# Epidemic Spread Modeling for COVID-19 Using Cross-Fertilization of Mobility Data

Anna Schmedding, Riccardo Pinciroli, Lishan Yang, *Member, IEEE*, and Evgenia Smirni, *Fellow, IEEE*

***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 for their limitations, we use detailed mobility data from Berlin, Germany after observing that mobility of individuals is surprisingly similar in both Berlin and Seoul. Using information from the Berlin mobility data, we cross-fertilize the KCDC Seoul data set and use it to parameterize an agent-based simulation that models the spread of the disease in an urban environment. After validating the simulation predictions with ground truth infection spread in Seoul, we study the importance of each input parameter on the prediction accuracy, compare the performance of our model to state-of-the-art approaches, and show how to use the proposed model to evaluate different *what-if* counter-measure scenarios.**

***Index Terms*—Data analysis, simulation models, individual-centric models, COVID-19, disease spread modeling, cross-fertilization.**

## I. Introduction

On March 11, 2020, the WHO[1] declared COVID-19 the first pandemic caused by a coronavirus. Since then, a tremendous amount of data has been collected to help public policy decisions that limit the spread of COVID-19. For example, Google[2] provides time-series data of infections at a coarse granularity (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 article, we use the data provided by the Korean Center for Disease Control (KCDC) and local governments during the first wave of the disease in South Korea. In contrast to the Google data, the KCDC data focus on *individual patients* and allow the development of an individual-centric model of the COVID-19 epidemic. Infected individuals are monitored[3] and their movements are logged using CCTV, cellphones, and credit card transactions. The KCDC records patient movements in plain text (i.e., natural language) without any unified rule. These logs 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) to extract geographical coordinates (i.e., latitude and longitude) and other data. The parsed logs are made publicly available [1] and being collected by KCDC are deemed trustworthy.

To the best of our knowledge, the KCDC logs are the only publicly available data that contain patient-centric information in great detail: they report on the patient mobility, i.e., traveled distance and the sequence of locations visited on a daily basis, the date of the onset of symptoms, whether and when the patient got in contact with other patients that are also diagnosed. This leads to our first research question, *RQ1*: What statistical information can be extracted by the KCDC mobility data to parameterize an agent-based simulation that models the spread of the disease? The KCDC logs are a valuable resource for studying the spread of COVID-19, yet they have limitations:

- The last version of the KCDC data set contains data collected up to May 31, 2020 (the KCDC data set has not been updated since then). By that date, approximately 11,500 COVID-19 cases were confirmed in South Korea [2], but only 35% of them have been logged into the data set.
- Some locations visited by patients (e.g., locations where people live) 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 when patients belong to the same household.
- 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 information for only a portion of the patients. Patient movement has been logged only for the 15% of all confirmed cases by May 31.

Manuscript received 8 July 2022; revised 16 January 2023; accepted 14 February 2023. Date of publication 24 February 2023; date of current version 1 September 2023. This work was supported in part by the National Science Foundation under Grants IIS-2130681 and IIS-1838022, and in part by the Ministry of University and Research under the PRIN project SEDUCE 2017TWRCNB and the PNRR project VITALITY (ECS00000041) Spoke 2 Astra. Recommended for acceptance by S. Mohanty. *(Corresponding author: Riccardo Pinciroli.)*

Anna Schmedding and Evgenia Smirni are with the Computer Science Department, William and Mary, Williamsburg, VA 23185 USA (e-mail: akschmedding@email.wm.edu; esmirni@cs.wm.edu).

Riccardo Pinciroli is with the Computer Science Department, Gran Sasso Science Institute, 67100 L'Aquila, Italy (e-mail: riccardo.pinciroli@gssi.it).

Lishan Yang is with the Computer Science Department, George Mason University, Fairfax, VA 22030 USA (e-mail: lyang28@gmu.edu).

Digital Object Identifier 10.1109/TBDATA.2023.3248650

[1] https://bit.ly/3izwIdL

[2] https://bit.ly/3H5YU1V

[3] https://bit.ly/3VMQvVm

- 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.

To compensate for the lack of information in the KCDC logs, we also analyze data sets detailing human mobility in German cities and districts [3]. These data sets contain detailed information on the routes of individuals, such as distance travelled, unique locations visited, and overlapping routes. The KCDC and German data sets still have several key differences. The KCDC logs contain information on COVID-19 cases, whereas the German data only contains information on healthy individuals. On the other hand, the German data sets contain detailed information on important factors that affect the disease spread, e.g., household size and time spent at a location by individuals. These observations lead to our second research question, *RQ2*: Can the Seoul data sets be cross-fertilized with German data by leveraging parallels between the two logs?

We illustrate that such cross-fertilization across the Seoul and Berlin logs is possible. Further, we show that cross-fertilized data can be fed into GeoSpread [4], an extended version of GeoMason [5] that leverages agent-based models (ABM) and geographic information systems (GIS), and showcase the benefit of using *inferential statistics* (i.e., using samples to make predictions about a population) for studying disease outbreaks. We validate the results of the simulations with the ground truth derived from the KCDC logs. GeoSpread offers a flexible model based on real-world COVID-19 spread information and can be used to facilitate evaluation of different mitigation measures to reduce the spread of the disease. GeoSpread needs only data distributions to simulate the spread of SARS-CoV-2. Here, we use distribution data in the form of histograms (and make them available to the community [6]). GeoSpread is the focus of our last research question, *RQ3*: Does an ABM, parameterized using only data distributions, accurately predict the spread of COVID-19 and the efficiency of possible counter-measures?

Contributions and outline of this paper are:

- *Data Discovery*: We analyze and connect data from various KCDC logs to extract information on patient movements (Sections II and III).
- *Statistical Analysis*: We provide statistical analysis of population movements and habits in the form of histograms for Seoul, Berlin, Dusseldorf, Kelheim (district), and Leipzig. This information is extracted using only *descriptive statistics* (i.e., the quantitative description of attributes).
- *Cross-fertilization:* We investigate similarities between the KCDC and German data sets seeking for common humna movement patterns in these urban environments (Section IV). Leveraging this information, we cross-fertilize to incorporate useful information from the Berlin data set which are unavailable in the Seoul data (e.g., travel speed, transportation means, household size).
- *GeoSpread*: We parameterize an agent-based model using the cross-fertilized data as input, see Section V, and outline its flexibility to capture a variety of conditions. The simulation tool, GeoSpread, and processed data is open sourced [6].
- *Model Validation with Real Data*: GeoSpread is validated, is compared to state-of-the-art approaches, and is used to analyze the effect of different mitigation measures (i.e., border lockdown, stay-at-home advisory, and vaccination) in Section VI. Its usage and limitations are discussed in Section VII.

## II. The KCDC Data Set

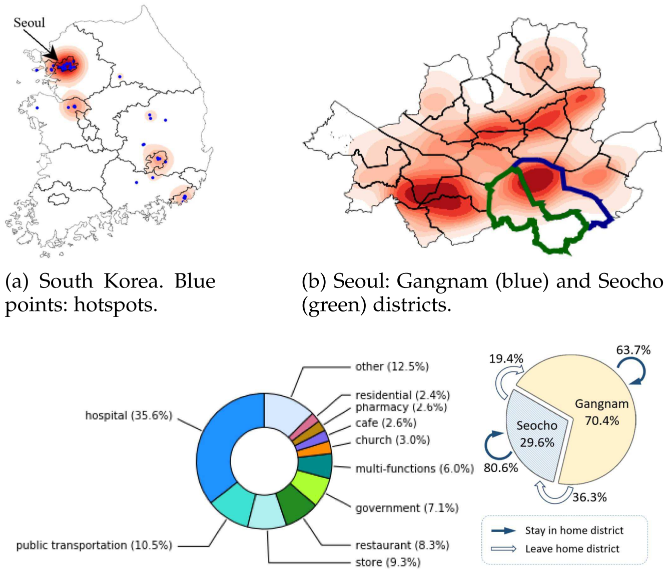
The data sets [1] used in this paper contain data collected by the KCDC and local governments from January 20, 2020, to May 31, 2020. PatientInfo and PatientRoute contain information and routes of COVID-19 patients in Seoul, respectively. The number of (healthy and sick) people moving across Seoul districts are provided in the SeoulFloating data set and has been collected using the Big Data Hub of SK Telecom, a Korean wireless telecommunications operator.

*PatientInfo Data Set.* This data set provides epidemiological data of COVID-19 patients. It contains 4,004 different entries, each entry represents a different patient identified by a unique ID (*patient_id*). Other attributes include their gender and age, their provenance (*country*, *province*, and *city*), whether they have been infected in a known case (*infection_case*, e.g., overseas inflow or contact with patient) 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*).

*PatientRoute Data Set.* This data set contains 8,092 entries, each one reporting a visit (to one of 2,992 unique locations) of 1,472 (out of 4,004) unique South Korean COVID-19 patients logged in the PatientInfo data set. A location is unequivocally 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 [1]). We manually look for their type using their geographical coordinates and OpenStreetMap[4] to compensate for this lack of data. 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. Data are collected from January 1 to May 31, 2020, by SK Telecom. 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 20 s through 70 s. 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

[4] https://www.openstreetmap.org/



(a) South Korea. Blue points: hotspots.

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

(c) Top-10 most visited locations in Seoul.

(d) Travels between Gangnam and Seocho.

Fig. 1. Most visited locations (and their type) in Seoul. Movements between Gangnam and Seocho districts.

on the *entire* Seoul population, not just the COVID-19 patients, and only considers those with cell phones.

## III. Data Discovery: KCDC Data

Although the information contained in the KCDC data sets is not as accurate as one would like, it still allows for the analysis of patient movements and interactions with high accuracy. In this section, we discuss information and statistical data that 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 [7].

