Infection Risk Score: Identifying the risk of infection propagation based on human contact

Rachit Agarwal
rachitag@iitk.ac.in
IIT Kanpur, India

Abhik Banerjee
abanerjee@swin.edu.au
Swinburne University, Australia

ABSTRACT

A wide range of approaches have been applied to manage the spread of global pandemic events such as COVID-19, which have met with varying degrees of success. Given the large-scale social and economic impact coupled with the increasing time span of the pandemic, it is important to not only manage the spread of the disease but also put extra efforts on measures that expedite resumption of social and economic life. It is therefore important to identify situations that carry high risk, and act early whenever such situations are identified. While a large number of mobile applications have been developed, they are aimed at obtaining information that can be used for contact tracing, but not at estimating the risk of social situations. In this paper, we introduce an infection risk score that provides an estimate of the infection risk arising from human contacts. Using a real-world human contact dataset, we show that the proposed risk score can provide a realistic estimate of the level of risk in the population. We also describe how the proposed infection risk score can be implemented on smartphones. Finally, we identify representative use cases that can leverage the risk score to minimize infection propagation.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **Mobile computing**; • **Networks** → *Peer-to-peer protocols.*

KEYWORDS

Infection risk score, Contact Tracing, Mobile Computing, Internet of Things, Mobile Health

ACM Reference Format:
Rachit Agarwal and Abhik Banerjee. 2020. Infection Risk Score: Identifying the risk of infection propagation based on human contact. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3423459.3430754

1 INTRODUCTION

21st century has already been witness to multiple pandemics in the first two decades, with the biggest being COVID-19 caused by the SARS-CoV-2 virus. The unprecedented spread of the COVID-19 has led to global efforts by governments to contain the pandemic and to limit the impact of the virus on human society. As with any other infectious disease, the efforts to contain the virus largely focus on *(i)* minimizing human-to-human contact by enforcing people to maintain a certain distance with others (also known as ***social distancing***), *(ii)* minimizing economic activity (also known as ***lockdown***) for a certain period in geographical regions, such as cities, states or even entire countries, and *(iii)* by performing ***contact tracing***, which involves tracking the disease spread by identifying the contacts of the confirmed cases. However, these efforts have been met with varying degrees of success, and the authorities have been trying to use technology as much as possible to elevate their efforts [13].

With the rise of the Internet of Things (IoT) and mobile health [33] (also referred to as *mHealth*), there has been a growth in the number of possibilities related to not only understanding the environment but also detecting diseases early. With regards to the COVID-19 pandemic in particular, governments around the world have looked to leverage the use of smartphone applications for limiting the spread of the disease, given the ubiquity of smartphone usage. While many of these applications focus on providing up-to-date information about the spread of the disease, other applications aim to notify users in real-time when they come in contact with an infected person [11]. These infection tracking applications use a variety of sensors embedded in a smartphone to help detect the transmission in real-time. A common type of sensor used is Bluetooth Low Energy (BLE), which can be used for proximity detection. Multiple applications that leverage BLE for the purpose of monitoring the growth of the COVID-19 pandemic have been introduced in various countries. In India, the Aarogya-Setu Application [20] informs how many infected people are within a certain distance of a person using the application by matching with national database of the infected people. In Australia, the COVIDSafe application [2] provide notifications to the users if their contact is detected with a confirmed infected person. Similar applications have been developed by the governments of many other countries. Further, a collaboration between Apple Inc. and Google has led to the development of an Exposure API that enables developers to build various applications using which application users can know if they came into contact with other infected people [6]. Apart from smartphone applications, other types of technologies are also used to help in the cause of containing pandemic. These include the use of SwipeSense technology to track use of medical equipment and to track whether hospital staff wash their hands regularly[1].

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430754

[1] https://www.cnbc.com/2020/08/02/hospitals-tracking-covid-19-with-badge-sensors-swipesense-technology.html

Despite the technological innovations and advancements, the use of applications, such as those described above, and technologies for managing and controlling the spread of the infection is challenging due to multiple reasons. Firstly, a person carrying the disease may not show any symptoms for a long period (e.g. when the infection is in the incubation period). As a result, any close contacts with other people would not be detected as being risky by infection tracking applications, and hence would not help in containing the spread of the infection. Indeed, in the case of COVID-19, a significant fraction of cases have been identified as asymptomatic for the entire duration of infection [8]. These also contribute to community spread of the disease, which can lead to exponential growth in the number of infections. Secondly, once someone is confirmed as infected, he/she is typically isolated and is not allowed to get involved in any social activities until fully recovered. Thirdly, existing methods for managing infection spread are primarily reactive. Counter-measures are often taken after a person is confirmed to be infectious. Subsequently, authorities proceed with counter measures such as lockdown of the specific geographical region. Thus, currently, the scope for detection of the infection spread and its management is limited. Therefore, there is a need for an early estimate of the potential risk in a geographical region to enable authorities to act quickly.

For risk estimation to be effective, it needs to identify people who have greater exposure to the infection, quarantine the exposed people, identify regions with potentially high exposures, and declare a region as hot-spot even before the outbreak happens in that region. Additionally, the risk estimation measure should also be able to identify situations which are likely to lead to transmissions even when there are not any confirmed presence of the known infections. Finally, any such risk estimation needs to be adaptable to a wide range of technology platforms. While a notion of risk score has been introduced as part of the Exposure API by Apple Inc and Google, its main drawbacks is that it only provides a risk measure based on confirmed exposures to infections.

In this paper, we present a risk score that can be used to assess the risk for individuals based on contact events identified using smartphones. A key novelty of the proposed risk score is that it estimates the risk propagation, unlike existing literature that only assess immediate risk. The proposed risk score can be used to assess the level of risk within geographical regions, enabling authorities to act early to contain a potential outbreak. Further, monitoring the risk score can also help individuals take actions.

In particular, the key contributions of this paper are as follows:

- ***Infection risk score***: We introduce a risk score that estimates infection propagation by monitoring contact events among individuals.The risk score takes into consideration factors such as the contact proximity, transmission likelihood and vulnerability to a disease.
- ***Evaluation using realistic dataset***: We evaluate the infection risk score using a real-world human contact dataset that has previously been using to study infection propagation. Our results show that potentially risky situations are well captured using the infection risk score.
- ***Adaption of risk score using smartphones***: We provide detailed description on how smartphones can be used to implement the infection risk score to track infections.

Finally, we also discuss how the accuracy of infection risk score can be improved by incorporating contextual information, and also present a discussion on potential use cases of the risk score for managing infection spread.

The rest of the paper is organised as follows. Section 2 provides detailed survey of related techniques and existing metrics used to quantify risk and exposure. Section 3 provides details of the proposed risk model. In Section 4 we evaluated the model using real data. Section 5 provides the details on the propose version of the smartphone application. This is followed by perspective uses of the risk score in section 6. We finally conclude in Section 7.

2 RELATED WORK

Recent studies focusing on containing pandemic can be mainly classified into three broad groups: survey based studies, IoT based studies and epidemic model based studies.

In survey based studies, in [19], authors report that factors such as contact with infected person, work overload, medical history of the person, and if the person wore Personal Protective Equipment (PPE) or not play an important role in determining the risk of infection transmission of COVID-19. Similarly, to identify potential exposure, WHO uses risk assessment forms to determine the risk of exposure. Here they ask questions related to if the person wore the PPE as recommended or not [34].

In IoT based studies, there is increased focus on smartphone based infection detection. Many applications and IoT Devices are available that perform contact tracing using proximity checks. A survey of some of these application is present in [11]. We do not survey these applications again and instead present, in brief, new applications and devices that have come-up since the publication of [11]. Recent applications and devices includes EasyBand, a wearable device that vibrates when a marked (infected) Easyband comes in close proximity [27]. Nonetheless, it has issues related to centralized control and communication. In [12], authors used magnetometer based proximity detection, while in [22] authors used multiple sensors to improve the distance estimation accuracy. Such techniques fail in the case when a smartphone lacks certain required sensor. Further, these applications achieve *privacy by architecture* and not *privacy by design*. Many recent application and IoT devices claim to follow privacy guidelines such as those mentioned in [27]. These applications and devices include: *(i)* Pan European Privacy-Preserving Proximity Tracing (PEPP-Pt)[2] that uses anonymized ID for communication, *(ii)* TraceSecure that uses secret sharing technique to identify proximity [4], and *(iii)* proximity-based privacy-preserving contact tracing (P^3CT) that uses ambient signature protocol [21]. Again, while these applications are privacy preserving, they achieve privacy by architecture. In summary, these studies model risk using factors such as distance [4, 12, 20, 22, 27] and duration [6]. These works mainly use either BLE or magnetometer to estimate distance from neighbor. Nonetheless, these works do not quantify the risk, and instead, just provide an estimate of whether a person was in contact with some other person or not.

[2]https://www.pepp-pt.org

On the other hand, from epidemic modeling point of view, there are many studies that quantify risk using different parameters such as: size of cough droplets, rate of cough, volume of particles generated, concentration of pathogens, max distance covered by pathogen in air, pathogen particles lost due to temperature and humidity, time an infected person stayed at a given location, duration of contact with susceptible person, and his pulmonary rate [26]. Most of these factors, until now, cannot be estimated using smartphone. Instead, disease specific average values for these factors can be used as constants while modeling risk score. In [26], authors estimated risk as an aggregation of risk score for both when a person comes in direct contact with other person and when a person gets infected indirectly (a case of ***community spreading***).

3 INFECTION RISK SCORE

In this section, we present the *infection risk score* that quantifies risk of catching an infection. Our score considers exposure to pathogen and context in a social network setting. For convenience, *infection risk score* is referred by the term *risk score* in the remainder of the paper.

3.1 Network model: Modeling the population as a temporal network

For any geographical area, we consider the population to be represented by a temporal graph G such that $G(V_t, E_t)$ is a temporal snapshot at time t that is created by individuals in a given area A_a. For the purpose of this paper, we consider that the risk score computation for individuals is done using mobile apps, and hence, each individual is represented using smartphones. Here V_t is the set of smartphones communicating and active at time t and E_t is the set of edges that exists between smartphones in V_t. Let Δt be the time difference between two consecutive temporal snapshots of G. For our model we assume that if two people are in contact, for say 10 epochs, then the edge between them is persistent over $\frac{10}{\Delta t}$ snapshots of the graphs. Each person $i \in V_t$ has a location, l_t, marked by latitude and longitude pair such that $l_t = (la_{i,t}, lo_{i,t})$. Given the interactions, at time t, each person i has a neighborhood, $N_{i,t}$ where each person $j \in N_{i,t}$ has an edge (in E_t) to the person i and is $d_{i,j,t}$ distance apart. Here $d_{i,j,t} < \theta_d$ i.e., i and j are within communication range and at maximum θ_d distance apart.

3.2 Risk score parameters

In this section, we identify the key factors that impact infection propagation.

(1) **Exposure caused by a neighbor**: Communicable diseases such as COVID-19 generally spread when a person $i \in V_t$ comes in close proximity with a infected person (person j) or touches the surface that infected person has touched [26]. In such a case, the person i is exposed to pathogens from the infected person, which can lead to infection spread. The exposure to a neighboring individual is a key factor determining the likelihood of a transmission event from a neighbor, and we term this as the **neighbor exposure**. For the scope of the current paper, we limit our discussion to the the exposure caused when an infected person come in close proximity, although this may easily be extended to include other modes of propagation.
To determine how the neighbor exposure impacts the spread of infection, we consider that an infected neighbor j exhales $n_{i,j,t} \in \mathbb{R}^+$ pathogens and these pathogens are homogeneously distributed within the permissible θ_d distance. Further, we consider the following assumptions: *(i)* there is no loss in pathogens, *(ii)* each time same number of pathogens are exhaled, and *(iii)* between two consecutive temporal snapshots of the graph (i.e., $G(V_t, E_t)$ and $G(V_{t-\Delta t}, E_{t-\Delta t})$), a person i stays in contact with person j for the Δt time. In such a scenario, the exposure to an infectious disease of the person i at time t with respect to a particular neighbor j is given by $E_{i,j,t} = \Delta t \times n_{i,j,t}$.
In ideal conditions, if a neighbor j is not infected, i.e., he/she does not cough, and wears proper protective gears such as face mask or face shields, $E_{i,j,t} = 0$ because there are no pathogens exhaled by j. In such a case, the whole idea of maintaining social distancing even when people are not infected would fail and susceptible people would be deemed harmless. On the other hand, if some neighbor is infected and coughing badly, $E_{i,j,t} >> 0$. In this case, other people would ideally limit from meeting the infected person. In such a situation also, barring the infected person, other susceptible people would continue their physical social activities. Let $r_{j,t-\Delta t}$ be the risk score of the neighbor at time $t - \Delta t$. To account for above mentioned aspects and ensure that social distancing is enforced between susceptible people also, we add the previous instance risk score of the neighbor to the exposure caused due to the neighbor, i.e., $E_{i,j,t} = \Delta t \times n_{i,j,t} + r_{j,t-\Delta t}$.

(2) **Neighbor weight**: We define the **neighbor weight** as the likelihood that an individual in the vicinity is infectious. Since we aim to estimate the risk even in situations where confirmed infections are not known, the neighbor weight can be estimated based on multiple contextual parameters. For instance, in the case of communicable diseases such as COVID-19, if a neighbor is from a hot-spot area or has a history of the disease then the risk of getting infection from the neighbor is high because the neighbor is coming from a containment zone. Further, impact of diseases like COVID-19 is high on people who have a weak immunity either due to age or have chronic diseases like kidney failure and diabetes. On top, if a person is staying indoor, with a poor ventilation chances of spreading the disease and getting infected increases manyfold [28, 32]. In [28], authors recommend that proper ventilation indoor can reduce infections up-to 60%. Nonetheless, for COVID-19, different countries have different statistics, for example, India having relatively younger population, middle age people are more infected while more older people have died. Let $w_{j,t}$ be the ***weight*** such that $w_{j,t} \in [0, 1]$ that identifies such contextual information of the neighbor. Summed over all the neighbors of the person i at time t, the total exposure of i from its neighbors $j \in N_{i,t}$ is

thus given by equation 1.

$$E_{i,t} = \sum_{j \in N_{i,t}} w_{j,t} \times (E_{i,j,t} + r_{j,t-\Delta t}) \quad (1)$$

3.3 Risk score formulation

In addition to the neighbor weight and exposure, we define **vulnerability** as the likelihood that an individual exposed to risky situations continues to be at risk. At any time t the risk score of an individual i is the dependent on the risk score at $t - \Delta t$, his vulnerability, and the exposure from the neighbors at t. The total risk, thus, is given by equation 2.

$$r_{i,t} = \frac{v_{i,t} \times r_{i,t-\Delta t} + \sum_{j \in N_{i,t}} w_{j,t} \times (E_{i,j,t} + r_{j,t-\Delta t})}{1 + \sum_{j \in N_{i,t}} w_{j,t}} \quad (2)$$

Here the denominator is the normalization factor. A disease usually has a period between when the person i gets infected from the disease and time when he becomes an active spreader of the disease. For example, for COVID-19, the median incubation period is around 5 to 6 days[3]. In our scenario, even if a person comes in contact with a person for whom the disease is still in incubation period, the risk exposure is equally high as compared to meeting a person who is an active spreader. Thus, our model does not consider the incubation period.

From the equation (2), the value of $r_{i,t} \in \mathbb{R}^+$. If the person is taken into isolation (i.e., no interaction with neighbors) after getting infected, his risk score will decrease with a factor $v_{i,t}$ and will eventually decay and reach minimum in $t = \lceil \frac{r_{i,t-\Delta t}}{v_{i,t}} \rceil$ time instances. This accounts for the fact that risk to and from such people is minimized when they are in isolation. For simplicity, at $t = 0$ (or the initial condition) for all people we assign them as susceptible and their risk score to $r_{i,0} = 1$. As the actual infection state of a person is unknown, the idea of social distancing mandates to maintain a certain distance even if the person is susceptible. Maintaining social distancing reduces the possibility of getting infected. We assign a non zero value to $r_{i,0}$ to ensure that social distance is maintained and our model captures it. For simplicity, let $r_{i,0} = 1$. As and when a person is officially tagged infected, we assign $r_{i,t} = 2$. Note that, a low value of $r_{i,t}$, is achieved when all the neighbors are susceptible. For a new person joining in, we assume that he is a susceptible person.

Our method only considers ego network of a person for the calculation of the risk score. This enables all the smartphones involved to compute their individual risk scores simultaneously.

4 EVALUATION AND VALIDATION

In this section we provide an evaluation and validation of our risk model using a real-world dataset.

4.1 Dataset

While there are many datasets which have previously been used to study epidemic spread, specially smartphone based datasets that use Call Detail Records (CDRs) and GPS location information, [5], they are *(i)* not widely used [23], and *(ii)* mostly generated from a random population sample which do not reflect true neighborhood size. Instead, we use a dataset of 789 individuals (including students and teachers) obtained on a single day in an American high school that has 158 rooms [25], which has previously been used to study spread of infectious diseases [28]. Here each point of interest (POI) is considered to be a room in the school. The dataset is mainly used to study human contact network for infectious disease transmission. The dataset is collected between 6AM to 4:30PM at an interval of 20 seconds. The granularity of positioning information available is at the level of rooms, and hence, each individual is geo-tagged with the room ID they are in at a particular epoch. We consider that contact events occur between individuals whenever they are in the same room, and all individuals present in a particular room at a given epoch are connected to each other.

The temporal distribution of individuals in the dataset is shown in Fig. 1. Fig. 1(a) shows the heatmap of number of people present in a room at different epochs. The white color represents that nobody was present in a room at the particular epoch. Fig. 1(b) presents total number of people in a school at a given epoch. A sudden increase and a sudden drop in the number of people accounts for the beginning of the school in the morning when people arrive, and the end of the day, when they went back from school. Fig. 1(c) presents the maximum number of rooms occupied by people. Note that at maximum only ≈62% rooms are occupied. Fig. 1(d) presents ratio between number of people in the school and rooms occupied at a given time. The maximum average density of people in a room is 9. A sudden increase at the end of the day is because most of the people were present in a single room. Fig. 1(e) presents number of times a given room was occupied during the data collection period. From the figure we infer that (i) some rooms were always empty and nobody went to those rooms, (ii) the entire population is concentrated in only a few rooms and after certain time period there is an exponential decrease in the population indicating the end of classes in the school, (iii) during the day, rooms gradually start to fill up and there is an exponential rise in the population size.

4.2 Dynamics of epidemic spread on the contact network

We evaluate the proposed risk score using both SI (Susceptible-Infected) and SIS (Susceptible-Infected-Susceptible) models. Here, we note that, since the risk score looks at contact events only, we do not evaluate an SEIR (Susceptible-Exposed-Infected-Recovered) model, as identification of "exposed" and "recovered" states are not the focus of the model.

Let $S_{i,t}$ be the fraction of people that are susceptible in a region i at time t, $I_{i,t}$ be the fraction of population that is infected in a region i at time t, $N_{i,t} = 1$ be the faction of total population in the region at time t, β_i be the infection rate in the region i, and γ_i be the recovery rate in the region i. Note that at any given point of time $N_{i,t} = S_{i,t} + I_{i,t}$ because we consider only two states, susceptible and infected. The change in the fraction of susceptible and infected people over time is given by equation (3) [10]. Here the underlying assumptions are that there is a homogeneous mixing of the population and no birth and death happens (the total population is fixed).

[3] https://www.mohfw.gov.in/pdf/DGSOrder04of2020.pdf

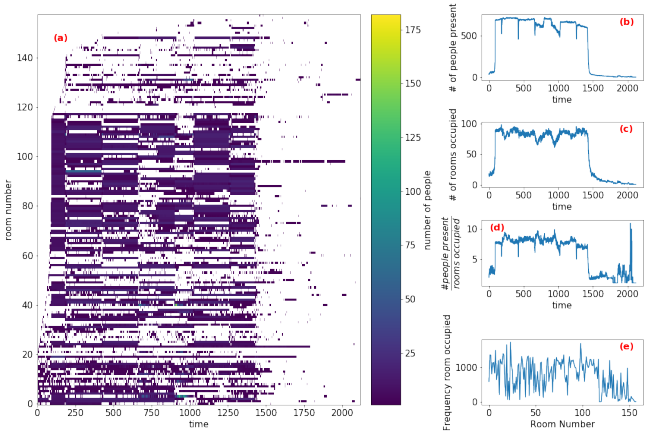


Figure 1: Temporal distribution of people in the rooms. (a) Heatmap showing number of people in each room over time. The white color represents empty room at the particular epoch. (b) Total number of people present in the school over time. (c) Total number of room occupied in the school over time. (d) average density of each room. (e) Number of times a room is occupied.

$$\begin{aligned} \frac{dS_{i,t}}{dt} &= -\frac{\beta S_{i,t} I_{i,t}}{N_{i,t}} + \gamma I_{i,t} \\ \frac{dI_{i,t}}{dt} &= -\frac{dS_{i,t}}{dt} \end{aligned} \tag{3}$$

4.3 Results

Currently, the exact behavior of exposure and vulnerability parameters for pandemics such as COVID-19 is not known. Further, as the dataset is POI based, actual distances are also not available in the dataset. Thus, we assume that the exposure parameter for each person is normally distributed with $\mu = 0.5$ and $\sigma = 0.1$. Further, the vulnerability parameter is also normally distributed with $\mu = 0.5$ and $\sigma = 0.2$.

For our analysis we study following three aspects using different initial condition, infection rate, and recovery rate. First, we identify fraction of people who are identified infected using the SIS and SI epidemic models. This helps us understand the infection spread over time in the population and understand the *dynamics on* the contact network. Second, we measure the ratio between the median risk scores of infected people and susceptible people. A ratio more than one indicates that the risk score of infected people is more, as intended. A higher ratio implies that the risk score can be used to better identify people who are exposed to infection and have high probability to get infected. When there are no infections in a neighborhood, this value tends to 0. Third, we study the fraction of people that are alerted using our model.

To test and study the above-mentioned aspects, as an initial condition, the values for $I_{i,0}$, β_i and γ_i used are $I_{i,0} \in \{0.0, 0.01, 0.5\}$, $\beta_i \in \{0.0, 0.5, 1.0\}$ and $\gamma_i \in \{0.0, 0.75\}$. $I_{i,0} = 0.0$ states that there are no initial infections in the region. $\beta_i = 0.0$ states there are no transmission happening and the disease does not spread via contact. On the other hand, $\beta_i = 1.0$ would state that the disease is highly contagious. Similarly, $\gamma_i = 0.0$ would state that there is no recovery which is equivalent to SI type epidemic model. $\gamma_i = 0.75$ would mean that the recovery rate is 75% (i.e. similar to the recovery rate of COVID-19 patients in India[4]). The results presented here are averaged over 50 simulations runs and conducted using python.

Figures 2 and 3 present results obtained for the above-mentioned aspects when different values of $I_{i,0}$, β_i and γ_i are used for SIS and SI models respectively. The β_i and γ_i values are assumed to not vary across rooms. Fig. 2 is obtained when $\gamma = 0.75$ while Fig. 3 is obtained when $\gamma = 0.0$. From the Fig. 2, as per SIS model, when there is no infection, dissemination of infection does not occurs because subsequent $dI_{i,t}/dt = 0$ (see fig. 2(a)). This lead to ratio of median risk scores (represented as $r_{m,Infected}/r_{m,Susceptible}$) to be 0 as there are no infected people (see Fig. 2(b)). The inset Fig. 2(b') shows the median risk scores of susceptible people ($r_{m,Susceptible}$) and indicates that, even when there are no confirmed infections, crowded situations which carry high risk can be identified as having high risk scores. As our risk model is not dependent on β_i, γ_i, and initial infection, via risk score, we are able to detects potential risky situations (see fig. 2(c), 2(f), and 2(i)) which epidemic models such as SIS model are not able to detect. For cases when $I_{i,0} \geq 0.01$ and $\beta_i \in \{0.0, 0.5, 1.0\}$, we observe that infections either die off (for $\beta_i = 0.0$) or achieve stability (for $\beta_i \in \{0.5, 1.0\}$, see fig. 2(d) and 2(g)). The reason for reduction in infections is the recovery rate, while for stability it is the low number of people present when $epoch > 1500$. The ratio of median risk scores for different β_i and $I_{i,0} \geq 0.01$ is shown in fig. 2(e) and 2(h)) where we observe that after few epochs the ratio is < 2 and even reaches < 1 in short duration. This behavior is because (a) the median value of infected identified by SIS model is less than the median value that of susceptible people and (b) the number of infected is less that number of susceptible. From the fig. 2(f) and 2(g) we also see that, irrespective of the epidemic state, most of the people are at high risk.

On the other hand, from Fig. 3, we see that when there is no recovery (i.e., $\gamma_i = 0.0$) and when $I_{i,0} = 0$, the behavior is similar to previous scenario (see fig. 3(a), 3(b), 3(b'), 3(c), 3(f), and 3(i)). Nonetheless, when $I_{i,0} > 0.0$ and $\beta_i \neq 0.0$, the infections eventually reach entire population which is true as there is no recovery (see fig. 3(d) and 3(g)). Further, in this case, due to the above-mentioned reason, ratio of median risk scores is also high (see fig. 3(e) and 3(h)). Ratio equal to 1 is achieved when $\beta_i = 0$.

As the results presented here show, the growth in risk score values increases when there is increased contact of susceptible individuals with infected individuals. Further, even in situations where infections are unknown, the risk score values are shown to grow when there are more crowded situations and population movement among those. Thus, the risk score values can be used for managing infection spread, even before confirmed infections are identified.

5 RISK SCORE IMPLEMENTATION USING SMARTPHONES

In this section, we demonstrate how the proposed risk score can be implemented as part of a smartphone based infection tracking applications. There are two main components required for estimation

[4]https://www.financialexpress.com/lifestyle/health/indias-covid-19-recovery-rate-nears-75-case-fatality-rate-one-of-the-lowest-globally-at-1-86/2063108/

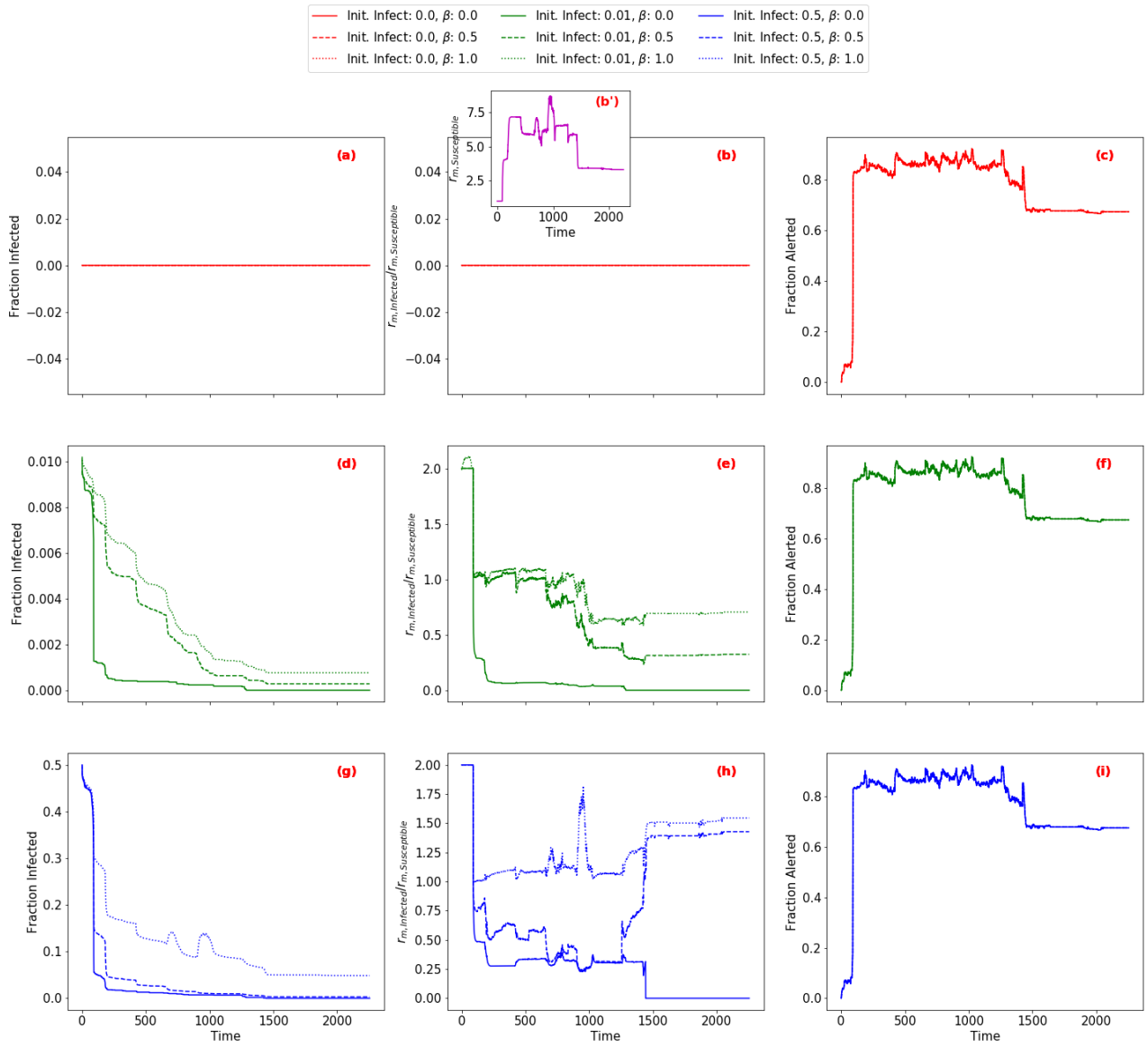


Figure 2: Results for SIS epidemic model, $\beta \in \{0.0, 0.5, 1.0\}$, $\gamma_i = 0.75$ and $I_{i,0} \in \{0.0, 0.01, 0.5\}$

of the risk score on a smartphone - *(a)* aggregation of risk scores of neighboring smartphones, and *(b)* computation of the risk score of the smartphone itself. Similar to the infection tracking applications used for COVID-19, we consider that estimation of the exposure is done using BLE. However, unlike existing applications which use centralized data repositories to obtain risk scores of neighboring smartphones (i.e. if they are confirmed to be infected), using our approach, each smartphone can *(a)* periodically broadcast its own risk score, by embedding this value in the BLE advertising packets, and *(b)* periodically update its own risk score by aggregating the risk scores of all other smartphones in its neighborhood. Such an approach has the following advantages:

- The risk score computation does not need to depend on a centralized database containing information about infected individuals, which might be outdated.
- The risk score reflect encounters not just with confirmed individuals, but also present environments that are risky from the perspective of infection spread.
- Our approach is better suited for privacy preservation, since no information pertaining to the identity of individuals is stored or communicated.

Next, we provide details on the design of the BLE advertising packet as well as how risk score computation can be done individually by a smartphone application. For the purpose of this discussion, we refer to the smartphone performing the risk computation as

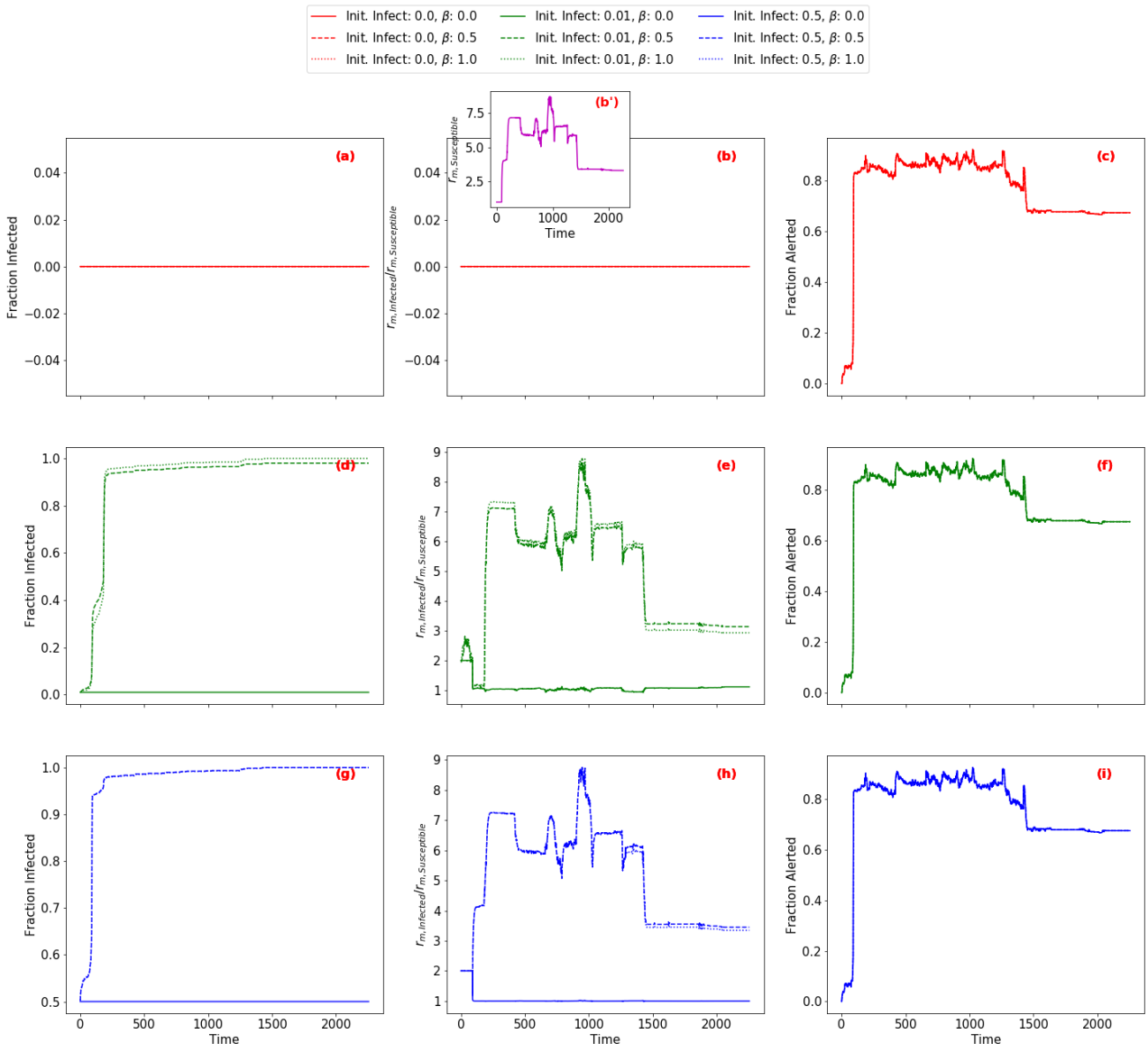


Figure 3: Results for SI epidemic model, $\beta \in \{0.0, 0.5, 1.0\}$, $\gamma_i = 0.0$ and $I_{i,0} \in \{0.0, 0.01, 0.5\}$

Figure 4: BLE Packet Format

the ego node, and all other smartphones in its vicinity as neighbor nodes.

