Epidemic Spread Modeling for COVID-19 Using Mobility Data

Anna Schmedding
akschmed@cs.wm.edu
William and Mary
Williamsburg, VA, USA

Lishan Yang
lyang11@cs.wm.edu
William and Mary
Williamsburg, VA, USA

Riccardo Pinciroli
riccardo.pinciroli@gssi.it
GSSI
L'Aquila, Italy

Evgenia Smirni
esmirni@cs.wm.edu
William and Mary
Williamsburg, VA, USA

ABSTRACT

We present an individual-centric model for COVID-19 spread in an urban setting. We first analyze patient and route data of infected patients from January 20, 2020, to May 31, 2020, collected by the Korean Center for Disease Control & Prevention (KCDC) and discover how infection clusters develop as a function of time. This analysis offers a statistical characterization of mobility habits and patterns of individuals at the beginning of the pandemic. While the KCDC data offer a wealth of information, they are also by their nature limited. To compensate, we use detailed mobility data from Berlin, Germany after observing that mobility of individuals is surprisingly similar in Berlin and Seoul. Using information from the Berlin mobility data, we cross-fertilize the Seoul data set and use it to parameterize an agent-based simulation that models the spread of the disease in an urban environment. We then validate the simulation predictions with ground truth infection spread in Seoul.

1 INTRODUCTION

On March 11, 2020, the WHO declared COVID-19 the first pandemic caused by a coronavirus [4]. Since then, a tremendous amount of data has been collected to guide public health policy decisions. For example, Google provides time-series data of infections at a coarse granularity [5] (i.e., as a function of the area's population, no information is provided at the granularity of single individuals). Epidemiological simulation and mathematical models have been used to predict the spread of the disease. Typically, model effectiveness is tied to its input parameterization.

In this paper, we use data provided by the Korean Center for Disease Control (KCDC) during the first wave of the disease in South Korea [21]. In contrast to the Google data, the KCDC data focuses on *individual patients* and allows the development of an individual-centric model of the COVID-19 epidemic. Infected individuals are monitored and their movements are logged using CCTV, cellphones, and credit card transactions [14]. The KCDC records patient movements in plain text which are parsed through automated code and rule-based methods to extract keywords that are then used with web mapping service APIs (e.g., Google Maps [1]) to extract geographical coordinates and other data.

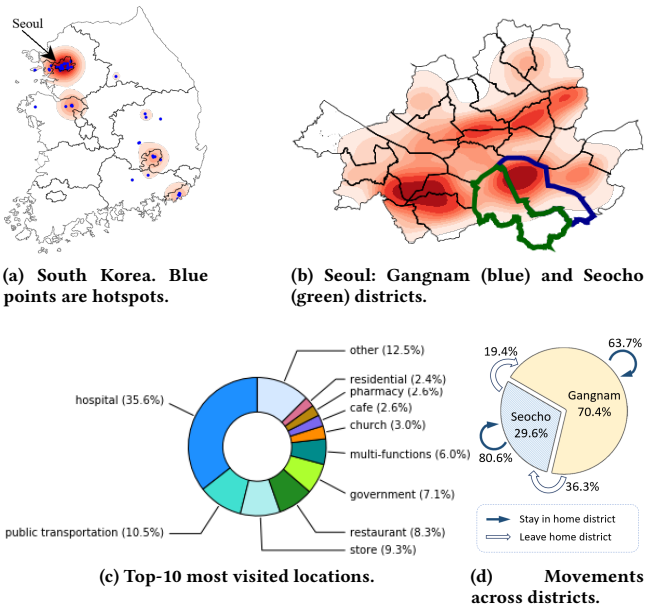
To the best of our knowledge, the KCDC logs are the only available data that contain patient-centric information in great detail: they report on patient mobility, i.e., traveled distance and the sequence of locations visited daily, the date of the onset of symptoms, whether and when the patient got in contact with other patients that are also diagnosed. The KCDC data set is a valuable resource for studying the spread of COVID-19, yet it presents limitations:

- The KCDC data sets contain data collected up to May 31, 2020 (i.e., the first wave) and have not been updated since then. By May 31, 2020 approximately 11,500 COVID-19 cases were confirmed in South Korea [14, 23].
- Some locations visited by patients are not recorded due to privacy concerns. Consequently, patient infection information and route data do not always coincide. For example, there are patients that infect each other even if their routes do not cross. This may happen for patients of the same household (locations where people live are rarely logged).
- Patient and route data may be incomplete (i.e., some attributes are occasionally missing, such as the type of locations visited by some patients) and require manual completion before analyzing the data set.
- There is route data for only a portion of the patients. Patient movement has been logged only for the 15% of all confirmed cases by May 31.
- The KCDC logs do not contain a complete picture of all different factors affecting the disease spread. For example, these logs have no information on the number of people living in a single residence, or on behaviors of healthy individuals. The length of time a patient spends at a particular location in their route is also not recorded.

We adopt different data discovery strategies to address the above challenges. We manually retrieve certain missing attributes: in the case of patient routes with missing location type (e.g., store, school, hospital, airport), we use the provided geographical coordinates to retrieve the visited location and identify its type. Regretfully, some missing data are not possible to recover.

Provided that the mobility of only the 15% of confirmed patients are logged in detail, we assume that the mobility of all patients is independent and identically distributed to the patients with detailed logs. We contend that while detailed logs provide data of statistical significance, their usage introduces some unavoidable bias towards the percentage of patients who voluntarily shared more information than others. Statistical information derived from histograms (i.e., processed data) fill-in the gaps of missing information and can be used as input of patient activity in the simulation.

Because there is still much information unavailable in the KCDC logs that may better help us understand the spread of the disease in an urban environment, we also analyze data sets detailing human mobility in Berlin, Germany [25]. Both data sets contain detailed information on the routes of individuals, such as distance travelled, unique locations visited, and overlapping routes showing potential contact. These data sets still have several key differences. The KCDC data sets contain information on COVID-19 cases, whereas the Berlin data has information on healthy individuals. On the other hand, the Berlin data sets contain detailed information on important factors that affect disease spread, such as household size and time spent at a location. Using the parallels between these datasets, we examine the opportunity to cross-fertilize the Seoul data with Berlin data.



(a) South Korea. Blue points are hotspots.

(b) Seoul: Gangnam (blue) and Seocho (green) districts.

(c) Top-10 most visited locations.

(d) Movements across districts.

Figure 1: Seoul: heat maps of most visited locations, most visited location types, and movement between two districts.