### A. Visited Locations

Fig. 1(a) and (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. Heat maps in Fig. 1 also show the South Korean cities for which movement data are recorded. Visibly, 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 head-quarters are located in the south-west part of the city. The south-east region corresponds to the Gangnam and Seocho districts, outlined in blue and green in Fig. 1(b), respectively. 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 KCDC 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 20 s through 70 s. The scarcity of logged residential facilities is due to privacy concerns. Fig. 1(d) illustrates the movement of population between two neighboring districts, Gangnam and Seocho that we use later in our model.

### B. Patient Connections

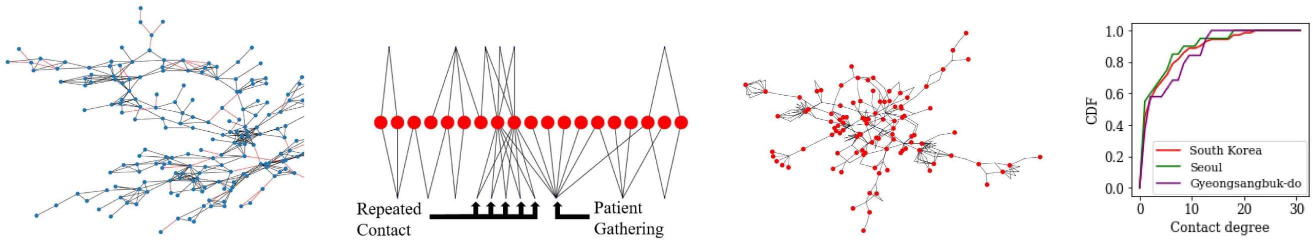
Fig. 2(a) presents a subgraph of patient connections discovered by linking the PatientRoute and PatientInfo data sets. To improve visibility, we only present a small portion of the entire graph. Here, nodes depict patients, black edges connect patients that visited the same place during the same day 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 in a *hypergraph* (i.e., a generalization of a graph where an edge can capture common relationships between two graphs and offer insights on the relationship between the graphs that have common hyperedges). Here, we use hypergraphs to connect information on two graphs, i.e., patients and locations, to discover how many times patients come into contact and at what locations. A small 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 with one another. Clusters of cases in Seoul can be seen in the hypergraph in Fig. 2(c).

Fig. 2(d) shows a summary view of patient connections: the contact degree cumulative distribution function (CDF) [7] 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. Further analysis is required to determine whether or not a patient is a super spreader.

### C. Super Spreaders

Fig. 3 illustrates a subset of patients where the *infected_by* relationship (i.e., patient A is infected by patient B) is *known* from the PatientInfo data set. The entire graph contains 1,052 patient nodes and 822 edges representing the known infection spread. For the sake of visibility, we present just a subset of the entire graph. Red nodes correspond to individuals with available route information who are known to have infected others, green nodes correspond to individuals who infected others but have no



(a) Patient connections (partial) (b) Hypergraph of connections (partial) (c) Hypergraph cluster (partial) (d) Contact degree CDF

Fig. 2. Patient contacts in Seoul (a)–(c) and contact degree at different levels of governance (d).

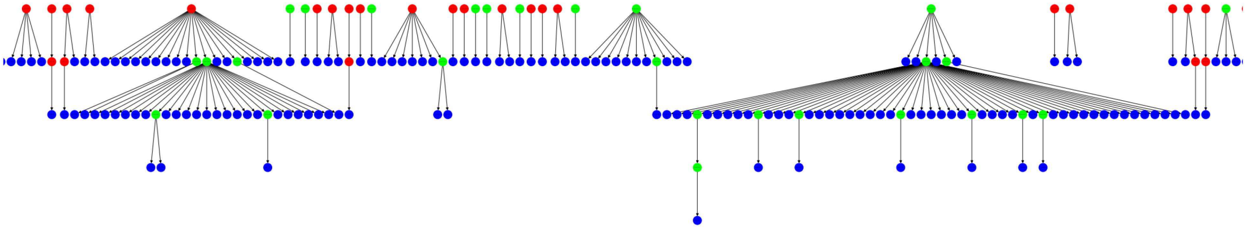
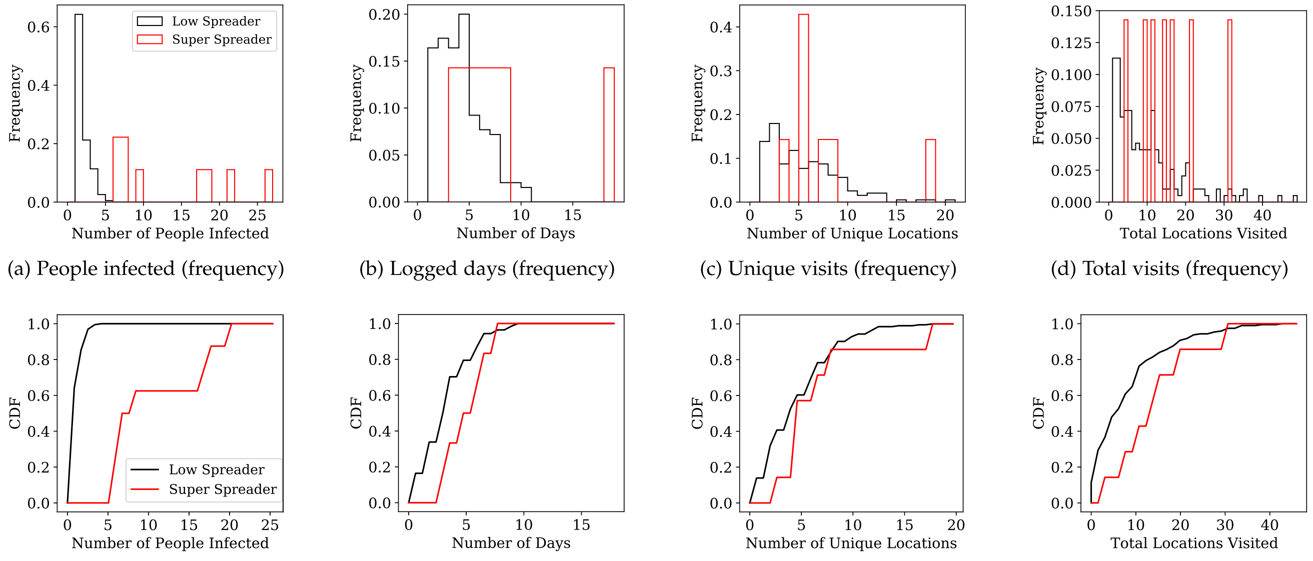


Fig. 3. Infection spread subgraph: Red nodes are patients with route information who infected others. Green nodes are patients who infected others but do not have route information. Blue nodes are patients who did not infect anyone else.



(a) People infected (frequency) (b) Logged days (frequency) (c) Unique visits (frequency) (d) Total visits (frequency)

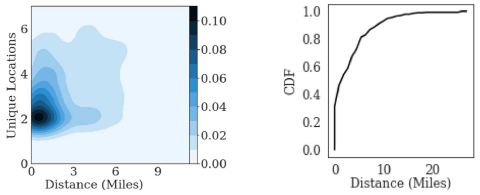
(e) People infected (CDF) (f) Logged days (CDF) (g) Unique visits (CDF) (h) Total visits (CDF)

Fig. 4. Super spreader analysis in Seoul.

available route information, and blue nodes correspond to patients who are not known to have infected others. This particular subset shows a mix of super spreaders (i.e., people who infected more than six people) and low spreaders, who infected six or fewer people.[5] The large "fans" in this figure are indicative of super spreaders. Using this classification of patients based on the number of people they infect, we discover different behaviors of super/low spreaders, shown in Fig. 4. Super spreaders account for 3.59% and low spreaders account for the remaining 96.41% of patients.

Fig. 4 presents frequencies (first row) and their respective CDFs (second row) for different attributes of low- and super-spreaders. Frequencies (a)–(d) show how likely low and super spreaders infect a specific number of people, appear in the logs for a given number of days, and visit a specific number of unique or total locations, respectively. CDFs (e)–(h) indicate that, in general, super spreaders tend to be active for more days, visit more unique locations, and have longer routes than low spreaders. Overall, these figures show that all super spreaders in the data set are active for three or more days and visit three or

[5] We define a "super spreader" as someone who infects at least 6 people. This allows us to divide the data set to obtain the most noticeable difference in patient behavior (number of locations, number of days, number of records).



(a) Density heat map (b) Distance CDF

Fig. 5. Daily traveled distance and unique locations visited in Seoul.

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.

### D. Daily Traveled Distance

Fig. 5(a) plots the density heat map of distance traveled by patients in Seoul and the number of locations visited in a day, two important features due to the vital nature of patient movement to spread COVID-19. The darker the area, the more patients have the same traveled distance and visited locations. 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. 5(b).