5.1 BLE advertising packet

Existing infection tracking techniques record the BLE Media Access Control (MAC) addresses of nearby smartphones [14] and compare them with a centralized database of infected individuals. Instead, we discuss how we use the BLE advertising packet to communicate risk score values.

The BLE advertising packet allows including optional payload of up to 31 bytes [17]. We use these available bytes for broadcasting the risk score. Our payload includes:

(1) A 128-bit unique identifier (**UUID**) which is a fixed value used to identify the service, enabling each smartphone to filter out all nearby beacons broadcasting the risk score.
(2) A 6 bytes long **Risk score** which includes the risk score value rounded to two decimal places and prefixed by "**r**".
(3) A 5 bytes long **weight of the neighbor** which includes the neighbor weight value rounded to two decimal places and prefixed by "**w**".

RSSI	Neighbor weight
> -55 dbm	0.8
> -63 & ≤ -55 dbm	0.5
> -75 & ≤ -63 dbm	0.1
≤ -75 dbm	0.0

Table 1: Neighbor exposure estimation from RSSI values of BLE advertisements received from neighbors

Fig. 4 shows the payload format of the BLE advertising packet.

5.2 Risk score computation

In addition to the risk scores obtained from the BLE advertisements from the neighbors, the weight of the neighbors, and exposure caused by the neighbors are also required for the purpose of risk score computation, along with the vulnerability of the node itself. Note that here "node" means the smartphone.

5.2.1 Neighbor exposure. The exposure from a neighbor is an estimation of the likelihood of a transmission event from a neighbor. For infectious diseases such as COVID-19, the likelihood of transmission increases with close contacts. While BLE signal characteristics, such as received signal strength indication (RSSI) and attenuation, can be used for distance estimation, they are known be noisy estimators [18]. Hence, for the purpose of estimation of exposure from the neighbor, we use a coarse grained mapping, similar to those used in the Exposure API [7]. Based on the existing studies, Table 1 shows how the exposure values can be mapped from the RSSI values [15]. A higher RSSI values maps to a higher exposure from a neighbor.

5.2.2 Neighbor weight. The neighbor weight is an estimate of the likelihood of a neighboring node to be infectious, which can depend on a range of factors, such as the prevalence of preexisting diseases, age, etc. If such information is available, the derived neighbor weight is included in the BLE advertising packet. However, while such information may not always be available at an individual level, approximate measures are often available at a population level, which can be used as fixed values for all smartphones in a geographic region. For instance, neighbor weight may be derived from the basic reproduction number (R_0) [9] value for a particular epidemic for a given geographical region.

5.2.3 Vulnerability. The vulnerability of the ego node is an estimate of how quickly an individual can recover when exposed to infection, and as with the neighbour weight, this depends on a range of factors such as preexisting conditions, age, etc, as well as the nature of the disease itself [3]. When available, such information is incorporated in the computation of the risk score by the ego node.

Currently, the only data shared between the smartphones are the neighbor weights and risk score values. These values are computed on individual smartphones and shared with the neighbors. As no other parameter is shared other than computed values of neighbor weights and risk score and no other information about the neighbor is shared, we enable privacy by design. As a proof of concept implementation, we can also provide an alpha version of a smartphone application upon request for the readers to test.

6 FUTURE DIRECTIONS

While the risk score proposed here focuses on parameters that can be readily measured using smartphones, it can be developed further, both in terms of increasing it's accuracy towards risk estimation, as well as applying it to individual use cases, which we highlight in this section.

6.1 Increasing accuracy of risk score

The accuracy of the risk score proposed in this paper can also be increased by incorporating additional contextual information, where available. Some examples of this are:

- **Indoor and Outdoor location detection:** The likelihood of infection spread has been known to be higher in indoor environments compared to outdoor [28]. This can be incorporated into the risk score by first, automatically detecting the indoor/outdoor context [1], and secondly, by incorporating it into the risk score itself.
- **Identification of exposure context:** As outlined previously in section 5.2, by identification of the infection context in real-time, the risk score computation can be made more accurate. This can include detection of respiratory symptoms to better estimate the exposure [16, 29].

In addition to the points above, a general challenge with all infection tracking applications is that they do not cater to the entire population, since people may not always have access to smartphones and other IoT devices.

6.2 Use cases

The proposed risk score is applicable towards monitoring and managing the spread of infection for population groups, such as over a geographical region, as well as for individuals.

(1) **Risk score at different spatial scale:** The proposed risk score, in addition to computing score of an individual can be used to compute the risk score at any spatial scale (i.e., a country, a city, a building, a house, a room). For instance, considering A_a be the area for which risk score has to be computed, such as a district, and let L_a be the group of people in that region at time t. The risk score of region A_a at time t is defined as equation 4.

$$r_t^{A_a} = \frac{\sum_{\forall i \in L_a} r_{i,t}}{||L_a||} \tag{4}$$

Here, $||L_a||$ represents the number of people in A_a. Consider a region, R to be comprised of many As, the total population at time t, N^t is thus $\sum_{i \in R} ||L_i||$. Some examples include:

(a) ***Monitoring of geographical regions by government authorities:*** As evidenced by the COVID-19 pandemic, the infection spread often starts from small geographical regions, which can grow exponentially if early actions are not taken. Our proposed risk score can be used to obtain an early estimate of the likelihood of infection transmissions within a geographical region. Subsequently, preventative actions, such as increased testing, can be taken, without

even resorting to lockdowns that have economic and social impacts.

(b) ***Monitoring of individual buildings:*** An important aspect of managing the spread of infections is to reduce the likelihood of spread in controlled environments, such as office buildings, hotels, hospitals, etc. In such scenarios, risk score can be used to monitor behavior of individuals within such a region, and take quick actions even before any infection is confirmed. Some examples of such use cases are:

(i) *Hotels:* Guest movements and interactions among guests at hotels can have significant consequences to the infection spread in a pandemic, as has been seen in the case of COVID-19 [30]. The risk score can be used to act quickly by enforcing close monitoring of the individuals who are found to be in risky situations.

(ii) *Hospitals:* In order to handle increasing case loads during a pandemic, hospitals typically have dedicated wards. In such cases, it is critical to minimize the likelihood of transmission from such dedicated wards to other wards in the hospital [24], which can be done through monitoring of the risk scores of patients, doctors and other hospital staff.

(iii) *Office buildings:* Managing the recovery from a pandemic is equally important to managing it's spread, and the risk score can be used as a part of the plans used for businesses and office buildings [31].

Similar scenarios may be envisioned for other closed environments, such as residential buildings, supermarkets, shopping malls, airports, etc.

(2) **Individual monitoring:** Risk score can also be used to provide real-time alerts to individuals to take action. For instance, it can be used to provide prompts to wear mask if one is detected to move from a less risky region to more risky one. Further, risk score can be used to provide personalized alerts for individuals. For instance, vulnerable people (i.e. who are likely to be affected more due to pre-existing conditions), can be alerted early by using a lower alert threshold.

7 CONCLUSION

In this paper, we introduced a risk score that estimates infection propagation by leveraging the neighborhood of an individuals at a given time. On top, our risk score also takes into consideration factors transmission likelihood and vulnerability to a disease. Our results show that our risk score is able to capture potential risky situations. To further leverage our risk score we demonstrate how our risk score can be implemented in a contact tracing applications and as a proof of concept make it available upon request. Nonetheless, as future directions, we provide use cases and potential parameters that can be included in the risk score to make it more robust.

ACKNOWLEDGEMENTS

The authors contributed equally in the research.

REFERENCES

[1] R. Agarwal, S. Chopra, V. Christophides, N. Georgantas, and V. Issarny. 2019. Detecting Mobile Crowdsensing Context in the Wild. In *20th IEEE International Conference on Mobile Data Management* (Hong Kong). IEEE, 170–175.

[2] Australian Government Department of Health. 2020. COVIDSafe App. https://www.health.gov.au/resources/apps-and-tools/covidsafe-app (Accessed 06/09/2020).

[3] BBC. 2020. Coronavirus: How long does it take to recover? https://www.bbc.com/news/health-52301633 (Accessed on 09/22/2020).

[4] J. Bell, D. Butler, C. Hicks, and J. Crowcroft. 2020. TraceSecure: Towards Privacy Preserving Contact Tracing. *arXiv* (April 2020), 1–22. arXiv:2004.04059 [cs.CR]

[5] V. Blondel, A. Decuyper, and G. Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4, 1 (Aug. 2015), 1–55.

[6] J. Clover. 2020. Apple's Exposure Notification System: Everything You Need to Know. https://www.macrumors.com/guide/exposure-notification/ (Accessed 13/07/2020).

[7] Google. 2020. Define meaningful exposures: Google API for Exposure Notifications. https://developers.google.com/android/exposure-notifications/meaningful-exposures (Accessed on 09/19/2020).

[8] J. He, Y. Guo, R. Mao, and J. Zhang. 2020. Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. *Journal of Medical Virology* Early Access (2020), 1–11.

[9] J. M. Heffernan, R. J. Smith, and L. M. Wahl. 2005. Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface* 2, 4 (Sept. 2005), 281–293. https://doi.org/10.1098/rsif.2005.0042

[10] H. Hethcote. 2000. The Mathematics of Infectious Diseases. *SIAM Rev.* 42, 4 (2000), 599–653.

[11] M. Islam, I. Islam, K. Munim, and A. Islam. 2020. A Review on the Mobile Applications Developed for COVID-19: An Exploratory Analysis. *IEEE Access* 8 (Aug. 2020), 145601–145610.

[12] S. Jeong, S. Kuk, and H. Kim. 2019. A Smartphone Magnetometer-Based Diagnostic Test for Automatic Contact Tracing in Infectious Disease Epidemics. *IEEE Access* 7 (Jan. 2019), 20734–20747.

[13] O. Khazan. 2020. The Most American COVID-19 Failure Yet. https://www.theatlantic.com/politics/archive/2020/08/contact-tracing-hr-6666-working-us/615637/ (Accessed 12/09/2020).

[14] P. Kindt, T. Chakraborty, and S. Chakraborty. 2020. How Reliable is Smartphone-based Electronic Contact Tracing for COVID-19? *arXiv* (May 2020), 1–13.

[15] D. Leith and F. Stephen. 2020. Coronavirus Contact Tracing: Evaluating the Potential Of Using Bluetooth Received Signal Strength for Proximity Detection. *arXiv* (May 2020), 1–11. arXiv:2006.06822 [eess.SP]

[16] D. Liaqat, R. Wu, T. Son, A. Gershon, H. Alshaer, E. de Lara, and F. Rudzicz. 2018. Towards Ambulatory Cough Monitoring using Smartwatches. In *C41 Health Services Research in Pulmonary Disease: Thematic Poster Session.* American Thoracic Society, A4929–A4929.

[17] Joakim Lindh. 2016. Bluetooth Low Energy Beacons (Rev. A). https://www.ti.com/lit/an/swra475a/swra475a.pdf. (Accessed on 09/24/2020).

[18] A. Mackey, P. Spachos, L. Song, and K. Plataniotis. 2020. Improving BLE beacon proximity estimation accuracy through bayesian filtering. *IEEE Internet of Things Journal* 7, 4 (April 2020), 3160–3169.

[19] M. Mhango, M. Dzobo, I. Chitungo, and T. Dzinamarira. 2020. COVID-19 Risk Factors Among Health Workers: A Rapid Review. *Safety and Health at Work* 11, 3 (Sept. 2020), 262–265.

[20] National Informatics Center, India. 2020. Aarogya Setu Mobile App. https://www.mygov.in/aarogya-setu-app/ (Accessed 06/07/2020).

[21] P. Ng, P. Spachos, S. Gregori, and K. Plataniotis. 2020. Epidemic Exposure Notification with Smartwatch: A Proximity-Based Privacy-Preserving Approach. *arxiv* (July 2020), 1–11. arXiv:2007.04399 [cs.CR]

[22] K. Nguyen, Z. Luo, and C. Watkins. 2020. Epidemic contact tracing with smartphone sensors. *Journal of Location Based Services* 14, 2 (Sept. 2020), 92–128.

[23] N. Oliver, B. Lepri, H. Sterly, et al. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances* 6, 23 (April 2020), 1–6.

[24] T. Oraby, M. Tyshenko, H. Balkhy, et al. 2020. Analysis of the Healthcare MERS-CoV Outbreak in King Abdulaziz Medical Center, Riyadh, Saudi Arabia, June-August 2015 Using a SEIR Ward Transmission Model. *International journal of environmental research and public health* 17, 8 (2020), 2936.

[25] M. Salathé, M. Kazandjieva, J. Lee, P. Levis, M. Feldman, and J. Jones. 2010. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* 107, 51 (Dec. 2010), 22020–22025.

[26] M. Shahzamal, R. Jurdak, R. Arablouei, M. Kim, K. Thilakarathna, and B. Mans. 2017. Airborne Disease Propagation on Large Scale Social Contact Networks. In *Proceedings of the 2nd International Workshop on Social Sensing* (Pittsburgh, PA, USA). ACM, 35–40.

[27] M. Shukla, M. Rajan, S. Lodha, G. Shroff, and R. Raskar. 2020. Privacy Guidelines for Contact Tracing Applications. *arXiv* (April 2020), 1–10. arXiv:2004.13328 [cs.LG]

[28] T. Smieszek, G. Lazzari, and M. Salathé. 2019. Assessing the Dynamics and Control of Droplet- and Aerosol-Transmitted Influenza Using an Indoor Positioning System. *Scientific Reports* 9, 1 (Feb. 2019), 1–10.

[29] X. Sun, Z. Lu, W. Hu, and G. Cao. 2015. SymDetector: Detecting sound-related respiratory symptoms using smartphones. In *Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan). ACM, 97–108.

[30] J. Taylor. 2020. Hotel quarantine linked to 99% of Victoria's Covid cases, inquiry told. https://www.theguardian.com/australia-news/2020/aug/18/hotel-quarantine-linked-to-99-of-victorias-covid-cases-inquiry-told (Accessed on 09/23/2020).

[31] Business Victoria. 2020. COVIDSafe Plan. https://www.business.vic.gov.au/coronavirus-covid-19/covid-safe-business/covid-safe-plan (Accessed on 09/23/2020).

[32] V. Vuorinen, M. Aarnio, M. Alava, and Others. 2020. Modelling aerosol transport and virus exposure with numerical simulations in relation to SARS-CoV-2 transmission by inhalation indoors. *Safety Science (Early Access)* 130 (Oct. 2020), 1–23.

[33] C. Wood, M. Thomas, J. Budd, T. Mashamba-Thompson, K. Herbst, D. Pillay, R. Peeling, A. Johnson, R. McKendry, and M. Stevens. 2019. Taking connected mobile-health diagnostics of infectious diseases to the field. *Nature* 566, 7745 (Feb. 2019), 467–474.

[34] World Health Organization. 2020. *Health workers exposure risk assessment and management in the context of COVID-19 virus: interim guidance, 4 March 2020.* Technical Report. World Health Organization. 1–8 pages. https://apps.who.int/iris/handle/10665/331340

Sensitivity Analysis for COVID-19 Epidemiological Models within a Geographic Framework

Zhongying Wang
zhongyin@usc.edu
University of Southern California
Los Angeles, California

Orhun Aydin
oaydin@usc.edu
University of Southern California
Los Angeles, California

Abstract

Spatial sciences and geography have been integral to the modeling of and communicating information pertaining to the COVID-19 pandemic. Epidemiological models are being used within a geographic context to map the spread of the novel SARS-CoV-2 virus and to make decisions regarding state-wide interventions and allocating hospital resources. Data required for epidemiological models are often incomplete, biased, and available for a spatial unit more extensive than the one needed for decision-making. In this paper, we present results on a global sensitivity analysis of epidemiological model parameters on an important design variable, time to peak number of cases, within a geographic context. We design experiments for quantifying the impact of uncertainty of epidemiological model parameters on distribution of peak times for the state of California. We conduct our analysis at the county-level and perform a non-parametric, global sensitivity analysis to quantify interplay between the uncertainty of epidemiological parameters and design variables.

CCS Concepts

• **Information systems → Geographic information systems**.

Keywords

Sensitivity Analysis, Epidemiological Model, COVID-19, Spatial-temporal Analysis, Uncertainty

ACM Reference Format:

Zhongying Wang and Orhun Aydin. 2020. Sensitivity Analysis for COVID-19 Epidemiological Models within a Geographic Framework. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3423459.3430755

1 Introduction

COVID-19 is a severe acute respiratory syndrome (SARS) caused by the SARS-CoV-2 virus [7]. On March 11, 2020, COVID-19 is declared to be a pandemic with 12,552,795 infected persons and 561,617 deaths globally as of July 12, 2020 [12]. In the United States, the number of total cases is at 4,974,959 with 161,284 deaths [12], making the COVID-19 pandemic a national problem.

Spatial analysis has played an essential role during the COVID-19 crisis in terms of spatial analysis of transmission and the number of new cases [2, 6, 11, 13, 20, 21], and mapping susceptible populations [10]. The SARS-CoV-2 is contracted from person-to-person via respiratory droplets [18], making public places where people are in close contact likely places for high transmission rates [2, 6].

Epidemiological models are used within a geographic context to map the spread of the novel SARS-CoV-2 virus and to make decisions regarding state-wide interventions and allocating hospital resources [13]. Data required for epidemiological models are often incomplete, and biased, making uncertainty quantification a necessity for decision-making. The spatial resolution of currently available curated data for United States COVID-19 case and death statistics is at the county-level. Thus, it is important to understand the impact of epidemiological model parameters on actionable variables within a geographic setting.

In this research paper, a sensitivity analysis on decision variables is conducted, and implications of parameter uncertainty on decision variables used by public health officials are showcased for California. We use the Sobol sensitivity [17] to model the impact of epidemiological variables on the spatial and space-time patterns of new COVID-19 hospitalizations at the county level in the state of California. We use the CHIME model (COVID-19 Hospital Impact Model for Epidemics) [5] from University of Pennsylvania to define COVID-19 hospitalization projections. The spatiotemporal series for predicted new COVID-19 hospitalizations is summarized temporally and spatially with time to peak demand, and the Moran's I statistic, respectively. The impact of epidemiological parameters of the CHIME model on the space-time patterns of modeled hospital demand are quantified with the Sobol sensitivity.

2 Data & Methodology

2.1 Data

CHIME model requires several parameters, including population, the number of currently hospitalized COVID-19 patients, doubling time, social distancing effect, infectious days, and optional including hospital resource parameters (number of beds, intensive care units (ICUs) and ventilators) to forecast future COVID-19 hospitalizations and its impact on the hospital resources.

The population data used in the model is from ESRI's 2019 Updated Demographics [1]. This data updates annually based on several sources of data, including a full-time series of intercensal and vintage-based county estimates from the US Census Bureau and a time series of county-to-county migration data from the Internal Revenue Service. Projections are necessarily derived from current events and past trends, which is calculated from previous census counts provided by the American Community Survey (ACS). COVID-19 related data, including incidence, confirmed cases and death are all from JHU CSSE COVID-19 Data [2][4].

COVID-19, November 3, 2020, Seattle, WA, USA

[1] https://doc.arcgis.com/en/esri-demographics/reference/methodologies.htm
[2] https://coronavirus.jhu.edu/

2.2 The CHIME Model

Predictive Healthcare at Penn Medicine initiated the tool Hospital impact model to assist hospitals and public health officials with capacity planning, including daily increase, peak hospitalized census, ICU admissions, number of patients requiring ventilators and timeline prediction.

CHIME model is one of many customized models based on SIR (Susceptible, Infected, Recovered) model [1], which is a commonly used epidemiological model to forecast the number of infected people from a disease in a closed population over time. The main idea of this model is dividing the population into compartments throughout the progression of the disease, such as susceptible, infected and recovered population. The model dynamics are defined by the following equations:

$$S_{t+1} = S_t - \beta S_t I_t \tag{1}$$

$$I_{t+1} = I_t + \beta S_t I_t - \gamma I_t \tag{2}$$

$$R_{t+1} = R_t + \gamma I_t \tag{3}$$

where β represents the effective contact rate, which can be computed as the transmissibility τ multiplied the average number of people exposed c: $\beta = \tau \times c$. The transmissibility is the basic property related to the virulence of the pathogen, but the number of people exposed is the parameter can be changed by policies, like social distancing or mask wearing. γ is the inverse of the mean recovery time, and recovery time indicates the period of infection getting cleared and varies for the severity of the symptoms. For COVID-19, the average is normally considered as 1/14. The basic reproduction number (R_0) is an indicator of the contagiousness or transmissibility of infectious and parasitic agents and represents average number of people can be infected by any given infected person without immunity from past exposures or vaccination [3]. It is defined as $R_0 = \beta/\gamma$. The disease is supposed to spread if R_0 is > 1 and the larger the number is, the faster it will spread. Since the transmissibility and social contact rates are hard to compute, this parameter can be replaced by doubling times. Since the rate of new infections in the SIR model g can be computed with doubling time T_d: $g = \beta S - \gamma$, β can be computed with the initial population size of susceptible individuals as $\beta = (g + \gamma)$.

ESRI has developed a toolbox for the CHIME model and parameters used in sensitivity analysis and their explanations are shown in **Table 1**.

2.3 Sobol Sensitivity

Sobol sensitivity analysis quantifies the impact of total-effect indices and higher-order interactions and has no limit for the preparation of the input sample, and such characters enable it to deal with auto-correlated spatial input [8].

The Sobol method is one of the variance-based methods, which can compute sensitivity indices regardless of the linearity or monotonicity, or other assumptions on the underlying model. In variance based method, the fractional contribution of each input to the variance V of the model is estimated and the total variance V of the model output is decomposed to calculate the sensitivity indices for every independent X_i.

$$V = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<m} V_{ijm} + \cdots + V_{12\ldots k} \tag{4}$$

where V_i is the share of the output variance explained by the ith model input, and indicates the sensitivity of Y to X_i. V_{ij} is the share of the output variance explained by the interaction of the ith and jth model inputs, and indicates the sensitivity of Y to the interaction of X_i and X_j. k is the total number of the model inputs.

The first-order sensitivity computes the contribution to the output variance of the main effect of X_i and is defined with conditional variances as

$$Z_i = \frac{V_i}{V} = \frac{Var[E(Y|X_i)]}{Var(Y)} \tag{5}$$

where the inner expectation of the numerator is conditional on X_i taking a value X_i* within its range of uncertainty, while the outer variance is calculated over all possible values of X_i. If the variance of the conditional expectation $E(Y|X_i = x_i*)$ for some particular value $X_i = x_i*$ is relatively large when compared to the total variance, and all the effects of the X_j, $j \neq i$, then factor X_i can be considered as an influential one. Similarly, $Z_{ij} = \frac{V_{ij}}{V}$ indicates the sensitivity indices of the interaction effect of X_i and X_j [17].

According to $\sum_{i=1}^{k} Z_i + \sum_{i=1}^{k} Z_{ij} + \cdots + \sum_{i=1}^{k} Z_{ij\ldots k} = 1$, total-order index Z_{Ti}, which measures the contribution to the output variance of X_i including all variance caused by its interactions, of any order, with any other input variables can be defined as

$$Z_{Ti} = 1 - \frac{V_{-i}}{V} = 1 - \frac{Var[E(Y|X_{-i})]}{Var(Y)} \tag{6}$$

Sobol sensitivity quantifies the contribution of variance from a set of explanatory variables on the variation of target variable of interest. Thus, it provides a statistical framework within which the impact of a model parameters can be assessed marginally and jointly.

2.4 Experimental Design

The method of Sobol sensitivity analysis computes the indices by using the decomposition of the output variance in Eq.1. Capturing representative variance requires rigorous design of experiments. In this work, experiments are designed using the Saltelli sampling scheme [15]. Steps of defining experiments are elaborated below:

(1) Choose an integer N as the size of the base sample.
(2) Generate a sample matrix $(N, 2k)$ of the input factors by using the Saltelli sampler, where k is the number of input factors. Divide the matrix into two and define each part as A and B, which contain half of the sample data.
(3) Duplicate the matrix A and replace the i-th column with the same column from matrix B, then define it as D_i. The matrix C_i is the duplicate of matrix B, except that the $i - th$ column is replaced with the $i - th$ column in matrix A.
(4) Compute the model output for all the input values in the sample matrices and then use the Eq.5 and 6 to compute the sensitivity indices.

Sampling scheme above defines experiments to model variance of response variables without increasing the computational load by employing full factorial design.

2.5 Spatiotemporal Sensitivity Analysis

The output response for every experiment is a spatiotemporal series of CHIME model output. We summarize the predicted number of hospitalizations time series at every county with time to peak hospitalizations.

Table 1: Parameters in the CHIME model

Parameter	Explanation
Doubling Time in Days	The number of days that the number of infected individuals to double without interventions.
Social Distancing	The quantitative estimation of social contact reduction in each catchment area.
Infectious Days	The number of days an infected person has the ability to infect others.
Hospitalization Rate	The percentage of all infected cases that will need hospitalization.
Average Days of Hospital Stay	The average number of days COVID-19 patients have needed to stay in a hospital.
ICU Rate	The percentage of all infected cases which will need to be treated in an ICU.
Average Days in ICU	The average number of days COVID-19 patients have needed ICU care.
Ventilated Rate	The percentage of all infected cases that need mechanical ventilation.
Average Days on Ventilator	The average number of days with ventilation needed for COVID-19 patients.

$$t_{peak}^{(i)} = \max_{1 \le t \le t_{max}} H_{new}^{(i)} \tag{7}$$

In Eq. 7, the time to peak hospitalizations at location i, $t_{peak}^{(i)}$, is the time at which the number of predicted hospitalizations $H_{new}^{(i)}$ reaches its maximum value. In cases where projections decline, $t_{peak}^{(i)}$ is assumed to be 0. $t_{peak}^{(i)}$ reduces the time series into a spatial distribution of time to peak hospitalizations, denoted as $T_{peak} = t_{peak}^{(1)}, t_{peak}^{(2)}, ..., t_{peak}^{(N)}$.

The spatial distribution of time to peak hospitalizations are summarized using the Moran's I statistic, that quantifies the spatial patterns of time to peak hospitalizations.

$$I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} (t_{peak}^{(i)} - \bar{T}_{peak})(t_{peak}^{(j)} - \bar{T}_{peak})}{W \sum_{i=1}^{n} (t_{peak}^{(i)} - \bar{T}_{peak})^2} \tag{8}$$

In Eq. 8, w is the geographic weight, and W is the sum of all weights, $W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}$. A positive and statistically significant I indicates spatial clustering, and a significant and negative I indicates dispersion of $t_{peak}^{(i)}$.

3 Results

The decision variable we investigate is the spatial clustering of time until the peak number of cases, T_{peak}, are observed at a given county.A short time to peak indicates that the hospitals in that county are about to receive a high number of COVID-19 patients. Spatial clustering of T_{peak} indicates that similar volumes in hospital demand will exist at neighboring counties.

We conducted 800 simulations by varying three model parameters. Our choice behind these parameters are due to high uncertainty associated with them. According to epidemiology analysis, the (R_0) ranges from about 2 to 6 based on initial estimates of the early dynamics of the outbreak in Wuhan, China [14]. The doubling time is computed with this uncertainty range. According to CDC and other researches, 88% and 95% of specimens no longer yielded replication-competent virus after 10 and 15 days, but recovery of replication-competent virus between 10 and 20 days after symptom onset has been documented in some persons with severe symptoms [19]. The uncertainty of infectious days is then chosen from 10 to 20. Experiment parameters and their ranges are presented in Table 2.

Table 2: Epidemiological Experiment Variables

Epidemiological Experiment Variables	
Doubling Time in Days	[2.27,10.05]
Social Distancing (%)	[0,50]
Infectious Days	[10,20]

We present the response surface for the average number of days to peak in the state of California with respect to uncertain model parameters. The response surface is depicted in Figure 1.

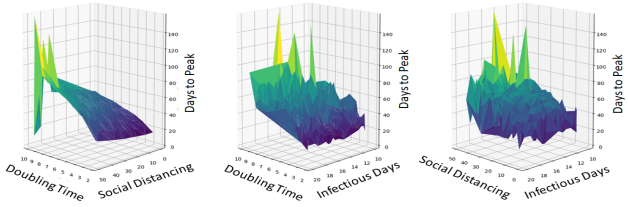


Figure 1: Surface Plot of Sensitivity Analysis

In some of our simulations, the peak is not observed within our simulation time span (180 days). These simulations correspond to peaks in the response surface. Our results indicate that for increasing doubling time and social distancing, the peak is delayed. Response surfaces with respect to infectious days indicate a more complex relationship that points to a high amount interaction between infectious days and other epidemiological variables.

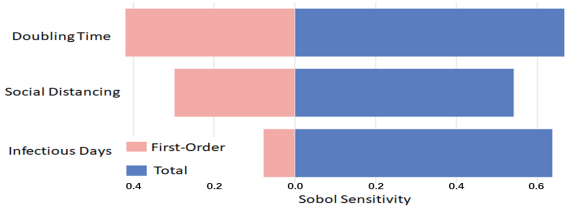


Figure 2: Tornado Plot for Sobol Sensitivity

Sensitivity of spatial patterns of T_{peak} is showcased in Figure 2. Doubling time is the most first-order sensitive variable, followed by social distancing and infectious days. The variables are ranked with respect to their first-order sensitivity. Note that all three model variables have high total sensitivity. This indicates strong interactions between these variables and the spatial patterns of hospital demand.

Next, we depict spatial patterns of T_{peak} with four extreme realizations from our set of 800 simulations. The spatial distribution of T_{peak} for different cases are shown in Figure 2:

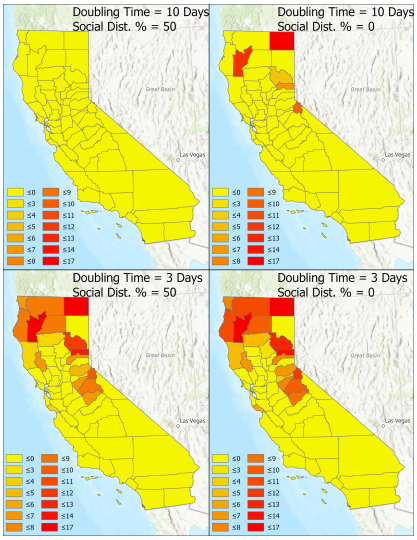


Figure 3: Time to Peak Cases for Different Model Parameters. Basemap courtesy of Esri and its partners

Figure 3 elaborates the importance of epidemiological model parameters on the spatial patterns of potential hospital surges. For short doubling times (fast transmission) and low social distancing, we observe significantly high peaks that occur within a week of our simulation start time in most of North California. Zero days to peak implies that a significant peak in cases is not observed. In our simulations, these locations with zero time to peak also have low number of new hospitalizations. Thus, for high doubling time and high levels of social distancing (top-left) no significant surges are observed. Figure 3 motivates the importance of incorporating uncertainty pertinent to epidemiological parameters for decision making.

4 Discussions

For the experimental design, asymptomatic and pre-symptomatic patients should be considered to improve the accuracy of the model. $SARII_qS_q$ model is developed based on SIR model, which takes asymptotic or mildly symptomatic, isolated infected and quarantined susceptible individuals into consideration[16]. As the proportion of COVID-19 transmission due to asymptomatic or pre-symptomatic infection compared to symptomatic infection is unclear, further study on virology need to be conducted [9].

5 Conclusions

Geographical sensitivity analysis shows complex interactions between uncertain epidemiological parameters and spatial patterns of COVID-19 incidence. In particular, doubling time and social distancing are shown to have a considerable impact on the spatial patterns of surges in the number of hospitalizations. Experimental design and associated simulations of time to peak hospitalization depicts the impact of uncertainty on decision variables. Sobol sensitivity analysis reveals high order interactions between model parameters, quantified by high total order sensitivity terms for all model parameters. Study showcases the importance of accurate understanding of COVID-19 drivers for spatial planning as drastic ranges for hospital surge times are observed in different simulations. Our results indicate that overall peak for new hospitalizations show spatial clustering, meaning nearby counties are likely to experience hospital surges at similar times. This points to the importance of resource planning ahead of time as our simulations show that directing patients during a surge to nearby counties may not be possible.

References

[1] Roy M Anderson, B Anderson, and Robert M May. 1992. *Infectious diseases of humans: dynamics and control*. Oxford university press.

[2] Leon Danon, Ellen Brooks-Pollock, Mick Bailey, and Matt J Keeling. 2020. A spatial model of CoVID-19 transmission in England and Wales: early spread and peak timing. *MedRxiv* (2020).

[3] Paul Delamater, Erica Street, Timothy Leslie, Y. Yang, and Kathryn Jacobsen. 2019. Complexity of the Basic Reproduction Number (R0). *Emerging Infectious Diseases* 25 (01 2019), 1–4. https://doi.org/10.3201/eid2501.171901

[4] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20, 5 (2020), 533–534.

[5] Penn Healthcare. 2020. COVID-19 Hospital Impact Model for Epidemics (CHIME). https://penn-chime.phl.io/. Accessed: 2020-08-23.

[6] Dayun Kang, Hyunho Choi, Jong-Hun Kim, and Jungsoon Choi. 2020. Spatial epidemic dynamics of the COVID-19 outbreak in China. *International Journal of Infectious Diseases* (2020).

[7] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* (2020).

[8] Linda Lilburne and Stefano Tarantola. 2009. Sensitivity analysis of spatial models. *International Journal of Geographical Information Science* 23, 2 (2009), 151–168.