We use logs and histograms to feed our tool, GeoSpread, an extended version of GeoMason [32]. GeoMason is a tool that uses agent-based models (ABM) and geographic information systems (GIS) and has been used to study disease outbreaks (e.g., cholera is studied using this tool in [12]). We simulate interactions of thousands of people in the Gangnam and Seocho districts of Seoul on roads and in buildings to investigate the COVID-19 outbreak in the largest metropolis of South Korea. We validate the results of the simulations with the ground truth derived from the KCDC logs. The GeoSpread tool offers a flexible model based on real-world COVID-19 spread information and can be used to facilitate evaluation of different mitigation measures and patient behaviors. Here, we use this processed data in the form of histograms (and also make them available to the community) [30]. Our contributions are:

- **Data Discovery**: We analyze and connect data from various KCDC logs to extract information on patient movements (Sections 2). Missing information is manually retrieved, when possible.
- **Statistical Analysis**: We provide statistical analysis of population movements and habits in the form of histograms.
- **Cross-fertilization:** We investigate similarities between the Seoul and Berlin data sets seeking common patterns. Leveraging this information, we cross-fertilize to incorporate useful information from the Berlin data sets which are unavailable in the Seoul data (e.g., travel speed, transportation means, household size).
- **GeoSpread**: We parameterize an ABM that uses the cross-fertilized data as input, see Section 4, and outline its flexibility to capture a variety of conditions. The simulation tool, GeoSpread, and processed data are open sourced [30].
- **Model Validation with Real Data**: The simulation model is validated and discussed in Section 5.

2 THE KCDC DATA

The data sets [21] used in this paper contain data collected by the KCDC from January 20, 2020, to May 31, 2020. The PatientInfo and PatientRoute data sets contain information and routes of COVID-19 patients in Seoul, respectively. The number of (healthy and sick) people moving across Seoul districts are also provided in the SeoulFloating data set. This data has been collected using the Big Data Hub of SK Telecom, a Korean wireless operator.

PatientInfo data set. This data set provides epidemiological data of COVID-19 patients. Each entry provides the *patient_id*, their gender and age, their provenance (*country*, *province*, and *city*), whether they have been infected in a known case (*infection_case*) and the ID of the patient that infected them (*infected_by*), the number of people that the patient came in contact with (*contact_number*), and the date of their first symptoms (*symptom_onset_date*). This data set is also described in [20].

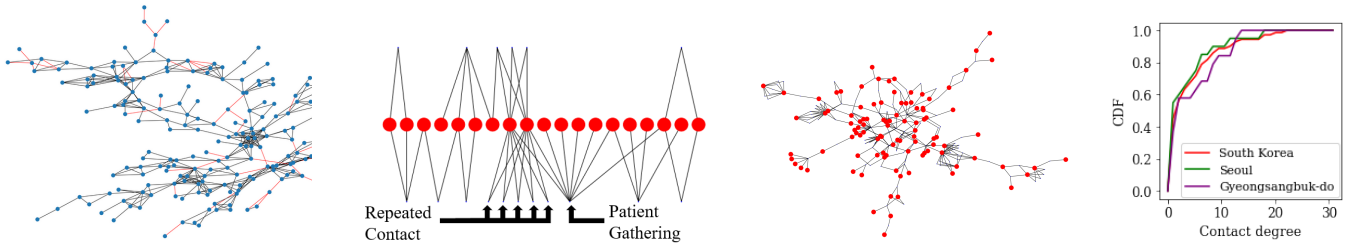
PatientRoute data set. This data set contains entries of unique South Korean COVID-19 patients. A location is identified by its *latitude* and *longitude*. *province*, *city*, and *type* (e.g., airport, hospital, store) of each location are also provided. The attribute *type* of almost 30% of entries is set to *etc* (i.e., locations that cannot be identified using the rule-based approach of [21]). We manually look for their type using their geographical coordinates and OpenStreetMap [2]. Each entry also contains the patient (identified by *patient_id*, the same as in the PatientInfo data set, and by *global_num*, another ID used only in this data set) that visited the location on a specific *date*. The time spent in the location is not available. Locations visited by a patient in a single day are logged in chronological order.

SeoulFloating data set. This data set provides hourly data of people moving across Seoul districts. Collected data are grouped by *gender*, *age*, and *district* and allow visualizing the movement of people in Seoul during this period. Age is provided at the decade granularity for people in their 20s through 70s. No information is provided for children or for people who are 80 or older. As a result, it is not possible to conclude on infections at education facilities or directly model mitigation measures that include school closings. This data set reports data on the *entire* Seoul population, not just the COVID-19 patients, and only considers those with cell phones.

2.1 Data Discovery

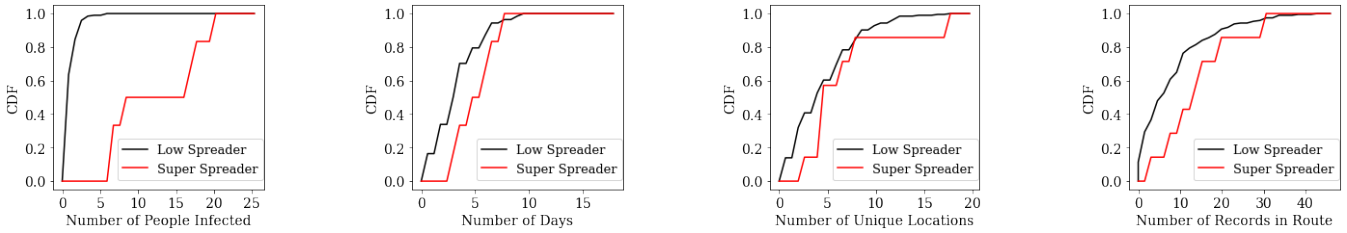
Here, we discuss what we extract from the data sets and how it is used to parameterize GeoSpread. All input parameterization data for GeoSpread is given in the form of distributions.

Visited Locations. Figs. 1(a) and 1(b) depict heat maps of the most visited locations in South Korea and Seoul, respectively, showing where COVID-19 outbreaks are more likely to happen. Seoul is the city with the most visited locations. Within Seoul, the south-west and south-east areas are those with more patient routes. The financial district and company headquarters are located in the south-west part of the city. The south-east region corresponds to the Gangnam district, outlined in blue in Fig. 1(b). Many shopping and entertainment centers are located in Gangnam. Fig. 1(c) shows the ten most visited facilities in Seoul, with *Hospital* being the first one. This is mainly due to the Korean data set being obtained during the COVID-19 pandemic by monitoring sick people. No information about schools is available since this data set monitors only people in their 20s through 70s. The scarcity of logged residential facilities is



(a) Patient connections (partial). (b) Partial hypergraph of patient contact. (c) Partial hypergraph cluster. (d) Contact degree CDF.