### E. Patient Mobility

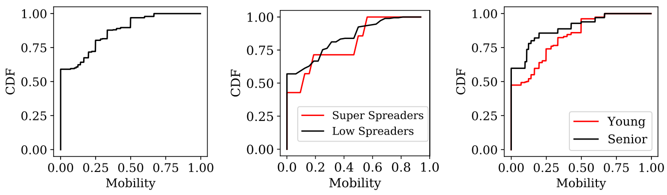
Patient mobility is another important attribute to consider. Intuitively, the more places a patient visits, the higher their mobility is. Analyzing the mobility of patients in the KCDC data set, there are days where individuals exhibit high mobility and days where they move significantly less. This leads us to a more usable definition of mobility as a function of different time periods (days). Considering how many unique locations are visited by all patients each day, we observe that a typical patient visits 1–3 locations in the 88.9% of days, and more than 3 locations in the remaining 11.1% of 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, depicted in the following equation.

$$Mobility = \frac{\#\ High\ Mobility\ Days}{Total\ Active\ Days}. \quad (1)$$

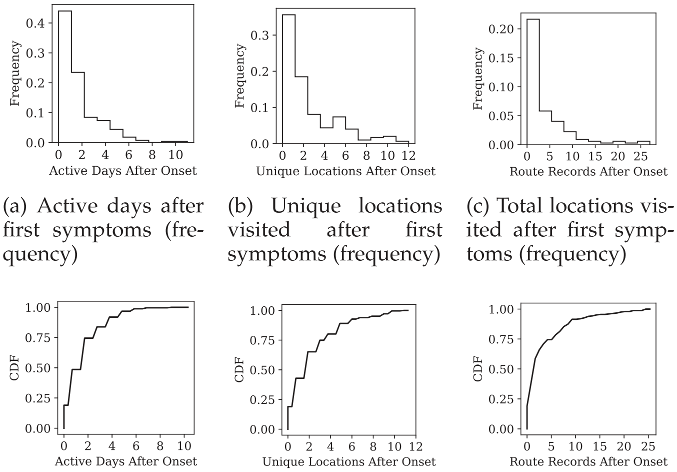
Note that this is not the only way to define mobility. For simulation purposes (see Section V), this definition provides a practical way to capture mobility with a probability. Based on the analysis of the KCDC data set, days with $L \leq 3$ are considered of low mobility. The CDF of patient mobility using the above definition is depicted in Fig. 6(a). The figure shows that the mobility of 57.6% of patients is 0, i.e., those patients never visit more than $L = 3$ unique locations in a day since $\#\ High\ Mobility\ Days = 0$, see (1).

Different classes of patients have different mobility. Fig. 6(b) shows the difference in mobility between low and super spreaders, while Fig. 6(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.



(a) All patients (b) Different spreaders (c) Young vs. seniors

Fig. 6. Patient mobility in Seoul.



(a) Active days after first symptoms (frequency) (b) Unique locations visited after first symptoms (frequency) (c) Total locations visited after first symptoms (frequency)

(d) Active days after first symptoms (CDF) (e) Unique locations visited after first symptoms (CDF) (f) Total locations visited after first symptoms (CDF)

Fig. 7. Irresponsible behavior of sick patients in Seoul.

### F. Irresponsible Behaviors

Patients may 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 *all* sick people continue to show mobility after exhibiting symptoms, see Fig. 7. The figure shows that only 20% of patients stop moving and isolate immediately after initial symptoms are observed. Many patients, see Fig. 7(a)–(c), may go to a pharmacy or hospital after showing symptoms, indicating that a few movements after onset is not necessarily irresponsible. Some patients, however, keep moving for more than a week after the onset of symptoms, see Fig. 7(d). They also visit many locations; Fig. 7(e) and (f) show the number of unique and total locations that sick patients visit after initial symptoms are observed.

Summarizing, the answer to *RQ1* is as follows.

TABLE I
AREA AND POPULATION OF SEOUL AND FOUR GERMAN CITIES CONSIDERED FOR COMPARISON AND CROSS-FERTILIZATION

| City | Country | Area (sq mi) | Population (12/2021) |
|---|---|---|---|
| Seoul | SK | 233.67 | 9,443,722 |
| Berlin | DE | 344.10 | 3,677,472 |
| Dusseldorf | DE | 83.94 | 619,477 |
| Kelheim (district) | DE | 412.00 | 123,899 |
| Leipzig | DE | 114.81 | 601,866 |

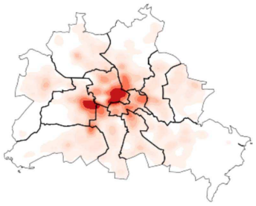


Fig. 8. Heat map of the most visited Berlin locations.

**RQ1: Information from the KCDC Logs**
We analyze movement habits of Seoul patients applying statistical analysis and descriptive statistics to the KCDC data sets. Patient connections, super spreaders, and irresponsible behaviors are examples of information that is not directly provided in the data sets, but can be obtained by manipulating the available data. These distributions are used as input to GeoSpread.

## IV. The Berlin Data Set

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, with distributions from German cities and districts (i.e., Berlin, Dusseldorf, Kelheim, and Leipzig) with different *areas* and *population*, see Table I. We focus on commonalities in movements of individual in Seoul and in German cities that can be used as a basis. After determining the German city (i.e., Berlin) whose population behavior better matches the one of people in Seoul, we extract new information to cross-fertilize the statistical data of the KCDC data set. Cross-fertilization across data sets is common in the broader systems area, where similarities across data sets are explored to fill-in missing data. In the following, we describe in detail the Berlin data set [3] that we use to cross-fertlize the KCDC log. Note that data sets of other German cities are similar to the Berlin one. The Berlin data set contains 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 transportation used by people for their movement, 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. 8 shows the most active district of Berlin, i.e., areas of the city that appear more frequently in the Berlin data set.

*EventWeekdays Data Set.* People's movements over 30 hours are logged in this data set, where almost 6 million activities are 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% (i.e., 3,919,990 entries) 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.

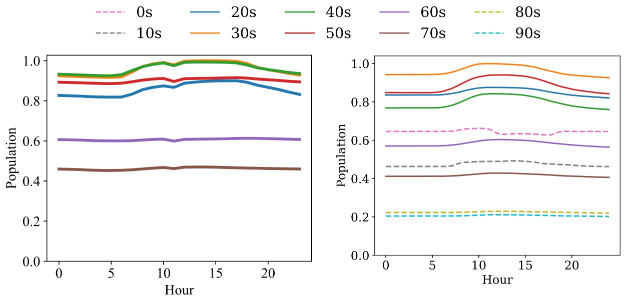
*Demographic Data Set.* This data set contains information about each person (i.e., more than 1.2 million people) whose activities have been logged in the EventWeekdays data set. Specifically, *age* and *gender* for all people is provided as well as their *home_district*, *home_id*, and home *coordinates*. The *home_district* attribute contains one of the 401 administrative districts of Germany. Here, since we focus just on Berlin, a metro area similar to 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 their home-place, while the *coordinates* attribute allows placing each building on a map with an accuracy of 500 meters.

*Facility Type 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) are 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.

*Public Transport Data Set.* This data set records vehicles used for public transportation. An *id* and a *type* (e.g., bus, metro, tram) are associated to each vehicle. Many people use public transportation: 1,791,061 movements are logged in this data set.

### A. Similarities of KCDC and German Data Sets

KCDC and German (i.e., Berlin, Dusseldorf, Kelheim, and Leipzig) data sets allow retrieving information and attributes (e.g., Age Group, Travel Distance, Unique Locations, and Contact Degree) that can be used for comparing movement habits of Seoul patients to those of German inhabitants. Besides visual and statistical (i.e., mean value and standard deviation) analyses, three widely used [8], [9], [10] statistical hypothesis tests (Mann–Whitney or MW, Pearson's chi-squared or CS, and Kolmogorov-Smirnov or KS) are considered to evaluate



(a) Seoul

(b) Berlin

Fig. 9. Daily presence of different age groups in Seoul and Berlin. The $y$-axis of both figures is normalized over the total number of people monitored in each city.

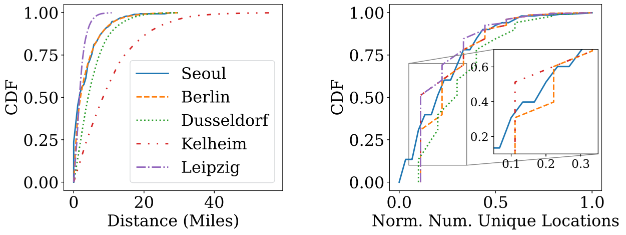
TABLE II
AVERAGE DAILY PRESENCE OF DIFFERENT AGE GROUPS IN SEOUL AND BERLIN

| Data set | Age Groups | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0s | 10s | 20s | 30s | 40s | 50s | 60s | 70s | 80s | 90s |
| Seoul | – | – | 0.86 | 0.96 | 0.96 | 0.90 | 0.60 | 0.46 | – | – |
| Berlin | 0.64 | 0.48 | 0.85 | 0.97 | 0.81 | 0.90 | 0.59 | 0.42 | 0.23 | 0.21 |

the goodness-of-fit of KCDC and German movement attributes defined by their CDFs. We use these tests to determine which German city is the most similar to Seoul in terms of movement habits, i.e., travel distance, unique locations, and contact degree attributes. For all these tests, the null/alternative hypothesis is that the two models are defined by identical/different distributions. The Mann–Whitney test is not affected by outliers since it evaluates the center of the distributions. The Pearson's chi-squared test evaluates similarities along the whole distributions by considering sample frequencies. The Kolmogorov-Smirnov test considers the CDFs of both groups and their maximum distance. We further evaluate the similarity of KCDC and German data sets using the Kullback-Leibler divergence test (KL or *relative entropy*), i.e., a statistical distance measure used in the literature [11]. The analysis of the *Age Group* parameters is only visual since no distribution is provided for this attribute. Moreover, the Berlin data set is the only one providing enough data to carry out such analysis. In the following, similarities and differences of KCDC and German attributes are analyzed and described in detail.

*Age Group.* Fig. 9 depicts Seoul and Berlin population floating during a business day. Data is grouped based on people's age with decade granularity. For the sake of comparison, since the number of observations in the two data sets is different, all values are normalized over the maximum number of people monitored in each city. Table II reports the average daily presence observed for each age group to highlight similarties and differences between the KCDC and Berlin data sets. No comparison between Seoul and Dusseldorf, Kelheim, or Leipzig population is given since population age is not reported in the data sets of these German cities and districts.