[9] Yang Liu, Li-Meng Yan, Lagen Wan, Tian-Xin Xiang, Aiping Le, Jia-Ming Liu, Malik Peiris, Leo LM Poon, and Wei Zhang. 2020. Viral dynamics in mild and severe cases of COVID-19. *The Lancet Infectious Diseases* (2020).

[10] Gregorio A Millett, Austin T Jones, David Benkeser, Stefan Baral, Laina Mercer, Chris Beyrer, Brian Honermann, Elise Lankiewicz, Leandro Mena, Jeffrey S Crowley, et al. 2020. Assessing differential impacts of COVID-19 on Black communities. *Annals of Epidemiology* (2020).

[11] Abolfazl Mollalo, Behzad Vahedi, and Kiara M Rivera. 2020. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of The Total Environment* (2020), 138884.

[12] World Health Organization et al. 2020. Coronavirus disease (COVID-19): situation report, 182. (2020).

[13] David O'Sullivan, Mark Gahegan, Dan Exeter, and Benjamin Adams. 2020. Spatially-explicit models for exploring COVID-19 lockdown strategies. *Transactions in GIS* (2020).

[14] Mark Channels Read. 2020. EID: High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis* 26 (2020).

[15] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. 2004. *Sensitivity analysis in practice: a guide to assessing scientific models*. Vol. 1. Wiley Online Library.

[16] Kankan Sarkar, Subhas Khajanchi, and Juan J Nieto. 2020. Modeling and forecasting the COVID-19 pandemic in India. *Chaos, Solitons & Fractals* 139 (2020), 110049.

[17] Ilya M Sobol. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation* 55, 1-3 (2001), 271–280.

[18] Kelvin Kai-Wang To, Owen Tak-Yin Tsang, Cyril Chik-Yan Yip, Kwok-Hung Chan, Tak-Chiu Wu, Jacky Man-Chun Chan, Wai-Shing Leung, Thomas Shiu-Hong Chik, Chris Yau-Chung Choi, Darshana H Kandamby, et al. 2020. Consistent detection of 2019 novel coronavirus in saliva. *Clinical Infectious Diseases* (2020).

[19] Jeroen JA van Kampen, David AMC van de Vijver, Pieter LA Fraaij, Bart L Haagmans, Mart M Lamers, Nisreen Okba, Johannes PC van den Akker, Henrik Endeman, Diederik AMPJ Gommers, Jan J Cornelissen, et al. 2020. Shedding of infectious virus in hospitalized patients with coronavirus disease-2019 (COVID-19): duration and key determinants. *medRxiv* (2020).

[20] Charlie H Zhang and Gary G Schwartz. 2020. Spatial disparities in coronavirus incidence and mortality in the United States: an ecological analysis as of May 2020. *The Journal of Rural Health* 36, 3 (2020), 433–445.

[21] Ruizhi Zheng, Yu Xu, Weiqing Wang, Guang Ning, and Yufang Bi. 2020. Spatial transmission of COVID-19 via public and private transportation in China. *Travel Medicine and Infectious Disease* (2020).

Analysis of the Impact of COVID-19 on Education Based on Geotagged Twitter

Zhu Wang
ADVIS Lab
Department of Computer Science
University of Illinois at Chicago
zwang260@uic.edu

Isabel F. Cruz
ADVIS Lab
Department of Computer Science
University of Illinois at Chicago
ifcruz@uic.edu

ABSTRACT

More than 150 colleges have reported hundreds of COVID-19 confirmed cases over all the states as the campuses have reopened and the schools have resumed in-person classes, after switching overnight to online teaching in the spring. We conduct a large scale study on education by using a geotagged Twitter dataset, which spans the whole U.S. during parts of the spring, summer, and fall terms of 2020. We analyze the temporal and spatial patterns of COVID-19 cases. Then, we conduct content and sentiment analysis to discover which topics and which thoughts people located at U.S. colleges and universities are communicating.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; **Online analytical processing**; *Social tagging systems*;

KEYWORDS

Geospatial Data integration, Semantics, Uncertainty, Range Queries

ACM Reference Format:
Zhu Wang and Isabel F. Cruz. 2020. Analysis of the Impact of COVID-19 on Education Based on Geotagged Twitter. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3423459.3430756

1 INTRODUCTION

The COVID-19 pandemic has had a widespread impact on all aspects of people's life in the last months. In particular, the education domain has been affected in the United States due to school and university closures and a subsequent move to online platforms. This move, which in many circumstances occurred practically overnight, has been a challenge to students, parents, and faculty. However, new challenges have been directly connected with the pandemic itself. University campuses became the new hot spots for COVID-19 in the fall semester [6], because many schools are providing in-person classes since August.

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430756

In a study, "203 'college town' counties where students comprise at least 10% of the population found that about half had experienced their worst weeks of the pandemic as students returned in August, and about half of those were experiencing peak infections" in September [6, 22]. What are the attitudes displayed and emotions felt by people in those towns, counties, and regions as they witness directly new infection peaks?

Because of the social distance practice in the era of COVID-19, social media platforms such as Twitter and Facebook can potentially be a conduit of choice to express and share opinions, emotions, and other life aspects related to the pandemic for those involved in academic life and for citizens in general. In this paper our focus is on Twitter. In fact, there is a wealth of millions of tweets with different times, locations, and users, in what constitutes a stream of snapshots throughout the pandemic, which we analyze in this paper, following the lead of so many others who have analyzed tweets in a variety of domains [3, 4, 7, 9, 10, 12, 13, 15, 16, 21, 23].

The emergence of Twitter brought important changes to how health and urban information are discovered and transmitted. When Twitter started in 2006, community health centers and healthcare providers in Chicago were communicating flu cases by *fax* and New York City had started their 3-1-1 call service for citizens to communicate non-emergency situations to local government, just three years earlier. Nowadays, researchers and healthcare entities rely on Twitter to detect flu trends and conduct disease surveillance [4, 9]. As for the communication of emerging urban situations, the 311 service can now be accessed via Twitter in many cities across the U.S. and is recognized as an important citizen participation tool [5]. Because tweets have an associated time and often spatial information, they lend themselves well to exploring the sentiments of people in different regions as the disease evolves in time. In the case of education, time and space are critical because different schools may have different semester or quarter periods, henceforth named *terms*, and distinct policy and regulations in each region or state.

Because tweets have an associated time and often spatial information, they lend themselves well to exploring the sentiments of people in different regions as disease evolves in time. In the case of education, time and space are critical because different schools may have different semester or quarter periods and distinct policy and regulations, in each region or state. The IEEE data port [8] provides a dataset of geotagged tweets including over 6 billions of COVID-19 related tweets up to now, 5% being education related worldwide. We extracted two time periods of tweets in states across the U.S. as follows. The former is from March to July including the spring and first summer terms, and the latter from August 24 to September 5 covering either the first week or two of the fall term.

In this paper, we collect and analyze at a large scale dataset of geotagged tweets to explore the reactions of people to the COVID-19 pandemic in the education domain, containing 673,601 tweets in the above two time periods. We study the temporal and spatial patterns and compare the density of tweets with the confirmed COVID-19 cases at U.S. colleges and universities. Then, we conduct content and sentiment analysis to discover which topics and which thoughts people are communicating.

This paper is organized as follows. Section 2 introduces data collection and pre-processing for our dataset. In Section 3, we demonstrate the methods applied in this study, including spatio-temporal patterns analysis, sentiment analysis, and content analysis. In Section 4, we present briefly some of the existing techniques that analyze social media (such as Yelp or Twitter). Then, we conclude the paper and discuss future research in Section 5.

2 DATASET

In this section we describe our dataset and how we process the dataset prior to the analytic steps.

2.1 Data collection

We use the Twitter API to hydrate tweets from the Coronavirus dataset that are written in English and occur in the U.S. from March to July and from August 24 to September 5. Also, we will be using tweets that are either geotagged or have otherwise associated with them the users' location, as provided by their profile, either in location or description. Then, we select tweets related to the education using a pre-defined keyword set: `school`, `course`, `class`, `student`, `teach`, `exam`, `educate`, `education`, `campus`, `university`, `college`, `tuition`, `learn`, `study`, `quiz`, `midterm`, `homework`, and `assignment`. That is, if a tweet contains one or more words from this set, they will be included in our study. Finally, we collected 560,780 tweets from March to July and 112,821 tweets from August 24 to September 5 from 265,654 users.

2.2 Pre-processing

We start by cleaning the dataset by organizing the geo-information, for example, we add to the users' displayed city or state information regarding latitude and longitude or vice versa. We use the Google Map API to get both the information on the coordinates and on the cities.

Raw data from the text of the tweets are noisy, thus leading to a sparse vector space and an increase in run time and storage. To address this problem, the text pre-processing tasks involve several steps: (1) standardizing the corpus to change all upper case letters to lower cases, (2) removing non-English words, URLs, special characters and stopwords, including the stopwords both from the dictionary and from a personalized set, in our case common words related to COVID-19, such as `covid` and `virus`, thus reducing the sparsity of the vector space, (3) tokenizing sentences or less structured text into a word level corpus. The NLTK toolkit[1] is used to perform the pre-processing of the text.

[1] http://www.nltk.org/

3 METHODS

In this section, we introduce the explanatory data analysis we will be performing in our study. First, we summarize the characteristics of the Twitter users, which users state their roles as parents, students, teachers or professors, and official accounts of the schools in their profiles and tweets, while others are not specified (see Table 1). Then, we study the spatio-temporal patterns of the tweets in our dataset, such as their hourly distribution in different days and their density and daily distributions in different states. Last, we study deeply the texts in the tweets to classify users' attitudes as positive or negative in different regions, and to extract the topics that were discussed on Twitter.

3.1 Spatio-temporal patterns

We explored the daily distribution of tweets in the various periods of the semesters from the different states, because we noticed that different schools have different academic calendars. Next, we display the hour distribution in each day of the week. Then, we compare the density of tweets on the map with the COVID-19 college and university cases tracker using the subset of college and university data.

Figures 1 and 2 show the different peaks of tweet counts in various states. California, Texas, New York and Florida generated the largest number of tweets in the two time periods. Each state has a similar increasing/decreasing pattern over time. The first peak in Figure 1 was around the start of the outbreak when many schools began closing their campus and moving to online classes. Other peaks in both periods correspond to the beginning and end of each term.

Hourly tweet distributions are shown in Figures 3 and 4. Figure 3 is about the spring and summer terms from March to July and Figure 4 is about the fall term. We converted the UTC time zone to local time zone for each tweet according to the spatial information given by the tweet. The interesting finding from the slight difference between the two terms is that the percentage of tweets from 1am to 3am in the spring term is larger than in the fall term, which we explain based on the intense workload of the final weeks of the term in comparison with the first couple of weeks in the fall term.

We extract the city and coordinate information of the tweets corresponding to the locations of colleges and universities, and display the volume of tweets on the U.S. map in Figure 5. According to the New York Times, a large number of new confirmed cases continues to emerge on college campuses [6, 22]. The Times provides an updated tracker map with the confirmed COVID-19 cases in colleges and universities in Figure 6. We found that the location distributions and density of tweets display a similar pattern to the confirmed cases in our college map, which attest to the relationship

Table 1: Summary of users' characteristics.

Roles	Counts	Percentage(%)
Parents	436	0.16
Student	141,568	53.29
Teacher/Professor	262	0.09
Official account	36	0.01

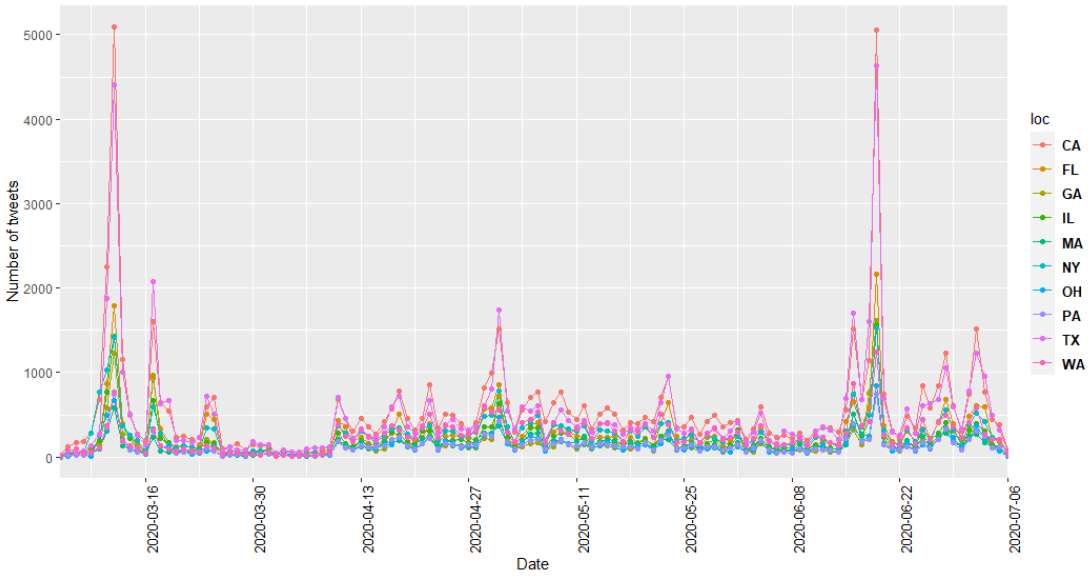


Figure 1: State-level tweet counts from March to July in the 10 states that generate the largest counts.

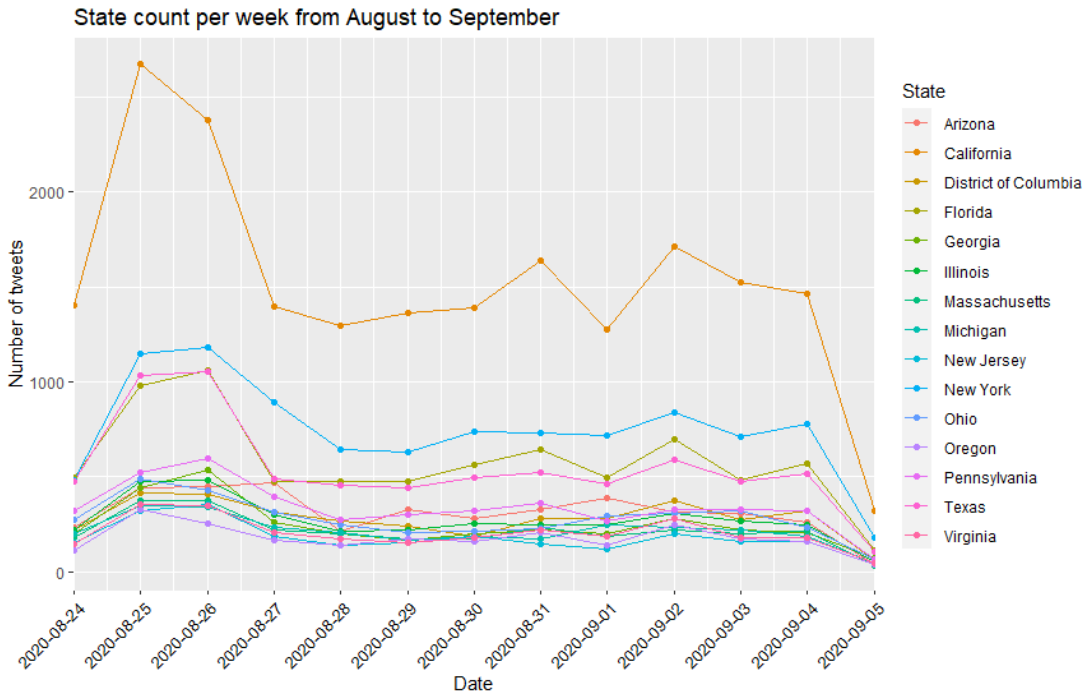


Figure 2: State-level tweet counts from August 24 to September 5 in the 15 states generating the most tweets.

between intense reactions and confirmed cases. A better comparison would be between the average of the number of cases per person compared with the average of the number of tweets per person. Otherwise one may wonder if the number of tweets is higher because the density of the population is also higher. However, the latter average is difficult to determine as a person may have different Twitter accounts and each person can post any number of tweets even if only using one account.

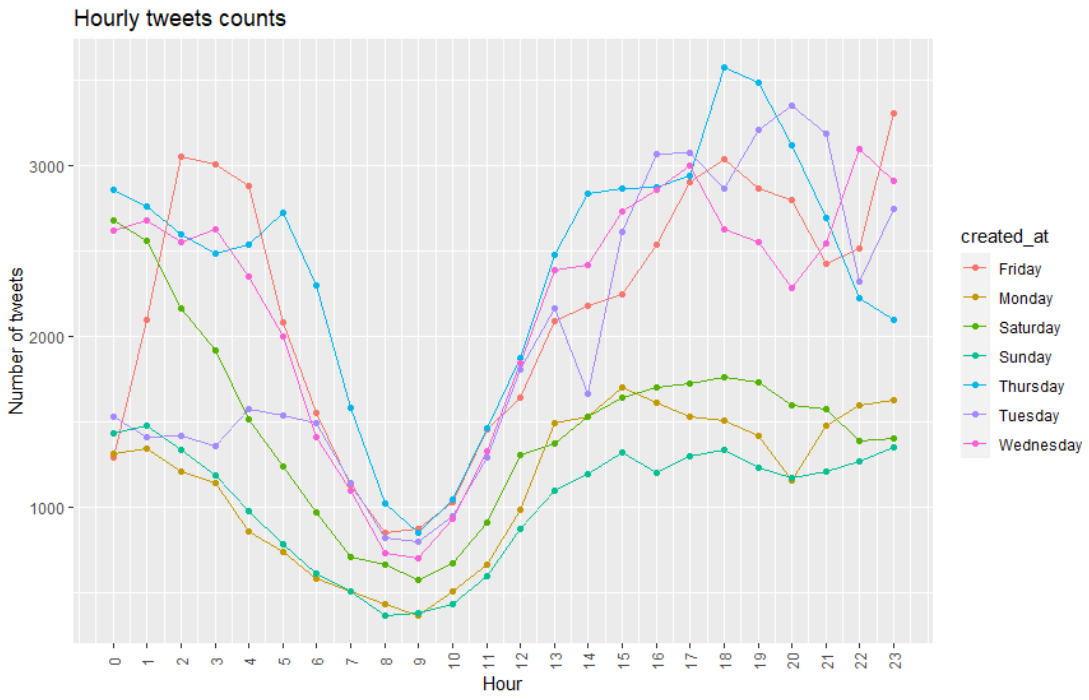


Figure 3: Hourly tweet counts in spring.

3.2 Sentiment analysis

People express their opinions about their daily life and events on micro-blogs such as Twitter, therefore, leveraging sentiment analysis of micro-blogs is beneficial to understand the users' attitudes towards the COVID-19 pandemic, and how it affects different domains, in our case the education domain. In this study, we deploy a dictionary based on sentiment method without having a ground truth corpus available. The sentiment dictionary we use is the *Language Assessment by Mechanical Turk, labMT* [3].

The polarity of a tweet, that is the emotion it expresses, can be determined using the labMT dictionary, which provides a happiness score for each word from 1 (`sad`) to 9 (`happy`). By computing the numeric average of the happiness score of all the words in a tweet, we determine its sentiment category as positive, neutral, or negative. We examine the convergent validity of negative tweets against the COVID-19 confirmed cases from colleges and universities. As a result, the scores obtained using labMT correlate positively with the confirmed positive cases with value 0.45.

Figure 7 illustrates the number of positive, neutral, and negative tweets in the 15 states that have the largest number of confirmed COVID-19 cases in colleges. Unexpectedly the number of positive tweets exceeds, for all states, the number of neutral and the number of negative tweets. In some states, the number of positive tweets even exceeds the sum of the neutral and negative tweets. Since we applied dictionaries to detect sentiment, a word like `positive` is a strong indicator of a positive sentiment. It also happens to be one of the words that contributes the most to the number of positives tweets in Figure 7, even if the word `positive` appears in `positive cases` or `test result is positive`, which are far from expressing a positive sentiment.

Moreover, we observe the daily sentiment—positive, neutral, or negative—using the percentage of the number of tweets for each sentiment with respect to the total number of tweets in the first time period—which is depicted in Figure 8. There are outliers in the different periods. The first two peaks (high values) of negative sentiment appear at the beginning of the pandemic when many schools start cancelling classes. From July 18 to July 20, there is an even higher negative peak. A possible explanation is that many states went into reopening phase around these dates, but students and faculty still did not feel safe to return to the campus.

To better understand the attitudes and opinions of users, we list the 10 most frequent words in positive and negative tweets in California and Illinois in Figures 9(a) and 9(b). There are many common words in these two states. The word `debt` appears frequently, which seems to indicate that students are concerned with their loans, since they need to pay the same tuition for online classes. Indeed, there are many discussions about tuition and expenses.

3.3 Content analysis

In this section, to further understand the social and psychological meaning associated with tweets, we use the *Linguistic Inquiry and Word Count (LIWC2015)* dictionary [2] preceded by *Latent Dirichlet Allocation (LDA)* [1] to infer coherent topics.

Since ground truth data are not available, we apply topic modeling as an unsupervised probabilistic method to discover latent topics. We treat the text of each tweet as a document to create

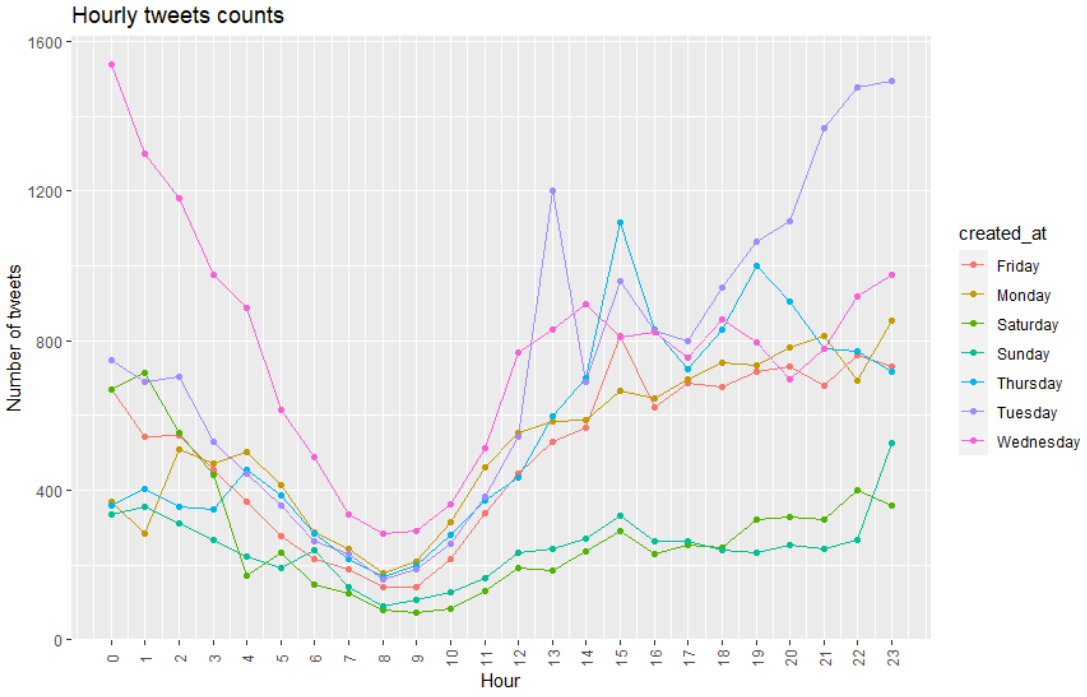


Figure 4: Hourly tweet counts in fall.

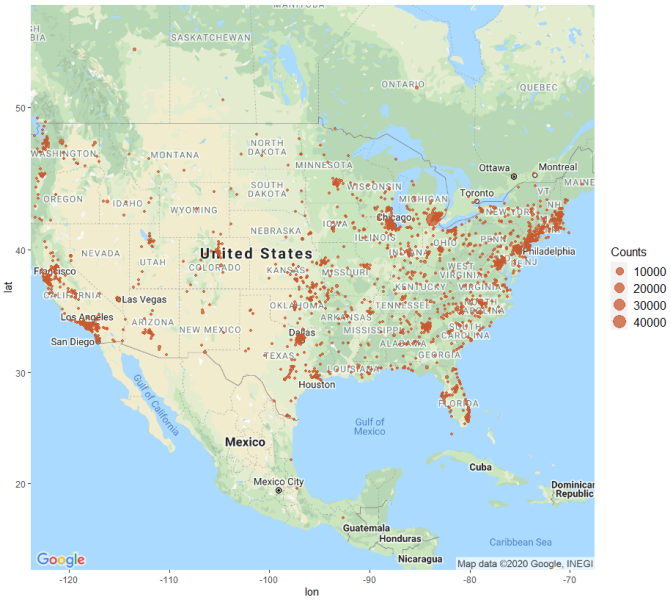


Figure 5: Density of city-level tweets.

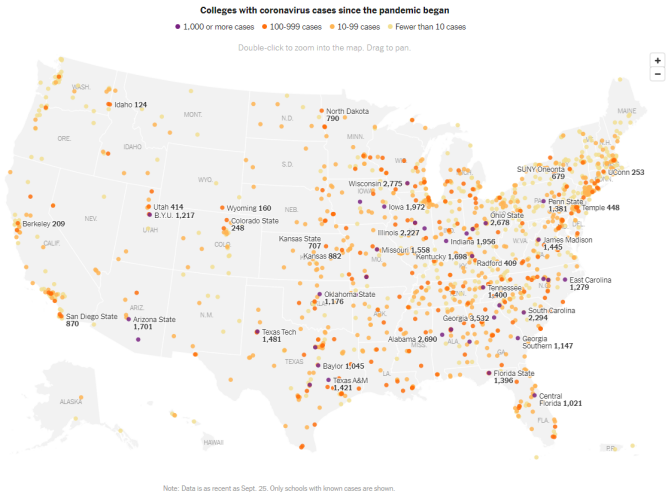


Figure 6: Colleges with confirmed COVID-19 cases [20].

the corpus, and we apply lemmatization to improve the model performance by keeping nouns, verbs, adjectives and adverbs. In this study, we mainly focus on the education domain, so we have fewer latent topics than a Twitter dataset that includes many other topics. We determine the optimal number of topics, n, topic coherence, an important metric in LDA [11], with $n = 4$ yielding the highest coherence of 0.53. We examine the produced topics and the associated keywords using the pyLDAvis package [18], which displays LDA results in an interactive chart, to verify that the discovered topics are reasonable. In Figure 10, each circle represents a topic in the inter-topic distance map. Our discovered topics are indeed reasonable because the topic circles are fairly big and non-overlapping.

Table 2 shows a sample of the topic results containing the top 10 most probable words in each topic. Topics appear coherent even

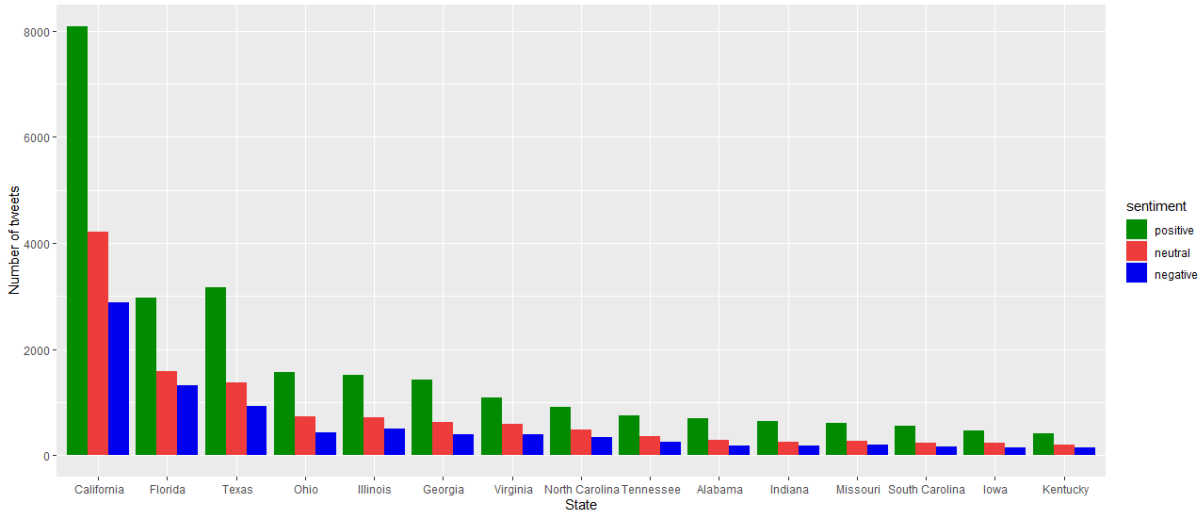


Figure 7: Sentiment analysis by state.

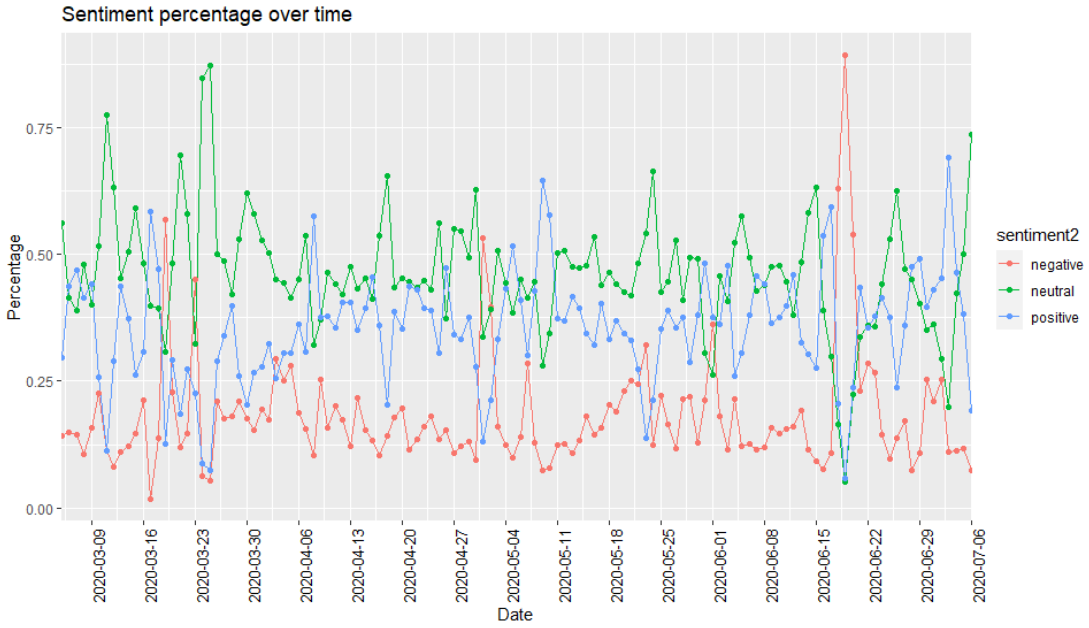


Figure 8: Daily sentiment in all states.

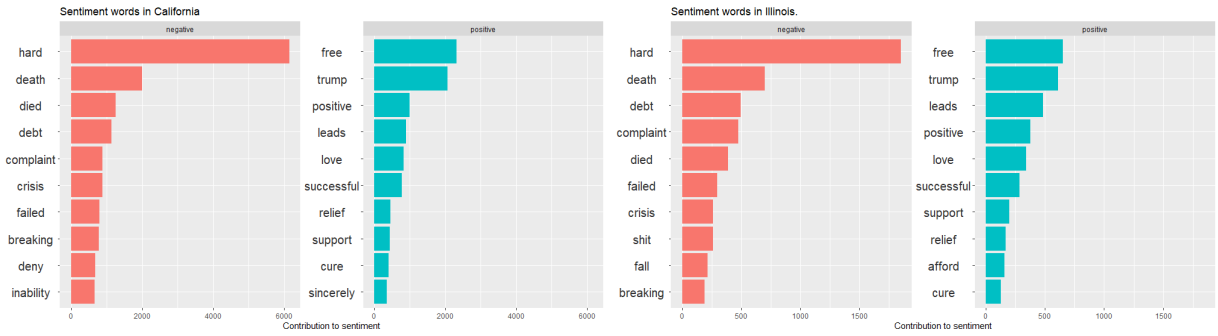
though some of words seem somewhat random, such as `free` and `today`. Topic 1 contains K12 and high school related words, such as `kid, child` and `high`, which appear in the group with high probability scores assigned by LDA.

High-level education is discussed in Topic 2, so we infer that college students are concerned with funding and tuition issues, as already explained in Section 3.2.

To go beyond the classification of tweets as positive, neutral, or negative, we use the 2015 Linguistic Inquiry and Word Count (LIWC15) to link daily word use to psychologically meaningful categories [19]. LIWC is able "to show attention focus, emotion, social relationships, thinking styles, and individual differences" [19]. We call these categories *language features*, and concentrate on `anxiety`, `positive emotion`, `family`, and `friend` (see Figure 11).

For each feature, r, the LIWC dictionary gives us a set of words $w_1, w_2, \ldots, w_j$ associated with that feature. For example, the words `nervous, afraid, tense` are a subset of the words associated with `anxiety` [17]. We can compute the score of r for the corpus of tweets T [2] in each region (for example, state, city), as an average:

$$score(r) = \frac{\sum_{k=1}^{|T|} \frac{s_{t_k}}{|t_k|}}{|T|} \quad (1)$$



(a) California (b) Illinois

Figure 9: Top sentiment words in California and Illinois.

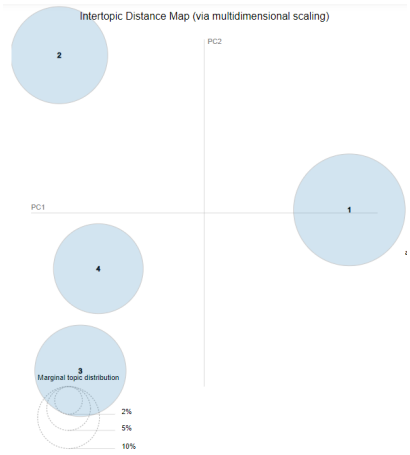


Figure 10: Intertopic distance map of the LDA results as obtained using pyLDAvis [18].

Table 2: Topics resulting from LDA.