Figure 2: Seoul Patient contacts.



(a) People infected. (b) Logged days. (c) Unique locations visited. (d) Total locations visited.

Figure 3: Super spreader analysis in Seoul.

due to privacy concerns. Finally, Fig. 1(d) illustrates the movement of population between two neighboring districts, Gangnam and Seocho that we use later.

Patient Connections. Fig. 2(a) presents a subgraph of patient connections discovered by linking the PatientRoute and PatientInfo data sets. To improve visibility, we only present a small portion of the entire graph. Here, nodes depict patients, black edges connect patients that visited the same place during the same day from the PatientRoute data set, and red edges represent the virus spreading information obtained from the PatientInfo data set (i.e., *infected_by* attribute). Some red edges do not overlap with black edges. This means that, even if one of the two nodes connected by the red edge infected the other, no connections (i.e., visits to the same location during the same day) have been recorded in the data set. The node degree in Fig. 2(a) shows the contact degree among patients and illustrates visually the complexity of the problem.

Patient connections can also be visualized as a hypergraph to capture how many times patients come into contact and at what locations. An example can be seen in Fig. 2(b) where a node represents a patient and a hyperedge represents the connection between any number of patients who met at a specific location on a specific date. Visually, a hyperedge is shown as an edge that branches to connect two or more patients. This allows us to look at gatherings of groups of people, rather than just the binary relationship of whether or not two individuals came into contact. Clusters of cases in Seoul can be seen in the hypergraph in Fig. 2(c).

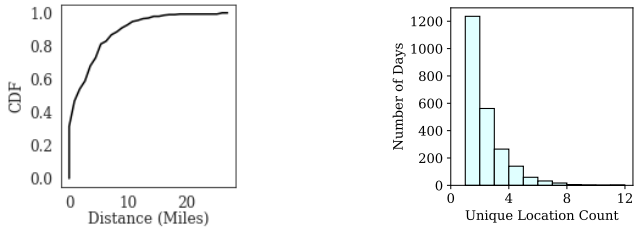
Finally, Fig. 2(d) shows a summary view of patient connections: the contact degree CDF of all patients for the entire dataset. Three CDFs are shown: one for the whole South Korea, one for Seoul, and another one for the Gyeongsangbuk-do province. Interestingly, all CDFs have a similar shape. High contact degrees indicate potential super spreaders (i.e., patients that infect many other people). People who come into contact with many others are not necessarily super spreaders since it is unknown whether they were sick or healthy when contact occurred. Because of this, further analysis is required to determine whether or not a patient is a super spreader.

Super Spreaders. The data set shows a mix of super spreaders (i.e., people who infected more than six people) and low spreaders, who infected six or fewer people[1]. Using this classification of patients based on the number of people they infect, we discover different behaviors of super/low spreaders, shown in Fig. 3. Super spreaders account for 3.59% and low spreaders account for the remaining 96.41% of patients.

Fig. 3 presents CDFs of the number of people infected by an individual, the number of days in the log that the individual appears, the unique visited locations, and the total number of visited locations. The CDFs in this figure indicate that, in general, super spreaders tend to be active for more days, visit more unique locations, and have longer routes than low spreaders. The figure shows that all super spreaders in the data set are active for three or more days and visit three or more unique locations. Some of these super spreaders are active for up to 19 days and visit up to 18 unique locations with route lengths of up to 31 locations.

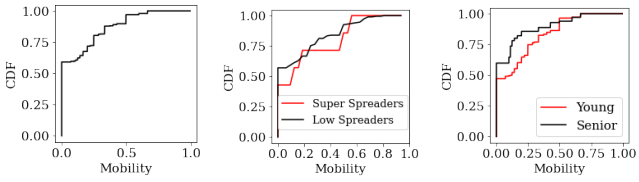
Daily Distance and Patient Mobility. With some exceptions, people mostly travel short distances and visit only a few locations each day. The CDF of the daily traveled distance is shown in Fig. 4(a). Intuitively, the more places a patient visits, the higher their mobility is. Fig. 4(b) shows the day count of unique locations reached by the patients in the data set: for 2,063 days (88.9% of days) a typical patient visits 1–3 locations, while for 258 days (11.1%) more than 3 unique locations are visited. Looking at the mobility of individual patients, there are days where they exhibit high mobility and days

[1]We define a "super spreader" as someone who infects at least 6 people in order to obtain the most noticeable difference in patient behavior (number of locations, number of days, number of records).



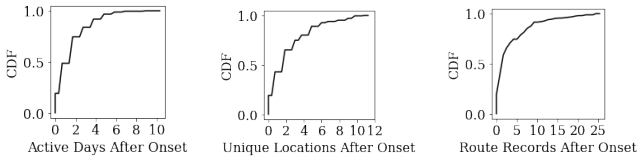
(a) Distance CDF. (b) Unique locations.

Figure 4: Daily distance and visited locations in Seoul.



(a) Mobility CDF. (b) Low/super spreaders. (c) Young vs. seniors.

Figure 5: Patient mobility in Seoul.



(a) Active days. (b) Unique locations. (c) Total locations.

Figure 6: Behavior after first symptoms in Seoul.

where they move significantly less. This leads us to a more usable definition of mobility as a function of different time periods (days).

Defining a *high mobility day* as a day during which a patient visits at least L locations, the *mobility of a patient* is given as the ratio of the patient high mobility days to all logged days for this specific individual. Note that this is not the only way to define mobility. For simulation purposes (see Section 4), this definition provides a practical way to capture mobility with a probability. Based on the histogram shown in Fig. 4(b), days with $L \leq 3$ are considered of low mobility. The CDF of patient mobility using the above definition is depicted in Fig. 5(a). The figure shows that 57.6% of patients never visit more than 3 locations in a day.

Different classes of patients have different mobility. Fig. 5(b) shows the difference in mobility between super spreaders and low spreaders, while Fig. 5(c) illustrates mobility by age groups. Super spreaders and young people have higher mobility compared to low spreaders and seniors, respectively. For higher percentiles, the low spreaders have higher mobility than super spreaders due to the small number of super spreader agents in the KCDC data set.

Irresponsible Behaviors. Patients behave irresponsibly when they keep moving after the onset of their first COVID-19 symptoms, which facilitates the diffusion of the disease. We present how long sick people continue to show mobility after exhibiting symptoms, see Fig. 6. The figure shows that only the 20% of patients stop moving and isolate immediately after initial symptoms are observed. Some patients keep moving for more than a week after the onset of symptoms, see Fig. 6(a). They also visit many locations; Figs. 6(b) and 6(c) show the number of unique and total locations that sick patients visit after initial symptoms are observed.