The SeoulFloating data set monitors people that are in their 20 s through 70 s for both healthy and sick individuals. As a result, this data set is valuable for comparison to the Berlin data set. We investigate the population habits from January 1, 2020, to May 31, 2020 by age group, see Fig. 9(a). Fig. 9(b) provides information on movements of people living in Berlin. Differently from the KCDC data set, in this case also people younger than 20 or older than 79 are monitored, see dashed lines. Overall, Seoul and Berlin experience similar people floating dynamics, probably due to both cities being the capital and the main economic center of their country. 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 since the number of working population in this age range is limited. Adults and young-adults of both cities show also similar dynamics, with the only exception of people in their 40 s and 50 s. 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 average number of people in their 50 s that live in Seoul and Berlin is the same (i.e., 0.9), although the two data sets present slightly different trends. Looking at the Berlin data, it is also possible to observe that there are not many people older than 80 and that their number does not change during the day. The only age group whose population decreases in the morning and increases in the evening is the one representing kids younger than 10.



(a) Daily distance

(b) Unique locations

Fig. 10. Movement habits of Seoul patients and Berlin inhabitants. The number of unique locations (b) is normalized over the total number of visited locations in each data set.

*Daily Traveled Distance.* Fig. 10(a) plots the CDF of daily traveled distance (in miles) for people living in Seoul and the considered German cities and districts (i.e., Berlin, Dusseldorf, Kelheim, and Leipzig). CDFs of Seoul and Berlin populations match closely meaning that Korean patients and Berlin inhabitants travel the same distance on a daily basis. Specifically, 75% of people move less than 5 miles and only a small percentage of the population travels more than 15 miles.

People living in Dusseldorf and Kelheim travel more than Seoul and Berlin inhabitants, possibly due to facilities and businesses more spread on the territory. Instead, the Leipzig population moves less than 10 miles every day. Table III reports mean value and standard deviation for all data sets and shows that the Berlin data set is the one whose average travel distance is closer to the one observed in the KCDC data set. All considered statistical tests accept the null hypothesis (i.e., samples are drawn from the same distribution) with 95% confidence (i.e., p-value $> 0.05$) only when comparing the distance traveled by Berlin and Seoul inhabitants. The divergence test further confirms the similarity between these attributes.

*Unique Locations.* Fig. 10(b) depicts the daily number of unique locations visited by all monitored people in Seoul and

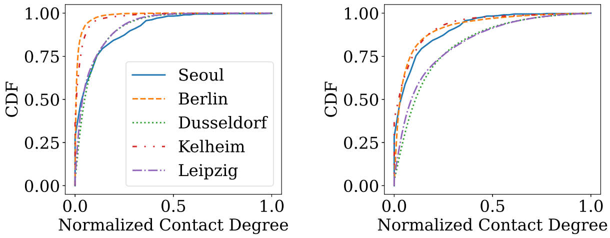
TABLE III

Statistical Analysis, Hypothesis Tests, and Divergence Test for Parameters Shared by the KCDC and German Data Sets. Column 1 Reports Shared Parameters; Columns 2 and 4 Show the Mean Value and Standard Deviation for the KCDC and German Data Sets, Respectively; Column 3 Reports the Considered German Data Sets; Columns 5–10 Show Results (i.e., Statistic and P-Value) From Three Well Known Hypothesis Tests, Mann-Whitney (MW), Chi-Squared (CS), and Kolmogorov-Smirnov (KS); Column 11 Shows Results From the Kullback-Leibler (KL) Divergence Test. For Each Parameter, the German Data Set That is More Similar to the KCDC One is Highlighted Using *Italic*. The Best Results for Each Test is Also Highlighted Using *Italic*

| **Parameter** | **Mean ± StD.**<br>KCDC | **German**<br>**Data Set** | **Mean ± StD.** | **MW** | | **CS** | | **KS** | | **KL** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | statistic | p-value | statistic | p-value | statistic | p-value | |
| Travel Distance | 2.79 ± 2.91 | *Berlin* | *2.90 ± 2.75* | *4833* | *0.68* | *1.55* | *0.82* | *0.17* | *0.81* | *0.01* |
| | | Dusseldorf | 4.86 ± 3.28 | 3491 | 2.28e-4 | 23.38 | 1.06e-4 | 0.40 | 0.02 | 0.08 |
| | | Kelheim | 9.63 ± 6.46 | 1802 | 5.51e-15 | 26.84 | 2.14e-5 | 0.63 | 5.80e-6 | 0.13 |
| | | Leipzig | 1.84 ± 1.11 | 6321 | 1.25e-3 | 40.51 | 3.30e-8 | 0.40 | 0.02 | 0.13 |
| Unique Locations | 0.25 ± 0.14 | Berlin | 0.28 ± 0.12 | 4527 | 0.25 | *7.31* | *0.12* | *0.18* | *0.40* | *0.04* |
| | | Dusseldorf | 0.33 ± 0.15 | 3508 | 2.67e-4 | 11.8 | 0.02 | 0.30 | 0.02 | 0.12 |
| | | *Kelheim* | *0.26 ± 0.14* | *5037* | *0.93* | *7.31* | *0.12* | *0.18* | *0.40* | *0.04* |
| | | Leipzig | 0.22 ± 0.12 | 5584 | 0.15 | 40.75 | 3.03e-8 | 0.30 | 0.02 | 0.08 |
| Contact Degree (w/ outliers) | 0.12 ± 0.13 | Berlin | 0.02 ± 0.03 | 7016 | 8.29e-7 | 15.39 | 3.95e-3 | 0.60 | 2.37e-5 | 0.14 |
| | | Dusseldorf | *0.13 ± 0.10* | 4731 | 0.51 | 2.96 | 0.57 | 0.17 | 0.81 | 0.04 |
| | | Kelheim | 0.03 ± 0.06 | 6974 | 1.18e-6 | 43.07 | 1.00e-8 | 0.53 | 2.93e-4 | 0.23 |
| | | *Leipzig* | 0.10 ± 0.10 | *4900* | *0.81* | *1.54* | *0.82* | *0.13* | *0.96* | *0.02* |
| Contact Degree (w/o outliers) | 0.12 ± 0.13 | *Berlin* | *0.10 ± 0.18* | *4861* | *0.73* | *4.00* | *0.41* | *0.17* | *0.81* | *0.02* |
| | | Dusseldorf | 0.19 ± 0.22 | 3233 | 1.52e-5 | 11.56 | 0.02 | 0.50 | 9.00e-4 | 0.13 |
| | | Kelheim | 0.08 ± 0.17 | 5409 | 0.31 | 4.67 | 0.32 | *0.17* | *0.81* | *0.02* |
| | | Leipzig | 0.18 ± 0.23 | 3502 | 2.46e-4 | 10.87 | 0.03 | 0.43 | 6.55e-3 | 0.13 |

Germany. To compare observations from different data sets, the attribute is normalized over the maximum number of unique visits for each city. Differences between Seoul and Dusseldorf population are noticeable when looking at Fig. 10(b), with inhabitants of the German city visiting in a day more unique locations than Seoul patients. Such differences are less visible when considering other German cities (i.e., Berlin and Leipzig) or districts (i.e., Kelheim), with Berlin and Kelheim having very similar CDFs except for $0.25 \leq \text{CDF} \leq 0.6$, see the box inside Fig. 10(b). Therefore, we leverage statistical analysis and hypothesis tests (i.e., Table III) to determine which German city better matches Seoul habits when considering this attribute. Specifically, Seoul and Kelheim populations visit the same number of unique locations on average, with Berlin and Leipzig showing similar average values. The three considered tests do not reject the null hypothesis with 95% confidence when comparing Seoul observations to Kelheim and Berlin ones, whereas the null hypothesis is accepted only by the Mann-Whitney test (i.e., the test which evaluates only the center of distributions) when the Seoul and Leipzig CDFs are compared. The divergence test shows smaller relative entropy when comparing Seoul unique locations to those of Berlin and Kelheim. These similarities might be due to the area of Seoul, Berlin, and Kelheim (i.e., all larger than 200 square miles), see Table I.

*Contact Degree.* The analysis of how many people are met by each person logged in KCDC and German data sets (i.e., contact degree) allows discovering relationships that can 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 the same facilities on the same day. People's movements in German data sets 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. The contact degree of inhabitants of Seoul and German cities is normalized over the maximum number of contacts for each city and compared in Fig. 11(a). Dusseldorf and Leipzig are the German cities whose contact degree follows a distribution similar to the Seoul CDF, whereas Berlin and Kelheim show large differences with respect to Seoul. This is due to a few individuals living in Berlin and Kelheim meeting many other people, i.e., the tail of the CDFs is long. As expected, when all monitored individuals are considered, the statistical tests reject the null hypothesis when the KCDC data set is compared to the Berlin or Kelheim ones. Instead, Dusseldorf and Leipzig show promising results, with the contact degree of Leipzig population being more similar to the one of Seoul patients.



