Poster Abstract: Simulating COVID-19 Containment Measures Using the South Korean Patient Data

Lishan Yang
lyang11@email.wm.edu
William & Mary
Williamsburg, VA, USA

Anna Schmedding
akschmedding@email.wm.edu
William & Mary
Williamsburg, VA, USA

Riccardo Pinciroli
rpinciroli@wm.edu
William & Mary
Williamsburg, VA, USA

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

ABSTRACT

As the COVID-19 outbreak evolves around the world, the World Health Organization (WHO) and its Member States have been heavily relying on staying at home and lock down measures to control the spread of the virus. In last months, various signs showed that the COVID-19 curve was flattening, but the premature lifting of some containment measures (e.g., school closures and telecommuting) are favouring a second wave of the disease. The accurate evaluation of possible countermeasures and their well-timed revocation are therefore crucial to avoid future waves or reduce their duration. In this paper, we analyze patient and route data collected by the Korea Centers for Disease Control & Prevention (KCDC). We extract information from real-world data sets and use them to parameterize simulations and evaluate different *what-if* scenarios.

CCS CONCEPTS

• **Computing methodologies → Agent / discrete models**; • **Information systems → Geographic information systems**; **Data analytics**; • **Mathematics of computing → Exploratory data analysis**.

KEYWORDS

COVID-19, SARS-CoV-2, Coronavirus, Data Analysis, Agent-Based Model (ABM), Geographic Information System (GIS), Simulation

1 INTRODUCTION

The first human cases of COVID-19 were observed in Wuhan, China, at the end of December 2019. Since then, COVID-19 spread in 114 countries all over the world. On March 11, 2020, the WHO declares that COVID-19 is the first pandemic caused by a coronavirus. Several measures to flatten the infection curve include wearing face masks, social distancing, and human movement restrictions. Simulation and mathematical models can be used for evaluate the

effectiveness of difference measures, but their effectiveness is tied to their parameterization with realistic data to obtain a clear picture of expected alternative measures.

In this paper, we analyze data of COVID-19 patients (and their routes) [2] collected by the Korean CDC and extract information on movement habits (e.g., travel speed and distance) of people living in Seoul, South Korea, to provide a useful tool to study spread of the virus. Such data is fed to GeoMason [3], a tool that uses agent-based models (ABM) and geographic information systems (GIS) to study disease outbreaks (e.g., cholera [1]). This way, we simulate interactions of people in Seoul roads and buildings to investigate the COVID-19 outbreak in the largest metropolis of South Korea and evaluating different *what-if* scenarios. A demo of our simulation can be found in this youtube video (https://youtu.be/H3qYZ47O6wU).
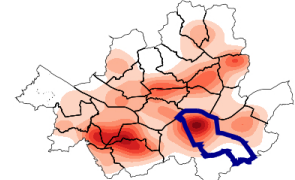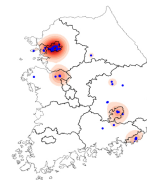
(a) South Korea. Blue points indicate hotspots.

(b) Seoul. Gangnam district is outlined in blue.

Figure 1: Heat maps of most visited locations.

2 WORKLOAD CHARACTERIZATION

The data sets [2] used in this paper contain data collected by the KCDC and local governments from January 20, 2020, to May 31, 2020. They monitor infected people (PatientInfo) and log their movements (PatientRoute) using CCTV, cellphones, and credit card transactions.

PatientInfo data set. It provides epidemiological data of 4004 COVID-19 patients in South Korea. Each entry represents a patient that is identified by an ID (*patient_id*). Other attributes are gender and age of patients (*sex* and *age*), their provenance (*country*, *province*, and *city*), whether they have been infected in a known case (*infection_case*), the ID of the patient that infected them (*infected_by*), the number of contacts with other people (*contact_number*), and the symptoms appearance date (*symptom_onset_date*).

PatientRoute data set. There are 8092 entries in this data set. Each entry provides the location visited during a specific *date* by the patient (*patient_id*) in PatientInfo. Locations are unequivocally identified by their *latitude* and *longitude*. *Province*, *city*, and *type* (e.g., airport, hospital, store) of each location are also provided. The visiting time is not available, but locations visited by a patient in a single day are logged in chronological order. Although the information is not as accurate as one would like, they still allow for analyzing patient movements and interactions with high accuracy.
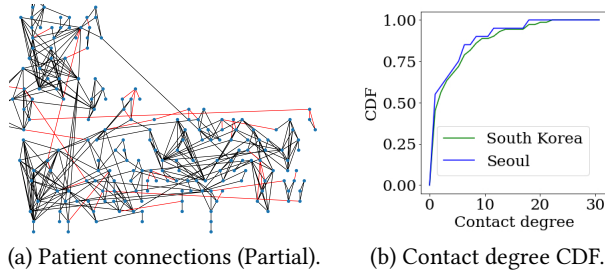
(a) Patient connections (Partial).   (b) Contact degree CDF.

Figure 2: Patient contacts.