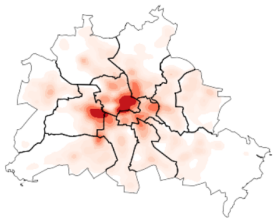


Figure 7: Heat map of the most visited Berlin locations.

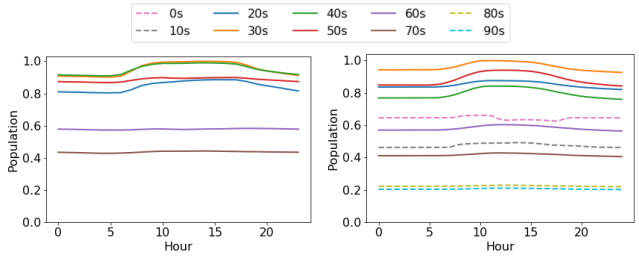
3 THE BERLIN DATA

In spite of the detailed data provided in the KCDC data sets, there is still a lot of unavailable information which is necessary for understanding how COVID-19 spreads in an urban environment. In this section, we compare distributions of different characteristics of human mobility from Seoul, South Korea with distributions from Berlin, Germany to determine whether similarities exist that allow for cross-fertilization. First, we focus on commonalities in movements of individual in the two urban environments.

The German data sets [25] contain movement logs obtained by monitoring people that visited Berlin *before* the COVID-19 pandemic, during business days and weekends. It provides demographic data of all monitored people, the public transport vehicles used by people for their movements, and the type and capacity of all visited facilities. Here, we consider movement logs collected during business days by observing people whose actions are located only in Berlin. Fig. 7 shows the most active district of Berlin, i.e., areas of the city that appear more frequently in the German data set.

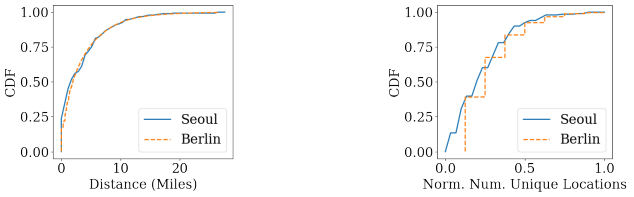
EventWeekdays data set. People's movements over 30 hours are logged in this data set, where almost 6 million activities have been recorded from start to finish. For each entry, the *timestamp* (in seconds) is provided as well as the *type* of entry (i.e., *start* for activities that begin or *end* for activities that are completed) and the *person* to which the activity is associated. For this analysis, we use only logs from people that never leave Berlin during the observation period, i.e., 67% (3,919,990) of this data set. All activities in this data set represent a visit to a facility or the usage of public transport. In the former case, *facility_id* and *link_id* allow associating the entry to a venue, while the *actType* attribute specifies the type of activity performed in that location (e.g., home, school, work). When an entry refers to a transport activity, it provides the *vehicle* attribute with the ID of the vehicle that is used for moving.

Demographic data set. This data set contains information about each person (i.e., more than 1.2 million people) whose activities have been logged in the EventWeekdays data set. Specifically, *age* and *gender* for all people is provided as well as their *home_district*, *home_id*, and home *coordinates*. The *home_district* attribute contains one of the 401 administrative districts of Germany. Here, since we focus just on Berlin, a metropolitan city like Seoul, we consider people who do not leave Berlin during the observation period. Therefore only 55% (671,256) of the original data set is analyzed. The *home_id* attribute associates each person in the data set to



(a) Seoul (b) Berlin

Figure 8: Daily presence of different age groups, with the y-axis normalized for easy comparison.



(a) Daily distance (in miles). (b) Number of unique locations.

Figure 9: Distance and location count comparisons.

their home-place, while the *coordinates* attribute allows placing each building on a map with an accuracy of 500 meters.

FacilityType data set. This data set contains all 631,290 facilities visited in the EventWeekdays data set. The 75% (476,572) of these venues are located in Berlin. Univocal *id* and *link_id* attributes are associated to all entries of this data set for the identification of each facility. *Coordinates* (using the *EPSG:25832* coordinate reference system) are also associated to each venue. This allows placing each venue on a map. *Functions* (e.g., home, school, work) might also be associated to each facility depending on the activities that are carried out within that venue. Note that multiple functions can be associated to the same building. For each function of a facility, a *capacity* attribute (i.e., the maximum number of people that can occupy the facility doing the same activity) is also provided.

PublicTransport data set. This data set records vehicles in public transport. An *id* and a *type* (e.g., bus, metro, tram) are associated to each vehicle. Although almost 1 million vehicles are recorded, only 42,725 of them (i.e., 4%) are located in Berlin. However, many people use public transport for moving in Berlin and 1,791,061 movements are completed using one of these vehicles.

3.1 Similarities Across the Two Data Sets

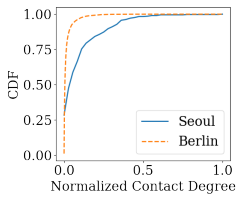
Both KCDC and German data sets allow retrieving information and attributes that can be used for comparing movement habits of people living in Seoul and Berlin. In the following, these features are analyzed and described.

Population Age. Fig. 8 depicts Seoul and Berlin population floating during a business day. Data is grouped based on people's age with decade granularity. The SeoulFloating data set in the KCDC data sets monitors people that are in their 20s through 70s for both healthy and sick individuals. As a result, this data set is valuable for comparison to the German data sets. We investigate the population habits from January 1, 2020, to May 31, 2020 by age group for comparison to movements in Berlin, see Fig. 8(a). Fig. 8(b) provides information of people living in Berlin that are also younger than 20 or older than 79 (see dashed lines) as well. Since the number of observations in the two data sets is different, the population of Seoul and Berlin is normalized over the maximum number of people observed in both cities. Overall, Seoul and Berlin experience similar people floating dynamics. Specifically, the normalized number of people that are between 60 and 79 is similar in both cities and it tends to be flat during the day. Adults and young-adults of both cities show also similar dynamics, with the only exception of people in their 40s and 50s. The normalized number of people that are between 40 and 49 is larger in Seoul than in Berlin, but they float similarly in both cities, i.e., they increase around 6 AM and decrease after 3 PM. The normalized number of people in their 50s that live in Seoul and Berlin is similar (i.e., 0.85). While such a number increases during the morning in Berlin (then it decreases in the evening), it does not change much in Seoul. Looking at the Berlin data, we observe that there are not many people older than 80 and that number does not change during the day. The only age group whose number decreases in the morning and increases in the evening is the one representing kids younger than 10.