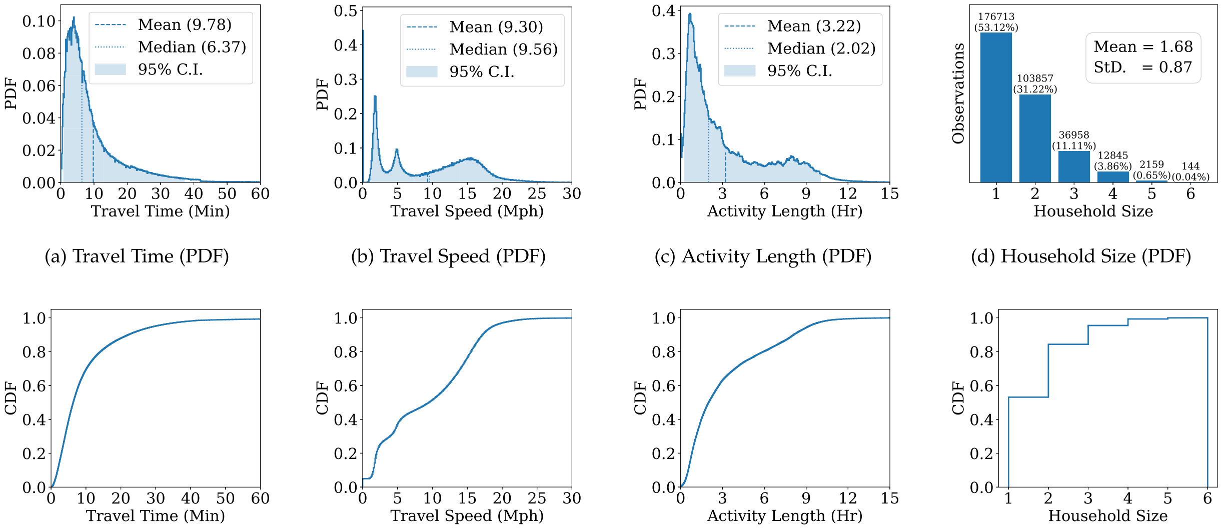
(a) Comparison with outliers (b) Comparison without outliers

Fig. 11. Contact degree of Seoul patients and Berlin inhabitants normalized (over the maximum value) for comparison. Outliers are considered in (a) and discarded in (b).

To further investigate how outliers (i.e., few people that meet many others) impact the goodness-of-fit of Seoul and German data sets, we also consider the contact degree of German cities up to the 99th percentile to exclude possible outliers from the analysis. Results are shown in Fig. 11(b), where the contact degree observed in Berlin and Kelheim is now closer to the one of Seoul patients. Such results are confirmed by analyzing distributions of these data sets with statistical tests. The null hypothesis is not rejected for Berlin or Kelheim by any of the considered tests with 95% confidence (i.e., p-value $> 0.05$).

TABLE IV
MAIN ATTRIBUTES, PARAMETERS, AND INFORMATION THAT CAN(NOT) BE EXTRACTED FROM THE KCDC AND BERLIN DATA SETS

| Parameters | Definition | KCDC data set | Berlin data set | Simulation |
|---|---|---|---|---|
| Age Group (20–79) | Daily presence of people in their 20s through 70s | ✓ | ✓ | |
| Travel Distance | Travel distance between two places | ✓ | ✓ | ✓ |
| Unique Locations | Number of unique locations visited by an agent | ✓ | ✓ | |
| Contact Degree | Number of encountered people | ✓ | ✓ | |
| Facility Type | Type of facility visited by an agent | ✓ | | ✓ |
| Super/Low Spreader | Type of spreaders | ✓ | | ✓ |
| Mobility | Probability of leaving a building | ✓ | | ✓ |
| Irresponsible Patients | People that move around even if infected | ✓ | | ✓ |
| Regional Habits | Probability of visiting different districts | ✓ | | ✓ |
| Age Group (19−, 80+) | Daily presence of people younger than 20 or older than 80 | | ✓ | |
| Activity Type | Type of activity carried out by an agent | | ✓ | |
| Minors | Minors' habits | | ✓ | |
| Activity Length | Total time spent on staying in a building | | ✓ | ✓ |
| Public Transport | Types of vehicles used by agents | | ✓ | |
| Travel Habits | Travel time and speed | | ✓ | ✓ |
| Household Size | Number of family members | | ✓ | ✓ |
| Population | Number of simulated agents | | | ✓ |
| Infection Rate | Probability of a healthy agent to be infected | | | ✓ |
| Caution Level | Agent willpower to leave their house | | | ✓ |



(a) Travel Time (PDF) (b) Travel Speed (PDF) (c) Activity Length (PDF) (d) Household Size (PDF)

(e) Travel Time (CDF) (f) Travel Speed (CDF) (g) Activity Length (CDF) (h) Household Size (CDF)

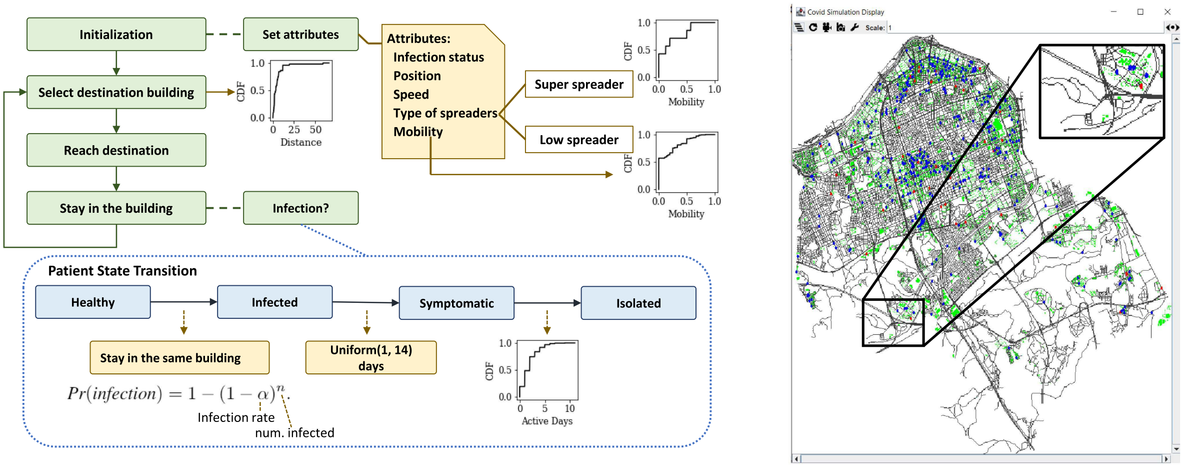
Fig. 12. PDFs and CDFs of unique features from the Berlin data set. (a)–(c) depict the mean value, the median, and the 95% confidence interval of the continuous distributions. (d) reports the mean value and the standard deviation of the discrete distribution. (e)–(h) depict CDFs of the four Berlin features.

Comparing movement habits of Seoul patients to habits of inhabitants of German cities (i.e., Berlin, Dusseldorf, and Leipzig) and districts (i.e., Kelheim) with different areas and populations, we identify the Berlin data set as a good candidate to cross-fertilize the KCDC data set. Besides close similarities among movement habits of Berlin and Seoul (that are confirmed by visual and statistical analysis, as well as hypothesis and divergence tests), the Berlin data set provides more information than other German data sets, i.e., Dusseldorf, Kelheim, and Leipzig data sets come without any information about population age and floating.

## B. Unique Characteristics of the Berlin Data Set

The prior analysis of KCDC and German data sets show that Seoul and Berlin share many attributes, summarized in the first section of Table IV. In addition to this, both data sets also contain a wealth of unique characteristics. Unique distributions pertaining to the KCDC data set are summarized in the second section of Table IV, and unique distributions pertaining to the Berlin data set are summarized in the third section of Table IV.

While both data sets contain information about distance traveled, the Berlin data set contains additional information about travel time and speed. The probability density function (PDF) [7] of these attributes are depicted in Fig. 12(a)–(b). The time spent for each travel is skewed towards small values, each movement takes less than 10 minutes on average, and the 95% of travels is completed in less than 40 minutes. The PDF of speed shows a tri-modal distribution: the first and second peaks may represent people walking at two different speeds (i.e., between 1.8 and 5 mph). The last peak might be people using a vehicle to move. In this



(a) Life cycle of an agent

(b) GeoSpread view: Gangnam and Seocho

Fig. 13. Simulation life cycle and visualization.

case, people might move at a reduced speed (i.e., around 15 mph) due to the typical traffic of metropolitan cities. One notable drawback of the KCDC data set 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 Berlin data set has detailed time stamps and records of the amount of time spent performing a specific activity (e.g., shopping or working). Fig. 12(c) shows the PDF of activity lengths, from which it is visible that 50% of activities last only 2 hours and, on average, activities are completed within 3.22 hours.

Since 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 Berlin data sets. This information is shown in Fig. 12(d). More than 50% of households are made of only one person, while the average household size is less than 2. This might help to limit the spread of COVID-19 through a household. 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. CDFs of travel time, travel speed, activity length, and household size are depicted in Fig. 12(e)–(h) and fed to GeoSpread in Section VI to study the spread of COVID-19 in Seoul.

The answer to *RQ2* can be summarized as follows.

**RQ2: Cross-fertilization of Data Sets**
Attributes of Seoul and Berlin data sets (i.e., one of the available German data sets) generally follow similar distributions. Moreover, the Berlin data set provides information that are not contained in other German data sets (e.g., daily presence of different age groups). Therefore, the KCDC data set is enriched with Berlin data to provide more information in the input of GeoSpread.

## V. Agent-Based Model

In this section, we show how to parameterize a simulation based on GeoSpread [6]. The attributes, life cycle, and states of an agent are shown in Fig. 13(a). The following attributes are set during the initialization phase:

1) *Infection status.* One or more random agents are selected as the initial case(s).
2) *Position.* Agents are randomly placed on a road in the simulated area.
3) *Speed.* Speed determines how fast an agent moves from one location to another and is selected according to a distribution. Specifically, we sample from the speed distribution from the Berlin data set characterization to select an agent's speed, see Fig. 12(f).
4) *Type of spreaders.* We define two classes of spreaders: 3.59% of patients are super spreaders and 96.41% are low spreaders (see Section III-C).
5) *Mobility.* We use the mobility of super spreaders and low spreaders depicted in Fig. 6(b) to model different types of 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. 12(h).