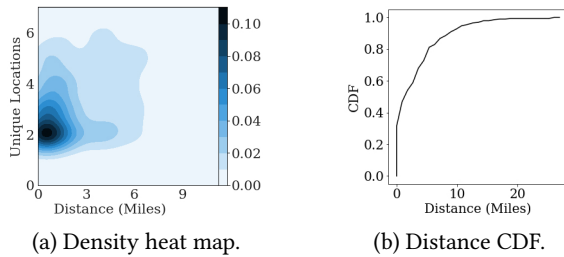(a) Density heat map.   (b) Distance CDF.

Figure 3: Daily traveled distance and visited locations.

Figures 1(a) and (b) depict a heat map of the most visited locations in South Korea and Seoul, respectively, showing where COVID-19 outbreaks are more likely to happen. In Seoul, the south-west and south-east (i.e., Gangnam district) areas are those with more patient routes.

Figure 2(a) presents a subset of patient connections. To improve visibility, we only present a small subset of the graph of patient connections. In this figure nodes depict patients, black edges connect patients that visited the same place during the same day, and red edges represent the virus spreading information obtained from the PatientInfo data set (i.e., *infected_by* attribute). The node degrees in Figure 2(a) illustrate the contact degree among patients. Figure 2(b) shows the CDF of contact degree of all patients (two CDFs, one for the entire South Korean and one just for Seoul). The figure shows that super spreaders (i.e., patients that tie many of the outbreaks together) exist. Such patients contribute to the large groups of connected patients in Figure 2(a).

Figure 3(a) plots the density heat map of daily distance traveled by patients in Seoul and the number of visited locations, two important features due to the vital nature of patient movement to spread COVID-19. With some exceptions, people mostly travel short distances and only visit a few locations each day. The CDF of the daily traveled distance is shown in Figure 3(b).

Patient mobility is another important attribute describing how much patients move. When exploring the data set, we observe some patients who visit many places during some days, and only one or two places on other days. Defining a *high mobility day* as a day during which a patient visit at least L locations, the mobility of a patient is computed as the ratio of the patient high mobility days to all logged days for this specific patient. The CDF of patient mobility (for $L = 4$) is shown on the right of Figure 4, where shows that 60% of patients never visit more than 4 locations in a day.

3 SIMULATION AND PRELIMINARY RESULTS

In this section, we show how to parameterize a simulation based on a patched version of GeoMason [3] using the characterization in Section 2.
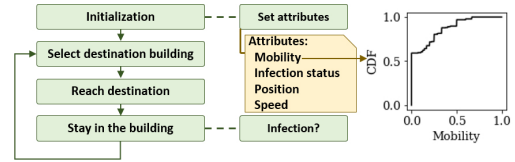

Figure 4: Life cycle of an agent.

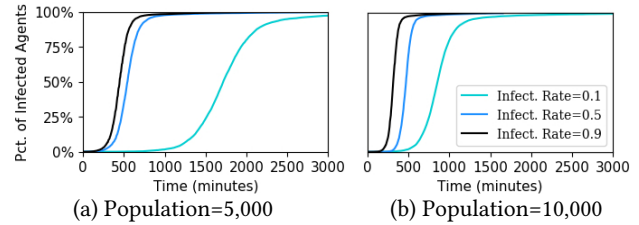(a) Population=5,000   (b) Population=10,000

Figure 5: CDF of infected agents within 3,000 minutes.

The attributes and life cycle of an agent are shown in Figure 4. The following attributes are set during the initialization phase:

(1) *Mobility.* It is obtained from the characterization in Section 2 and its CDF is shown on the right of Figure 4.
(2) *Infection status.* A random agent is selected as patient zero in the considered area.
(3) *Position.* Agents are randomly placed in the simulated area.
(4) *Speed.* There are two types of agents: 50% of agents walk at a speed of 3 MPH; other agents drive a vehicle and their speed is uniformly distributed between 10 and 25 MPH.

The simulation time is defined by cycles. At each 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. If multiple agents are inside the same building, they may infect each other. 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 in the same building to be infected is:

$$Pr(infection) = 1 - (1 - \alpha)^n. \qquad (1)$$

The cycle length is set to 5 minutes and the simulation stops when all agents are infected. We simulate the COVID-19 outbreak in the Gangnam district, i.e., the sub municipality of Seoul with the most hotspots, see Figure 1(b). 7043 buildings of this area are considered. Figure 5 shows the time to infect the entire population for different infection rates and population sizes. Containment measures such as partial lockdowns of specific areas, geographic movement restriction, curfews, and other protective measures to reduce the transmission rate can be simulated directly using this simulator and are subject of ongoing work.

ACKNOWLEDGEMENT

REFERENCES

[1] Andrew Crooks and Atesmachew Hailegiorgis. 2014. An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software* 62 (2014), 164–177.
[2] Jihoo Kim and JoongKun Lee. 2020. Data Science for COVID-19 (DS4C). https://www.kaggle.com/kimjihoo/coronavirusdataset. [Accessed on 2020-09-25].
[3] Keith Sullivan, Mark Coletti, and Sean Luke. 2010. *GeoMason: Geospatial support for MASON.* Technical Report. Department of Computer Science, George Mason University.