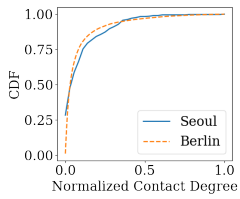
Daily Traveled Distance. Fig. 9(a) plots the cumulative distribution function (CDF) of daily traveled distance (in miles) for people living in Seoul and Berlin. The two CDFs match closely implying that Korean patients and Berlin inhabitants travel similar distance daily. Specifically, 75% of people move less than 5 miles and only a small percentage of the population travels more than 15 miles.

Unique Locations. Fig. 9(b) depicts the daily number of unique locations visited by all monitored people in Seoul and Berlin. The two distributions (i.e., the x-axis) in Fig. 9(b) are normalized over the maximum number of unique visits for each city. Since the KCDC data set monitors patients' movements for different days, the daily number of unique locations visited by each patient in Seoul is averaged over their number of active days. For this reason, non-integer values are also observed when looking at the unique location distribution of Seoul in Fig. 9(b) (i.e., blue line). The number of unique locations in Berlin is not averaged over the number of active days since people are monitored for only 30 hours in the German data set. Hence, the number of unique locations visited by Berlin people is an integer value and the distribution (i.e., yellow line) is discrete. Nevertheless, the distributions of unique locations visited by people living in Seoul and Berlin show similar trends.

Contact Degree. The analysis of how many people are met by each person logged in the two data sets (i.e., contact degree) allows discovering relationships that might facilitate the spread of the virus. Intuitively, the more people a COVID-19 patient meets, the faster the virus can spread. In the KCDC data set, no data is provided about the time a patient visits a facility, only the date is known. For this reason, their contact degree is computed as the number of other people that visit their same facilities on the same day. People's movements in the German data set are provided with their exact time. This enables a more precise evaluation of the contact degree since we can determine who is in the same facility during the same period. Due to the large number of people monitored in the German data set (i.e., more than 1.2 million), the contact degree in Berlin happens to be greater than 12,000 for a small percentage of individuals. This makes impossible to compare the Seoul and Berlin contact degrees, see Fig. 10(a), even after normalizing both

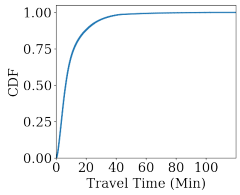


(a) With outliers.

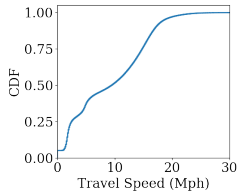


(b) Without outliers.

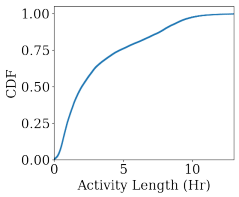
Figure 10: Contact degree (normalized) of Seoul patients and Berlin inhabitants. Outliers are discarded in (b).



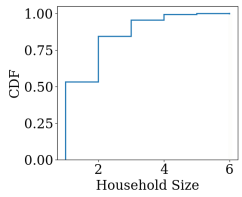
(a) Travel Time



(b) Travel Speed



(c) Activity Length



(d) Household Size

Figure 11: Unique characteristics of the Berlin Data set.

distributions over the maximum number of contacts. For this reason, we consider Berliners with a very large contact degree to be outliers and discard their interactions by considering only people whose contact degree is within the 99th percentile. To compare the Seoul contact degree with the Berlin one, in the latter case we do not account for contacts on public transportation since this makes the contact degree significantly larger. Results are depicted in Fig. 10(b), where it is visible that normalized contact degrees of people living in Seoul and Berlin match. This implies that, the chances for the virus to spread in Seoul and Berlin are similar.

3.2 Unique Attributes of the Berlin Set

The prior analysis of the KCDC and German data sets show that the two cities share many characteristics, however, both data sets also contain a wealth of unique characteristics. While both data sets contain information about distance travelled, the German data sets contain additional information about travel time and speed. These distributions are seen in Fig. 11(a) and Fig. 11(b). One notable drawback of the KCDC data sets is the lack of fine-grained time stamps on patient routes. The KCDC logs only contain the date and the order in which locations were visited by that patient on that date. The German data has detailed time stamps and records of the amount of time spent performing a specific activity (e.g., shopping, work, etc.). Fig. 11(c) shows the distribution of activity lengths in the German data sets. Because the KCDC data sets only contain information about individuals with COVID-19, and route information is often incomplete due to privacy concerns, no information can be extracted about the number of people living together. On the other hand, household size is available in the German data sets. This information is shown in Fig. 11(d). These unique characteristics have the potential to cross-fertilize the information extracted from the KCDC data sets, and aid us in modeling and understanding different factors of human mobility that affect virus spread.

4 AGENT-BASED MODEL

In this section, we show how to parameterize a simulation based on GeoSpread [30], our extended version of GeoMason [32] using the characterization presented in Sections 2 and 3. The following attributes are set during initialization:

(1) *Infection status.* One or more random agents are selected as the initial case(s).
(2) *Position.* Agents are randomly placed on a road.
(3) *Speed.* Speed determines how fast an agent moves from one location to another and is selected according to a distribution: we sample from the speed distribution from the Berlin data set characterization to select an agent's speed.
(4) *Type of spreaders.* We define two classes of spreaders: 3.59% are super spreaders and 96.41% are low spreaders.
(5) *Mobility.* We use the mobility of super spreaders and low spreaders depicted in Fig. 5(b) to model patient mobility.
(6) *Home district and home building.* We assign agents a home building within their home district based on Fig. 1(d). Agents select destination buildings in the simulation depending on how agents move between these districts, see Fig. 1(d).
(7) *Family size and family members.* Agents are assigned family members who all live together in a home building. While at home, agents are able to infect family members they are in contact with. The number of individuals in a family is determined by sampling from the household size distribution in Berlin described in Fig. 11(d).

In addition to the mobility distribution of super spreaders and low spreaders, the CDF of daily traveled distance in Fig. 4 is also used to determine the distance to a destination. The location type an agent will travel to is determined by Fig. 1(c). The amount of time agents spend at a location is determined according to Fig. 11(c). Simulation time is defined by cycles. In each simulation cycle, agents outside a building move along the road towards their destination; agents inside a building can choose to stay or leave, based on their mobility. Agents with high mobility have a high probability to leave the building and visit many others. Note that agents stay in a building for at least 15 minutes in order to meet the definition of close contact [9]. If multiple agents are inside the same building, they may infect each other with a certain probability.