In addition to the mobility distribution of super spreaders and low spreaders, the CDF of daily traveled distance in Fig. 5(a) 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. 12(g). Simulation time is defined by cycles.

In each simulation cycle, agents outside a building move along the road toward 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.[6] 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, as the state transition shown in Fig. 13(a). We assume the outdoor infection probability to be negligible. Given the probability of infection inside a building, $\alpha$, 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

$$Pr(infection) = 1 - (1 - \alpha)^n. \tag{2}$$

Note that the probability of infection defined by (2) is nominal. Any model can be used here 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 [3] for examples on how to adjust (2).

It takes 1–14 days for patients to show symptoms after infection according to the WHO.[7] GeoSpread supports any distribution (e.g., uniform, log-normal) to define the transition of an individual from infected to symptomatic. This allows capturing different scenarios and model future variants of SARS-CoV-2 or different pathogens.

Since there exist patients who continue to move even after showing symptoms, as seen in Fig. 7, we use the CDF in Fig. 7(d) 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 [12], 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.[8]

A summary of all distributions used for simulations is recorded in the last column of Table IV. Note that we do not directly incorporate patient age due to lack of detailed infection-spread data, and we do not directly use the contact degree and unique number of locations visited due to the individual-centric nature of the simulations. Contact degree and the number of unique locations visited are used for validation since these are not explicitly used as parameters. Tunable simulation parameters are listed in the last section of Table IV.

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[9] has 11,438 road intersections and 7,043 buildings. 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 VI.

A screenshot of the GeoSpread simulation execution can be seen in Fig. 13(b). Black lines are roads that agents travel on and green areas are buildings where agents stop. Agents only have two states in terms of infection, i.e., healthy (blue dots) or infected (red dots). The box in the top-right corner zooms on a detail of the GeoSpread view.

## VI. Model Validation and Case Study

After presenting the generic GeoSpread tool in Section V, we showcase the flexibility of this simulation model. We first validate the simulation using the ground truth and show that GeoSpread can efficiently predict the temporal evolution of COVID-19 cases in a given place. We investigate the effect of each data distribution on the prediction accuracy. Then, we compare GeoSpread to two state-of-the-art approaches, i.e., mathematical [13] and an agent-based model [3]. Hence, we use GeoSpread to simulate different mitigation measures (i.e., stay-at-home advisory, border lockdown, and vaccination) and assess their effectiveness.

### A. Validation

We focus on agents moving between Seocho and Gagnam. Fig. 1(d) shows the percentage of residents in these two districts that have been infected, the figure also illustrates the frequency of residents visiting buildings in their home district, as well as visiting the other district. 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. After initialization, when selecting destination buildings, the probability of a resident staying or leaving their home district follows Fig. 1(d).

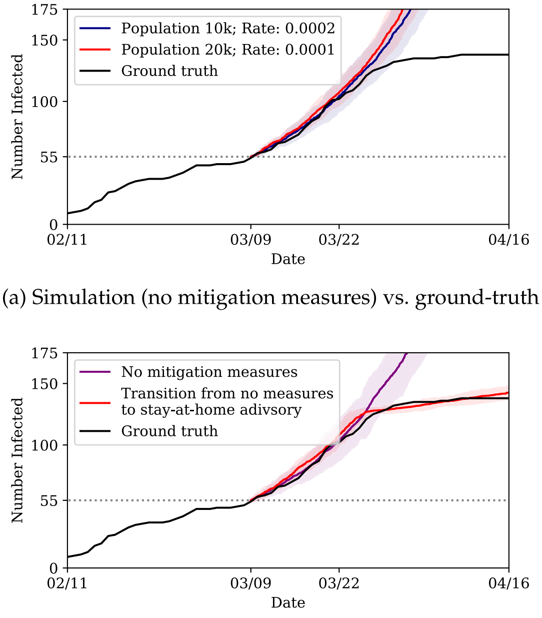
Since two districts are considered in this simulation, starting with only one infected agent in one of the two areas could bias the results. Here, we start the simulation with 55 infected agents, i.e., the number of infections observed from the data set on March 9, 2020, proportionally assigned to agents in the two districts (29.6% in Seocho, 70.4% in Gangnam). We selected March 9,

[6] https://bit.ly/3FkLHRn

[7] https://bit.ly/3EXl2Jh

[8] In this simulation, we do not explicitly model agent recovery: a recovered agent that resumes its mobility is considered immune and non-contagious, therefore does not contribute to the disease spread. The simulation can be trivially extended to model recovered agents re-entering the simulation cycle.

[9] https://bit.ly/3gWMD5g



(a) Simulation (no mitigation measures) vs. ground-truth

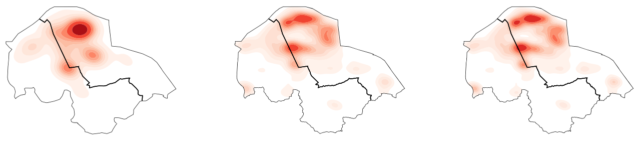
(b) Simulation (stay-at-home advisory) vs. ground truth

Fig. 14. Validation. Results are presented with 95% confidence intervals (shaded areas).

2020 because mitigation efforts in Seoul have yet to produce a noticeable effect on disease spread, while also allowing us to clearly see trends. Simulations starting at any time earlier or around March 9, result in similar infection trends.

Fig. 14(a) depicts the number of COVID-19 cases in the Gangnam and Seocho districts observed from the data set (black line) and simulation (red and blue lines). The ground truth 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 (i.e., stay-at-home advisory) 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. 14(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. 14(a) (whose 95% confidence interval is represented 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. Different population and infection rate values can be adopted, e.g., using the approach proposed in [14] to estimate dynamic parameters from real epidemic trends. The integration of dynamic parameters with GeoSpread is left for future work.



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

Fig. 15. Hotspots in the data set (ground truth) and model.

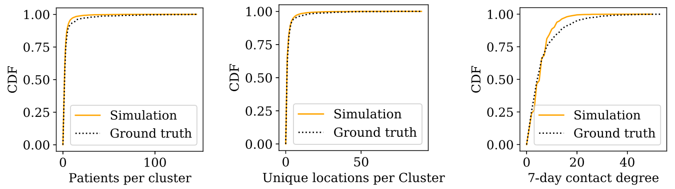
We note in Fig. 14(a) an interesting relationship between population and infection rate: when the population is doubled, dividing the infection rate in half gives similar simulation outcomes. This observation also meets the results in the generic simulation that higher population leads to faster spreading of the COVID-19 virus, while lowering the infection rate slows down the virus spreading. We conclude that we can use a “limited” population with an adjusted infection rate to efficiently (yet accurately) model the expected behavior of 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. 14(b) depicts the results of these simulations (with 95% confidence interval) against the ground truth. In this simulation case, we begin with no mitigation measures and apply a stay-at-home advisory once we reach a certain threshold number of infections. Here, we select this threshold based on the number of infections in the ground truth data when the Strong Social Distancing campaign was enacted, however, this threshold is a parameter and we can choose to transition between no measures and a stay-at-home advisory at any given number of infections. After applying the stay-at-home advisory mid-simulation, the simulation results also exhibit a flattening trend, which is consistent with the ground truth. This further highlights the ability of the model to capture what-if scenarios of different patterns of population movement.

Next, we focus on hotspot locations. In Fig. 15(a), we present the heat map of most visited locations in the Gangnam and Seocho districts from the data set (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, which results in higher traffic areas. Fig. 15(b) and (c) show the heat map of visits in the first week for simulated populations of 10,000 and 20,000, accordingly. From both simulations, we observe similar hotspots, consistent with the ground truth heat map.

Additionally, we examine properties of clusters (i.e., outbreaks) in the ground truth KCDC logs and the simulations. Fig. 16(a) depicts the number of patients seen in infection clusters in a 7-day sliding window. Fig. 16(b) shows the number of unique locations visited by patients in infection clusters in a 7-day sliding window. Finally, we can see the contact degree between patients over seven days in Fig. 16(c). The similarity of these curves further validates the accuracy of the simulation.

Here, we consider the effects of the different distributions on the simulation accuracy. To this end, we either remove an



(a) Patients per cluster over 7-day sliding window (b) Unique Locations per cluster over 7-day sliding window (c) 7-day contact degree between individuals who are infected

Fig. 16. Validation of clusters and contact degree.

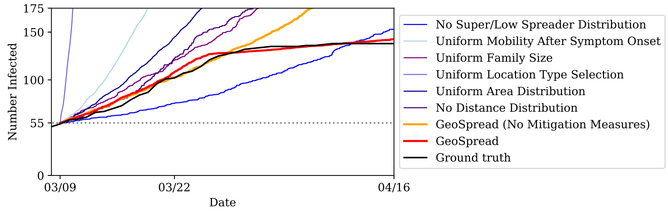


Fig. 17. Effects of removing parameters or using simpler distributions.

