distributed platform for global-scale agent-based models of disease transmission. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 22(1):1–25, 2011.

[33] J. Pesavento. Foot-traffic informed covid-19 simulation and mitigation. https://jpes707.github.io/safegraph-simulation, Accessed Setember 10, 2020.

[34] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[35] C. M. Schneider, C. Rudloff, D. Bauer, and M. C. González. Daily travel behavior: lessons from a week-long survey for the extraction of human mobility motifs related information. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pages 1–7, 2013.

[36] C. Siettos, C. Anastassopoulou, L. Russo, C. Grigoras, and E. Mylonakis. Modeling the 2014 ebola virus epidemic–agent-based simulations, temporal analysis and future predictions for liberia and sierra leone. *PLoS currents*, 7, 2015.

[37] P. C. Silva, P. V. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, and R. C. Silva. Covid-abs: An agent-based model of covid-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals*, page 110088, 2020.

[38] E. Taskesen. distfit. urlhttps://github.com/erdogant/distfit, 2019.

[39] X. Teng, G. Trajcevski, J.-S. Kim, and A. Züfle. Semantically diverse path search. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 69–78. IEEE, 2020.

[40] X. Teng, J. Yang, J.-S. Kim, G. Trajcevski, A. Züfle, and M. A. Nascimento. Fine-grained diversification of proximity constrained queries on road networks. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pages 51–60. ACM, 2019.

[41] C. Văduva, I. Gavăt, and M. Datcu. Latent dirichlet allocation for spatial analysis of satellite images. *IEEE Transactions on Geoscience and Remote sensing*, 51(5):2770–2786, 2012.

[42] S. Venkatramanan, B. Lewis, J. Chen, D. Higdon, A. Vullikanti, and M. Marathe. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, 22:43–49, 2018.

[43] A. D. Williams, I. M. Hall, G. J. Rubin, R. Amlôt, and S. Leach. An individual-based simulation of pneumonic plague transmission following an outbreak and the significance of intervention compliance. *Epidemics*, 3(2):95–102, 2011.