When infection happens, the agent state changes from healthy to infected. We assume the outdoor infection probability to be negligible. Given the probability of infection inside a building, α, and the number of infected agents in the building, n, the probability of a healthy agent to be infected by a contact within the building is $Pr(\mathit{infection}) = 1 - (1 - \alpha)^n$. Note that this equation for the probability of infection is nominal. Any model can be used to capture the viral load: the total number of people in the location, the duration of interaction among individuals, the square footage of the room, its air circulation, wearing a mask or not, see [24].

It takes 1–14 days for patients to show symptoms after infection according to the WHO [35]. We therefore use a Uniform distribution between 1 and 14 days to transition from infected to symptomatic. A uniform distribution is again nominal here, one could easily use any distribution, e.g., a lognormal distribition with its peak set to 5 to capture a more realistic scenario consistent with hard data. Since some patients continue to move even after showing symptoms, as seen in Fig. 6, we use the CDF in Fig. 6(a) to determine the number of active days after their first symptoms. After each infected person exhausts their active days after infection, they are isolated.

Consistent with infectious disease simulation studies [22], we set the simulation cycle to 5 minutes. The simulation stops either when all agents are infected or after a number of cycles defined by the user. Contact degree and the number of unique locations visited are used for validation since these are not input parameters.

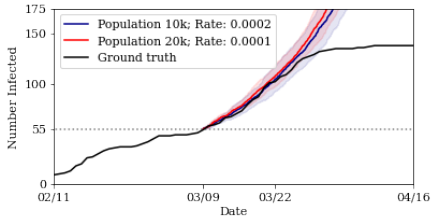
We simulate the COVID-19 outbreak in the Seocho and Gangnam districts, i.e., the region of Seoul with the most hotspots, see Fig. 1(b). This area has 11,438 road intersections and 7,043 buildings. Roads and buildings are placed in the simulated area as described in [3], a collection of GIS data with regard to Seoul. GeoSpread loads the GIS data (e.g., roads, road intersections, buildings) stored in a shapefile format, i.e., a file that stores geometric locations and their attribute information. Although the longest distance we observe in the PatientRoute data set in Seoul is 30 miles, the longest distance between two buildings in the simulated Gangnam district is 7.06 miles. Therefore, we normalize the maximum distance to 3.53, which is half of the longest distance in the simulated area, to ensure a valid building selection as the agent's destination. In the Gangnam district there are 604,586 people and a total of 7,043 buildings. We do not have any information on building stories, entries, or number of rooms. This information is crucial, especially for apartment buildings, where multiple people can be inside the same building at the same time without contact. To address this lack of information, we limit the population in our simulations. We validate parameter choices against ground truth data in Section 5.

5 MODEL VALIDATION

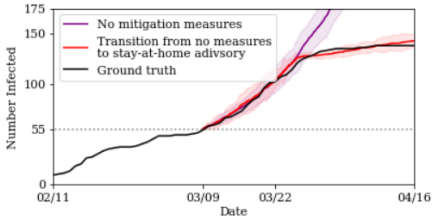
Fig. 1(d) shows the residents in Seocho and Gangnam that have been infected, the figure also illustrates the movement between the two districts. We use this information to parameterize the simulation. During the initialization phase, we separate the agents into Gangnam residents (70.4% of the population) and Seocho residents (29.6% of the population). Next, we retrieve the distributions of agent mobility and spreader types from the data set for residents of each district to set their attributes. The probability of a resident staying or leaving their home district follows Fig. 1(d).

Since two districts are considered here, starting with only one infected agent in one of the two areas could bias results. Here, we start the simulation with 55 infected agents, i.e., the number of infections observed from the data set on March 9, 2020, proportionally assigned to agents in the two districts (29.6% in Seocho, 70.4% in Gangnam). Simulations starting at any time earlier or around March 9, result in similar trends.

Fig. 12(a) depicts the number of COVID-19 cases in the Gangnam and Seocho districts observed from the data set (black line, ground truth) and simulation (red and blue lines). The ground truth



(a) Infections: simulation vs. ground truth.



(b) Validation of mitigation measures. Strong Social Distancing campaign starts on March 22.

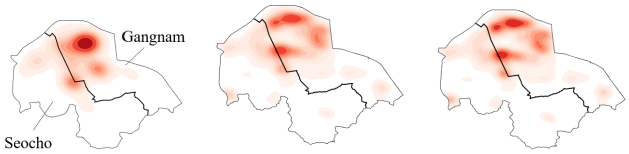
Figure 12: Validation. Results are presented with 95% confidence intervals (shaded areas).

line illustrates the COVID-19 outbreak in the two districts. At the beginning of April, the curve flattens. This is likely due to effective counter-measures executed in Seoul, especially the Strong Social Distancing Campaign which began on March 22. Consistent with the COVID-19 incubation timeline, the effectiveness of the Strong Social Distancing Campaign does not show immediately, but after the beginning of April. Our simulation in Fig. 12(a) does not model the effect of social distancing campaign so it is expected not to capture the knee of the ground truth curve.

We align the beginning of simulation data to the time of 55 infection cases in the ground truth, since this is the starting point of the simulation. The two simulation lines in Fig. 12(a) (their 95% confidence intervals are given by the shaded areas) closely follow the ground truth: the simulation of population 10,000 with infection rate 0.004 and the simulation of population 20,000 with infection rate 0.002 are in excellent agreement with the ground truth from March 26, 2020 to April 5, 2020, when the effects of any counter-measures are not discernible yet. The overlap of two simulation cases with the ground truth validates the simulation.

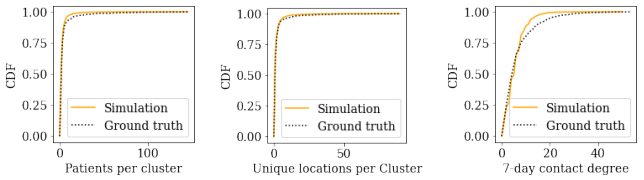
We note in Fig. 12(a) an interesting relationship between population and infection rate: when population is doubled, dividing the infection rate in half gives similar simulation outcomes. This observation meets the results in the generic simulation that higher population leads to faster spreading of the COVID-19 virus, while lowering the infection rate slows down virus spreading. We conclude that we can use a "limited" population with an adjusted infection rate to efficiently (yet accurately) model larger populations.

As further validation, we simulate the effects of applying a stay-at-home advisory mid-simulation in order to capture the effects of the mitigation measures taken in Seoul on March 22 – the Strong Social Distancing Campaign. Fig. 12(b) depicts the simulation results (with 95% confidence intervals) against ground truth. In this experiment, we begin with no mitigation measures and apply a stay-at-home advisory when the Strong Social Distancing campaign is enacted. After applying the stay-at-home advisory mid-simulation,



(a) Ground truth. (b) Simulation: 10K. (c) Simulation: 20K.