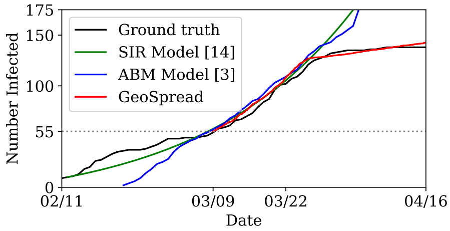
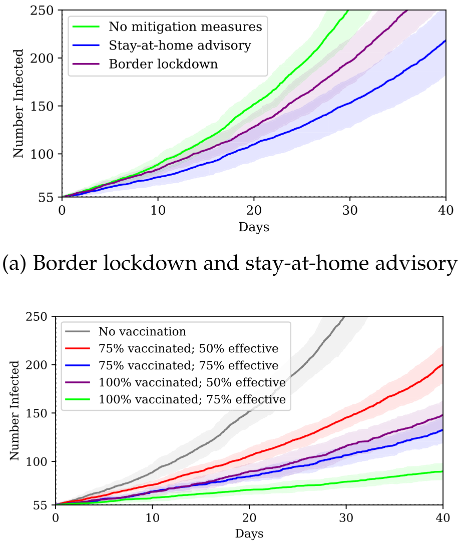


Fig. 18. Results from a state-of-the-art ABM [3] and a mathematical model (SIR) [13]: comparisons with ground truth and GeoSpread.

input parameter from the simulation or utilize a more generic distribution (i.e., Uniform) for sampling. When using the Uniform distribution, we assume to know the approximate minimum and maximum values, but no further information. Results are presented in Fig. 17. The simulation matches the ground truth closely only when all data distributions and mitigation measures are considered. Larger errors are detected when (i) low and super spreaders are not considered, (ii) when irresponsible behaviors, i.e., mobility after symptoms onset, are neglected, and (iii) when the location type is selected from a Uniform distribution.

In Fig. 18, we compare GeoSpread to two other models: the Kermack-McKendrick Susceptible-Infected-Recovered (SIR) model [13] as well as a state of the art agent-based model [3]. While both models achieve reasonable accuracy, it is important to highlight that both have shortcomings for our particular case. The SIR model cannot take into account spatial information and this cannot be used to simulate situations such as border lockdowns (see Section VI-B for using GeoSpread to evaluate this scenario). Additionally, it is not suited for analyzing different classes of patients, such as super spreaders. On the other hand, the state-of-the-art ABM is able to perform this kind of analysis, however, this model requires synthetic traces to achieve its results. The results shown here are based on a synthetic mobility trace constructed using GeoSpread, and achieves high accuracy since GeoSpread accurately captures patient mobility.



(a) Border lockdown and stay-at-home advisory

(b) Vaccination

Fig. 19. Comparison of mitigation measures.

## B. Applying Mitigation Measures

We now turn to the evaluation of the effectiveness of countermeasures. We first consider stay-at-home advisory that allows for only essential activity outside of the agent's domicile. On average, agents stay at home for longer periods time under the advisory, but are are permitted to leave periodically. The probability of leaving home is set to 20% of the agent's mobility. This can be tuned to simulate a stricter (or more relaxed) stay-at-home advisory. Once the agent arrives at the destination building, the probability of leaving the building is defined by the mobility without any additional scaling (i.e., the time spent outside the domicile is not affected).

In addition to this counter-measure, we also consider strict district border control between the Gangnam and Seocho districts, i.e., forbid movements between these two areas entirely. With a strict border control between these two districts, agents can only stay in their home district: the probability of leaving their home district is set to 0. We simulate these two mitigation measures under population 10,000, see Fig. 19(a) where all results are shown with 95% confidence intervals. First, the application of a stay-at-home advisory decreases the rate of virus spread in comparison to the baseline scenario where no counter-measures are applied. The strict border control offers a mild mitigation measure comparing to the baseline scenario.

Other mitigation measures (e.g., [15], [16], [17]) can be evaluated by tuning available parameters to simulate different behaviors of the population. For example, the effect of counter-measures that limit the transmission of the virus (e.g., face masks) can be studied by changing the parameter *Infection Rate*, see Table IV.

Finally, we consider the effects of vaccination. We consider the situations where 75% and 100% of the population are vaccinated, with the vaccine being 50% and 75% effective. The four variations are compared against simulations with no mitigation measures in Fig. 19(b). We see that a more effective vaccine is very important to slowing the spread of the virus when much of the population receives the vaccine. All cases see notable improvement over the baseline case. The answer to *RQ3* is given in this section and is summarized below.

**RQ3: Efficiency and Accuracy of GeoSpread**

Despite its simple input parameters, GeoSpread predicts accurately the spread of COVID-19 in two Seoul neighborhoods. Other mathematical and agent-based models may achieve similar accuracy but they have shortcomings: they cannot assess mitigation measures and need an actual synthetic trace as input. GeoSpread uses as input empirical distributions extracted from available data sets and can evaluate an array of counter-measures (e.g., stay-at-home advisory, border lockdown, and vaccinations).

## VII. Discussion and Limitations

The proposed model captures the spread of COVID-19 in an urban setting. Although the model is validated using ground truth, incomplete and/or missing data may limit its generalization and make it far from being the definitive COVID-19 spreading model. Main limitations of our approach include:

*First Wave Data.* This data is from the first wave in 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. In order to help address this problem, we examine distributions from the Berlin data set regarding household size, but this problem still limits the scenarios that can be analyzed, e.g., the impact of school closures.

*Transportation Assumptions.* The KCDC data set does not show the transportation mode of patients. We overcome this limitation by extracting such information from the Berlin data set. The Seoul and Berlin data sets present comparable attributes and can be used for cross-fertilization.

*Data Set Volume.* Despite more than 9.5 million people lived in Seoul in 2020,[10] movements of only 4004 unique patients are logged in the KCDC data set. This might give rise to doubts on the representativeness of the data set. Although data sets with more information about movements of Korean people are not available, we verify that the information extracted from the KCDC data set (i.e., distributions presented in Section III) is emblematic of the population of a metropolitan city by comparing it with the information from the *German data sets* [3] which collects the movements of millions of individuals in Germany before the COVID-19 pandemic. We compare movements and habits of Seoul patients with those of the Berlin, Dusseldorf, Kelheim (district), and Leipzig. These data sets are suited for this analysis since they monitor movements of millions of people through their cellphones (i.e., provide person-centric data) but cannot be used to validate GeoSpread since they do not include information about the virus spread.

## VIII. Related Work

The COVID-19 pandemic has been studied extensively in recent months due to its disruptive effects. Different approaches have been adopted to increase our knowledge on the pandemic. Bao et al. [18] propose COVID-GAN, a framework that allows generating human mobility traces when different real-world conditions apply (e.g., local policies and disease severity). Pung et al. [19] interview COVID-19 patients in Singapore to collect epidemiological/clinical data to study the spread of the virus in three different Singapore *clusters*, this approach by its nature can be applied to populations of a small scale only. Blockchain is used to deploy a contact tracing system [15] and to predict the pandemic evolution from real-time data [16], Internet-of-Medical-Things is adopted to limit the contagion while gradually lifting restrictions [17]. A co-location model is used in [20] to study the spread of SARS-CoV-2 with limited data. Contreras et al. [21] use a numerical simulation to evaluate the efficiency of a test-trace-and-isolate strategy in containing the COVID-19 pandemic in Germany. An ML-based framework is proposed in [14] to estimate dynamically changing values (i.e., contact, recovery, and mortality rates) of a SIRD epidemiological model (acronym of Susceptible, Infected, Recovered and Deceased individual) starting from available mobility data and epidemic trends. Epidemiological models study how an infection spreads on a larger scale and are either mathematical or agent-based.

*Mathematical models* are defined by a set of equations that allow describing the evolution of the disease [22]. Kermack et al. [13] develop a SIR model based on differential equations to study the spread of diseases. SIR models are widely adopted in the literature. Since they do not consider spatial attributes, the analysis of space-related scenarios is not supported. Bi et al. [23] use conditional logistic regression to study the transmission of COVID-19 in Shenzhen, China. Using data from contact-based surveillance and accurate infector-infectee relationships, they confirm that, on average, COVID-19 has an incubation period of less than a week and a long clinical course. Rader et al. [24] evaluate the socio-economic and environmental aspects of a region affect the spreading of COVID-19 but do not focus on the actual virus spread.

Pejó and Biczók [25] use game theory to evaluate the efficiency of face masks and social distancing in limiting the spread of COVID-19 when some selfish patients do not use any counter-measures. Bhattacharyya and Bauch [26] use game theory to study the efficiency of protective vaccines, i.e., the safest way to achieve herd immunity [27].

*Agent-based models (ABMs)* are a simulation-based alternative of mathematical models that incorporate human

[10] https://bit.ly/3AZhe99

interactions [28]. ABMs are widely used in the literature to successfully model the spread of diseases [29].

Ferguson et al. [30] model the spread of influenza in British and American households, schools, and workplaces. Their simulations are parameterized using census and land use data as well as air travel patterns. Note that the above work considers only large scale (international) population movements. ABMs parameterized by census data have been used to capture the spread of COVID-19 in Australia [31]. Using census and age-distribution data from Germany and Poland, Bock et al. [32] investigate the efficiency of mitigation strategies by accounting for interactions within households where it is hard to social distance. Almagor et al. [33] use an ABM to evaluate the effectiveness of contact tracing app to limit the spread of COVID-19. Kim et al. [12] use synthetic, location-based social network data to study outbreaks and evaluate the effectiveness of different mitigation strategies, especially how social behaviors affect the virus spread. ABMs are used also to model the spread of SARS-CoV-2 in small areas, e.g., supermarkets [34]. *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. [3] use an ABM parameterized with mobility traces from mobile phone data for public transportation applications to study the COVID-19 outbreak in Berlin. This work is the closest to the one presented here but does not have any detailed statistics on agent mobility during the pandemic.

*Here, we extract human movement habits and dynamics from the KCDC data set of real COVID-19 patients. Statistics on patient mobility, traveled distance, and visited locationare used to tune GeoSpread and model 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. GeoSpread allows the investigation and identification of mitigation strategies.*

## IX. Conclusion

Information and routes of South Korean COVID-19 patients are analyzed to study the disease outbreak in the Gangnam and Seocho districts of Seoul. We enrich this analysis by analyzing detailed mobility data of four German cities and districts, i.e., Berlin, Dusseldorf, Kelheim (district), and Leipzig. Movement habits in South Korea are extracted from available data sets (i.e., the KCDC dataset cross-fertilized with the Berlin one) to parameterize simulations in GeoSpread, our tool based on ABM and GIS, and to study interactions among people. Simulation results are in excellent agreement with ground truth and show that this model can be used to flexibly examine and evaluate a wide variety of different scenarios based on different human mobility patterns from real-world data. While we do not claim that it is a definitive COVID-19 spread model, GeoSpread can be used to investigate useful *what-if* scenarios. We plan to (i) expand the simulation model to capture more details on a wide variety of mitigation measures, (ii) extract additional information from the data in [3] to investigate the impact of public transport and minors' movement habits on the COVID-19 pandemic, and (iii) include dynamic parameter computation in GeoSpread.