Topic 1	Topic 2	Topic 3	Topic 4
school	learn	test	case
go	campus	help	time
get	online	even	day
kid	college	change	start
year	classroom	return	see
take	university	reopen	good
child	job	free	feel
high	tuition	try	share
teacher	fund	grow	home
risk	lab	believe	today

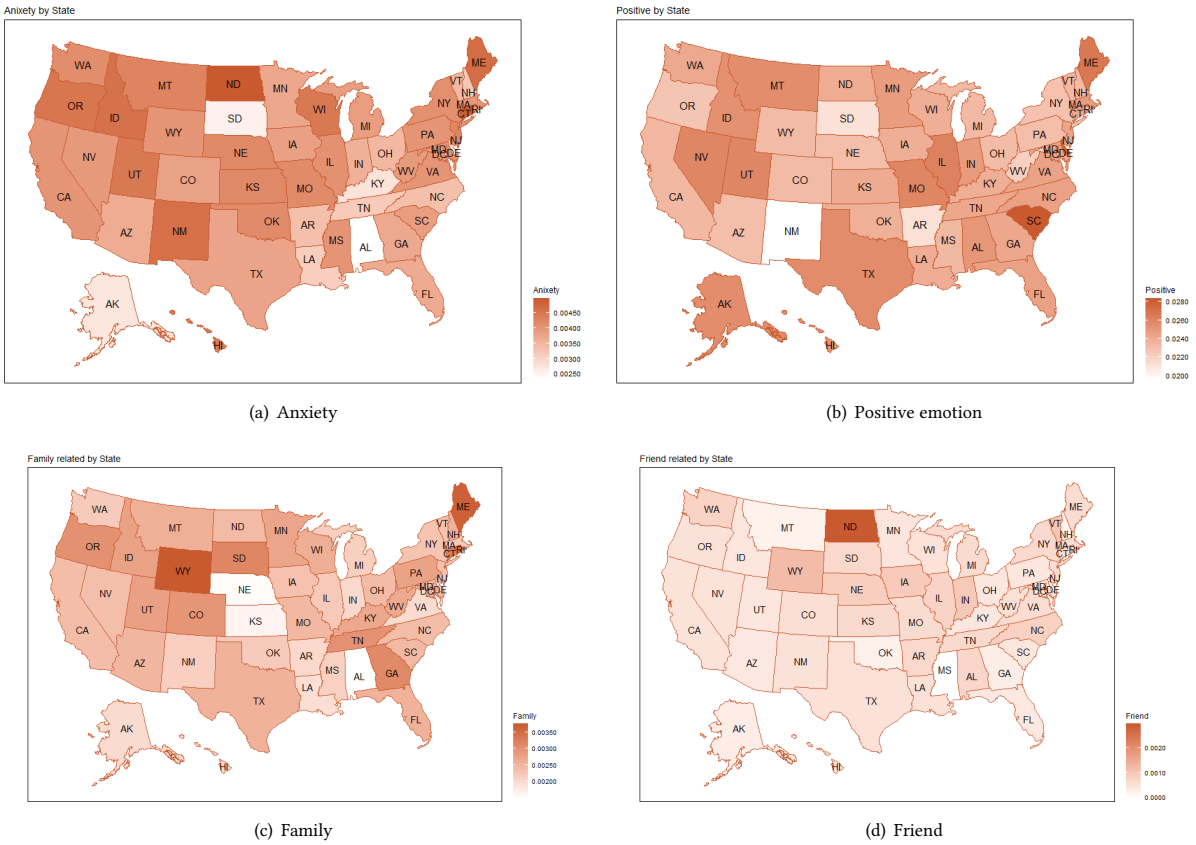
where $|t_k|$ is the number of words in tweet t_k, and s_{t_k} is the number of occurrences of all the words $w_1, w_2, \ldots, w_j$ in tweet t_k.

Figure 7 shows the language features at the state level. From Figure 11(a), we conclude that the levels of anxiety of the people in North Dakota and Maine are the highest in this study. A sample tweet says that `Going #backtoschool can be stressful, especially with the added uncertainty surrounding the #COVID19 pandemic`. However, Maine is also a state with relatively high positive emotions as shown in Figure 11(b). The states with many COVID-19 confirmed cases such as Georgia and North Carolina show less positive emotions. As shown in Figure 11(c), Maine is also one of the states where tweets relate more to family while North Dakota, as shown in Figure 11(d), is the state where tweets relate more to friends. Overall, tweets seem to be more about family than about friends as we compare the intensity of the color shades in Figures 11(c) and 11(d) possibly as the result of stay at home orders. For example, `I stay at home with my family to take online`.

4 RELATED WORK

In recent years, social media platforms have rapidly grown. Twitter is one of the most important platforms and gives away significant spatial and temporal information about the users and their posts. Users express and share details about their life, including health information and opinions on emerging events. For example, Twitter-based approaches can be used to detect late breaking news [16] or local news in spite of data scarcity [23]. The study of public health using Twitter encompass health monitoring and surveillance for early prediction of disease outbreaks. Dredze et al. [4] describe a system, called Carmen, which utilizes geocoding tools for influenza surveillance and considers geo-information from both tweets and users. Lee et al. [9] have collected over 6 million of flu-related tweets, with which they can analyze influenza rates in real time using a novel flu surveillance system.

A system by Paul et al. [13], called Compass, applies neural network methods for sentiment modeling and a dictionary for text classification to capture democratic vs. republican sentiment for the 2016 U.S. presidential election, at the county and state levels. A study of neighborhood happiness, diet, and physical activity has been performed by Nguyen et al. [12]. It applies sentiment analysis to regions, namely neighborhoods, based on geotagged tweets and socio-demographic characteristics, as provided by census data. Martinez et al. examine the use of and perceptions about e-cigarettes



(a) Anxiety

(b) Positive emotion

(c) Family

(d) Friend

Figure 11: Language features.

in the U.S. [10]. They use a sample of tweets that they manually geocoded and interpreted, which has the advantage of creating a detailed categorization of the tweets.

Sadilek et al. [15] and Wang et al. [21] perform studies to detect and predict foodborne illness using respectively Twitter and Yelp. They analyze the tweets using language features and language models. Zhao et al. [24] compared the content of Twitter with a typical traditional news medium, the New York Times, using unsupervised topic modeling. They develop a new Twitter-LDA model that is effective for short tweets. Jaidka et al. [7] want to monitor well-being at large scale. They find that text-based methods that use language dictionaries including labMT and LIWC2015, which we used, can easily work at a large scale. However, they found that supervised data-driven methods are more robust, when compared with a gold standard. Similarly to our paper, Paul and Dredze [14] use an unsupervised model, in their case called Ailment Topic Aspect Model (ATAM), and LDA to learn how users express their illnesses and ailments in tweets.

5 CONCLUSIONS AND FUTURE WORK

While Twitter has been used to analyze and predict other diseases, in the case of COVID-19, everything is new: the disease itself, its extent, how it propagates and affects several sectors, including education, and which emotions it brings out. Our large scale study analyzes 673,601 tweets over two time periods in 2020.

Our study is about several aspects, including psychological categories [19], which capture emotion, social relationships, and individual differences. We analyze individual data, and aggregate that information, as a function of time or of state. We illustrate the impact of COVID-19 on education by means of spatio-temporal patterns, sentiment and content analysis applied to a large geotagged Twitter dataset, and capture the topics of greatest concern such as funding and tuition. We found many similarities such as daily tweet numbers (Figure 1) and common patterns such as sentiment percentage in different states (Figure 7).

All in all, we converted a large dataset of tweets into meaningful geospatial information. There are many interesting facts that can be observed. We also verify the correlation of the results that are obtained from sentiment analysis with the number of confirmed positive cases.

As we mentioned in Section 3.2, there is ambiguity when we encounter the word `positive`, which usually denotes a positive sentiment. However, in the case of `positive result`, that tweet is the opposite of being positive, while a `negative result` is great news, and definitely *positive*. The problem arises because we use bags of words. In future work, we will be looking at capturing better

the semantics of such cases. It is to our advantage that tweets are focused on COVID-19 only, thus limiting the number of possible ambiguous word associations. Another possibility is to manually label a random subset of tweets so as to apply machine learning algorithms for sentiment analysis. At this time (October of 2020), it seems that the pandemic will last several more months. If that is the case, then we will have a baseline to compare results for 2021 with those we obtain in this paper.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This work was partially supported by NSF award III-1618126.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022. http://jmlr.org/papers/v3/blei03a.html

[2] C. K. Chung and J. W. Pennebaker. 2018. What Do We Know When We LIWC A Person? Text Analysis As An Assessment Tool for Traits, Personal Concerns and Life Stories. In *The SAGE Handbook of Personality and Individual Differences*, V. Zeigler-Hill and T. K. Shackelford (Eds.) (Eds.). Sage Reference, 341–360. https://doi.org/10.4135/9781526451163.n16

[3] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PloS One* 6, 12 (2011), e26752.

[4] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health. In *AAAI Workshop on Expanding the Boundaries of Health informatics using AI (HIAI)*, Vol. 23. Citeseer, 45.

[5] Xian Gao. 2017. Networked Co-Production of 311 Services: Investigating the Use of Twitter in Five U.S. Cities. *International Journal of Public Administration* 41 (03 2017), 1–13. https://doi.org/10.1080/01900692.2017.1298126

[6] Shawn Hubler and Anemona Hartocollis. 2020. How Colleges Became the New Covid Hot Spots. (September 11, 2020). Retrieved October 18, 2020 from https://www.nytimes.com/2020/09/11/us/college-campus-outbreak-covid.html (Updated October 2, 2020).

[7] Kokil Jaidka, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2020. Estimating Geographic Subjective Well-being from Twitter: A Comparison of Dictionary and Data-driven Language Methods. *Proceedings of the National Academy of Sciences* 117, 19 (2020), 10165–10171.

[8] Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset. (2020). https://doi.org/10.21227/781w-ef42

[9] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time Digital Flu Surveillance using Twitter data. In *2nd Workshop on Data Mining for Medicine and Healthcare*.

[10] Lourdes S. Martinez, Sharon Hughes, Eric R. Walsh-Buhi, and Ming-Hsiang Tsou. 2018. Okay, We Get It. You Vape: An Analysis of Geocoded Content, Context, and Sentiment Regarding E-cigarettes on Twitter. *Journal of Health Communication* 23, 6 (2018), 550–562.

[11] David Newman, Edwin V. Bonilla, and Wray L. Buntine. 2011. Improving Topic Coherence with Regularized Topic Models. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems (NIPS) 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (Eds.). 496–504. http://papers.nips.cc/paper/4291-improving-topic-coherence-with-regularized-topic-models

[12] Quynh C Nguyen, Suraj Kath, Hsien-Wen Meng, Dapeng Li, Ken R Smith, James A VanDerslice, Ming Wen, and Feifei Li. 2016. Leveraging Geotagged Twitter Data to Examine Neighborhood Happiness, Diet, and Physical Activity. *Applied Geography* 73 (2016), 77–88.

[13] Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, and Richie Frost. 2017. Compass: Spatio Temporal Sentiment Analysis of US Election: What Twitter Says!. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 1585–1594. https://doi.org/10.1145/3097983.3098053

[14] Michael J. Paul and Mark Dredze. 2011. You are What You Tweet: Analyzing Twitter for Public Health. In *Fifth International AAAI Conference on Weblogs and Social Media*. Citeseer.

[15] Adam Sadilek, Henry A. Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio. 2017. Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. *AI Mag.* 38, 1 (2017), 37–48. https://doi.org/10.1609/aimag.v38i1.2711

[16] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in Tweets. In *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2009, November 4-6, 2009, Seattle, Washington, USA, Proceedings*. ACM, 42–51. https://doi.org/10.1145/1653771.1653781

[17] "Kovach Computing Services". 2007. LIWC Dictionary (Linguistic Inquiry and Word Count). (2007). https://www.kovcomp.co.uk/wordstat/LIWC.html Accessed September 26, 2020.

[18] Carson Sievert and Anemona Kenny Shirley. 2015. pyLDAvis. (April 5, 2015). Retrieved October 21, 2020 from https://github.com/bmabey/pyLDAvis

[19] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[20] New York Times. 2020. Tracking Covid at U.S. Colleges and Universities. (2020). https://www.nytimes.com/interactive/2020/us/covid-college-cases-tracker.html Accessed September 25, 2020.

[21] Zhu Wang, Booma Sowkarthiga Balasubramani, and Isabel F. Cruz. 2017. Predictive Analytics Using Text Classification for Restaurant Inspections. In *ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, Huy T. Vo and Bill Howe (Eds.). ACM, 14:1–14:4. https://doi.org/10.1145/3152178.3152192

[22] Sarah Watson, Shawn Hubler, Danielle Ivory, and Robert Gebeloff. 2020. A New Front in America's Pandemic: College Towns. (September 6, 2020). Retrieved October 18, 2005 from https://www.nytimes.com/2020/09/06/us/colleges-coronavirus-students.html

[23] Hong Wei, Jagan Sankaranarayanan, and Hanan Samet. 2020. Enhancing Local Live Tweet Stream to Detect News. *GeoInformatica* 24, 2 (2020), 411–441. https://doi.org/10.1007/s10707-019-00392-9

[24] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *European Conference on Information Retrieval*. Springer, 338–349.

Corona Games: Masks, Social Distancing and Mechanism Design

Balázs Pejó*
CrySyS Lab
Dept. of Networked Systems and Services
Budapest Univ. of Technology and Economics
pejo@crysys.hu

Gergely Biczók*
CrySyS Lab
Dept. of Networked Systems and Services
Budapest Univ. of Technology and Economics
biczok@crysys.hu

Abstract

Pandemic response is a complex affair. Most governments employ a set of quasi-standard measures to fight COVID-19 including wearing masks, social distancing, virus testing and contact tracing. We argue that some non-trivial factors behind the varying effectiveness of these measures are selfish decision-making and the differing national implementations of the response mechanism.
In this paper, through simple games, we show the effect of individual incentives on the decisions made with respect to wearing masks and social distancing, and how these may result in a sub-optimal outcome. We also demonstrate the responsibility of national authorities in designing these games properly regarding the chosen policies and their influence on the preferred outcome. We promote a mechanism design approach: it is in the best interest of every government to carefully balance social good and response costs when implementing their respective pandemic response mechanism.

CCS Concepts

• **Computing methodologies** → **Modeling methodologies**; **Agent / discrete models**.

Keywords

Masks, Social Distancing, Game Theory, Mechanism Design

ACM Reference Format:
Balázs Pejó and Gergely Biczók. 2020. Corona Games: Masks, Social Distancing and Mechanism Design. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3423459.3430757

1 Introduction

The current coronavirus pandemic is pushing individuals, businesses and governments to the limit. People suffer owing to restricted mobility, social life and income, complete business sectors face an almost 100% drop in revenue, and governments are scrambling to find out when and how to impose and remove restrictions. In fact, COVID-19 has turned the whole planet into a "living lab" for human and social behavior where feedback on response measures employed is only delayed by around two weeks (the incubation period). From the 24/7 media coverage, all of us have been introduced to a set of quasi-standard measures introduced by national and local authorities, including wearing masks, social distancing, virus testing, contact tracing and so on. It is also clear that different countries have had different levels of success employing these measures as evidenced by the varying normalized death tolls and confirmed cases[1].

We believe that apart from the intuitive (e.g., genetic differences, medical infrastructure availability, hesitancy, etc.), there are two significant factors that have not received enough attention. First, the *individual incentives* of citizens, e.g., "is it worth more for me to stay home than to meet my friend?", have a significant say in every decision situation. While some of those incentives can be inherent to personality type, clearly, there is a non-negligible rational aspect to it, where individuals are looking to maximize their own utility. Second, countries have differed in their specific *implementation* of response measures, e.g., whether they have been distributing free masks (affecting the efficacy of mask wearing in case of equipment shortage) or providing extra unemployment benefits (affecting the likelihood of proper self-imposed social distancing). Framing pandemic response as a mechanism design problem, i.e., architecting a complex response mechanism with a preferred outcome in mind, can shed light on these factors; what's more, it has the potential to help authorities (mechanism designers) fight the pandemic efficiently. The objective of this paper is to show that both individual incentives and the actual design and implementation of the holistic pandemic response mechanism can have a major effect on how this pandemic plays out.

Contribution. In this paper we model decision situations during a pandemic with game theory where participants are rational, and the proper design of the games could be the difference between life and death. Our main contribution is two-fold. First, regarding decisions on wearing a mask, we show that i) the equilibrium outcome is not socially optimal under full information, ii) when the status of the players are unknown the equilibrium is not to wear a mask for a wide range of parameters, and iii) when facing an infectious player it is almost always optimal to wear a mask even with low protection efficiency. Furthermore, for social distancing, using current COVID-19 statistics we showed that i) going out is only rational when it corresponds to either a huge benefit or staying home results in a significant loss, and ii) we determined the optimal duration and meeting size of such an out-of-home activity. Second, we take a look at pandemic response from a mechanism design perspective, and demonstrate that i) different government

*Although in the same institute, authors have collaborated remotely.

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430757

[1]Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html

policies influence the outcome of these games profoundly, and ii) individual response measures (sub-mechanisms) are interdependent. Specifically, we discuss how contact tracing enables targeted testing which in turn reduces the uncertainty in individual decision making regarding both social distancing and wearing masks. We recommend governments treat pandemic response as a mechanism design problem when weighing response costs vs. the social good.

Organization. The remaining of the paper is structured as follows. Section 2 briefly describes related work while Section 3 recaps some basics of game theory. Section 4 develops and analyzes the Mask Game adding uncertainty, mask efficiency and multiple players to the basic model. Section 5 develops and analyzes the Distancing Game including the effects of meeting duration and size. Section 6 frames pandemic response as a mechanism design problem using the design of the two games previously introduced as examples. Finally, Section 7 outlines future work and concludes the paper.

2 Related Work

In this section we review some well-known epidemic spreading models and game-theoretic works in relation to pandemics.

COVID-19 have been modelled using different models: for instance using SIR [4], SEIQR [28], and SIDARTHE [12]. Which model suits the ongoing epidemic best is still undetermined. Besides the model, the input data instantiating the model may be imperfect as well, thus some efforts are also made to account for potential inaccuracies in the reported data [14]. An orthogonal extension of these models is proposed in [23], which discusses how factors such as hospital capacity, test capacity, demographics, population density, vulnerable people and income could be integrated into these models. In contrast with the previous models, the one in [16] takes into consideration the networked structure of human interconnections and the locality of interactions, without attempting a mean-field approach. In the following we briefly review some related research efforts in the intersection of epidemics and game theory. For a comprehensive survey we refer the reader to [5].

Some researchers modeled the behavioral changes of people to a pandemic: for instance in [21] authors used evolutionary game theory, and showed that slightly reducing the number of people an individual was in contact with could make a difference regarding the spread of disease. Another group showed that there was a critical level of concern, i.e., empathy, by the infected individuals above which the disease is eradicated rapidly [9]. Others focused on the mobility habits of people traveling between areas affected unevenly by the disease, and found conflict between the Nash Equilibrium (individually optimal strategy) and the Social Optimum (optimal group strategy) only under specific changes in economic and epidemiological conditions [29]. In [1] an optimization problem was formalized by accommodating both isolation (modeled by how far individuals are from home) and social distancing (how far individuals are from each other). Authors also provided incentives for maintaining social distancing to prevent the spread of COVID-19 (i.e., making "staying home" the Nash Equilibrium). Moreover, social distancing was also shown to be able to delay the epidemic until a vaccine becomes widely available [22].

Several studies focused on how the availability of vaccines affects human behaviour. A model was introduced in [3] where vaccine delayers relied on herd immunity and vaccine safety information generated by early vaccinators. Consequently, the Nash Equilibrium was "wait and see". Another study concerning this vaccination dilemma proposed a model with incentives for individuals to choose the prevention strategy according to risks and expenses in the epidemic campaign [2]. Similarly, researchers in [27] showed the optimal use of anti-viral treatment by individuals when they took into account the direct and indirect costs of treatment. The game-theoretic model in [25] focused on the various level of drug stockpiles in different countries, and found controversial results: sometimes there was an optimal solution with a central planner (such as the WHO), which improved on the decentralized equilibrium, but other times the central planner's solution (minimizing the number of infected persons globally) required some countries to sacrifice part of their population.

The exact dynamics of demand and supply for medical resources at different phases of a pandemic was also studied [7]. Predicting such dynamics would provide a quantitative basis for mechanism designers (e.g., decision makers of healthcare systems) to understand the potential imbalance of supply and demand. The authors extended the concepts of reserving and capital management in the classical insurance literature and aimed to provide a quantitative framework for quantifying and assessing pandemic risk, and developed optimal strategies for stockpiling spatio-temporal resources.

The Centers for Disease Control and Prevention created a policy review of social distancing measures for pandemic influenza in non-healthcare settings [11]. They identified measures to reduce community influenza transmission such as isolating the sick, tracing contacts, quarantining exposed people, closing down school, changing workplace habits, avoiding crowds, and restricting movement. The impact of several of these (and wearing masks) was studied in [24] in which the authors model the pandemic by emulating people, business and government. Other researchers demonstrated that early school and workplace closure, and restriction of international travel are independently associated with reduced national COVID-19 mortality [20]. On the other hand, lock-down procedures could have devastating impact on the economy. This was studied in [6] with a modified SIR model and time-dependent infection rate. The authors found that, surprisingly, in spite of the economic cost of the loss of workforce and incurred medical expenses, the optimum point for the entire course of the pandemic is to keep the strict lock-down as long as possible.

As detailed above, related work has mostly studied narrowly focused specifics of epidemic modelling such as the intricate behaviour of individuals in relation with vaccines or the preferred actions of mechanism designers such as healthcare system operators. In contrast, our work takes a step back, and focuses on the big picture: we model decision situations during a pandemic as games with rational participants, and promote the proper design of these games. We highlight the responsibility of mechanism designers such as national authorities in constructing these games properly with adequately chosen policies, taking into account their interdependent nature.

3 Preliminaries

In this section we shortly elaborate on the main game theoretical notions used in this paper, to enable the conceptual understanding of the implications of our results.

Game theory [13] is "the study of mathematical models of conflict between intelligent, rational decision-makers". Almost every multi-party interaction can be modeled as a game. In relation to COVID-19, decision makers could be individuals (e.g., whether to wear a mask), cities (e.g., whether to enforce wide-range testing within the city), governments (e.g., whether to apply contact tracking within the country), or companies (e.g., whether to apply social distancing within the workplace). Potential decisions are referred to as strategies; decision makers (players) choose their strategies rationally so as to maximize their own utility.

The Nash Equilibrium (NE) — arguably the most famous solution concept — is a set of strategies where each player's strategy is a best response strategy. This means every player makes the best/optimal decision for itself as long as the others' choices remain unchanged. NE provides a way of predicting what will happen if several entities are making decisions at the same time where the outcome also depends on the decisions of the others. The existence of a NE means that no player will gain more by unilaterally changing its strategy at this unique state. Another game-theoretic concept is the Social Optimum, which is a set of strategies that maximizes social welfare. Note, that despite the fact that no one can do better by changing strategy, NEs are not necessarily Social Optima (we refer the reader to the famous example of the Prisoner's Dilemma [13]). In fact, it is a central problem in game theory how much a distributed outcome (NE) is worse than a centrally planed social optimum.

If one knows the NE they prefer as the outcome of a game, e.g., everybody wearing a mask, and they have the power to instantiate the game accordingly, i.e., fixing the structure, game flow and any free parameters, then we talk about mechanism design [17]. In a way, mechanism design is the inverse of game theory; although a significant share of efforts within this field deals with auctions, mechanism design is a much broader term widely applicable to any mechanism, e.g., optimal organ matching for transplantation or school-student allocation, aimed at achieving a given steady state result.

4 The Mask Game

Probably the most visible consequence of COVID-19 are masks: before their usage was mostly limited to some Asian countries, hospitals, constructions and banks (in case of a robbery). Due to the coronavirus pandemic, an unprecedented spreading of mask-wearing can be seen around the globe. Policies have been implemented to enforce their usage in some places, but in general, it has been up to the individuals to decide whether to wear a mask or not based on their own risk assessment. In this section, we model this decision situation via game theory. We assume that there are several types of masks, providing different level of protection.

- **No** Mask corresponds to the behavior of using no masks during the COVID-19 (or any) pandemic. Its cost is consequently 0; however, it does not offer any protection against the virus.
- **Out** Mask is the most widely used mask (e.g., cloth mask or surgical mask). They are meant to protect the environment of the individual using it. They work by filtering out droplets when coughing, sneezing or simply talking, therefore they limit the spreading of the virus. They do not protect the wearer itself against an airborne virus. The cost of deciding for this protection type is noted as $C_{out} > 0$.
- **In** Mask is the most protective prevention gear designed for medical professionals (e.g., FFP2 or FFP3 mask with valves). Valves make it easier to wear the mask for a sustained period of time, and prevent condensation inside the mask. They filter out airborne viruses while breathing in, however the valved design means they do not filter the while air breathing out. Note that CDC guidelines[2] recommend using a cloth/surgical mask for the general public, while valved masks are only recommended for medical personnel in direct contact with infected individuals. The cost of this protection type is $C_{in} >> C_{out}$.

Besides which mask they use (i.e., the available strategies), the players are either susceptible or infected[3]. The latter has some undesired consequence; hence, we model it by adding a cost C_i to these players' utility (which is magnitudes higher than even C_{in}, i.e., $C_i >> C_{in} >> C_{out}$). We summarize all the parameters and variables used for the Mask game in Table 1.

Variable	Meaning
C_{out}	Cost of playing **out**
C_{in}	Cost of playing **in**
C_i	Cost of being infected
C_{use}	Cost of playing **use**
ρ	Prob. of being infected
p	Prob. of using a mask
a	Protection Efficiency
b	Spreading Efficiency

Table 1: Parameters of the Mask Games

Using these states and masks, we can present the basic game's payoffs where two players with known health status meet and decide which mask to use. The payoff matrix in Table 2 corresponds to the case when both players are susceptible. Note, that in case both players are infected, the payoff matrix would be the same with an additional constant C_i. Table 3 corresponds to the case when one player is infected while the other is susceptible.

	no	out	in
no	$(0, 0)$	$(0, C_{out})$	$(0, C_{in})$
out	$(C_{out}, 0)$	(C_{out}, C_{out})	(C_{out}, C_{in})
in	$(C_{in}, 0)$	(C_{in}, C_{out})	(C_{in}, C_{in})

Table 2: Payoff matrices when both players are susceptible

[2] Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

[3] We simplify the well-known SIR model [8] since in case of COVID-19 it is currently unclear if and for how long an individual is resistant after recovery.

In Table 2 it is visible that both players' cost is minimal when they do not use any masks, i.e., the Nash Equilibrium of the game when both players are susceptible is (**no**, **no**). This is also the social optimum, meaning that the players' aggregated cost is minimal. The same holds in case both players are infected, as this only adds a constant C_i to the payoff matrix.

	no	out	in
no	(C_i, C_i)	$(0, C_{out} + C_i)$	$(C_i, C_{in} + C_i)$
out	$(C_{out} + C_i, C_i)$	$(C_{out}, C_{out} + C_i)$	$(C_{out} + C_i, C_{in} + C_i)$
in	(C_{in}, C_i)	$(C_{in}, C_{out} + C_i)$	$(C_{in}, C_{in} + C_i)$

Table 3: Payoff matrices for the cases when only one player is susceptible.

When only one of the players is susceptible as represented in Table 3, using no mask is a dominant strategy for the infected player[4], since it is a best response, independently of the susceptible player's action. Consequently, the best option for the susceptible player is **in**, i.e., the NE is (**in**, **no**). On the other hand, the social optimum is different: (**no**,**out**) would incur the least burden on the society since $C_{out} << C_{in}$.

In social optimum, susceptible players would benefit, through a positive externality, from an action that would impose a cost on infected players; therefore it is not a likely outcome. In fact, such a setting is common in man-made distributed systems, especially in the context of cybersecurity. A well-fitting parallel is defense against Distributed Denial of Service Attacks (DDoS) attacks [15]: although it would be much more efficient to filter malicious traffic at the source (i.e., **out**), Internet Service Providers rather filter at the target (i.e., **in**) owing to a rational fear of free-riding by others.

4.1 Bayesian Game

Since in the basic game no player plays **out**, we simplify the choice of the players to either **use** a mask or **no** (hence, we note the cost of a mask with C_{use}). To represent the situation more realistically, we introduce ambiguity about the status of the players: we denote the probability of being infected as ρ. We know from the basic game that if both players are infected (with probability ρ^2) or susceptible (with probability $(1-\rho)^2$) they play (**no,no**), while if only one of them is infected (with probability $2 \cdot \rho \cdot (1-\rho)$) the infected player plays **no**, while the susceptible plays **use**. Hence, the players play **no** in most of the cases (e.g., with probability $1 - (\rho \cdot (1-\rho))$).

On the other hand, this is not the case if we do not assume that the players know their statuses. Consequently, with uncertainty we must minimize the costs of the players: if both players are infected with equal probability, the payoff for Player 2 is Equation 1 where p_n is the probability that Player n plays **use** (otherwise she plays **no**). The payoff for the other player is similar since the game is symmetric. In more detail, the first two lines correspond to the case when Player 2 is not infected (hence the multiplication with $1-\rho$ at the beginning), while the last line captures when she is infected. Either way, she plays **use** with probability p_2, which incurs a cost of C_{use}. Otherwise she plays **no**, which has no cost except when Player 1 is infected and she plays **no** as well.

$$\begin{aligned} U_2 = (1-\rho) \cdot [(1-\rho) \cdot [p_2 \cdot C_{use} + (1-p_2) \cdot 0] + \\ \rho \cdot [p_2 \cdot C_{use} + (1-p_2) \cdot [(1-p_1) \cdot C_i + p_1 \cdot 0]]] + \\ \rho \cdot [p_2 \cdot (C_i + C_{use}) + (1-p_2) \cdot C_i] \end{aligned} \quad (1)$$

Since this formula is linear in p_2, its extreme point within [0,1] is situated exactly at the boundary. We take its derivative to uncover the function steepness: the condition for the function to be decreasing (i.e., higher probability for using a mask corresponds to lower cost) is seen below. Consequently, the only scenario which might admit wearing a mask with non-zero probability corresponds to the availability of sufficiently cheap masks.

$$\frac{\partial U_2}{\partial p_2} < 0 \Leftrightarrow \frac{C_{use}}{C_i} < \rho \cdot (1-\rho) \cdot (1-p_1) \leq 1 \quad (2)$$

Example. Lets assume Alice is going to meet Bob after a long time without any correspondence. Consequently, she does not know whether Bob has been exposed to SARS-CoV-2 recently. Actually, Alice herself could have been exposed as well without her knowledge, as up to 80% of the infectious cases could be asymptomatic.[5] For this reason, without taking into account any available spatial data, she estimates that they could be infectious with $\rho = 50\%$: either yes or no. She also does not have any information about Bob's mask wearing habits, so she guesses $p_1 = 0.5$ as well.

Alice is tested at her workplace every day, and she is sent to a 1-week quarantine without payment if tested positive. If we represent Alice as an average American, she earns approximately 1000 USD per week[6], hence, we set $C_i = 1000$. Substituting these into the right side of Equation (2), she decides to wear a mask only if it costs less than 125 USD, which does hold as of September 2020.

4.2 Efficiency Game

In the basic game we assumed **in** provides perfect protection from infected players, while **out** protects the other player fully. However, in real life, these strategies only mitigate the infection by decreasing its probability (i.e., ρ) to some extent. For this reason, we define $a, b \in [0,1]$ in a way that the smaller value of the parameter corresponds to better protection; a measures the protection efficiency of the protection strategy, while b captures the efficiency of eliminating the further spread of the disease. Consequently, a and b was set in the previous cases to $a_{\textbf{out}} = 0, a_{\textbf{in}} = 1$ (**in** prevents further spreading, while **out** does not), $b_{\textbf{out}} = 1$ (**out** has no effect on protecting the player) $b_{\textbf{in}} = 0$ (**in** perfectly protects the player).

We simplify the action space of the players as we did in the Bayesian game: **in** and **out** is merged into **use** Obviously, **no** corresponds to $a_{no} = b_{no} = 0$. We abuse the notion a and b to represent a_{use} and b_{use} respectively. We set $b = \frac{2}{3}$, as surgical masks on the infectious person reduce cold & flu viruses in aerosols by 70% according to [19]. Parameter a is much harder to measure. It should be $a \leq b$ since any mask keeps the virus inside the players more efficiently than stopping the wearer from getting infected. For the

[4]Note that the payoffs does not take into account the legal consequences of a deliberate infection such as in https://www.theverge.com/2020/4/7/21211992/coughing-coronavirus-arrest-hiv-public-health-safety-crime-spread.

[5]Centre for Evidence-Based Medicine. https://www.cebm.net/covid-19/covid-19-what-proportion-are-asymptomatic/

[6]Bureau of Labour Statistics. https://www.bls.gov/news.release/pdf/wkyeng.pdf

sake of the analysis we set $a = \frac{b}{2} = \frac{1}{3}$, but any other choice would be possible.

We are interested in the mask-wearing probability of a susceptible player when the other player is infected.[7] The utility in such a situation is shown in Equation (3), where for simplification we defined $p = p_1 = p_2$, i.e., both players play a specific strategy with the same probability. With such a constraint, we restrict ourselves from finding all the solutions; however, since the game is symmetric, an equilibrium of this reduced game is also an equilibrium when the players could use different strategy distributions.

$$\begin{aligned} U =& p^2 \cdot (C_{use} + C_i \cdot a \cdot b) + \\ & p \cdot (1-p) \cdot (C_{use} + C_i \cdot a) + \\ & (1-p) \cdot p \cdot (C_i \cdot b) + \\ & (1-p)^2 \cdot C_i \qquad (3) \\ \Rightarrow U =& p^2 \cdot (C_{use} + C_i \cdot 0.\dot{2}) + \\ & p \cdot (1-p) \cdot (C_{use} + C_i) + \\ & (1-p)^2 \cdot C_i \end{aligned}$$

From this we easily deduce that **use** corresponds to a smaller cost that **no** if $\frac{C_{use}}{C_i} < \frac{7}{9}$, which holds by default as $C_{use} \ll C_i$ (even for less efficient masks). Moreover, **use** (i.e., $p = 1$) is the best response most of the time because of the following.