Figure 13: Hotspots in the data set (ground truth) and model.



(a) Patients per cluster. (b) Unique locations. (c) 7-day contact degree.

Figure 14: Validation of clusters and contact degree.

GeoSpread also exhibits a flattening trend, which is consistent with ground truth. This highlights the ability of the model to capture what-if scenarios of mitigation measures.

Next, we focus on hotspot locations. In Fig. 13(a), we present the heat map of most visited locations in the Gangnam and Seocho (ground truth). The most visited areas are in the northern part of Gangnam and across the border between the two districts. These hotspots correspond to the density of commercial buildings in these areas. Fig. 13(b) and 13(c) show the heat maps of visits for simulated populations of 10, 000 and 20, 000, respectively, and are consistent with ground truth.

Additionally, we examine properties of clusters (i.e. outbreaks) in the ground truth KCDC logs and the simulations in 7-day sliding windows. Fig. 14(a) depicts the number of patients seen in infection clusters. Fig. 14(b) shows the number of unique locations visited by patients in infection clusters. Finally, we can see the contact degree between patients in Fig. 14(c). The similarity in these curves further validates the accuracy of the simulation.

5.1 Model Limitations

Although the model is validated using ground truth, incomplete and/or missing data limit its generalization. Limitations include:

First wave data. This data is from the first wave of the disease in South Korea. With South Korea having one of the best responses to the disease globally, the mobility patterns reflect inevitably cultural and demographic characteristics as well as policy decisions.

Privacy concerns. The KCDC data set is anonymized and no sensitive data of monitored patients can be retrieved. No data about the underage population is provided as well as movements of patients from/to their private homes. To address this, we examine distributions from the German data set regarding household size, but this still limits the scenarios that can be analyzed, e.g., the impact of school closures. Note also that the per-patient mobility information (and its statistics) are retrieved from the PatientRoute data set. We have no way to evaluate how mobility statistics changed during other waves of COVID-19.

Transportation. The KCDC data set does not show the transportation mode of patients. We overcome this limitation by using distributions from the German data set.

6 RELATED WORK

COVID-19 has been studied extensively in recent months. COVID-GAN [7] allows generating human mobility traces when different real-world conditions apply (e.g., local policies). Epidemiological/clinical data are collected in [28] via patient interviews to study the spread of the virus in three Singapore *clusters*, this approach by its nature is difficult to scale. A contact tracing system based on blockchain is proposed in [27]. A numerical simulation is adopted in [11] to evaluate the efficiency of a test-trace-and-isolate strategy in containing the pandemic in Germany. A co-location model is used in [37] to study the spread of SARS-CoV-2 with limited data.

Agent-based models (ABMs) are a simulation-based alternative of mathematical models that incorporate human interactions [19]. ABMs are typically used for modeling pedestrian movements, human mobility during rare events (e.g., natural disasters), resource usage, and to study the spread of diseases [12, 16, 26, 33]. The spread of influenza in British and American households, schools, and workplaces is modeled in [13] using census and land use data as well as air travel patterns. This work considers only international population movements. ABMs parameterized by census data have been used to capture the spread of COVID-19 in Australia [10, 29]. Using census and age-distribution data from Germany and Poland, Bock et al. [8] investigate the efficiency of mitigation strategies by accounting for interactions within households. Census ABM-based frameworks have been used to simulate the COVID-19 outbreak [17], evaluate the efficiency of contact tracing [6], face masks [18], and testing strategies [34]. Kim et al. [22] use synthetic, location-based social network data to study how social behaviors affect the virus spread. Geo-located data from social networks (i.e., Twitter) are used in [31] to identify hotspots that facilitate the spread of infectious diseases (i.e., Dengue). ABMs are used to model the spread of SARS-CoV-2 in small areas: crowded areas of supermarkets [36] and university campuses [15]. *Differently from our approach, no fine-grained movement data is used in any of the above works. The above models are parameterized using census or synthetic data while population movement habits are captured at a coarse granularity.*

Müller et al. [24, 25] use an ABM parameterized with synthetic mobility traces (originally generated from mobile phone data) to study the COVID-19 outbreak in Berlin and analyze the effect of mitigation measures. This work is the closest to ours but uses no detailed statistics on agent mobility during the pandemic.

7 CONCLUSIONS

In this paper, we extract human movement habits and dynamics of real COVID-19 patients from the KCDC data set. We enrich this analysis by analyzing and discussing detailed mobility data in Berlin, Germany. The mobility information and statistics are used to tune our ABM tool GeoSpread and investigate the COVID-19 outbreak in two districts of Seoul. Agent movements and behaviors are simulated using the statistics of actual human movements, other structures (e.g., networks or graphs) are not required. The proposed approach allows investigating scenarios under different circumstances to identifying mitigation strategies.

REFERENCES

[1] 2020. Google Maps. https://www.google.com/maps/. [Online; 2021-01-13].
[2] 2020. OpenStreetMap. https://www.openstreetmap.org/. [Online; 2021-01-13].

[3] 2020. OSM extracts for Seoul. https://download.bbbike.org/osm/bbbike/Seoul/. [Online; 2021-01-13].

[4] 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. [Online; 2021-01-13].

[5] 2021. COVID-19 Open Data. https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-open-data. [Online; 2021-10-29].

[6] Jonatan Almagor and Stefano Picascia. 2020. Exploring the effectiveness of a COVID-19 contact tracing app using an agent-based model. *Scientific reports* 10, 1 (2020), 1–11.

[7] Han Bao, Xun Zhou, Yingxue Zhang, Yanhua Li, and Yiqun Xie. 2020. COVID-GAN: Estimating Human Mobility Responses to COVID-19 Pandemic through Spatio-Temporal Conditional Generative Adversarial Networks. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 273–282.

[8] Wolfgang Bock, Barbara Adamik, Marek Bawiec, Viktor Bezborodov, Marcin Bodych, Jan Pablo Burgard, Thomas Goetz, Tyll Krueger, Agata Migalska, Barbara Pabjan, et al. 2020. Mitigation and herd immunity strategy for COVID-19 is likely to fail. *medRxiv* (2020).

[9] CDC. 2020. Public Health Guidance for Community-Related Exposure. https://www.cdc.gov/coronavirus/2019-ncov/php/public-health-recommendations.html. [Online; 2021-01-13].

[10] Sheryl L Chang, Nathan Harding, Cameron Zachreson, Oliver M Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature communications* 11, 1 (2020), 1–13.