## Acknowledgments

We would like to thank Prof. David Dowdy (Johns Hopkins, Bloomberg School of Public Health) for his constructive feedback on the manuscript. The authors also thank the Editor and the anonymous reviewers for their constructive comments and valuable feedback.

## References

[1] J. Kim and J. Lee, "Data science for COVID-19 (DS4C)," 2020. [Online]. Available: https://web.archive.org/web/20221202145858/https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset

[2] S. Kim and M. C. Castro, "Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020)," *Int. J. Infect. Dis.*, vol. 98, pp. 328–333, 2020.

[3] S. A. Müller et al., "Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data," *PLoS One*, vol. 16, no. 10, 2021, Art. no. e0259037.

[4] A. Schmedding, L. Yang, R. Pinciroli, and E. Smirni, "GeoSpread: An epidemic spread modeling tool for COVID-19 using mobility data," in *Proc. Conf. Inf. Technol. Social Good*, 2022, pp. 125–131.

[5] K. Sullivan, M. Coletti, and S. Luke, "GeoMason: Geospatial support for MASON," George Mason Univ., Fairfax, VA, Tech. Rep. GMU-CS-TR-2010-16, 2010.

[6] A. Schmedding, L. Yang, R. Pinciroli, and E. Smirni, "Replication package: GeoSpread: An epidemic spread modeling tool for COVID-19 using mobility data," 2022. [Online]. Available: https://github.com/akschmedding/GeoSpread

[7] K. S. Trivedi, *Probability and Statistics With Reliability, Queuing and Computer Science Applications, Second Edition*. Hoboken, NJ, USA: Wiley, 2002.

[8] H. Lyu, L. Chen, Y. Wang, and J. Luo, "Sense and sensibility: Characterizing social media users regarding the use of controversial terms for COVID-19," *IEEE Trans. Big Data*, vol. 7, no. 6, pp. 952–960, Dec. 2021.

[9] F. Kamran and J. Wiens, "Estimating calibrated individualized survival curves with deep learning," in *Proc. Conf. Artif. Intell.*, AAAI Press, 2021, pp. 240–248.

[10] Y. Kang, Y. He, J. Luo, T. Fan, Y. Liu, and Q. Yang, "Privacy-preserving federated adversarial domain adaptation over feature groups for interpretability," *IEEE Trans. Big Data*, to be published, doi: 10.1109/TBDATA.2022.3188292.

[11] P. Fang et al., "How to realize efficient and scalable graph embeddings via an entropy-driven mechanism," *IEEE Trans. Big Data*, vol. 9, no. 1, pp. 358–371, Feb. 2023.

[12] J.-S. Kim et al., "Location-based social simulation for prescriptive analytics of disease spread," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 53–61, 2020.

[13] W. O. Kermack, A. G. McKendrick, and G. T. Walker, "A contribution to the mathematical theory of epidemics," *Proc. Roy. Soc. London. Ser. A., Containing Papers Math. Phys. Character*, vol. 115, no. 772, pp. 700–721, 1927.

[14] V. L. Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "An epidemiological neural network exploiting dynamic graph structured data applied to the COVID-19 outbreak," *IEEE Trans. Big Data*, vol. 7, no. 1, pp. 45–55, Mar. 2021.

[15] Z. Peng, C. Xu, H. Wang, J. Huang, J. Xu, and X. Chu, "P $^{2}$ B-trace: Privacy-preserving blockchain-based contact tracing to combat pandemics," in *Proc. Int. Conf. Manage. Data*, 2021, pp. 2389–2393.

[16] B. S. Egala, A. K. Pradhan, V. Badarla, and S. P. Mohanty, "iBlock: An intelligent decentralised blockchain-based pandemic detection and assisting system," *J. Signal Process. Syst.*, vol. 94, no. 6, pp. 595–608, 2022.

[17] A. K. Tripathy, A. G. Mohapatra, S. P. Mohanty, E. Kougianos, A. M. Joshi, and G. Das, "EasyBand: A wearable for safety-aware mobility during pandemic outbreak," *IEEE Consum. Electron. Mag.*, vol. 9, no. 5, pp. 57–61, Sep. 2020.

[18] H. Bao, X. Zhou, Y. Zhang, Y. Li, and Y. Xie, “COVID-GAN: Estimating human mobility responses to COVID-19 pandemic through spatio-temporal conditional generative adversarial networks,” in *Proc. Int. Conf. Adv. Geographic Inf. Syst.*, 2020, pp. 273–282.

[19] R. Pung et al., “Investigation of three clusters of COVID-19 in Singapore: Implications for surveillance and response measures,” *The Lancet*, vol. 395, pp. 1039–1046, 2020.

[20] S. Zeighami, C. Shahabi, and J. Krumm, “Estimating spread of contact-based contagions in a population through sub-sampling,” *Proc. VLDB Endowment*, vol. 14, no. 9, pp. 1557–1569, 2021.

[21] S. Contreras et al., “The challenges of containing SARS-CoV-2 via test-trace-and-isolate,” *Nature Commun.*, vol. 12, no. 1, pp. 1–13, 2021.

[22] H. V. D. Parunak, R. Savit, and R. L. Riolo, “Agent-based modeling vs. equation-based modeling: A case study and users’ guide,” in *Proc. Int. Workshop Multi-Agent Syst. Agent-Based Simul.*, Springer, 1998, pp. 10–25.

[23] Q. Bi et al., “Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study,” *Lancet Infect. Dis.*, vol. 20, no. 8, pp. 911–919, 2020.

[24] B. Rader et al., “Crowding and the shape of COVID-19 epidemics,” *Nature Med.*, vol. 26, no. 12, pp. 1829–1834, 2020.

[25] B. Pejó and G. Biczók, “Corona games: Masks, social distancing and mechanism design,” in *Proc. ACM SIGSPATIAL Int. Workshop Model. Understanding Spread COVID-19*, 2020, pp. 24–31.

[26] S. Bhattacharyya and C. Bauch, ““Wait and see” vaccinating behaviour during a pandemic: A game theoretic analysis,” *Vaccine*, vol. 29, no. 33, pp. 5519–5525, 2011.

[27] G. D’Souza and D. Dowdy, “Rethinking herd immunity and the COVID-19 response end game,” 2021. [Online]. Available: https://web.archive.org/web/20221202151949/https://publichealth.jhu.edu/2021/what-is-herd-immunity-and-how-can-we-achieve-it-with-covid-19

[28] R. A. Kelly et al., “Selecting among five common modelling approaches for integrated environmental assessment and management,” *Environ. Modelling Softw.*, vol. 47, pp. 159–181, 2013.

[29] A. Crooks and A. Hailegiorgis, “An agent-based modeling approach applied to the spread of cholera,” *Environ. Modelling Softw.*, vol. 62, pp. 164–177, 2014.

[30] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, “Strategies for mitigating an influenza pandemic,” *Nature*, vol. 442, no. 7101, pp. 448–452, 2006.

[31] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, “Modelling transmission and control of the COVID-19 pandemic in Australia,” *Nature Commun.*, vol. 11, no. 1, pp. 1–13, 2020.

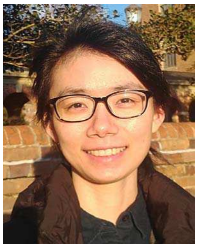
[32] W. Bock et al., “Mitigation and herd immunity strategy for COVID-19 is likely to fail,” medRxiv, 2020, doi: 10.1101/2020.03.25.20043109.

[33] J. Almagor and S. Picascia, “Exploring the effectiveness of a COVID-19 contact tracing app using an agent-based model,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020.

[34] F. Ying and N. O’Clery, “Modelling COVID-19 transmission in supermarkets using an agent-based model,” *PLoS One*, vol. 16, no. 4, 2021, Art. no. e0249821.

**Anna Schmedding** received the BS degree in computer science and mathematics from the York College of Pennsylvania, in 2016, the MS degree in mathematics from Syracuse University, in 2019, and the MS degree in computer science from William & Mary, in 2021. She is currently working toward the PhD degree with William & Mary Computer Science Department. Her research interests include reliability, workload characterization, and performance.

**Riccardo Pinciroli** received the MS and PhD degrees in computer engineering from the Politecnico di Milano, Italy, in 2014 and 2018, respectively. He is a postdoctoral fellow in computer science with Gran Sasso Science Institute, Italy. His research interests include stochastic modeling, performance evaluation, energy efficiency, and uncertainty propagation applied to cloud computing, data-centers, and cyber-physical systems.



**Lishan Yang** (Member, IEEE) received the BS degree in computer science from the University of Science and Technology of China (USTC), in 2016. She is currently working toward the PhD degree with Computer Science Department, William & Mary. Her research interests are in GPU architecture, reliability analysis, and performance analysis. She is a member of ACM.

**Evgenia Smirni** (Fellow, IEEE) is the Sidney P. Chockley professor of computer science with William and Mary, Williamsburg, Virginia. Her research interests include queuing networks, stochastic modeling, resource allocation, storage systems, cloud computing, workload characterization, and modeling of distributed systems and applications. She is an ACM distinguished scientist.