(1) The utility is a second order polynomial, hence it has one extreme point.
(2) This extreme point is a minimum due to $U'' = \frac{4}{9} \cdot C_i > 0$.
(3) The utility (i.e., cost) is decreasing on the left and increasing on the right of this minimum point.
(4) The utility's minimum point is at $p = \frac{9}{4} \cdot \frac{C_i - C_{use}}{C_i}$ due to $U' = C_{use} - C_i + \frac{4}{9} \cdot C_i \cdot p$.
(5) The minimum point is expected to be above 1 due to $C_{use} \ll C_i$.
(6) $p \in [0, 1]$ is on the left of the minimum point, hence, a higher p corresponds to a smaller cost.

4.3 Multi-Player Game

This game can be further extended by allowing more players to participate. In this extension — if we assume all players meet with probability 1 — with any number of infected players (who play **no** as we showed already) all the susceptible players should play **in**. This NE is the SO as well if the ratio of the infected (which is identical to the probability ρ of being infected) is sufficiently high: the accumulated cost when the susceptible players play **in** (and the infected play **no**) is less than the accumulated cost when the infected players play **out** (and the susceptible play **no**) if $\frac{C_{in}}{C_{out}} < \frac{\rho}{1-\rho}$. Although it is mathematically possible that the infected plays **no** in the SO, but it is doubtful: both the cost of **in** is significantly higher than **out**, and the infection ratio ρ is low (at least at the beginning of the pandemic).

[7] The Bayesian game combined with efficiency is left for future work due to the lack of space.

5 The Distancing Game

Another phenomenon most people has experienced during the current COVID-19 pandemic is social distancing. Here we introduce a simple Distancing Game; it is to be played in sequence with the previously introduced Mask Game: once a player decided to meet up with friends via the Distancing Game, she can decide whether to wear a mask for the meeting playing the Mask Game. To improve readability, we summarize all the corresponding parameters and variables in Table 4.

Variable	Meaning
C	Cost of staying home
B	Benefit of going out
m	Mortality rate
L	Value of Life
ρ	Probability of infection
p	Probability of meeting
t	Time duration of meeting
g	Group size of meeting

Table 4: Parameters of the Distancing Games

We represent the cost of getting infected with $m \cdot L$, i.e., the mortality rate of the disease multiplied with the player's evaluation about her own life.[8] Besides the risk of getting infected, going out or attending a meeting could benefit the player, denoted as B. On the other hand, staying home or missing a meeting could have some additional costs, denoted as C. The probability of getting infected is denoted as ρ. With these notations, the utility of the Distancing Game is captured on the left of Equation (4), where p is the probability of going out. Since this is linear in p, its maximum is either at $p = 0$ (stay home) or $p = 1$ (go out). Precisely, the player prefers to stay home if the right side of Equation (4) holds.

$$U = p \cdot (B - \rho \cdot m \cdot L) - (1-p) \cdot C \qquad \frac{B+C}{\rho \cdot m} < L \qquad (4)$$

Example. For instance, should a rational American citizen (e.g., Alice) go out based on how much she values her life? We estimate[9] $m = 0.034$ and $\rho = 0.0077$ as $0.028 \approx \frac{\#\{\text{deceased}\}}{\#\{\text{all cases}\}} < m < \frac{\#\{\text{deceased}\}}{\#\{\text{closed cases}\}} \approx 0.04$ while $\frac{\#\{\text{active cases}\}}{\#\{\text{population}\}} \approx 0.0077$.

Using these values, Alice should go out only if she values her life less than $3820 (= \frac{1}{0.034 \cdot 0.0077})$ times the benefit (of going out) and the loss (of staying home) together. According to [26], the value of a statistical life in the US was 9.2 million USD in 2013, which is equivalent to 11.3 million USD in 2020 (with 0.3% interest rate). This means, Alice should only meet someone if the benefit of the meeting (and thus the cost of missing out) would amount to more than USD $2,958 (= \frac{11.3M}{3820})$.

5.1 Number of Participants and Duration

One way to improve the above model is by introducing meeting duration and size. Leaving our disinfected home during a pandemic

[8] This is an optimistic approximation, as besides dying the infection could bear other tolls on a player.
[9] Data from https://www.worldometers.info/coronavirus/ (accessed September 2020)

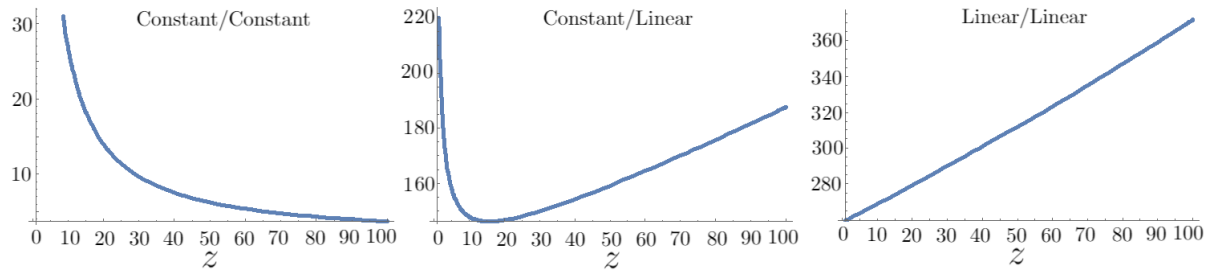


Figure 1: A few examples for various benefit and cost functions of the lower limit on the life value which would ensure that Alice (i.e., a rational American) would stay home (i.e.,left side of Equation (5)) with $m = 0.034$ and $\rho = 0.0077$.

is risky, and this risk grows with the time. Similarly, a meeting is riskier when there are multiple participants involved. In the original model, we captured the infection probability with $\rho = 1 - (1 - \rho)$. This ratio increases to $1 - (1 - \rho)^{g \cdot t}$ when there are g possible infectious sources for t time. Since g and t are interchangeable, we merge this two together under a common notation: $z = g \cdot t$.

This extended model can be used to determine the optimal duration and size of a meeting, once the player decided to go out according to the basic Distancing Game. We define $0 < z < 100$, as no player has infinite time or meeting partners. Moreover, the benefit and the loss of attending and missing a meeting should depend on this new parameter. For instance, staying home in isolation for a longer period might cause anxiety, which could get worse over time (i.e., increasing the cost); on the other hand, attending a meeting with many friends at the same time could significantly boost the experience (i.e., increase the benefit). Consequently, a rational person should leave her home only if Equation (5) holds which is the extension of the right side of Equation (4).

$$\max_{0<z<100} \left(\frac{B(z) + C(z)}{(1 - (1 - \rho)^z) \cdot m} \right) < L \tag{5}$$

In Figure 1, we present three use-cases of the formula inside the maximization above: the left one represents the case when both the benefit and the cost are constant, the right one corresponds to the case when both of them are linear. In the middle, there is a mixture of these two. Note that we needed to restrict z to be under a certain amount as it represents both the time and the size of a meeting.

6 Pandemic Mechanism Design

Pandemic response is a complex affair. The two games described above model only parts of the bigger picture.

6.1 The government as mechanism designer

We refer to the collection (and interplay) of measures implemented by a specific government fighting the epidemic in their respective country as *mechanism*. Consequently, decisions made with regard to this mechanism constitutes *mechanism design* [17]. In its broader interpretation, mechanism design theory seeks to study mechanisms achieving a particular preferred outcome. Desirable outcomes are usually optimal either from a social aspect or maximizing a different objective function of the designer.

In the context of the corona pandemic, the immediate response mechanism is composed of e.g., wearing a mask, social distancing, testing and contact tracing, among others. Note that this is not an exhaustive list: financial aid, creating extra jobs to accommodate people who have just lost their jobs, declaring a national emergency and many other conceptual vessels can be utilized as sub-mechanisms by the mechanism designer, i.e., usually, the government; we do not discuss all of these in detail due to the lack of space. Instead, we shed light on how government policy can affect the sub-mechanisms, how sub-mechanisms can affect each other and, finally, the outcome of the mechanism itself. We illustrate the importance of mechanism design applying different policies to our two games, and adding testing and contact tracing to the mix.

6.2 Policy impact on sub-mechanisms and the final mechanism

Here we analyze the impact of commonly seen policies: compulsory mask wearing, distributing free masks, limiting the amount of people gathering and total lock-down.

Compulsory mask wearing and free masks. If the government declares that wearing a simple mask is mandatory in public spaces (such as shops, mass transit, etc.), it can enforce an outcome (**out**,**out**) that is indeed socially better than the NE. The resulting strategy profile is still not SO, but it i) allocates costs equally among citizens; ii) works well under the uncertainty of one's health status; and iii) may decrease the first-order need for large-scale testing, which in turn reduces the response cost of the government. By distributing free masks, the government can reduce the effect of selfishness and, potentially, help citizens who cannot buy or afford masks owing to supply shortage or unemployment.

Limiting the amount of people gathering and total lock-down. If the government imposes an upper limit l for the size of congregations, this will put a strict upper bound on the "optimal meeting size" g^*, and the resulting group size will be $\min(l, g^*)$. Note that if $l < g^*$ then it creates an "opportunity" for longer meetings (larger t), as Equation (5) maximizes for $z = gt$. On the other hand, if the chosen restrictive measure is a total lock-down, both the Distancing Game and the Mask Game are rendered moot, as people are not allowed to leave their households.

Testing and contact tracing. It is clear that the Distancing and the Mask Games are not played in isolation: people deciding to meet up

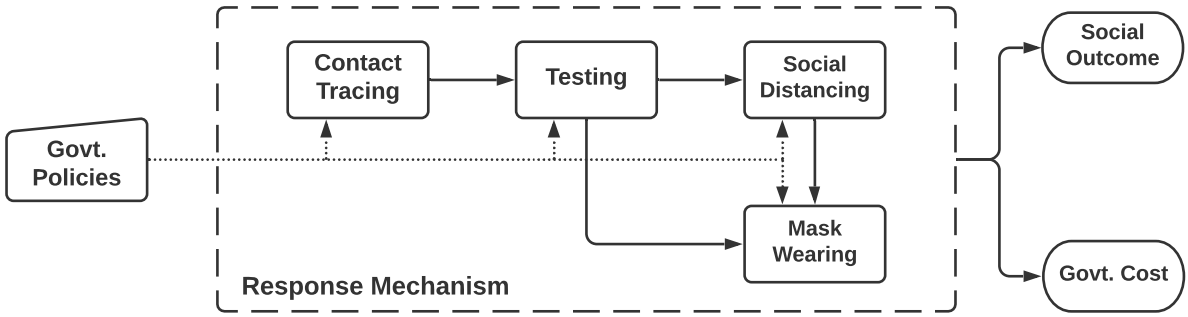


Figure 2: Pandemic response mechanism as influenced by government policy (dotted lines) and the interplay of sub-mechanisms (solid lines)

invoke the decision situation on mask wearing. On the other hand, so far we have largely ignored two other widespread pandemic response measures: testing and contact tracing.

With appropriately designed and administered coronavirus tests, medical personnel can determine two distinct features of the tested individual: i) whether she is actively infected spreading the virus and ii) whether she has already had the virus, even if there were no or weak symptoms. (Note that detecting these two features require different types of tests, able to show the presence of either the virus RNA or specific antibodies, respectively.) In general, testing enables both the tested person and the authorities to make more informed decisions. Putting this into the context of our games, testing i) reduces the uncertainty in Bayesian decision making, and ii) enables the government to impose mandatory quarantine thereby removing infected players. Even more impactful, mandatory testing (as in Wuhan[10]) completely eliminates the Bayesian aspect, essentially rendering the situation to a full information game: it serves as an exogenous "health oracle" imposing no monetary cost on the players. To sum it up, the testing sub-mechanism outputs results that serve as inputs to both the Distancing and the Mask Game.

Naturally, a "health oracle" does not exist: someone has to bear the costs of testing. From the government's perspective, mandatory mass testing is extremely expensive[11]. (Similarly, from the concerned individual's perspective, a single test could be unaffordable.) Contact tracing, whether traditional or mobile app-based, serves as an important input sub-mechanism to testing [10]. It identifies the individuals who are *likely* affected based on spatial proximity, and inform both them and the authorities about this fact. In game-theoretic terms, for such players, the benefit of testing outweigh the cost (per capita) with high probability. From the mechanism designer's point of view, contact tracing reduces the overall testing cost by enabling *targeted testing*, potentially by orders of magnitude, without sacrificing proper control of the pandemic. Another potential cost of contact tracing for individuals could be the loss of privacy. Note that mobile OS manufacturers are working on integrating privacy-preserving contact tracing into their platform to eliminate adoption costs for installing an app[12].

The big picture. As far as pandemic response goes, the mechanism designer has the power to design and parametrize the games that citizens are playing, taking into account that sub-mechanisms affect each other. After games have been played and outcomes have been determined, the cost for the mechanism designer itself are realized (see Figure 2). This cost function is very complex incorporating factors from ICU beds through civil unrest to a drop in GDP over multiple time scales [18]. Therefore, governments have to carefully balance the – very directly interpreted – social optimum and their own costs; this indeed requires a mechanism design mindset.

7 Conclusion

In this paper we have made a case for treating pandemic response as a mechanism design problem. Through simple games modeling interacting selfish individuals we have shown that it is necessary to take individual incentives into account during a pandemic. We have also demonstrated that specific government policies significantly influence the outcome of these games, and how different response measures (sub-mechanisms) are interdependent. As an example we have discussed how contact tracing enables targeted testing which in turn reduces the uncertainty from individual decision making regarding social distancing and wearing masks. Governments have significantly more power than traditional mechanism designers in distributed systems; therefore it is even more crucial for them to carefully study the tradeoff between social good and the cost of the designer when implementing their pandemic response mechanism.

Limitations and future work. Clearly, we have just scratched the surface of pandemic mechanism design. The models presented are simple and mostly used for demonstrative purposes. Also, the mechanism design considerations are only quasi-quantitative without proper formal mathematical treatment. In turn, this gives us plenty of opportunity for future work. A potential avenue is extending our models to capture the temporal aspect, combine them with epidemic models as games played by many agents on social graphs, and parametrize them with real data from the ongoing pandemic (policy changes, mobility data, price fluctuations, etc.). Relaxing

[10] New York Times. https://www.nytimes.com/2020/05/26/world/asia/coronavirus-wuhan-tests.html

[11] But not without precedence, e.g., in Slovakia (https://edition.cnn.com/world/live-news/coronavirus-pandemic-10-18-20-intl/h_beb93495fe9b83701023eafd5f28e39d)

[12] Apple. https://covid19.apple.com/contacttracing

the rational decision-making aspect is another prominent direction: behavioral modeling with respect to obedience, other-regarding preferences and risk-taking could be incorporated into the games. Finally, a formal treatment of the mechanism design problem constitutes important future work, incorporating hierarchical designers (WHO, EU, nations, municipality, household), an elaborate cost model, and analyzing optimal policies for different time horizons. If done with care, these steps would help create an extensible mechanism design framework that can aid decision makers in pandemic response.

Acknowledgements

This work was supported by the National Research, Development and Innovation Fund of Hungary in the frame of FIEK_16-1-2016-0007 (Higher Education and Industrial Cooperation Center) project.

References

[1] Anupam Kumar Bairagi, Mehedi Masud, Do Hyeon Kim, Md Munir, Abdullah Al Nahid, Sarder Fakhrul Abedin, Kazi Masudul Alam, Sujit Biswas, Sultan S Alshamrani, Zhu Han, et al. 2020. Controlling the Outbreak of COVID-19: A Noncooperative Game Perspective. *arXiv preprint arXiv:2007.13305* (2020).

[2] Chris T Bauch and David JD Earn. 2004. Vaccination and the theory of games. *Proceedings of the National Academy of Sciences* 101, 36 (2004), 13391–13394.

[3] Samit Bhattacharyya and Chris T Bauch. 2011. "Wait and see" vaccinating behaviour during a pandemic: A game theoretic analysis. *Vaccine* 29, 33 (2011), 5519–5525.

[4] Timoteo Carletti, Duccio Fanelli, and Francesco Piazza. 2020. COVID-19: The unreasonable effectiveness of simple models. *arXiv preprint arXiv:2005.11085* (2020).

[5] Sheryl L Chang, Mahendra Piraveenan, Philippa Pattison, and Mikhail Prokopenko. 2019. Game theoretic modelling of infectious disease dynamics and intervention methods: a mini-review. *arXiv preprint arXiv:1901.04143* (2019).

[6] Sung-Po Chao. 2020. Simplified model on the timing of easing the lockdown. *arXiv preprint arXiv:2007.14072* (2020).

[7] Xiaowei Chen, Wing Fung Chong, Runhuan Feng, and Linfeng Zhang. 2020. Pandemic risk management: resources contingency planning and allocation. (2020).

[8] O Diekmann and JAP Heesterbeek. 2000. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Vol. 5. John Wiley & Sons.

[9] Ceyhun Eksin, Jeff S Shamma, and Joshua S Weitz. 2017. Disease dynamics in a stochastic network game: a little empathy goes a long way in averting outbreaks. *Scientific reports* 7 (2017), 44122.

[10] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 368, 6491 (2020).

[11] Min W Fong, Huizhi Gao, Jessica Y Wong, Jingyi Xiao, Eunice YC Shiu, Sukhyun Ryu, and Benjamin J Cowling. 2020. Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—social distancing measures. *Emerging infectious diseases* 26, 5 (2020), 976.

[12] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. 2020. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine* (2020), 1–6.

[13] John C Harsanyi, Reinhard Selten, et al. 1988. A general theory of equilibrium selection in games. *MIT Press Books* (1988).

[14] Hyokyoung G Hong and Yi Li. 2020. Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic. *PloS one* 15, 7 (2020), e0236464.

[15] MHR Khouzani, Soumya Sen, and Ness B Shroff. 2013. Incentive analysis of bidirectional threat filtering in the internet. In *Workshop on Economics of Information Security*. Citeseer.

[16] A-R Lagos, I Kordonis, and GP Papavassilopoulos. 2020. Games of Social Distancing during an Epidemic: Local vs Statistical Information. *arXiv preprint arXiv:2007.05185* (2020).

[17] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

[18] M McDonald, KB Scott, WJ Edmunds, P Beutels, and RD Smith. 2008. The macroeconomic costs of a global influenza pandemic. In *Global Trade Analysis Project 11th Annual Conference on Global Economic Analysis, Future of Global Economy, Helsinki, June*.

[19] Donald K Milton, M Patricia Fabian, Benjamin J Cowling, Michael L Grantham, and James J McDevitt. 2013. Influenza virus aerosols in human exhaled breath: particle size, culturability, and effect of surgical masks. *PLoS Pathog* 9, 3 (2013), e1003205.

[20] Dimitris I Papadopoulos, Ivo Donkov, Konstantinos Charitopoulos, and Samuel Bishara. 2020. The impact of lockdown measures on COVID-19: a worldwide comparison. *medRxiv* (2020).

[21] Piero Poletti, Marco Ajelli, and Stefano Merler. 2012. Risk perception and effectiveness of uncoordinated behavioral responses in an emerging epidemic. *Mathematical Biosciences* 238, 2 (2012), 80–89.

[22] Timothy C Reluga. 2010. Game theory of social distancing in response to an epidemic. *PLoS computational biology* 6, 5 (2010).

[23] KC Santosh. 2020. COVID-19 Prediction Models and Unexploited Data. *Journal of medical systems* 44, 9 (2020), 1–4.

[24] Petrônio CL Silva, Paulo VC Batista, Hélder S Lima, Marcos A Alves, Frederico G Guimarães, and Rodrigo CP Silva. 2020. COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals* (2020), 110088.

[25] Peng Sun, Liu Yang, and Francis De Véricourt. 2009. Selfish drug allocation for containing an international influenza pandemic at the onset. *Operations Research* 57, 6 (2009), 1320–1332.

[26] Polly Trottenberg and RS Rivkin. 2013. Guidance on treatment of the economic value of a statistical life in US Department of Transportation analyses. *Revised departmental guidance, US Department of Transportation* (2013).

[27] Michiel van Boven, Don Klinkenberg, Ido Pen, Franz J Weissing, and Hans Heesterbeek. 2008. Self-interest versus group-interest in antiviral control. *PLoS One* 3, 2 (2008), e1558.

[28] Yuzhen Zhang, Bin Jiang, Jiamin Yuan, and Yanyun Tao. 2020. The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: A data-driven SEIQR model study. *medRxiv* (2020).

[29] Shi Zhao, Chris T Bauch, and Daihai He. 2018. Strategic decision making about travel during disease outbreaks: a game theoretical approach. *Journal of The Royal Society Interface* 15, 146 (2018), 20180515.

On Improving Toll Accuracy for COVID-like Epidemics in Underserved Communities Using User-generated Data

Hamada A. Aboubakr[a] Amr Magdy[b,c]
[a] Department of Veterinary Population Medicine, CVM
University of Minnesota - Twin Cities
[b]Department of Computer Science and Engineering [c]Center for Geospatial Sciences
University of California, Riverside
aboub006@umn.edu amr@cs.ucr.edu

ABSTRACT

This paper envisions using user-generated data as a cheap way to improve accuracy of epidemic tolls in underserved communities. The global widespread of COVID-19 pandemic has imposed several unprecedented challenges. One of these challenges is constantly monitoring the unprecedented epidemic widespread at a fine-granular spatial scale, so experts can model, understand, and prevent disease transmission and field personnel can reach and treat infected people. Unfortunately, the limited resources compared to the pandemic widespread has led to a significant number of unreported cases in underserved communities and developing countries, including a large number of severe cases.

We propose in this paper enhancing epidemic case reporting in underserved communities through exploiting the power of data that are posted by people on web. Our vision is building a data analysis pipeline that filters and categories use-generated data objects to provide informal estimates for tolls in unreachable regions and enhance estimates in other regions. The pipeline consist of five stages, that starts with filtering epidemic-specific data to visualize advanced aggregates to end users. We also discuss several technical challenges that face different stages of the pipeline.

CCS CONCEPTS

• **Information systems → Information systems applications**; *Information integration.*

KEYWORDS

COVID, user-generated data, big data, query processing

ACM Reference Format:
Hamada A. Aboubakr[a] Amr Magdy[b,c]. 2020. On Improving Toll Accuracy for COVID-like Epidemics in Underserved Communities Using User-generated Data. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3423459.3430758

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430758

1 INTRODUCTION

The global widespread of COVID-19 pandemic has clearly introduced unprecedented challenges to humanity at different fronts. In the front line of these challenges are the health-related challenges, including reaching out and providing appropriate medical care to infected people. However, this pandemic has a global widespread almost in every country, province, and village worldwide, which makes monitoring it a tremendously difficult task. With the limited resources, the health systems have to prioritize patients for care based on different factors [7, 14–16]. Unfortunately, the underserved communities, e.g., rural areas and slums in developed countries or small cities and villages in developing countries, are highly impacted by the consequences of this pandemic relative to other communities. This is because of their higher exposure to the causes of infection and their limited access to COVID testing and equipped medical care facilities [5, 10–12, 25]. Furthermore, failure to monitor and report cases is a growing concern particularly in developing countries because of the limited public health infrastructure, the weak health systems, insufficient laboratory capacity of diagnostic testing, and the poor surveillance systems for diseases [1, 18]. Therefore, the number of reported infections and deaths in underserved communities does not reflect the actual numbers almost everywhere [3, 23]. This leads to a very high cost in lives. For example, as of September 2020, more than 75% of children who have died of COVID-19 in the U.S. are minorities, though they account for just 41% of the overall youth population [28].

To improve access to underserved communities, we propose to use the power of people to mitigate reporting inaccuracy. The main idea is using user-generated data that flows on web around the clock to extract related information that helps in improving epidemic reporting to health officials. Such mitigation will have a great impact as it will enable reaching currently inaccessible cases. This helps health officials to provide appropriate medical care, surround infection foci, and control the situation faster especially in underserved communities that are highly impacted with limited reporting means and highly infectious environments.

Existing work on coronavirus-related social media data puts a particular focus on controlling spread of misinformation that are related to the pandemic symptoms, transmission modes, and other misleading information that could harm people's health [2, 4, 8, 9, 13, 19, 20, 22, 24]. Although this is a crucially important problem to address, it deals with extracting harmful information from user-generated data to prevent the negative aspects of spreading misinformation. On the contrary, our work deals with user-generated

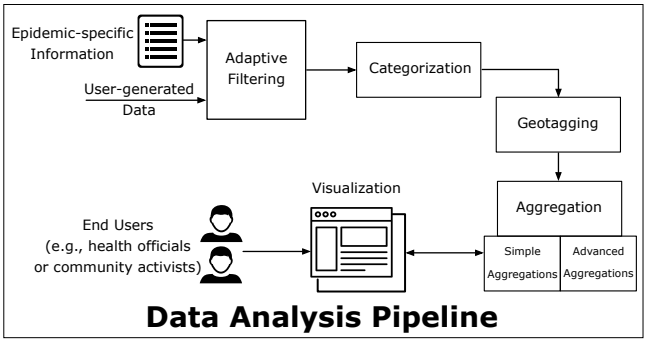


Figure 1: Proposed Pipeline Architecture

data positively as a source of important information that could help health experts. This is also related to orthogonal efforts that deal positively with coronavirus-related user-generated data, including sentiment analysis [27], integrating with IoT data [26], and modelling transmission [21].

The proposed data analysis pipeline consists of five main stages: adaptive filtering, categorization, geotagging, aggregation, and visualization. Each of these stages has different issues related to either lingual dependency, processing streams, or granularity. The rest of this paper outlines each analysis stage, discussing the technical issues and their implications.

2 DATA ANALYSIS PIPELINE

This section outlines the proposed data analysis pipeline. Figure 1 shows the proposed pipeline architecture. The pipeline consists of five ordered stages, namely, *adaptive filtering*, *categorization*, *geotagging*, *aggregation*, and *visualization*. The stages work in a sequential order, where the output of each stage is an input to the following stage. The first stage takes the input data and epidemic-specific information, while the last stage output visualized aggregates for epidemic cases grouped by spatial locations and temporal intervals. The main functionalities and distinguishing characteristics of each stage are briefly outlined below.

(1) **Adaptive filtering**: This stage takes two inputs: (a) A static dataset or a dynamic data stream of user-generated objects, e.g., tweets, posts, comments, or fusion of them. (b) Epidemic-specific characteristics; a set of seed keywords, optional locations of interest, and optional times of interest. Using the two inputs, an adaptive filter is employed to filter out any data object that does not satisfy the epidemic characteristic. Therefore, any data object that does not contain any of the keywords, lies outside the areas of interest, or posted outside the times of interest will not be considered for further processing. When neither locations nor times are provided, all locations and times are considered relevant, e.g., all locations are relevant for the global COVID-19 pandemic. However, this filter should be adaptive in terms of improving the filtering keywords while the filtration process goes on. To this end, when a relevant data is found based on the seed keywords, the adaptive filter should keep all other words of this data except stop words. Over time, the filter will discover more keywords that identify epidemic-related data adaptively, either by using frequent words or other keyword identification methods. This adaptation should also consider the type of input dataset, as static datasets are easier to discover new keywords compared to dynamic data streams.

(2) **Categorization**: This stage takes the set of relevant data objects, that are output of the first stage, to categorize them based on the epidemic case statuses. For example, for COVID-19 pandemic, three potential case categories are: a death case, a mild infection case, and a severe infection case. Such categorization is epidemic-specific in terms of number of categories and how to identify each category. One way is keyword-based categorization, where each category is defined by a set of keywords and the object is assigned based on the corresponding keywords. This way can be performed jointly with the adaptive filtering stage where the list of filtering keywords are categorized into multiple categories, or separately based on different keyword sets. Another way is using machine learning techniques that have shown effectiveness in document classification. Regardless the categorization method, data in each category will be used in the aggregation stage for improving miscounting accuracy.

(3) **Geotagging**: Another piece of information that is needed in data aggregation is the geographical location of each data object to map the epidemic case to a corresponding city, district, or village. Despite the widespread of mobile devices and mobile users of online platforms, automatic geotagging is still a limitation where majority of data comes either with very coarse spatial granularity or without any spatial information. A main reason is legal privacy concerns, where user-generated data platforms disable automatic geotagging by default to protect personal privacy and avoid legal problems. To overcome this limitation, this stage analyzes the data object's content and metadata to assign a primary relevant location. Geotagging has been studied in the literature for different settings and performance trade-off, including for short posts, long posts, etc. Among the recent work is [17] that uses deep learning to geotag tweets of any language. This type of work is the most relevant for user-generated data of epidemic analysis due to high percentage of short posts and popularity in different languages. This is also related to the cross-lingual issues that will be discussed in Section 3.

(4) **Aggregation**: After processing over the first three stages, the output data objects are ready to be aggregated into corresponding locations and time intervals. This spatio-temporal aggregation stage represents the main counting and analysis stage. Locations could be attached from the original data source or resulted from the geotagging stage. The object timestamp is attached from the original data source in majority of platforms. The aggregation could be either a simple counting aggregation grouped by location and time for all places and times, or advanced aggregation for a specific place or certain time intervals. We outline our vision for both below.

Simple aggregations. The simple spatio-temporal aggregation stage sums up data objects counts based on user-defined hierarchies for both spatial and temporal dimensions. For the spatial dimension, end users, e.g., health officials or activists, might, for example, define a hierarchy of <*city, county, state*> to count different categories of epidemic cases for each provided city, county, and state within the USA. Users should be able to control defining this hierarchy based on the needs and the different administrative region divisions around the world. Also, the provided regions are not necessarily

to be of predefined borders, but could be arbitrary, e.g., output of a regionalization algorithm, to enable exploring areas based different attributes, e.g., economic level, population density, or environmental factors. In all cases, the attached location information to data objects affects the count accuracy for this user-defined spatial hierarchy. For example, if the attached information provides city-level locations but not district-level, any hierarchy that includes districts will suffer form low counting accuracy. This is discussed among the technical issues in Section 3.

Unlike the spatial dimension, the temporal dimension is more deterministic and has clearer aggregation options, and in turn less issues. Users can still define temporal hierarchy for aggregations. For example, the user can define <*day,week,month*> hierarchy to count cases for each day, week, and month in each city, county or state. Unlike spatial information, attached temporal information are provided in fine granularity, e.g., second-level granularity, in majority of platforms. This makes temporal aggregation easier and of much better accuracy. By default, each level of the temporal hierarchy assumes disjoint time intervals, e.g., disjoint days, for ease of use due to popularity of this temporal aggregation model. However, users should be also able to define temporal hierarchies of overlapping time intervals. For example, if the analysis is performed on a continuous data stream, a sliding window of three days will be of interest for health officials to monitor in different places, which is by definition a set of overlapping time intervals. This could be also applicable to static datasets in certain analysis scenarios. So, allowing overlapping time intervals will be a useful aggregation feature to support.

Advanced aggregations. Beyond the simple count aggregations for all levels of spatial and temporal hierarchies, end users will be interested in more advanced aggregations that better show the situation in specific places and at certain times. For example, when Southern California appears as a region with high number of cases on the epidemic map, health officials will be interested in producing advanced aggregates for Southern California counties that show the absolute and relative increase in number of cases over the past seven days. Another example is finding areas that have the highest rate of increase over the past three days to mitigate most vulnerable regions. We can discuss endless examples that combine spatial, temporal, and counts in an advanced way to show a different information or insight. It is important to identify the most important blocks that are used in such advanced analysis based on the need of domain experts, making use of existing analysis frameworks as data analysis infrastructures.

(5) **Visualization**: The last stage is visualizing both simple and advanced spatio-temporal aggregates to end users, e.g., health officials or leading community activists, to enable them making use of these counts effectively. This stage should make use of the existing rich literature of visualization frameworks, such as UCI Cloudberry [6], to provide low-effort and effective visualization. Obviously, a geographical map will be an essential element in such visualization. Domain experts should be involved in collecting requirements for all needed visualization features, so they are effective for them as end users. For this context, fundamental visualization elements that should be supported are heatmaps that are either based on administrative borders or cross-borders, hover display boxes that shows cases counts in each spatial entity, filters that allow fragmenting the data based on location and time, and filters that allow fragmenting based on other important attributes such as case category, e.g., either death, mild infection, or severe infection of COVID-19. In addition, the traditional pan, zoom, linking, and brushing features of interactive geovisualization should be supported to enable effective display and exploration for both simple and advanced aggregates.

3 DISCUSSIONS

This section discusses some technical issues that should be addressed while developing the proposed data analysis pipeline. We discuss issues of *language dependency*, *real-time streams*, *granularity*, and *multi-locatable objects*.

Language dependency. One of the main challenges in supporting underserved communities for epidemic data applications is the language issue. Obviously, the language and its usage is highly variant from one underserved community to another, depending on the country and even the locality within that country. Orthogonal from differences in languages among countries, it is known that dialects could be very different within different parts of the same country. This issue affects the first three stages of our pipeline, adaptive filtering, categorization, and geotagging. Addressing this issue could take one of two forms. The first way is tailoring the developed pipeline for a certain underserved community, and hence use its specific language and dialects as input to process. This means tailoring the filtering and categorization keywords and using language-specific geotagging tool, e.g., place ontology. The alternative way is training machine learning models that uses blended datasets of several languages to adapt for a multi-lingual setting. This approach is used in the literature for different tasks. For example, the work in [17] uses this approach for cross-lingual geotagging.

Real-time streams. When data analysis is performed on a dynamic data stream that continuously receive data objects around the clock, different aspects of data analysis change including data storage, processing schemes, and query models. This has triggered the whole literature of streaming data management that is active for a couple of decades. For our proposed data analysis pipeline, analyzing streaming data will have impact on all stages. The least affected stage is geotagging as many of existing geotagging methods depend solely on the data object's content and metadata, without much dependency on previous or upcoming objects. A main problem for this stage will be geotagging efficiency in real time, however, using fast machine learning models could solve this problem [17]. For other stages, the impact is clearer. The filtering phase will be a driver stage as it will help to significantly reduce the streaming data size and output only relevant data objects, so the number of objects to be processed by the following stages are much smaller in size. This will eliminate one of the major overhead in stream processing, which is excessive data size. The other major overhead, which is incremental data processing, will clearly affect the other three stages, categorization, aggregation, and visualization. In categorization, incremental document classification has to be incorporated. If keyword-based categorization is employed, incremental processing will be straightforward to incorporate. Machine learning based categorization will be more challenging to handle.