[11] Sebastian Contreras, Jonas Dehning, Matthias Loidolt, Johannes Zierenberg, F Paul Spitzner, Jorge H Urrea-Quintero, Sebastian B Mohr, Michael Wilczek, Michael Wibral, and Viola Priesemann. 2021. The challenges of containing SARS-CoV-2 via test-trace-and-isolate. *Nature communications* 12, 1 (2021), 1–13.

[12] Andrew Crooks and Atesmachew Hailegiorgis. 2014. An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software* 62 (2014), 164–177.

[13] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. 2006. Strategies for mitigating an influenza pandemic. *Nature* 442, 7101 (2006), 448–452.

[14] Korea Centers for Disease Control & Prevention. 2020. Coronavirus Disease-19, Republic of Korea. http://ncov.mohw.go.kr/en/. [Online; 2021-01-13].

[15] Philip T Gressman and Jennifer R Peck. 2020. Simulating COVID-19 in a university environment. *Mathematical Biosciences* 328 (2020).

[16] Kathryn H Jacobsen, A Alonso Aguirre, Charles L Bailey, Ancha V Baranova, Andrew T Crooks, Arie Croitoru, Paul L Delamater, Jhumka Gupta, Kylene Kehn-Hall, Aarthi Narayanan, et al. 2016. Lessons from the Ebola outbreak: action items for emerging infectious disease preparedness and response. *EcoHealth* 13, 1 (2016), 200–212.

[17] Masoud Jalayer, Carlotta Orsenigo, and Carlo Vercellis. 2020. CoV-ABM: A stochastic discrete-event agent-based framework to simulate spatiotemporal dynamics of COVID-19. *arXiv preprint arXiv:2007.13231* (2020).

[18] De Kai, Guy-Philippe Goldstein, Alexey Morgunov, Vishal Nangalia, and Anna Rotkirch. 2020. Universal masking is urgent in the covid-19 pandemic: Seir and agent based models, empirical validation, policy recommendations. *arXiv preprint arXiv:2004.13553* (2020).

[19] Rebecca A Kelly, Anthony J Jakeman, Olivier Barreteau, Mark E Borsuk, Sondoss ElSawah, Serena H Hamilton, Hans Jørgen Henriksen, Sakari Kuikka, Holger R Maier, Andrea Emilio Rizzoli, et al. 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. *Environmental modelling & software* 47 (2013), 159–181.

[20] Jimi Kim, Seojin Jang, Woncheol Lee, Joong Kun Lee, and Dong-Hwan Jang. 2020. DS4C Patient Policy Province Dataset: a Comprehensive COVID-19 Dataset for Causal and Epidemiological Analysis. In *Advances in Neural Information Processing Systems*.

[21] Jihoo Kim and JoongKun Lee. 2020. Data Science for COVID-19 (DS4C). https://www.kaggle.com/kimjihoo/coronavirusdataset. [Online; 2021-01-13].

[22] Joon-Seok Kim, Hamdi Kavak, Chris Ovi Rouly, Hyunjee Jin, Andrew Crooks, Dieter Pfoser, Carola Wenk, and Andreas Züfle. [n.d.]. Location-based social simulation for prescriptive analytics of disease spread. *SIGSPATIAL Special* 12, 1 ([n. d.]).

[23] Sun Kim and Marcia C Castro. 2020. Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020). *International Journal of Infectious Diseases* 98 (2020), 328–333.

[24] Sebastian A Müller, Michael Balmer, William Charlton, Ricardo Ewert, Andreas Neumann, Christian Rakow, Tilmann Schlenther, and Kai Nagel. 2020. A realistic agent-based simulation model for COVID-19 based on a traffic simulation and mobile phone data. *arXiv preprint arXiv:2011.11453* (2020).

[25] Sebastian A Müller, Michael Balmer, William Charlton, Ricardo Ewert, Andreas Neumann, Christian Rakow, Tilmann Schlenther, and Kai Nagel. 2021. Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data. *medRxiv* (2021).

[26] Yanbo Pang, Kota Tsubouchi, Takahiro Yabe, and Yoshihide Sekimoto. 2020. Intercity Simulation of Human Mobility at Rare Events via Reinforcement Learning. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 293–302.

[27] Zhe Peng, Cheng Xu, Haixin Wang, Jinbin Huang, Jianliang Xu, and Xiaowen Chu. 2021. P^2B-Trace: Privacy-Preserving Blockchain-based Contact Tracing to Combat Pandemics. In *SIGMOD '21: International Conference on Management of Data*. ACM, 2389–2393.

[28] Rachael Pung, Calvin J Chiew, Barnaby E Young, Sarah Chin, Mark IC Chen, Hannah E Clapham, Alex R Cook, Sebastian Maurer-Stroh, Matthias PHS Toh, Cuiqin Poh, et al. 2020. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *The Lancet* (2020).

[29] Rebecca J Rockett, Alicia Arnott, Connie Lam, Rosemarie Sadsad, Verlaine Timms, Karen-Ann Gray, John-Sebastian Eden, Sheryl Chang, Mailie Gall, Jenny Draper, et al. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine* (2020), 1–7.

[30] Anna Schmedding, Lishan Yang, Riccardo Pinciroli, and Evgenia Smirni. 2022. GeoSpread. https://github.com/akschmedding/GeoSpread. [Online; 2022-03-30].

[31] Roberto CSNP Souza, Renato M Assunção, Daniel B Neill, and Wagner Meira Jr. 2019. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 359–368.

[32] Keith Sullivan, Mark Coletti, and Sean Luke. 2010. *GeoMason: Geospatial support for MASON.* Technical Report. Department of Computer Science, George Mason University.

[33] Srinivasan Venkatramanan, Bryan Lewis, Jiangzhuo Chen, Dave Higdon, Anil Vullikanti, and Madhav Marathe. 2018. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics* 22 (2018), 43–49.

[34] Yingfei Wang, Inbal Yahav, and Balaji Padmanabhan. 2020. Whom to Test? Active Sampling Strategies for Managing COVID-19. *arXiv preprint arXiv:2012.13483* (2020).

[35] WHO. 2020. Q&A on coronaviruses (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses. [Online; 2021-01-13].

[36] Fabian Ying and Neave O'Clery. 2021. Modelling COVID-19 transmission in supermarkets using an agent-based model. *Plos one* 16, 4 (2021).

[37] Sepanta Zeighami, Cyrus Shahabi, and John Krumm. 2021. Estimating Spread of Contact-Based Contagions in a Population Through Sub-Sampling. *Proceedings of VLDB Endowment* 14, 9 (2021), 1557–1569.