Although several existing techniques handle this setting, the result accuracy is expected to be lower compared to static datasets. The incremental aggregation will be easier and less impacted by the streaming nature. The reason is that all our aggregations depend on counting, which is easy to maintain incrementally. The visualization will consume the aggregation results as is. However, incremental results updates will need to be visualized incrementally to end users. However, in case of epidemics, even hourly updates are considered fast enough for most of the cases, and this can be adequately served by existing visualization platforms.

Granularity. At different stage of the proposed pipeline, granularity plays a role in trading off usability and processing overhead of analyzing user-generated data. For example, adaptive filtering could classify objects as *relevant* or *irrelevant* and output one type of relevant objects. It could also filter at a finer granularity and further classifies relevant objects into further types to distinguish epidemic-specific cases. This is clearer in the categorization stage that can provide coarse-granular or fine-granular categories with a wide variety of options. Finer granularity levels will provide better accuracy and more information, but it will come with further processing requirements. Granularity is also a trade off for geotagging, where accurate point geotagging consumes much larger processing overhead, while city-level or province-level geotagging is much faster. The granularity of aggregation over space and time will also introduce the same trade off, but it will add a storage trade off as well to decide how much data to store. In general, granularity is a cross-stage issue to consider while designing and developing the proposed pipeline, and it should consider the trade off between available computing resources and required functionality.

Multi-locatable objects. Some data objects might be attachable to multiple locations. Examples for sources of such phenomenon are location ambiguity, e.g., Alexandria is a city name in different countries, mentioning multiple locations either within the content or in both content and metadata, e.g., the user profile shows a city in USA and the post is about a city in India. Regardless the source of multiple locations, this represents a challenge as we cannot assume the case is replicated in multiple physical places unless the locations are nested, e.g., California and USA. However, for the general case where the attachable locations are different, it is essential to promote one of them as the primary location to be used in further analysis. Location selection could be rule-based or certainty-based. Rule-based location selection will apply some heuristic rules to promote the most probable location, e.g., favoring the content words over the user profile location or favoring the earliest mentioned location. Certainty-based location selection will depend on assigning a probability to each potential location, either based on a probabilistic model or a multi-class classifier. Then, locations can be considered or neglected based on these probabilistic values. This certainty-based model opens the door to consider more than one location in the aggregation by introducing uncertain query processing. However, we believe that might be confusing for non-expert end users. Another option is to consider all uncertain locations to contribute partially while distinguishing them from certain locations in both aggregation and visualization stages.

REFERENCES

[1] H. Aboubakr and S. Goyal. Involvement of Egyptian Foods in Foodborne Viral Illnesses: The Burden on Public Health and Related Environmental Risk Factors: An Overview. *Food and Environmental Virology*, 11(4):315–339, Sept. 2019.

[2] Social Media Firms Fail to Act on Covid-19 Fake News. https://www.bbc.com/news/technology-52903680, 2020.

[3] Two Charts Estimate the True Scope of the US's Coronavirus Infections and Deaths. https://www.businessinsider.com/us-coronavirus-cases-deaths-real-scale-estimates-charts-2020-7, 2020.

[4] J. S. Brennen, F. Simon, P. N. Howard, and R. K. Nielsen. Types, Sources, and Claims of COVID-19 Misinformation. *Reuters Institute*, 7:1–13, 2020.

[5] Health Equity Considerations and Racial and Ethnic Minority Groups. https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html, 2020.

[6] Cloudberry Big Data Visualization. http://cloudberry.ics.uci.edu/, 2020.

[7] P. D. N. et. al. Multi-Criteria Decision Analysis to Prioritize Hospital Admission of Patients Affected by COVID-19 in Low-resource Settings with Hospital-bed Shortage. *International Journal of Infectious Diseases*, 98:494–500, Sept. 2020.

[8] SOCIAL MEDIA STRUGGLES WITH CORONAVIRUS MISINFORMATION. https://www.rpc.senate.gov/policy-papers/social-media-struggles-with-coronavirus-misinformation-, 2020.

[9] Battling the pandemic of misinformation. https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods/, 2020.

[10] Coronavirus infection by race: What's behind the health disparities? https://www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-infection-by-race/faq-20488802, 2020.

[11] HHS Initiatives to Address the Disparate Impact of COVID-19 on Ethnic Minorities. https://www.hhs.gov/sites/default/files/hhs-fact-sheet-addressing-disparities-in-covid-19-impact-on-minorities.pdf, 2020.

[12] The Impact on Underserved Communities in Times of Crisis. https://www.himss.org/resources/impact-underserved-communities-times-crisis, 2020.

[13] A. N. Islam, S. Laato, S. Talukder, and E. Sutinen. Misinformation Sharing and Social Media Fatigue During COVID-19: An Affordance and Cognitive Load Perspective. *Technological Forecasting and Social Change*, 159:120201, Oct. 2020.

[14] California Sets Guidelines on Which Patients Are Prioritized if Hospitals Overwhelmed by Coronavirus. https://www.latimes.com/california/story/2020-04-21/california-healthcare-guidelines-shortages-coronavirus-treatment, 2020.

[15] Coronavirus: Italy Doctors Forced to Prioritise ICU Care for Patients with Best Chance of Survival. https://www.euronews.com/2020/03/12/coronavirus-italy-doctors-forced-to-prioritise-icu-care-for-patients-with-best-chance-of-s, 2020.

[16] The Heart-wrenching Choice of Who Lives and Dies. https://www.bbc.com/future/article/20200428-coronavirus-how-doctors-choose-who-lives-and-dies, 2020.

[17] M. Izbicki, V. Papalexakis, and V. J. Tsotras. Geolocating Tweets in any Language at any Location. In *CIKM*, pages 89–98, 2019.

[18] T. P. Mashamba-Thompson and E. D. Crayton. Blockchain and Artificial Intelligence Technology for Novel Coronavirus Disease-19 Self-Testing. *Diagnostics*, 10(4):198, Apr. 2020.

[19] Quantifying the COVID-19 Misinformation Epidemic on Twitter. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7152572/, 2020.

[20] Hoaxes Are Making Doctors' Jobs Harder. https://www.nytimes.com/2020/08/28/opinion/sunday/coronavirus-misinformation-faceboook.html, 2020.

[21] Z. Peng, R. Wang, L. Liu, and H. Wu. Exploring Urban Spatial Features of COVID-19 Transmission in Wuhan Based on Social Media Data. *ISPRS International Journal of Geo Information*, 9(6):402, 2020.

[22] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, June 2020.

[23] Actual Covid-19 Case Count Could be 6 to 24 Times Higher than Official Estimates. https://www.statnews.com/2020/07/21/cdc-study-actual-covid-19-cases/, 2020.

[24] S. Tasnim, M. M. Hossain, and H. Mazumder. Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, 53(3):171–174, May 2020.

[25] COVID-19 Risks Amplified for underserved Communities. https://www.mobihealthnews.com/news/covid-19-risks-amplified-underserved-communities, 2020.

[26] B. Wang, Y. Sun, T. Q. Duong, L. D. Nguyen, and L. Hanzo. Risk-Aware Identification of Highly Suspected COVID-19 Cases in Social IoT: A Joint Graph Theory and Reinforcement Learning Approach. *IEEE Access*, 8:115655–115661, 2020.

[27] T. Wang, K. Lu, K. Chow, and Q. Zhu. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access*, 8, 2020.

[28] Coronavirus Kills Far More Hispanic and Black Children Than White Youths. https://www.washingtonpost.com/health/2020/09/15/covid-deaths-hispanic-black-children/, 2020.

COVID-19 Risk Estimation using a Time-varying SIR-model

Mehrdad Kiamari
Viterbi School of Engineering,
University of Southern California
Los Angeles, USA
kiamari@usc.edu

Gowri Ramachandran
Viterbi School of Engineering,
University of Southern California
Los Angeles, USA
gsramach@usc.edu

Quynh Nguyen
Viterbi School of Engineering,
University of Southern California
Los Angeles, USA
quynhu@usc.edu

Eva Pereira
Office of the Mayor,
City of Los Angeles
Los Angeles, USA
eva.pereira@lacity.org

Jeanne Holm
Office of the Mayor,
City of Los Angeles
Los Angeles, USA
jeanne.holm@lacity.org

Bhaskar Krishnamachari
Viterbi School of Engineering,
University of Southern California
Los Angeles, USA
bkrishna@usc.edu

ABSTRACT

Policy-makers require data-driven tools to assess the spread of COVID-19 and inform the public of their risk of infection on an ongoing basis. We propose a rigorous hybrid model-and-data-driven approach to risk scoring based on a time-varying SIR epidemic model that ultimately yields a simplified color-coded risk level for each community. The risk score Γ_t that we propose is proportional to the probability of someone currently healthy getting infected in the next 24 hours based on their locality. We show how this risk score can be estimated using another useful metric of infection spread, R_t, the time-varying average reproduction number which indicates the average number of individuals an infected person would infect in turn. The proposed approach also allows for quantification of uncertainty in the estimates of R_t and Γ_t in the form of confidence intervals. Code and data from our effort have been open-sourced and are being applied to assess and communicate the risk of infection in the City and County of Los Angeles.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**.

KEYWORDS

Risk Modelling, COVID-19, SIR model

ACM Reference Format:
Mehrdad Kiamari, Gowri Ramachandran, Quynh Nguyen, Eva Pereira, Jeanne Holm, and Bhaskar Krishnamachari. 2020. COVID-19 Risk Estimation using a Time-varying SIR-model. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3423459.3430759

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430759

1 INTRODUCTION

The ongoing COVID-19 epidemic has forced governments and public authorities to employ stringent measures [6, 10], including closing business and implementing stay-at-home orders, to contain the spread. When making such decisions, policymakers require tools to understand in "real-time" how the virus is spreading in the community, as well as tools to help communicate the level of risk to citizens so that they can be encouraged to take appropriate measures and take the public health directives seriously.

One metric that has been found to be useful for authorities to assess the level of containment over time is the effective reproduction number [7]. The effective reproduction number, R_t, indicates on average how many currently susceptible persons can be infected by a currently infected individual. The epidemic grows if this measure is above one. It is desirable to keep this value as far below one as possible over time in order to contain and eventually, hopefully, eliminate the virus from the community.

While R_t is meaningful to understand the rate at which the epidemic is spreading and has been proposed previously (for example, see https://rt.live/), what has been missing in the public discourse is a risk metric that is more suitable for communication to a wider public. One key requirement for such a metric is that it be something that a citizen could relate to on an individual basis. Another requirement is that it needs to be easy to communicate to a wide audience. We address both these requirements in this work and make the following contributions.

First, we obtain the daily effective reproduction number R_t of a time-varying SIR model as well as the corresponding confidence Interval. The confidence interval reflects uncertainty in both the parameter of the underlying model and uncertainty in the data itself. Further, we present the mathematical derivation of the distribution of R_t.

Second, we propose a novel risk score Γ_t for a community that is proportional to the probability that an individual will get infected in the next 24 hours. We show that the risk score can be calculated given estimates of four quantities: a) an estimate of $I_{rep,new}(t)$, the most recently reported count of new confirmed infectious cases, b) an estimate of R_t as discussed above, c) an estimate of K, the ratio of true infectious cases to the number of confirmed cases, and d) an estimate of $S(t)$, the current number of susceptible individuals in the community. To make the score more meaningful, we normalize

the probability of infection by multiplying it by 10,000. Then, a risk score of x is an indication that there is, on average, a chance of x in 10,000 of an individual in the community becoming infected in the next 24 hours. We also propose to convert the numerical risk score, which has an intuitive meaning as indicated above, to a color-coded risk level based on suitably chosen thresholds[1]. We propose the use of four color-levels to indicate the corresponding risk level from very low to high: green, yellow, orange, and red.

Third, we have implemented software to estimate the risk level for any community and released it as open-source. The code requires only time-series data on confirmed new cases, the population of the community, and an estimate for the ratio of true to confirmed (detected) COVID-19 positive cases. This software is being used at USC to process the daily data of communities within Los Angeles County to estimate and generate maps of risk levels by community. The block diagram in figure 1 illustrates key elements of our system design. Our data parser is able to get the raw data from online data sources, clean them up and store them in machine-friendly (csv and json) formats. Our code for infection risk calculation uses this data in conjunction with a time-varying SIR-based Bayesian mathematical model to obtain risk estimates and prediction for different communities. The results are provided in CSV format and can be used to generate a heatmap-type visualization as well.

The risk scoring model we describe in this work is now being used by the City of Los Angeles, which in turn is working with the County of Los Angeles and other partners to develop a publicly accessible tool that can be used by individuals and communities to grow awareness and mitigate risk of infection. We believe that our risk estimation approach will be similarly of value to other communities around the world.

The rest of the paper is structured as follows: Section 2 reviews the related work. The novel risk calculation methodology is presented in Section 3. Section 4 discusses the implementation and evaluation of the proposed method in Los Angeles County. The key results are discussed in Section 5. Section 6 concludes the paper.

2 RELATED WORK

There have been a few recent works studying different transmission models for COVID-19 such [4] which developed an agent-based model to reproduce the characteristics of COVID-19 transmission or [1] which proposed a mobility-based model to measure COVID-19 growth rate ratio for a given day.

As noted above, the calculation of the risk score requires an estimate of R_t. We show how this can be estimated using a time-varying SIR model, a generalization of the well-known SIR compartmental model [3, 8] which consists of three states, namely the susceptible state, the infected state, and the recovered state. While traditionally this model is assumed to have a interaction rate / infection rate parameter that is constant, one recent work has used a time-varying SIR model to recover the time-varying effective reproduction number [5]. Going beyond that work, we also show how to derive a confidence interval for R_t in this work. Further, the authors of [5] make strong assumptions on the number of susceptible individuals by approximating it as a constant factor of the entire population. This assumption may not be accurate when the number of infected individuals are high compared to the total population of a community; we therefore take a more general approach.

Another recent work by Systrom [9] has presented a Bayesian prediction approach to obtain confidence intervals for R_t. However, Systrom's work builds on [2], where the definition of infection rate R_t is not based on a time-varying contact rate of the SIR model. Instead, their approach estimates infection rate probabilistically based on the number of new cases alone.

We are not aware of prior work that has proposed defining risk for COVID-19 or other epidemics in terms of an individual's probability of infection, which we argue is more meaningful for communicating risk to the public.

3 METHODOLOGY

Compartmental mathematical models for epidemic spreads including the well-known SIR model have been used since the work of Kermack and McKendrick in 1927 [3]. In the SIR model, each member of a given population is in one of three states at any time: susceptible, infectious, recovered. Any individual that is susceptible could become infected with some probability when they come into contact with an infected individual. Any individual that is infectious eventually recovers (in the context of COVID-19 when applying the SIR model, note that the category of recovered individuals will also include removed individuals due to deaths, which could be modeled as a constant fraction of all individuals in this category). In the classical SIR model, the number of susceptible individuals that become infected depends on the rate at which infected and susceptible individuals encounter each other and this rate is assumed to be constant. A well-known parameter in the classical SIR model is called R0, the effective reproductive number, which measures the average number of infections caused by infectious individuals at the beginning of the epidemic.

Time-Varying SIR model and R_t: In our work, we have extended the SIR model to a time-varying model, in which the rate of encounters and infection probability between individuals in the population is assumed to be time-varying. This better reflects the reality of our present epidemic where interventions such as stay-at-home have been put in place and relaxed and various times and compliance with recommendations such as wearing masks and maintaining physical density has also been time-varying. Based on this model, we are able to define and derive a new approach to calculating a time-varying version of the effective reproductive number, which we refer to as R_t.

A particularly innovative aspect of our model is that it is a Bayesian model that allows the incorporation of various sources of uncertainty in the model, including uncertainty in the actual numbers of infected individuals (due to not every infected individual having been tested, as studies [2] have shown), uncertainty in recovery times, and uncertainty in the choice of parameters for de-noising the empirical data. This allows us to generate not only an estimate of R_t, but also quantify confidence in the estimate from a rigorous statistical perspective.

In this section, we elaborate upon the SIR model in detail. The SIR model is one of the simplest and the most well-known epidemic model [3, 8] where each person belongs to one of the following three

[1]Thresholds for categorizing into very low-risk, low-risk, medium-risk, and high-risk levels are set from a medical perspective.

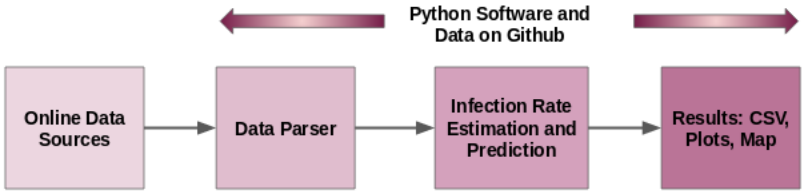


Figure 1: Overview of Our System.

states: the susceptible state, the infected state, and the recovered state. Regarding the susceptible state, individuals have not had the virus yet. However, they may get infected in case of being exposed to an infected individual. As far as the infected state is concerned, a susceptible person has the virus after being exposed to infected individuals. Finally, a person enters the recovered state in case of either the individual gets healed or dies. One important point about this model is that a recovered person will not be a susceptible one anymore. This is how the model is constructed, as in most cases it appears that COVID-19 has an extremely low re-infection rate, at time of writing this paper.

The SIR model follows the following differential equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\beta \frac{S(t)I(t)}{N} \\ \frac{dI(t)}{dt} &= \beta \frac{S(t)I(t)}{N} - \sigma I(t) \\ \frac{dR(t)}{dt} &= \sigma I(t) \end{aligned} \tag{1}$$

where $S(t)$, $I(t)$, and $R(t)$ respectively represent the number of susceptible, infected, and recovered people in a population size of N at time t. Regarding the parameter σ, it is the recovery rate after being infected and is equal to $\frac{1}{D_I}$ where D_I represents the average infectious days. Parameter β is known as the effective contact rate, i.e. the average number of contacts an individual have with others is β.

In analyzing whether any pandemic is contained, it is very crucial to obtain parameter β. We next show that how we can derive β from the aforementioned differential equations.

3.1 Obtaining β_t and R_t for the SIR Model

In the SIR model, we can express the number of susceptible individuals in terms of population size and the number of infected persons as $S(t) \approx N - I(t)$. By replacing $S(t)$ with $N - I(t)$ in the second differential equation of (1), we would have

$$\frac{dI(t)}{dt} = \beta \frac{\Big(N - I(t)\Big)I(t)}{N} - \sigma I(t). \tag{2}$$

We can rewrite (2) as follows:

$$\frac{dI(t)}{(\beta - \sigma)I(t) - \frac{\beta}{N} I^2(t)} = dt. \tag{3}$$

By taking definite integral from time t_1 to t_2 and assuming β to be constant in this time interval, we would have

$$\int_{t_1}^{t_2} \frac{dI(t)}{(\beta - \sigma)I(t) - \frac{\beta}{N} I^2(t)} = \int_{t_1}^{t_2} dt \tag{4}$$

which leads to

$$\frac{1}{\beta - \sigma}\Big(\log \frac{I(t_2)}{\beta - \sigma - \frac{\beta}{N} I(t_2)} - \log \frac{I(t_1)}{\beta - \sigma - \frac{\beta}{N} I(t_1)}\Big) = t_2 - t_1 \tag{5}$$

One can easily check (5) has a unique solution for β due to the fact that term $\frac{1}{\beta - \sigma}$ and log term have monotonic behaviors.

An epidemic happens in case of increase in the number of infected individuals, i.e. $\frac{dI(t)}{dt} > 0$, or consequently

$$\beta \frac{\Big(N - I(t)\Big)I(t)}{N} - \sigma I(t) > 0. \tag{6}$$

In the early stage of an epidemic, almost everyone are susceptible except very few cases. Therefore, $N - I(t) \approx N$ and as a result, condition (6) would turn into $\frac{\beta}{\sigma} > 1$.

The variable $R \triangleq \frac{\beta}{\sigma}$ is defined as the *effective reproduction number*. It is a useful metric to determine epidemic growth. In case of having $R > 1$, the epidemic is growing exponentially while $R < 1$ indicates the epidemic is contained and will decline and die out eventually.

For discrete-time cases such as daily reporting on number of infected cases, the time-variant effective contact rate β_t, which represents the contact rate for time slot t can be derived by solving the following equation:

$$\frac{1}{\beta_t - \sigma}\Big(\log \frac{I(t+1)}{\beta_t - \sigma - \frac{\beta_t}{N} I(t+1)} - \log \frac{I(t)}{\beta_t - \sigma - \frac{\beta_t}{N} I(t)}\Big) = 1 \ \forall t. \tag{7}$$

Therefore, the time-variant effective reproduction number would be defined as $R_t \triangleq \frac{\beta_t}{\sigma}$. Since it is difficult to write a closed form solution for β_t in (7), we take a simpler approximation to β_t by considering the following which is based on (2)

$$\beta_t \approx \frac{\sigma I(t) + \Big(I(t+1) - I(t)\Big)}{\big(1 - \frac{I(t)}{N}\big)I(t)}. \tag{8}$$

Then, we estimate R_t as $\frac{\beta_t}{\sigma}$.

3.2 Obtaining the Confidence Interval for R_t

Since there is uncertainty about parameter D_I (or equivalently σ) and the number of infected cases $I(t)$, we now provide the derivation of confidence interval for parameter R_t. Regarding modeling

the ambiguity in the number of the infected cases, we present the uncertainty about the actual number of infected cases as a factor of reported ones, i.e. $I_{rep}(t) \triangleq \frac{1}{K} I(t)$, and K is a constant greater than 1. The main intuition behind this factor is due to taking into account the following two phenomena, namely lack of sufficient number of tests (specially in the beginning of the pandemic) and asymptomatic cases (mild infections which might not even be noticed). To derive the confidence interval, we need to first find the marginal distribution of R_t. By considering $f_D(d)$ and $f_K(k)$ as the probability distribution function (pdf) for parameters D_I and K, respectively, the joint pdf of these parameters would be

$$f_{D,K}(d,k) = f_D(d) f_K(k) \tag{9}$$

due to the independence of D_I and K. We can derive the probability distribution function of R_t by performing the following transformation on parameters D_I and K and introducing auxiliary variable Z:

$$Z \triangleq K\ ,\ R_t \quad = \frac{1}{1 - \frac{K I_{rep}(t)}{N}} \Big(1 + D_I \frac{I_{rep}(t+1) - I_{rep}(t)}{I_{rep}(t)}\Big). \tag{10}$$

Since the transformation of (Z, R_t) to (D_I, K) is one-to-one, we have

$$K = Z\ ,\ D_I \quad = \frac{R_t(1 - Z a_t) - 1}{b_t}, \tag{11}$$

where $a_t \triangleq \frac{I_{rep}(t)}{N}$ and $b_t \triangleq \frac{I_{rep}(t+1) - I_{rep}(t)}{I_{rep}(t)}$, the joint pdf of Z and R_t would be $f_{Z,R_t}(z,r) = |J| f_{D,K}(d,k)$ with

$$J \triangleq \begin{bmatrix} \frac{\partial d}{\partial z} & \frac{\partial d}{\partial r} \\ \frac{\partial k}{\partial z} & \frac{\partial k}{\partial r} \end{bmatrix}. \tag{12}$$

By substituting the corresponding values of parameters and the Jacobian, we have:

$$f_{Z,R_t}(z,r) = |\frac{1 - z a_t}{b_t}| f_D(\frac{r(1 - z a_t) - 1}{b_t}) f_K(z). \tag{13}$$

The marginal pdf of R_t can be obtained by taking integral of (13) over parameter z, i.e.

$$f_{R_t}(r) = \int f_{Z,R_t}(z,r) dz = \int |\frac{1 - z a_t}{b_t}| f_D(\frac{r(1 - z a_t) - 1}{b_t}) f_K(z) dz. \tag{14}$$

Remark 1: Based on statistical experiments, one reasonable assumption regarding the pdf of parameters D_I and K is that both of them have Gaussian distributions. By considering $D_I \sim \mathcal{N}(\mu_D, \sigma_D^2)$ and $K \sim \mathcal{N}(\mu_K, \sigma_K^2)$, the pdf of R_t can be simplified as

$$\begin{aligned} f_{R_t}(r) = &\int_{-\infty}^{\frac{1}{a_t}} (\beta_0 + \beta_1 z) C \sqrt{2\pi\sigma_c^2} \phi_{\mu_c, \sigma_c^2}(z) dz \\ &+ \int_{\frac{1}{a_t}}^{\infty} (-\beta_0 - \beta_1 z) C \sqrt{2\pi\sigma_c^2} \phi_{\mu_c, \sigma_c^2}(z) dz, \end{aligned} \tag{15}$$

where $\phi_{\mu_c, \sigma_c^2}(.)$ indicates the pdf of a normal distribution with mean μ_c and variance σ_c^2 while

$$\begin{aligned} \beta_0 \triangleq \frac{1}{b_t}\ ,\ \beta_1 \triangleq \frac{-a_t}{b_t}, \\ \alpha_0 \triangleq \frac{(\frac{r-1}{b_t} - \mu_D)^2}{2\sigma_D^2} + \frac{\mu_K^2}{2\sigma_K^2}\ ,\ \alpha_1 \triangleq \frac{(-\frac{r a_t}{b_t})(\frac{r-1}{b_t} - \mu_D)}{\sigma_D^2} - \frac{\mu_K}{\sigma_K^2}, \\ \alpha_2 \triangleq \frac{(\frac{r a_t}{b_t})^2}{2\sigma_D^2} + \frac{1}{2\sigma_K^2}\ ,\ \mu_c \triangleq \frac{-\alpha_1}{2\alpha_2}\ ,\ \sigma_c^2 \triangleq \frac{1}{2\alpha_2}\ ,\ C \triangleq \frac{e^{-(\alpha_0 - \frac{\alpha_1}{4\alpha_2})}}{2\pi\sigma_D\sigma_K}. \end{aligned} \tag{16}$$

By taking integral through using change of parameters, (15) can be rewritten as follows

$$\begin{aligned} f_{R_t}(r) = &-2C\beta_1\sigma_c^2 e^{-\frac{(\frac{1}{a_t} - \mu_c)^2}{2\sigma_c^2}} + C\sqrt{2\pi\sigma_c^2}(\beta_1\mu_c + \beta_0)\Phi_{\mu_c,\sigma_c^2}(\frac{1}{a_t}) \\ &+ C\sqrt{2\pi\sigma_c^2}(-\beta_1\mu_c - \beta_0)(1 - \Phi_{\mu_c,\sigma_c^2}(\frac{1}{a_t})), \end{aligned} \tag{17}$$

where $\Phi_{\mu_c,\sigma_c^2}(.)$ represents the cumulative distribution function (cdf) of a normal distribution with mean μ_c and variance σ_c^2.

The confidence interval would belong to $(\bar{R}_t - \delta, \bar{R}_t + \delta)$ where $\bar{R}_t \triangleq \mathbb{E}[R_t] = \int r f_{R_t}(r) dr$ and δ can be derived by satisfying $\mathbb{P}(|R_t - \bar{R}_t| \le \delta) = \int_{\bar{R}_t - \delta}^{\bar{R}_t + \delta} f_{R_t}(x) dx = 1 - \epsilon$ for some small $\epsilon > 0$.

3.3 Estimating the Risk Score

We propose a novel risk score metric for a given community that is proportional to the probability of someone in that community becoming infected in the next time period (typically, 24 hours). The risk score can be derived as the average number of people in that community that are likely to get infected in the next 24 hours by the currently infectious people divided by the current number of susceptible individuals. We further normalize this probability by multiplying by 10,000, so that a score of 1 implies a 1 in 10,000 chance of getting infected, a score of 2 implies a 2 in 10,000 chance of getting infected, and so on. Mathematically, the risk score is defined as follows:

$$\Gamma_t = \frac{I(t) \cdot R_t}{D_I \cdot S(t)} \cdot 10000 \approx \frac{K \cdot I_{rep,new}(t) \cdot R_t}{N} \cdot 10000, \tag{18}$$

where $I_{rep,new}(t)$ indicates the most recently reported count of new confirmed infectious cases, K refers to the ratio of true cases to reported cases, R_t is the time-varying reproduction number, and N is the total population size of the community. The approximation follows from the fact that $I_{rep,new}(t)$ is approximately equal to $\frac{I(t)}{D_I \cdot K}$ and $S(t)$ the number of susceptible people in the community is approximately equal to N in the early stages of the epidemic. Confidence intervals for the risk score Γ_t could be obtained numerically using a similar process as described for R_t accounting also for uncertainty in K. Note that since K may not be known for a given community, it may be helpful to use the following normalized form of the risk score: $\frac{\Gamma_t}{K}$, which is still proportional to the probability of infection for an individual.

3.4 Color-coded Risk Levels

To further simplify the presentation of the risk score to a wider audience, we propose to classify the risk levels into four color-coded levels: (Green, Yellow, Orange, Red). The risk level is determined by evaluating the normalized risk score ($\frac{\Gamma}{K}$) with respect to three pre-specified threshold levels $\theta_1, \theta_2, \theta_3$, such that when $\frac{\Gamma}{K} < \theta_1$ the risk level is green, when $\theta_1 \le \frac{\Gamma}{K} < \theta_2$ the risk level is yellow, when $\theta_2 \le \frac{\Gamma}{K} < \theta_3$ the risk level is orange, and when $\frac{\Gamma}{K} \ge \theta_3$ the risk level is red.

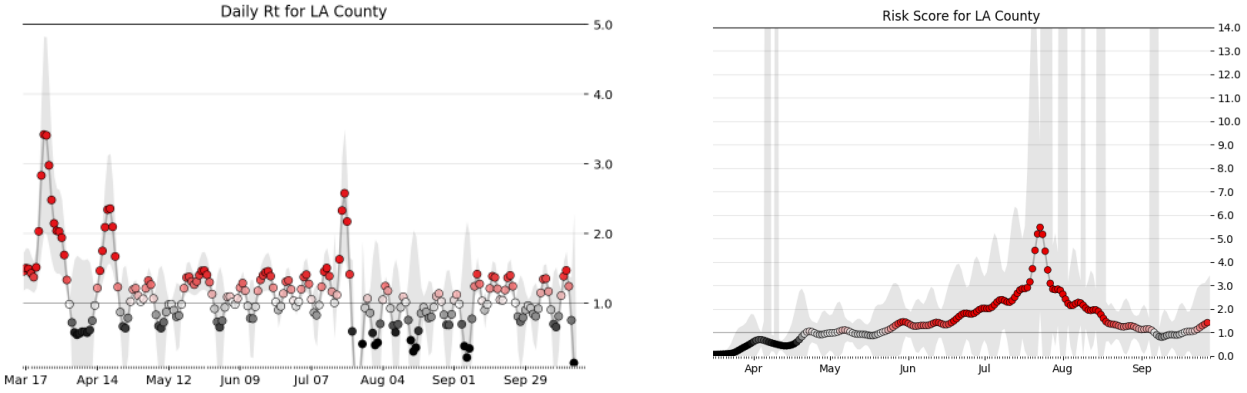


Figure 2: The left and right plots respectively represent the estimated effective reproduction number R_t and the risk score Γ_t over time for the entire county of LA considering $\mathbb{E}[D_I] = 7.5$, $Var[D_I] = 9$, $\mathbb{E}[K] = 3$, and $Var[K] = 0.44$. The gray area represents the 95% confidence interval in the estimates.

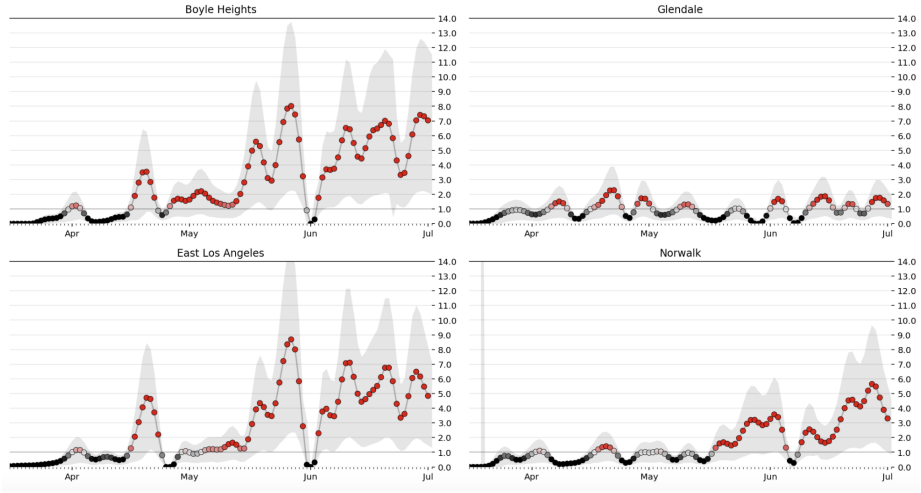


Figure 3: Estimate of risk score Γ_t over time for four representative communities in LA County: Boyle Heights, Glendale, East LA, and Norwalk. Regarding the settings, we considered the following $\mathbb{E}[D_I] = 7.5$, $Var[D_I] = 9$, $\mathbb{E}[K] = 3$, and $Var[K] = 0.44$. Our approach also yields uncertainty in the estimate, as shown in the form of confidence intervals (in gray).

4 IMPLEMENTATION AND EVALUATION IN LOS ANGELES COUNTY

The software for data collection, infection rate estimation and prediction has already been implemented and made available as open-source software (at the following repository: https://github.com/ANRGUSC/covid19_risk_estimation). The software is written in Python using standard data processing libraries such as NumPy and SciPy.

4.1 Data Sources

We have acquired COVID-19 case data from the LA County's Department of Public Health using a Python-based data parser we wrote (open-sourced at the following link: https://github.com/ANRGUSC/lacounty_covid19_data). We have been updating this repository regularly with the latest data every day since mid-march and also making available plots of the number of cases, number of fatalities, top 6 communities with the large number of cases, infection rate for the entire LA County, and the top 9 communities with the highest infection rate at the following link: http://anrg.usc.edu/www/covid19.html.

The following data sources are used for the infection rate and prediction:

- The COVID-19 case information was collected through LA County's daily press releases (Accessible through the following website: http://publichealth.lacounty.gov/media/Coronavirus/).
- Recovery information provided by the World Health Organization.

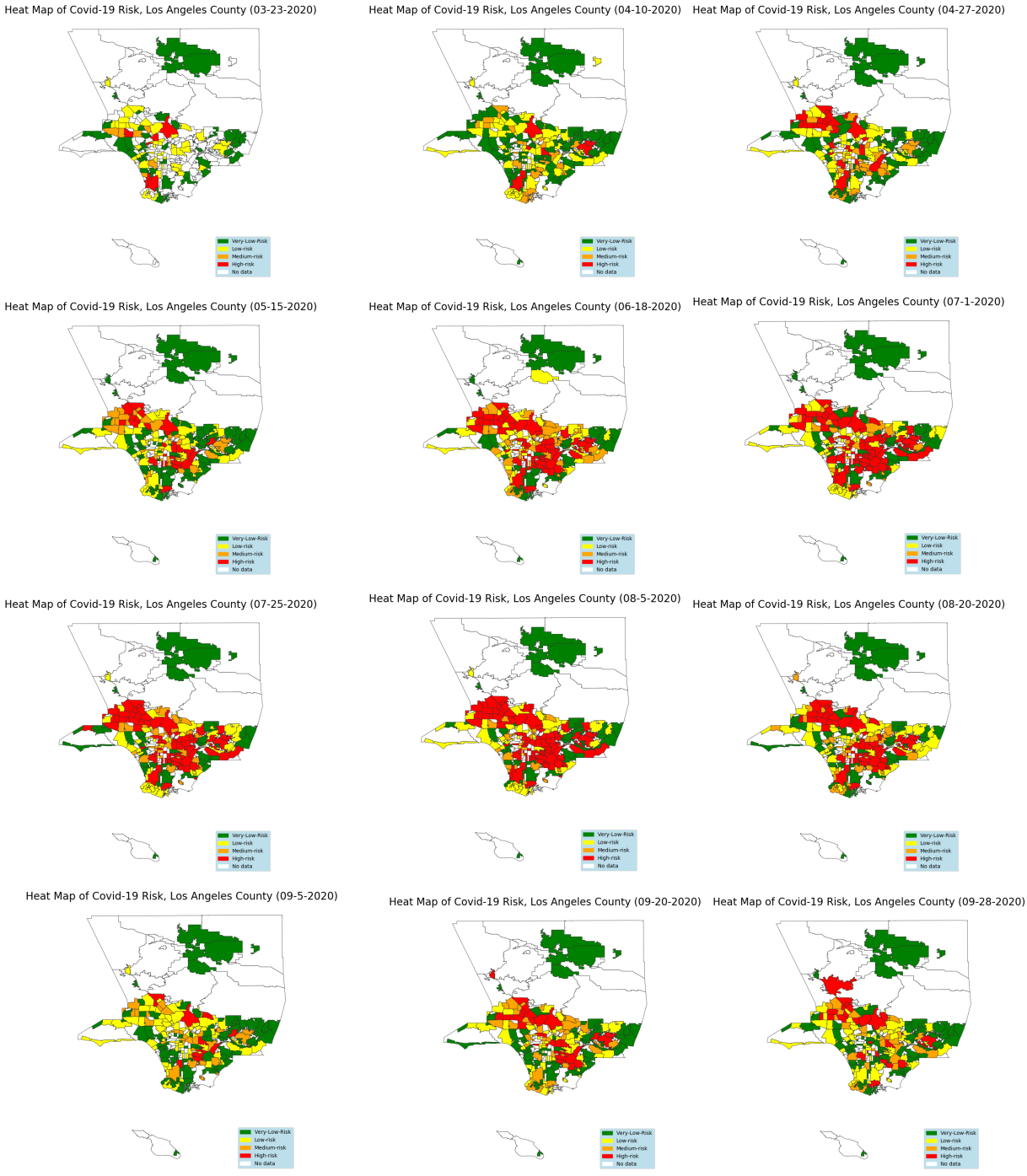


Figure 4: Maps showing the estimated risk score for different LA County Communities on different dates since mid-March 2020. Top row: March/23/2020, April/10/2020, April/27/2020; Second row: May/15/2020, June/18/2020, July/1/2020. Third row: July/25/2020, August/05/2020, August/20/2020. Last row: September/05/2020, September/20/2020, September/28/2020.

- The population data from LA County Census is available online (from lacounty.gov/government/geography-statistics/cities-and-communities/).

4.2 Real-world Usage

The City of Los Angeles is currently using the risk model described in this work that has been developed by researchers at USC, to help assess location-based risk for COVID-19 infection. The City is working with the County and other partners to develop a tool that is publicly accessible and can be used by individuals and communities to mitigate risk of infection. The goal is to change behaviors to reduce risk of infection and promote a greater understanding of factors that increase COVID risk. A color-coded COVID-19 threat level tool that can be used by citizens has also been unveiled by the Mayor of the City of LA, online at https://corona-virus.la/covid-19-threat-level. Besides, our risk score is made available to community members here https://grmds.org/risk/, wherein the users can enter their community name to understand their community's risk level. This prototype mapping site is helpful in developing options that are useful for people and continues to evolve.

5 EVALUATION RESULTS

We present plots from our analysis of LA County community case data using the estimation approach described in this work in Figure 2, Figure 3 and Figure 4. The code and the data used for generating these plots are available at the following GitHub repositories:

- **Risk model software:** - https://github.com/ANRGUSC/covid19_risk_estimation
- **COVID-19 case data for LA County:** - https://github.com/ANRGUSC/lacounty_covid19_data

Figure 2 shows plots of the estimated expected reproductive number R_t and the estimated risk score for the entire LA county. These plots are based on a 14-day moving average applied on the daily number of confirmed cases. In accordance with LA county daily press releases, there is a sharp jump in both R_t and risk score around the beginning of July. Note that the reason the risk score during the beginning of July is higher than the risk score during the last week of March, despite having the same R_t, is due to the fact that there are significantly more confirmed cases in July compared to March.

Figure 3 shows the risk score estimates over time for four representative communities within LA County. The case data continue to increase for some communities, while the number of cases remained somewhat study for a few communities.

Figure 4 shows the color-coded risk levels for communities in LA County for select dates over the past six months. The risk levels have gone up for many communities in the last week of July and the beginning of August, which is also visible in the county-wide risk score, as shown in Figure 2.

6 CONCLUSION

We have proposed a new risk metric Γ_t that can be used by individuals in any community to assess their probability of getting infected by COVID-19. The metric builds on the estimation of R_t, the average reproductive number, which is obtained from a time-varying extension of the classical SIR model. We show how to evaluate the uncertainty in both metrics as well. In future work, we plan to generalize the approach to the SEIR model, which also models an additional incubation period. We have released code to implement an estimation of the risk score that can be used for any community worldwide as long as time-series data for confirmed new cases and the population are known. We have also proposed the use of simple color-coded risk levels to inform and guide the public, as has been adopted in the City of Los Angeles. One open question to be investigated in future work is how residents respond to these color-coded levels and how to communicate the behaviors appropriate to each level.

7 ACKNOWLEDGMENTS

This work is supported by the USC Viterbi Center for Cyber-Physical Systems and the Internet of Things (CCI).

REFERENCES

[1] Marshall Maximilian Dong Ensheng Squire Marietta Badr Hamada, Du Hongru and Gardner Lauren. 2020. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases* (2020).
[2] Luís Bettencourt and Ruy Ribeiro. 2008. Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases. *PloS one* 3 (02 2008), e2185. https://doi.org/10.1371/journal.pone.0002185
[3] Fred Brauer. 2005. The Kermack–McKendrick epidemic model revisited. *Mathematical biosciences* 198, 2 (2005), 119–131.
[4] Sheryl L. Chang, Nathan Harding, Cameron Zachreson, Oliver M. Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. arXiv:q-bio.PE/2003.10218
[5] Yu-Chun Chen, Ping-En Lu, and Cheng-Shang Chang. 2020. A Time-dependent SIR model for COVID-19. *ArXiv* abs/2003.00122 (2020).
[6] Kai Kupferschmidt. 2020. The lockdowns worked—but what comes next?
[7] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine* (2020).
[8] M. Newman. 2010. Networks: An Introduction. In *Oxford University Press*.
[9] Kevin Systrom. 2020. The Metric We Need to Manage COVID-19. http://systrom.com/blog/the-metric-we-need-to-manage-covid-19/
[10] Aurelio Tobías. 2020. Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up. *Science of The Total Environment* (2020), 138539.

COVID-19 Joint Pandemic Modeling and Analysis Platform

Gautam Thakur, Kevin Sparks, Anne Berres, Varisara Tansakul, Supriya Chinthavali, Matthew Whitehead, Erik Schmidt, Haowen Xu, Junchuan Fan, Dustin Spears, Elton Cranfill
Oak Ridge National Laboratory
Oak Ridge, Tennessee
{thakurg,sparksa,berresas,tansakulv,chinthavalis,whiteheadmc,schmidteh,xuh4,fanj,Spears,spearsdl,cranfillej}@ornl.gov

ABSTRACT

The non-pharmaceutical intervention to reduce the impact and spread of COVID-19 requires the development of policies and guidance through a collaborative effort among government, academia, medicine, and citizens. To operationalize this effort, we have developed an all-encompassing situational awareness platform that can process multi-modal and multi-source data allowing informed decision making. Besides, showing the current spread of infection, the platform also captures the impact of human dynamics on the infection spread, location, and availability of critical infrastructure, prediction, and high-performance computing driven simulation. The platform is extensible, allowing third-party integration and services to consume the curated data and analytics in near real-time. We believe the platform will augment critical decision making for reducing the impact and spread of the pandemic.

CCS CONCEPTS

• **Computer systems organization → Real-time system architecture**.

KEYWORDS

real-time architectures, visualization, distributed systems

ACM Reference Format:
Gautam Thakur, Kevin Sparks, Anne Berres, Varisara Tansakul, Supriya Chinthavali, Matthew Whitehead, Erik Schmidt, Haowen Xu, Junchuan Fan, Dustin Spears, Elton Cranfill. 2020. COVID-19 Joint Pandemic Modeling and Analysis Platform. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3423459.3430760

1 INTRODUCTION

In January 2020, the Director-General of the World Health Organization (WHO) declared the novel coronavirus (COVID-19) outbreak a Public Health Emergency of International Concern (PHEIC), WHO's highest level of alarm. Since then, both pharmaceutical and non-pharmaceutical interventions sprung to action in an attempt to staunch the spread of infection. In the latter case, healthcare agencies, volunteers, non-profit organizations, and several others have put forward an effort of epic proportions to curate high-quality datasets on cases and available healthcare resources. Prediction and simulation models are built to demonstrate the likelihood of infection and its impact. However, to a large extent, these efforts are isolated and focus solely on one thing; for example, Johns Hopkins University (JHU) reports cases at the province level in China; at the city level in the USA, Australia, and Canada; and at the country level otherwise[7]. Policymakers need an all-encompassing view of the situation to make an informed and rational decision for maximum impact. This involves quick access to current situational awareness, future prediction, and ancillary information such as the location of critical infrastructure. Until recently, no such mechanism exists that provide everything under a single umbrella. There could be several reasons: i) it is difficult to conflate multi-variate data with varying spatial and temporal granularity, ii) in a continuously evolving situation, data variables and schema also evolves, making it impossible to converge on a stable data format, iii) the volume and velocity of data is enormous, requiring specialized data and compute architecture. Also, in a dynamic environment like this, the currency of the data and analytics is paramount. Thus, Real-time (RT) architectures capable of stream processing are vital to address the challenge arising from the currency of data insights. Developing scalable and operational RT architectures have a high upfront cost, sometimes making it difficult to build and run.

The objective of this work is to develop an integrated COVID-19 pandemic monitoring, modeling, and analysis capability that will include, - i) historical and current spatio-temporal trends of disease spread, ii) estimates of required hospital beds, ICU units, ventilators, etc., iii) needed testing capacity and where iv) quantify the effectiveness of implemented interventions and mitigation strategies. To address these challenges, we developed and operationalized an agile, online COVID-19 platform for integrating, synthesizing, analyzing, and visualizing geographically resolved data (collected as part of this effort) as well as conveying modeling and simulation results that anticipate future COVID-19 transmission dynamics.

The proposed platform is built by hybridizing the concepts of Lambda architecture and Hyperscaling to achieve real-time analytics and visualization of spatiotemporally disparate datasets through load-aware vertical and horizontal scaling of available infrastructure with zero downtime. Besides the architecture, the proposed platform offers two key application-level functions:

Multisource data integration data store: A critical component of the online platform is the ability to ingest and merge structured and unstructured data sources curated in support of the COVID-19 platform from multiple sources that include hospitals, regional governments, social media, and other crowd-sourced outlets related to COVID-19 infectious diseases spread.

Interactive analytics dashboard: A substantive system for merging and scientifically analyzing multiple, disparate open-source data

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430760

streams (e.g. COVID-19 cases, twitter content, quarantine maps, demographic context, and news feeds), physical data (e.g. temperature, precipitation) and modeling and simulation outcomes. The work leverages hyperscale architecture for data curation, analysis, and visualization of a range of curated and modeled data.

The remainder of the paper is structured as follows: section 2 modestly discusses related work and relevant background, the proposed architecture is discussed in section 3, while the operational workflow release plan is discussed in section 4. The end product of the platform is a suite of interactive dashboards, their design and development are discussed in section 5. Hotspot detection is discussed in section 6 and extending the platform to connect with third-party applications is discussed in section 7. Finally, the paper concludes in section 8 with a discussion on stress testing.

2 BACKGROUND

Ever since the onset of the COVID-19 pandemic, countless web-based dashboards have been developed to track the development of COVID-19 at the local, national or global level. Most governmental agencies (e.g., Centers for Disease Control and Prevention (CDC), state or county health agency) use web-based dashboards to release the latest information about COVID-19 to the public. It does not take long for researchers and decision makers to realize the need for a more comprehensive platform that can collect, aggregate, visualize, and predict the dynamics of COVID-19 using various scattered data sources.

One of the most widely referenced platforms is the web-based dashboard from Johns Hopkins University[8], which tracks the latest number of COVID-19 cases and deaths around the world, at different spatial granularities for different countries. It functions both as an authoritative COVID-19 data curator and a basic tracking tool with visualizations from ESRI. The New York Times[25] also developed a dashboard that offers similar functions. There are many other web-based platforms that visualize and track different aspects of the pandemic, e.g., the spread and evolution of different strains of SARS-CoV-2[14], real-time symptom and public behavior survey around the world[6, 11], online conversation and information spread[22], healthcare system capacity[2], and human mobility[9]. Most of these tools focus on a relatively narrow aspect of a large system, and lack a predictive analysis capability. In this work, we have developed a geospatial platform that provides real-time situational awareness, prediction, and simulation results generated by different laboratories. More importantly, it conflates disparate data sources to depict a story with context rather than mere statistics.

For all emergency response analytic platforms, context is a critical component of communication. As stated in [26], "Geography and history offer unique perspectives on context through study of the interconnectedness of phenomena, events, and places across multiple spatial and temporal scales through which situations are understood." For COVID-19, this means effectively communicating information beyond case counts and deaths. Providing geographic and historical information relating to the spread of the pandemic is important. Furthermore, context surrounding COVID-19 extends to supporting information like hospitalizations, where and how much people are traveling to public/commercial spaces, school closures, and more. To that end, the front-end visual portion of many COVID-19 dashboards contextualize the pandemic by including map views with multiple layers, supported by various graphs, charts, and tables all with historical and current data, with which users can interact.

The proposed platform is built on lambda architecture[17] allowing a way of processing massive quantities of data that provides access to batch-processing and stream-processing methods with a hybrid approach. The lambda architecture itself is composed of 3 layers: batch, serving, and stream. The platform benefits from this approach, when combining archive data with streaming data that the platform collects. In addition, the platform incorporates the features of hyper-scaling architecture[4] that can benefit from expanding both compute and storage power as required. Besides these, there exists a plethora of architectures such as kappa[18], derived from lambda architecture and less complex to deploy in real-world, Apache Hadoop[15], Apache Spark[27], among others. There are more architectures, we keep the details limited, and for more information review[21, 24].

Besides high-availability, another important measure of success for any scalable architecture is its ability to maintain low latency during I/O intensive operations. Distributed Caching Mechanisms (DCM) that stores large amounts of data in the memory of more than one machine offers to bridge the gap and improves the latency. Information-centric networking[28] is one of the best examples of a distributed cache implementation that focuses on location-independent content sharing across the planet. There are four types of DCMs, that includes cache aside, read-through cache, write-through cache and write-back. These approaches are application and scenario dependent on maximizing the application throughput. Open source implementation of DCMs are widely available, such as Hazelcast, Memcache, and Redis, among others[23]. These are data agnostic and their effectiveness depends on the implementation. The proposed platform uses a combination of write through caching for improving the performance of temporally located data.

3 ARCHITECTURE

This platform is built on the principles of lambda[17] and hyperscale[4] architecture to address the challenges of combining disparate data sources and dynamically scale to address computational challenges. The architecture benefits from the use of widely available off the shelf servers and computational equipment. The biggest benefit lies in the ability to scale the platform as a function of load and latency to accommodate additional users and requests. The architecture can be scaled both vertically and horizontally, maximizing in-built fault tolerance and cost-effectiveness.

The proposed architecture is shown in Figure-1 that includes data collection and processing, distributed data grid to expedite the data transfer and reduce latency, application server interface, machine learning, and data quality evaluation. The remainder of this section discusses these in more detail.

3.1 Data Collection

The first step in release planning is the collection and curation of high-quality authoritative data. This involves discovering relevant data source(s), sanitizing and transforming the new data, de-duplicating, and semantically conflating it with other existing data sources. The data collection's geographic coverage is the entire planet and the spatial granularity goes to the county or even census block group level. Besides, data gathering should be done in

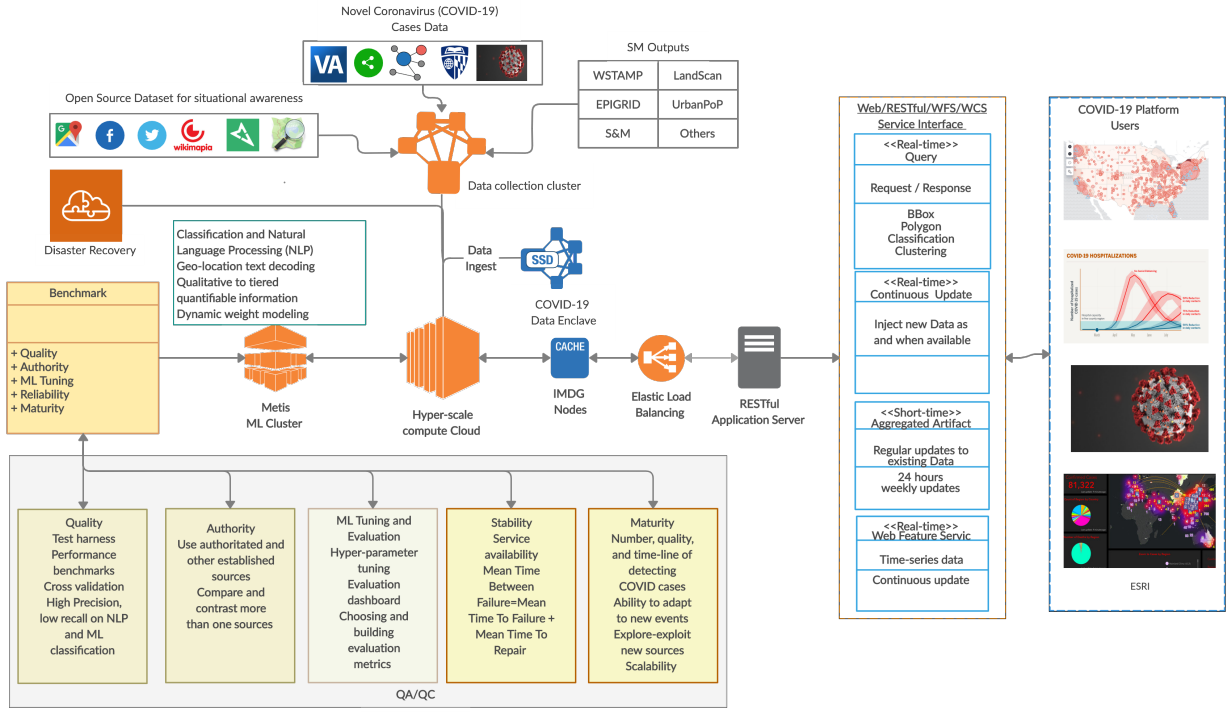


Figure 1: COVID 19 platform architecture

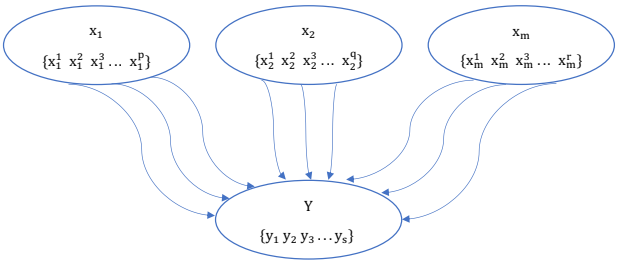


Figure 2: Unified mapping format for disparate data sources

real-time and continuously for the currency of insights. On average, over 100 data sources were searched daily and several million records were collected. At this scale, data curation efforts must be automated with a human in the loop only when necessary.

A universal attribution format was designed to unify disparate data sources ($X \in \{x_1, x_2, x_3, \ldots x_m\}$) and a translation engine (Θ) is developed to map attributes of disparate data sources to the universal format data ($Y \in \{y_1, y_2, y_3, \ldots y_s\}$). For each data source, a separate sequence of translations ($\Theta \in \{f_1, f_2, f_3, \ldots f_m\}$) were designed such that each uniquely maps to universal data format as,

$$f_i : x_i \rightarrow Y \tag{1}$$

where each data source x_i has k attributes, $x_i \in \{x_i^1, x_i^2, x_i^3, \ldots x_i^p\}$ that maps to Y s attributes, such that $p \leq q \leq r \leq s$, as shown in Figure 2. This was an important step to be performed so aggregate statistics and visualization from disparate data sources can be done seemliness manner.

3.2 Data Processing and Analytic

Developers use RESTful APIs to access the data for processing and generating analytics. Also, boilerplate templates are made available for developers to generate analytics via natural language processing and machine learning. Besides authoritative data, the platform also harvests social media data (e.g. twitter) to gather information about infection spread and response. Classification models can be built to evaluate communication and assess their impact on health. Rigorous benchmarks that include, quality control, tuning and maturity of results are evaluated before releasing the results.

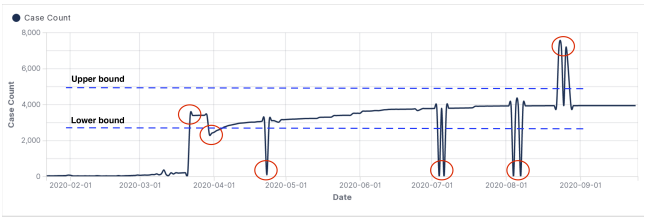


Figure 3: Missing data in time-series

3.3 Quality Assurance Quality Control

A suite of spatio-temporal and statistical data processing templates are developed to gather insights from the data. In addition, visualization packages such as Vega and custom ensemble visualization are integrated in the platform to allow the development of interactive dashboards and processing of reports. At times, the platform ingests direct results of simulation[10] or predictive models[16] for the purpose of visualization.

Algorithm 1: Algorithm for server-side HyperCache

Function `Remove`(*key*) *return value*
- **Data:** Distributed map
- **Result:** value of element to be removed
- *acquire_lock*();
- *value* ⟵ *remove*(*hmap*, *key*) ;
- **if** *value != null* **then**
  - return *value*;
  - *sync_cache*() ;
  - *wait*();
- **else**
  - return *null* ;
- *release_lock*();

Function `Update`(*key, value*) *return result*
- **Data:** Distributed map
- **Result:** value of element to be updated
- *acquire_lock*();
- *result* ⟵ *update*(*hmap*, *key*) ;
- **if** *result != 1* **then**
  - return *false*;
- **else**
  - return *true* ;
- *sync_cache*() ;
- *wait*() ;
- *release_lock*();

3.4 Distributed HyperCache

Developing elastic architectures that scale as a function of an evolving computing workload is essential for real-time applications. Significant advances have been made in hyper-scalable storage, data centers, and cloud computing infrastructure to accommodate the exponential increase of such workload. In this architecture, we utilize the distributed memory of nodes to improve storage latency that is processing and simultaneously retrieving a large amount of disparate data for real-time analysis. HyperCache is implemented in the form of an In-Memory Distributed Grid and is built on the top of the Direct-Attached Storage (DAS) computing cluster running simultaneous applications. These applications communicate with HyperCache via client-server architecture for maximum compatibility. A monitoring system is developed and deployed for performance bench-marking and providing essential support during exponential data compute workloads. A simple representation of adding/updating and removing the element in HyperCache is shown in Algorithm-1.

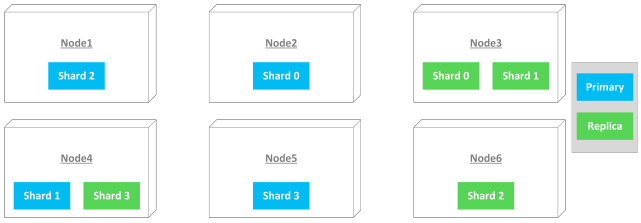


Figure 4: Replication of an index across six nodes.

3.5 Replicated Data Management

In this section, we discuss fault tolerance and approaches taken to ensure the platform remains accessible and robust when it is needed most. The platform manages its data integrity through replication across multiple servers. As shown in Figure 4, a database index is broken down in four primary shards (pieces) and four replicated shards. They are distributed across six nodes in such a way that if any machine is down or corruption, the database can still be rebuilt and no data loss occurs. The global consistency is maintained by frequently broadcasting updates and propagating them on all the nodes.

3.5.1 Serialization and Optimistic Concurrency. The management and consistency of aforementioned replicated data is achieved through serialization with optimistic concurrency control, such that the execution of a set of parallel data operations (transactions) must be equivalent to a serial execution of the same data operation[5]. Consider $\Gamma = \{t_0, t_1, ..., t_m\}$ is a set of parallel transactions. Then, for each transaction, t_i, let R_i is the read set and W_i is write set respectively for t_i. For example, in parallel when $t_i \rightarrow t_j$ occurs, then t_i must come before t_j equivalent in a serial transaction. Optimistic concurrency protocol ensures that any execution if not consistent is aborted based on the timestamp. A workspace is maintained for each transaction that later on executed to maintain long-term data consistency. This mean, sometimes user request response includes cached results that are not fully updated. Broadly, for three transactions t_0, t_1, t_2 such that their respective timestamps are $t_0 < t_1 < t_2$, the operations on a shared object occur in increasing order of the timestamp. Recent transactions (smaller timestamp) wait for older transaction (large timestamp) to finish to maintain data integrity. If an older transaction with larger timestamp (e.g. r_2) encounters a younger transaction (e.g. t_0), the previous dies and restarts with a smaller timestamp. This approach avoids potential deadlocks, as the execution of transactions are based on increasing order of timestamps.

The optimistic concurrency is broken down in three phases - i) execution phase, ii) validation phase, and iii) update phase. It begins by assigning each transaction t_i a timestamp TS_i at the start of the transaction and TV_i at the beginning of validation. It's read as assigned as R_i and written are assigned as W_i in this phase. In *execution phase*, a local workspace is created for each transaction with shadow copies of object to be updated. These objects are updated

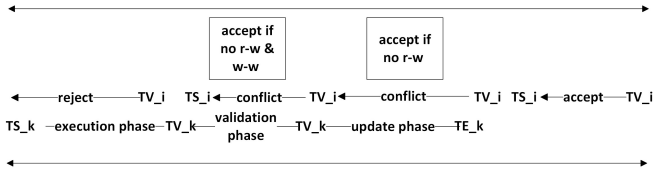


Figure 5: Optimistic Concurrency Control to store and visualize real-time data ingestion

locally and assigned a version number. In case of abort, the transaction is cancelled and the workspace is deleted. Otherwise, the transaction moves to next phase. In *validation phase*, mutual consistency among this and the distributed transactions is performed at remote locations to ensure serializability. The validation between two transactions (Say t_i and t_k) is achieved as follows-

- Validation of transaction t_i is not accepted if $TV_i < TV_k$.
- Validation of transaction t_i is accepted if it does not overlap with any t_k.
- The execution phase of t_i can overlap with update phase of t_k, given it completes its update phase before TV_i. Validation of t_i is accepted if $R_i \cap W_k = \emptyset$.
- The execution phase of t_i overlaps with the validation and update phase of t_k, and t_k completes its execution phase before TS_i. Validation of t_i is accepted if $R_i \cap W_k = \emptyset$ and $W_i \cap W_k = \emptyset$.

The details of these stages are shown in Figure 5. In update phase, the changes to the data objects are made permanent and propagated across the cluster and in persistent memory and storage.

3.5.2 Disaster Recovery Plan. Besides replicating the data across the cluster onsite, a remote disaster recovery site was also deployed to keep operations running in case of major failures (such as natural disaster). We utilized Google Cloud Platform as Disaster Recovery as a Service (DRaaS) to snapshot and backup the status of current database every four hours on google cloud. Since, the snapshots are not instantaneous, take time to complete and asynchronous with respect to time and data integrity. This approach uses incremental backup strategies to save change quickly and efficiently. The cloud only serves as a remote location to store the data and serves as a backup. When snapshot is in-progress, it is still possible to add new data and make other requests to the cluster.

3.6 Data and Analytic Quality Control

Data quality and analytics are evaluated to ensure insights are scientifically accurate. The data curation task uses authoritative sources (such as CDC, JHU, etc.), reducing issues related to accuracy. However, curation issues occur when data attribution format changes (e.g. new attribute is added), network connectivity (intermittent disconnection, synchronization issue because of latency), among other issues. An example of Not missing at random (NMAR) data[20] in univariate time series of COVID-19 case counts are shown in Figure 3. In the figure, the probability for a missing observation depends on the value of the observation itself (the observations are not recorded because of a network error)[13]. If needed, linear interpolation or arithmetic smoothing is used to rectify the missing data in time series. Besides, the process also benefits from review by subject matter experts (epidemiologists, geographers, statisticians) from time to time.

3.7 Application Server and End User

The application server holds the core deployment of the application. We have used a load balancer with the multi-instance deployment of an application server for fail-over and load distribution. The platform is deployed at https://covid19.ornl.gov can be accessed via a web browser or through integrating RESTful services. A user authentication mechanism is implemented to secure and limit access to authorized users only. In section 7, we demonstrate an approach to extend the platform connecting ESRI services and the development of the story-telling feature.

4 RELEASE PLAN

The on-set and rapid spread of COVID-19 created an immediate need to deploy a reliable and stable situational awareness platform accommodating inputs generated by several research entities. It was critically important that this platform should display key scientific findings for policy guidance and informed decision making within a given schedule, quality, and effort constraints. This led us to formalize a systematic release plan workflow for the selection and development of features and their incremental release at a regular interval. Each release addresses the production of meaningful insights and new features developed by the participating research entities. This also allowed all the moving parts of the collaboration to work in sync, work towards a common goal, and for the end-user base to anticipate the changes and new updates. The release plan workflow is shown in Figure 6. The platform also monitors the usage of its services, detection of spurious activities, report generation, and allocating resources to allow a third-party application to utilize analytics and data stream. The post-deploy release step is useful toward extending the platform and for measuring the use.

5 VISUALIZATION

The visualization technology we use to display the data in this platform is Kibana, an open source data visualization platform from the Elastic stack. It provides a variety of standard charts, time series graphs, geospatial visualizations, and support for Vega visualizations. In addition, we have developed custom visualizations that we integrate as plug-ins. The user-facing part of this platform is organized as a series of dashboards.

5.1 Dashboards

In an effort to effectively organize, explore, and reflect the different uses of a large and varied volume of data ingested into the platform, we created several dashboards. These dashboards include a Situational Awareness dashboard (displaying current and historical data from global, to US state and county spatial resolutions), a Predictive Analytics dashboard (displaying multiple predictive models at national, state, and county levels), and more. Some of these dashboard provide high-level overviews, others provide a deep-dive into a particular aspect, or even one specific model or data feature.

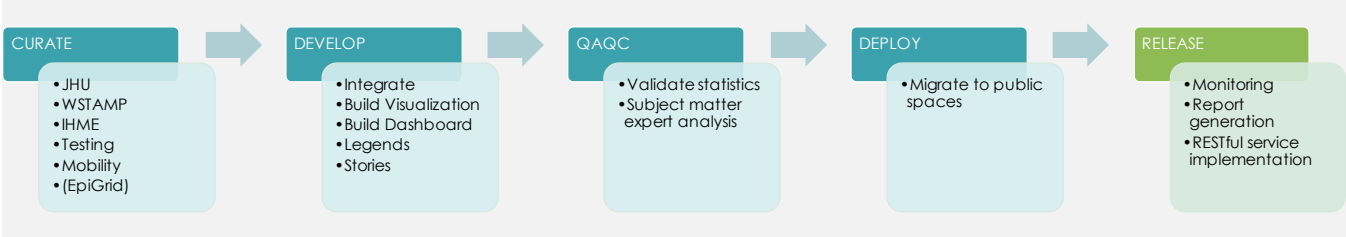


Figure 6: COVID 19 Release Plan Workflow

Incoming data's spatial resolution varies from national, state, county, and city. Temporal resolution ranges from daily, weekly, and monthly time steps. Data updates occur at a variety of times, from static-single upload data layers, to daily updates, to a few updates a month. As such, merging multiple data products into a simple, cohesive view was a challenge. Furthermore, designing visualizations to be responsive to global user input and querying when existing in different spatial and temporal resolutions required careful consideration during data processing, storage, and visualization.

When possible, we attempted to maintain thematic consistency for the purposes of intuitive user experience. Most dashboards consist of interactive maps, time series graphs, and basic charts. Map layers have zoom-dependent views from country, to state, to county. For each layer, we define an appropriate aggregation type (e.g. sum, maximum, etc) depending on the variable presented. To further create a sense of cohesion between dashboards, we attempt to use color as a guide: Similar data is shown in similar colors where possible. For categorical data, we use custom colormaps that represent the type of data appropriately.

In addition to the utilizing Kibana's standard visualizations, we also leverage its support for Vega, a declarative language for web-based visualizations. Vega allows us to create a wider variety of visualizations for the platform and gives us the flexibility to further customize visualizations. For example, in Figure 7 we use Vega to visualize R-naught estimates for the contiguous U.S., Alaska, and Hawaii at the county-level in a single map view. The diverging color scheme provides users with an at-a-glance view of counties experiencing reproduction rates above or below 1, and the hover-over tooltip functionality lets them quickly see the data behind the map. This and other Vega visualizations on the platform query the most recent data from Elasticsearch indices and are configured to respect global filters selected by the user, which allows for seamless and responsive integration with other charts in Kibana dashboards.

5.2 Visualization for Situational Awareness

To illustrate situational awareness (Figure 8), we prioritized the display of current and historical data and analytics regarding case counts, deaths, testing, and various mobility metrics. For example, using daily case counts, we displayed up to date cumulative cases, new case rates, cases per capita, results from transmission rate models, and more. We provided supporting data to illustrate variables that likely influence case counts and deaths. These largely included various measurements of social reaction to the pandemic, including

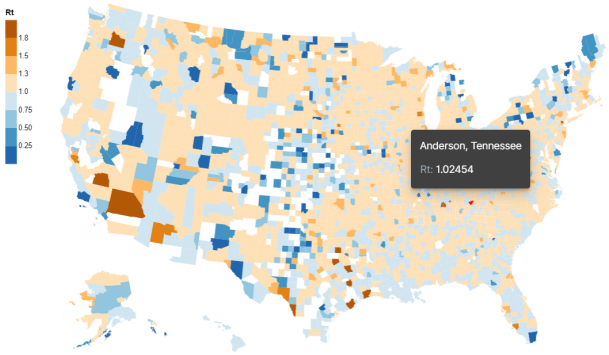


Figure 7: County-level map of R-naught estimates with tooltip functionality. Built with Vega.

dates of school closures, general mobility indices at the national, state and county scales, transportation intensity, and more. Possible user interactions include selectable map layers (country, state, and county scales), dynamically updating map layers based on zoom level, country/state/county term filters, and hover-over tooltips.

5.3 Forecasting and Predictive Visualization

We accumulated multiple predictive models from public (national laboratories) and private industries, that provided near future estimations on case counts, deaths, hospitalizations, intensive care patients, and more. Figure 9 shows filtered visualization of mechanistic model outputs [19] with different resolutions, including state and US metropolitan statistical area (MSA). For example, Figure 9a shows predictions for new cases in Alabama and Figure 9a shows predictions for new cases in Atlanta, Georgia. Note that Figure 9 show the results generated on September 6th, 2020. When the spatial scale and temporal resolution of predictive models were consistent with one another, attempts were made to group these model outputs into a single visualization. Otherwise, model outputs were placed side by side for comparison. All available model outputs were displayed in the dashboard, and users had the ability to filter results by state or county.

5.4 Visualization for Simulation Results

We created dashboards for EpiGrid and EpiCast [10] simulation results that were produced specifically in the context of this project. For each model, we developed a custom processing workflow which

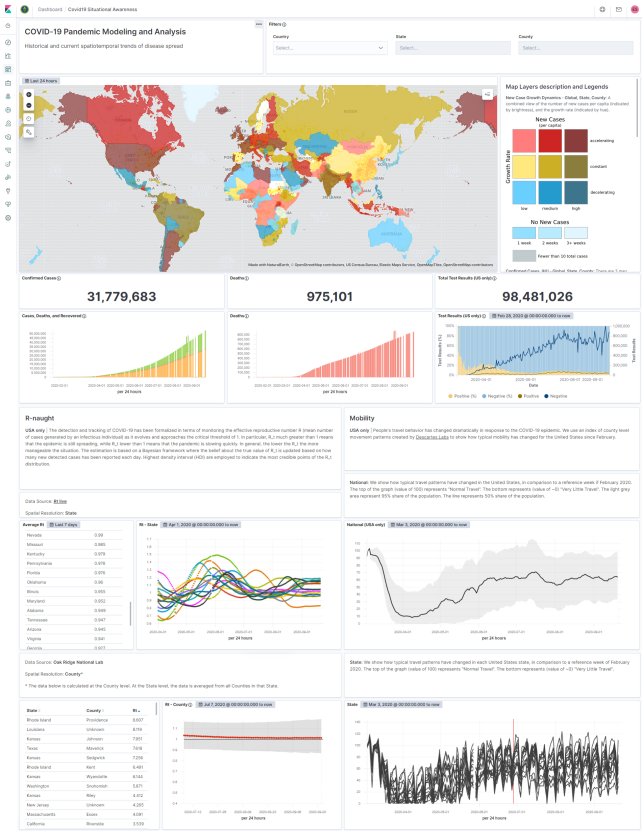
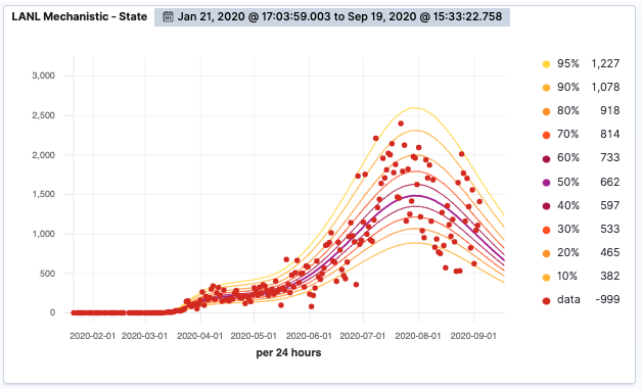
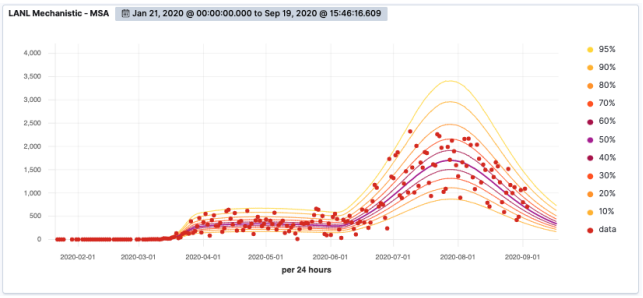


Figure 8: Situational Awareness Dashboard.



(a) New cases predictions for Alabama



(b) New cases predictions for Atlanta

Figure 9: Mechanistic model visualization.

converts the model outputs from their respective formats into a common format that is ingestible into ElasticSearch.

Since aggregations for map layers require incident data, but both models report cumulative infected individuals, we also computed the daily increase in cases as part of our workflow. This allows us to select a start date and end date and produce an accurate count of new cases that were predicted within the given timeframe.

EpiGrid focuses on different stages of infections from first exposure, infectiousness (with isolation status to account for quarantine), requiring hospitalization (with differentiation on whether or not they receive it to model healthcare availability), recoveries, and deaths. EpiCast focuses more on the severity of symptoms with more detailed outputs for hospitalized individuals. It provides outputs for symptomatic, hospitalized, in Intensive Care Unit (ICU), and requiring a ventilator. Unlike EpiGrid, it does not provide recoveries or deaths.

The dashboards for both models are similar, with general information about the model, interactive elements for filtering, a map, and a side bar with instructions, information about layers, and legends. The blue section in Figure 10 shows the top section of the EpiCast dashboard with a view of predicted county-level data of individuals in the ICU on the map. The green section of Figure 10 shows a sample of visualizations from the EpiGrid dashboard for a subset of states (Georgia, Kansas, New Mexico, and New York), curves for each case type, and a comparison with ground truth data from the New York Times dataset [25]. In the pink section of Figure 10 we show a comparison of different model output parameters using the same colormaps as other dashboards (cases = yellow/red, recovered = green, death = black/gray) for some counties in New York. The data for this model run does not cover all counties of each state, which reflects in the comparison chart (top right in green section): the predicted case number (blue) is much lower than the actual case number (red).

6 HOTSPOT DETECTION

Spatial hotspot analysis can identify clustering areas of a spatial phenomenon. To help decision makers better understand the geographic patterns of the COVID-19 in the US, we have developed a hotspot detection module in our platform. For this initial version, we have selected two metrics (*cases per 100k population* and *deaths per 100k population*) that capture the prevalence and seriousness of COVID-19 in a region. In the future, the hotspot detection module can easily be extended to detect hotspots for other types of metrics (e.g, positive rate of testing, hospitalization). The raw data for hotspot detection is collected from Johns Hopkins University (JHU) Data Repository, which provides daily update of confirmed positive cases, deaths for the US at county level. The raw data is cleaned and then joined with census data to provide population data, with US county shape file to provide spatial information. For each county, we then calculated the *confirmed case per 100k population* and *deaths per 100k population*.

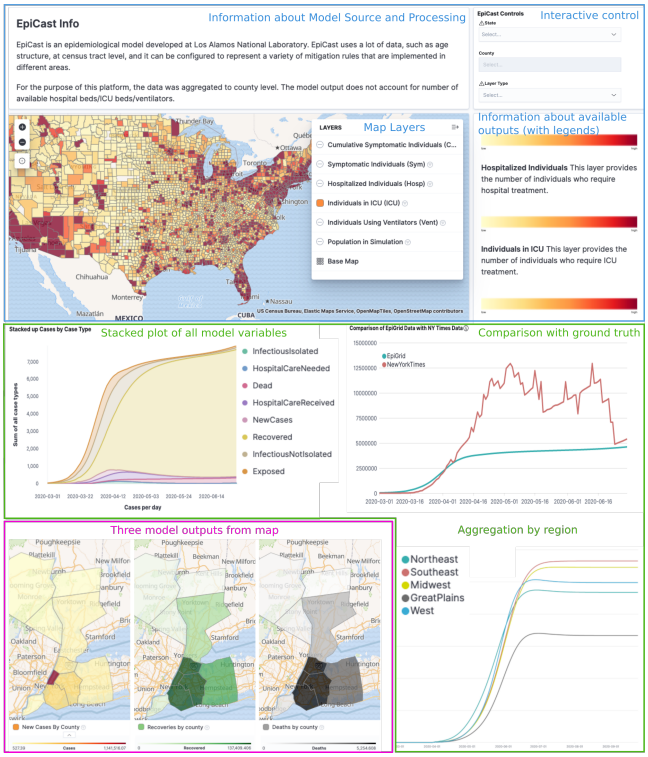


Figure 10: Simulation dashboard layout and some example visualizations from the EpiCast (blue section) and EpiGrid (green and pink sections) dashboards.

6.1 Hotspot detection algorithm

The hotspot detection algorithm uses the Getis-Ord Gi* statistic [12], which works by looking at each county within the context of neighboring counties as well as the national average, to determine whether a county is a hotspot or not.

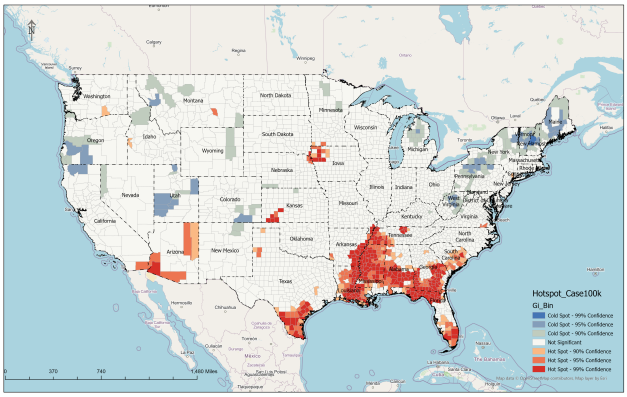
$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}} \quad (2)$$

where $\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$ and $S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$

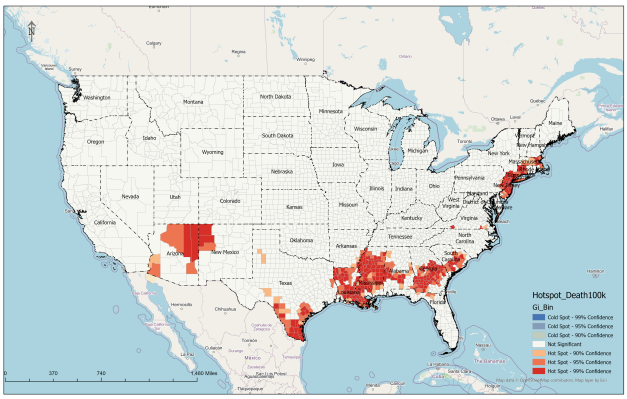
In equation 2, n is the total number of features, x_j is the attribute value of target feature and $w_{i,j}$ is the spatial weight between feature i and j. As we can see from equation 2, Gi* statistic accounts for both national average and neighborhood average. A county that has a high value and is surrounded by other counties with high values as well is a statistically significant hot spot.

6.2 Hotspot visualization

The Gi* statistic calculated for each county is a z-score. If the z-score is statistically significant, the larger the z-score is, the higher confidence we have about the clustering of high values (hot spot). Since the significance of each county is tested individually during the hotspot detection, there could be false positive due to multiple testing. Therefore, we have to calculate the corrected p-value cut-off to correct the bias of multiple testing [3]. The corrected p-value is used for hotspot visualization in our platform.



(a)



(b)

Figure 11: Hotspot visualization.(a) case per 100k population. (b) death per 100k population.

7 EXTENDING THE PLATFORM

In an effort to add extensibility to the platform and allow additional flexibility in web visualization, middle-ware was developed and added to the COVID-19 platform. The middle-ware is based off of the open source project Koop developed by Esri. Koop is a Node.js web server that translates GeoJSON stored in native formats and locations into RESTful Web services. Using the ElasticSearch plugin, Koop was integrated with the existing COVID-19 platform. The output from Koop is Web Feature Services (WFS) which are usable by many web mapping platforms and frameworks. In this case, the WFS generated from Koop are used in a deployment of ArcGIS Enterprise where visualization products including web maps and web mapping applications are created and deployed.

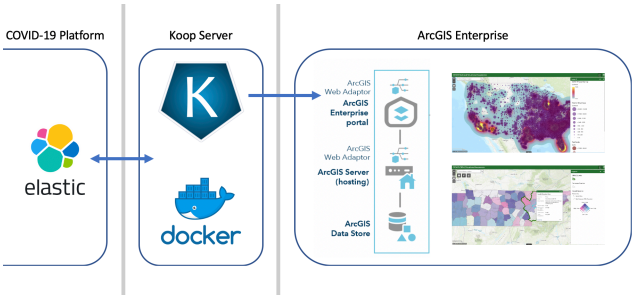


Figure 12: Koop architecture

7.1 Design: Extending Koop

As previously mentioned, Koop with the ElasticSearch plugin, was used to extend the COVID-19 platform, but additional functionality was also developed to customize Koop to fit specific needs within the platform. Koop expects the data being queried to follow the right-hand rule which is a GeoJSON standard that mandates coordinates of exterior rings of a polygon be formatted in a counterclockwise order. Some indices in the COVID-19 ElasticSearch do not follow this standard so functionality was added to reverse the order of the coordinates when necessary. In Figure 12, we show the architecture to connect with COVID-19 platform.

There was also an issue of indices having multiple documents representing the same geometry resulting in a WFS with stacked features. Functionality was developed to allow a service to be configured to only return a single geometry in cases of redundant geometries. The attribution to be returned is also configurable, but as it relates to the COVID-19 platform, the document with the most current date is returned in the case of stacked geometries.

Koop was also extended to account for cases where an index did not include and geometry but had geographic context (e.g county name, state name, FIPS). An additional feature was developed to allow joining the attribution of an index to the geometry of another index. The COVID-19 platform includes indices for both county and state geometries which were used to add geometry to non-spatial indices that had a geographic indicator, in most cases FIPS values.

7.2 Deployment

Because Koop is a simple and lightweight Node.js web server there are many options for deployment that offers flexibility and scalability. One limiting factor, however, is that Koop is single threaded and performance degrades as more services are being generated and used in front end applications. To counteract this, Docker was used to deploy multiple instances of Koop. The deployment strategy was to have a separate Docker container for each WFS output. In some cases, this included multiple containers and services for a single ElasticSearch index.

7.3 Use Case

The extended version of Koop deployed on an array of Docker containers was used to integrate the COVID-19 Platform to a deployment of ArcGIS Enterprise. ArcGIS Enterprise includes ArcGIS Portal which allows users to create and share maps and applications. The WFS generated from Koop that has been configured to the COVID-19 platform's ElasticSearch instance can be added directly into a web map or web mapping application in ArcGIS Portal seamlessly. An application was developed to give a national overview of the current state of the COVID-19 crisis. Also, a state level application was also created to provide a more detailed look at the situation. Both of these applications leverages openly available services, services generated from ArcGIS Enterprise, and services coming from Koop and the COVID-19 platform.

8 PERFORMANCE AND LOAD TESTING

The platform's stability and reliability is evaluated using a suite of non-functional tests that specifically evaluates the readiness of a system. Some examples include load testing, performance testing, availability testing, etc. This is achieved to determine a platform's behavior under both normal and at peak conditions. For this study, we perform load testing to evaluate simultaneous user access and measure network performance.

8.1 Basics

For load testing the COVID-19 platform, k6[1] was used, an open-source load testing tool. To visualize and track load testing metrics on the VM hosts, Prometheus and Grafana were deployed for collection and visualization of the metrics. To generate host metrics, Node Exporter provided a way to constantly expose metrics over a port number and Prometheus was then configured to scrape those metrics by providing the target IPs to Prometheus' configuration. The community dashboards available for Node Exporter provided visualization of the VM specific metrics. The official k6 Grafana dashboard provided details of the load test, including HTTP request durations, HTTP requests per second, etc. An InfluxDB instance was created as well, as k6 provides native support on sending metrics directly to InfluxDB.

8.2 Approach

To create the capable script to be used by k6, its recorder chrome plugin was used to record browser actions and convert them to the script. This included logging in and loading various layers in Kibana dashboards. After modifying the scripts to allow variable overrides, a Docker container was created that loaded the scripts vi CI/CD in GitLab. Then using a batch job deployment in Kubernetes, the load test could begin with various user count simulation and duration. Running this as a Kubernetes job offered the capability to scale the job, deploying multiple parallel running containers. This was used to generate load on the system with environment variable overrides selecting user count simulation and duration of the test. This also permitted the selection of various scripts available in the container for different dashboards within covid19.ornl.gov, without the need of multiple unique docker images.

8.3 Benchmark Results

Our test methodology was to generate load to certify that the COVID19 website was capable of supporting 1,000 simultaneous users. The load testing was run in four stages: 100, 250, 500, and 1,000 simultaneous users. In Figure-13, requests per second generated by 1,000 users is shown. The performance was as expected with low latency and higher throughput as show in Figure-14.

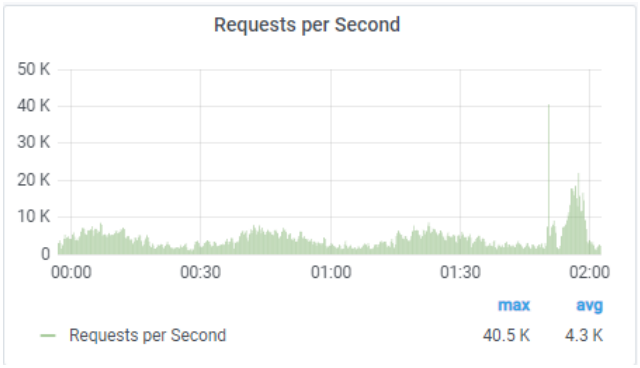


Figure 13: Loading testing for 1000 user shows number of request simulated per unit time.

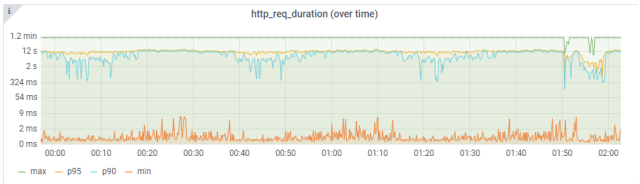


Figure 14: Request frequency and performance latency

We ran a peak test at 1500 users to determine how the system would behave and if performance would degrade beyond the 1000 user limit. Apache reverse proxy was unable to process more than 1,024 users and should be replaced with HAProxy or another proxy engine should the number of simultaneous users exceed 1000.

9 CONCLUSION

In this work, we discussed the development of an all-encompassing operational platform for non-pharmaceutical interventions to the COVID-19 pandemic. The platform is deployed at https://covid19.ornl.gov/ and accessible to authorized users. The underlying scalable architecture supports an end to end workflow for joint pandemic modeling and analysis towards policy guidance and decision making. Custom visualizations are added to display a complex relationship among various data set and the user-facing part of this platform is organized as a series of dashboards. We hope the integration of various datasets, predictions, and simulation results will provide a complete picture to decision-makers for policy guidance.

ACKNOWLEDGMENTS

Research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

REFERENCES

[1] [n.d.]. GitHub - loadimpact/k6: A modern load testing tool, using Go and JavaScript - https://k6.io. https://github.com/loadimpact/k6 Accessed 2020-09-23.
[2] Azavea Aaron Su, Anthropocene Labs Dave Luo, Azavea Hector Castro, Azavea Jeff Frankl, Lauren Moos, Azavea Matt McFarland, Azavea Rob Emanuele, Azavea Simon Kassel, and Development Seed Zhuangfang NaNa Yi. 2020. CovidCareMap – Open geospatial work to support health systems' capacity to effectively care for rapidly growing COVID19 patient needs. https://www.covidcaremap.org/
[3] Marcia Caldas de Castro and Burton H. Singer. 2006. Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association. *Geographical Analysis* 38, 2 (2006), 180–208. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0016-7363.2006.00682.x
[4] A. M. Caulfield, E. S. Chung, A. Putnam, H. Angepat, J. Fowers, M. Haselman, S. Heil, M. Humphrey, P. Kaur, J. Kim, D. Lo, T. Massengill, K. Ovtcharov, M. Papamichael, L. Woods, S. Lanka, D. Chiou, and D. Burger. 2016. A cloud-scale acceleration architecture. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 1–13.
[5] Randy Chow and Theodore Johnson. 1997. Distributed Operating Systems & Algorithms. , 569 pages. https://books.google.com/books?id=J4MZAQAAIAAJ&source=gbs_book_other_versions Accessed 2020-09-22.
[6] Delphi Research Group. 2020. COVIDcast - Real-time COVID-19 Indicators. , 10 pages.
[7] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet. Infectious diseases* 20, 5 (2020), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1
[8] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 20, 5 (2020), 533 – 534. https://doi.org/10.1016/S1473-3099(20)30120-1
[9] Facebook. 2020. COVID-19 Mobility Data Network. , 10 pages. https://visualization.covid19mobility.org/?date=2020-09-23&dates=2020-06-23_2020-09-23&region=WORLD Accessed 2020-09-25.
[10] Paul Fenimore, Benjamin McMahon, Nicolas Hengartner, Timothy Germann, and Judith Mourant. 2018. A Suite of Mechanistic Epidemiological Decision Support Tools. *Online Journal of Public Health Informatics* 10, 1 (may 2018), e1. https://doi.org/10.5210/ojphi.v10i1.8299
[11] Frauke Kreuter, Kathleen Stewart, Andres Garcia, Yao Li, Joe O'Brien, Junchuan Fan, Samantha Chiu, and Anil Kommareddy. 2020. COVID-19 World Symptom Survey. https://covidmap.umd.edu/ Accessed 2020-09-25.
[12] Arthur Getis and J. K. Ord. 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24, 3 (1992), 189–206.
[13] John Graham. 2012. *Missing data: Analysis and design. New York, NY: Springer.*
[14] James Hadfield, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. 2018. NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* 34, 23 (2018), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407
[15] Apache Hadoop. 2011. Apache hadoop. *URL http://hadoop. apache. org* (2011).
[16] IHME COVID-19 Team and Simon I Hay. 2020. COVID-19 scenarios for the United States. *medRxiv* (2020). https://doi.org/10.1101/2020.07.12.20151191
[17] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja. 2015. Lambda architecture for cost-effective batch and speed big data processing. In *2015 IEEE International Conference on Big Data (Big Data)*. 2785–2792.
[18] Jay Kreps. 2014. Questioning the Lambda Architecture – O'Reilly. , 10 pages.
[19] Yen Ting Lin, Jacob Neumann, Ely Miller, Richard G Posner, Abhishek Mallela, Cosmin Safta, Jaideep Ray, Gautam Thakur, Supriya Chinthavali, and William S Hlavacek. 2020. Daily Forecasting of New Cases for Regional Epidemics of Coronavirus Disease 2019 with Bayesian Uncertainty Quantification. (2020).
[20] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. 2015. Comparison of different Methods for Univariate Time Series Imputation in R. arXiv:1510.03924 [stat.AP]
[21] Valerio Persico, Antonio Pescapé, Antonio Picariello, and Giancarlo Sperlí. 2018. Benchmarking big data architectures for social networks data processing using public cloud platforms. *Future Generation Computer Systems* 89 (2018), 98 – 109. http://www.sciencedirect.com/science/article/pii/S0167739X17328303
[22] FBK Researcher. 2020. Covid19 Infodemics Observatory. https://covid19obs.fbk.eu/. (Accessed on 09/25/2020).
[23] Haytham Salhi, Feras Odeh, Rabee Nasser, and Adel Taweel. 2018. Benchmarking and Performance Analysis for Distributed Cache Systems: A Comparative Case Study. In *Performance Evaluation and Benchmarking for the Analytics Era*, Raghunath Nambiar and Meikel Poess (Eds.). Springer International Publishing.
[24] A. Sanla and T. Numnonda. 2019. A Comparative Performance of Real-time Big Data Analytic Architectures. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. 1–5.
[25] NY Times. 2020. Coronavirus in Texas: Map and Case Count - The New York Times. , 10 pages. https://www.nytimes.com/interactive/2020/us/texas-coronavirus-cases.html Accessed 2020-09-25.
[26] Brian Tomaszewski and Alan M MacEachren. 2012. Geovisual analytics to support crisis management: Information foraging for geo-historical context. *Information Visualization* 11, 4 (2012), 339–359.
[27] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, Ion Stoica, et al. 2010. Spark: Cluster computing with working sets. *HotCloud* (2010).
[28] Guoqiang Zhang, Yang Li, and Tao Lin. 2013. Caching in information centric networking: A survey. *Computer Networks* 57, 16 (2013), 3128–3141.

Using Animation to Visualize Spatio-Temporal Varying COVID-19 Data

Hanan Samet
University of Maryland
hjs@umd.edu

Yunheng Han
University of Maryland
yhhan@umd.edu

John Kastner
University of Maryland
kastner@umd.edu

Hong Wei
University of Maryland
hyw@umd.edu

ABSTRACT

CoronaViz (http://coronaviz.umiacs.io) is a research prototype developed by us to enable the dynamic map visualization of COVID-19 related variables including the number of confirmed cases, active cases, recoveries, and deaths all on a daily basis from the Johns Hopkins University web site at ter.ps/coronajhu, by allowing the underlying spatial region and the spanned time interval to vary. Any combination of the variables can be viewed. subject to a possibility of clutter which is avoided by the use of concentric circles (termed geo-circles) whose radius values correspond to the variable values. The variable values are provided both on cumulative and day-by-day bases. The visualization enables spatial and temporal variation.

ACM Reference Format:
Hanan Samet, Yunheng Han, John Kastner, and Hong Wei. 2020. Using Animation to Visualize Spatio-Temporal Varying COVID-19 Data. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3423459.3430761

1 INTRODUCTION

CoronaViz (http://coronaviz.umiacs.io) is a research prototype developed at the University of Maryland to enable the dynamic map visualization of COVID-19 related variables including the number of confirmed cases, active cases, recoveries, and deaths all on a daily basis from the Johns Hopkins University web site at ter.ps/coronajhu, by allowing the underlying spatial region and the spanned time interval to vary. Any combination of the variables can be viewed. subject to a possibility of clutter which is avoided by the use of concentric circles (termed geo-circles) whose radius values correspond to the variable values. The variable values are provided both on cumulative and day-by-day bases. Some like the number of confirmed cases and deaths are also reported as a result of normalization with respect to a measure such as per 100,000 inhabitants (termed an incidence rate). Others like the number of deaths are normalized with respect to the number of confirmed cases (termed a mortality rate). The visualization enables spatial and temporal variation.

CoronaViz was motivated by the continuing spread of COVID-19 which led to the desire to track its progress over time to be better prepared to anticipate its emergence in new regions. There exist numerous systems to monitor and map officially released numbers of cases [3], which are the current established means of keeping track of the progress of the virus. However, as we mentioned earlier, these systems do not necessarily paint a complete picture. For example, they are primarily mashups in that they do not support zooming in on the map in the sense that they just increase the resolution of the map but do not show the data for the additional units (e.g., states/provinces, counties, etc.) that have become visible as a result of the zoom. The visualization enables the comparison of disease-related variables pairwise or region-wise. Particular attention is paid to proper scaling of the disease-related variables so that we can visualize them even if they are all small values or large values in terms of magnitude. To run the system, preferably using the Google Chrome or Microsoft Edge browsers on a laptop or desktop, go to http://coronaviz.umiacs.io

The rest of this paper is organized as follows. Section 2 discusses the queries our system is able to support. Section 3 reviews related work by discussing existing disease monitoring systems. Section 4 describes the CoronaViz user interface, while Section 5 provides examples of the use of CoronaViz that highlight its utility. Section 6 contains concluding remarks and discusses directions for future work.

2 QUERIES

The values of all of the variables in CoronaViz are presented in a time-varying manner as time moves on with the aid of a time slider thereby leading them to be characterized as *dynamic* variables. This is in contrast to visualization tools where such variables are presented in a graph where time is the horizontal axis and the variable value is the vertical axis thereby leading them to be characterized as *static*. Thus we see that the presentation manner is the key to the characterization. It is not easy to present several static variables as they tend to clutter the display regardless of whether they are represented as one graph for the set of all variables or one graph per variable. The situation becomes more complex when values of the variables vary in a spatially-varying manner. In this case the only way to deal with the static variables is to repeat the graph at each location. This is OK when the data is spatially sparse but this is not something we can count on.

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430761

In contrast, CoronaViz deals with variables that are both time-varying and spatially varying by re-examining the dimensionality of the data in the sense that a time slider is a natural representation of one-dimensional data (i.e., time) while a two-dimensional map is a natural representation of two-dimensional data. The problem is how we represent the values of the variables. One possible solution is via a histogram but this leads to clutter on the display and is cumbersome when the data is not spatially sparse. Moreover, there may be a layout problem here in the sense that we cannot allow the histograms to overlap. An alternative common solution with the same overlap issues is to use solid concentric circles where the radii of the circles correspond to the value and the color corresponds to the identity of the variable. This type of visualization is known in cartography as a *proportional symbol map* [7]. The problem here is when we have multiple dynamic variables as is the case for our application, then only the one with the largest magnitude can be viewed. One solution is to vary the colors of the circles but if this method is used, then we must pay close attention to the order in which we display the circles so that the one with the largest radius is displayed first and the remaining circles be displayed in decreasing order of radius values (this is analogous to the "back-to-front" z-buffer display algorithm used in computer graphics). We can avoid the need to worry about the order in which we display the circles by using hollow concentric circles where again the color indicates the identity of the variable while the radius corresponds to a scaled variable magnitude. We use the term *geo-circle* to describe this approach.

The visual strain posed by having a large number of circles can be relieved by drawing the circles using broken lines of the same width. At times, the width of the broken lines can be increased with the goal of drawing attention to a particular set of concentric circles (i.e., a location whose variable values at a particular instance of time) which is of interest. We do this in the case of a hover operation while panning on the map to show the spatially closest location with nonzero variable values. This operation is common in computer graphics where it is known as a "pick" operation (e.g., see [5]). However, care must be exercised when implementing it in the sense that we don't always want the closest geo-circle. For example if we are hovering in Brazil, then we want the geo-circle of Brazil even though the geo-circles of Paraguay or Bolivia may be closer to the hover location in Brazil.

CoronaViz makes use of 7 dynamic variables comparing the number of confirmed cases. active cases, recoveries (although not reported by all jurisdictions), and deaths, as well as normalized variants which include the incidence rate (number of confirmed cases per 100,000 inhabitants), mortality rate (number of deaths divided by the number of confirmed cases), and the recovery rate (number of recoveries divided by the sum of the numbers of deaths and recoveries). No active rate is tabulated as the number of active cases is simply the number of confirmed cases minus the number of deaths and recoveries and thus the only possible rate measure is a normalized active cases value per 100,000 inhabitants which is similar to the incidence rate and thus we do not provide it. Concentric circles (i.e., geo-circles) drawn with broken lines are used for th 4 disease related variable values while concentric circles drawn with solid lines are used for the 3 disease related rates. They are drawn with different colors with the same color being used for the corresponding variable and rate.

The concentric circles make it easy to spot trends and similar values on the map by looking at the magnitude of the radii. Other observations of interest involve trends such as noting lower confirmed case and death counts over time as the circles get smaller. Another encouraging trend is when confirmed case counts become smaller than death counts. Of particular interest is the situation when concentric circles intersect and change their relative order. Of course this must be treated with caution as the magnitudes of the variables change). In particular, of a comparison is only meaningful when comparing variable values and not rates.

There are a number of ways of presenting the variable values. The default in our case is of a cumulative nature. However, it is possible to normalize the values over population, or even area. Normalizing over the area is of possible interest as it could be used to see if densely populated areas are more likely to lead to a higher number of confirmed cases of COVID-19 and deaths.

Our goal is to endow CoronaViz with a full compliment of queries that are consistent with its role as a spatiotemporal database. First of all, we have two types of queries:

(1) location-based: given a location or time, what are the values of certain variables and rates.
(2) feature-based: given a variable or rate value, where or when is its value present. This is also known as spatial data mining [2]. In CoronaViz we might be looking for locations or time instances where there are no deaths.

The location-based queries are supported by the ability to pan the map with a hover operation and always returning the variable values with the nearest location for which we have data. CoronaViz supports this query by a PR quadtree where we have one PR quadtree for each of the disease-related dynamic queries or variables or rates. Feature-based queries require the use of a pyramid-like data structure on each of the disease-related dynamic queries or variables or rates.

The animation window is a very important feature as it enables the execution of a range query where the range is temporal. Users can vary the start and end times of the query as well as the animation step size. In addition, users can specify what statistic is being computed for the temporal window. It can be cumulative, or a time period whose length can be in terms of days, weeks, months, or even years. Average values for the window can also be computed. This is particularly useful for the "reopen" discussion which is often based on a rolling weekly daily average computation involving the number of confirmed cases.

Spatial range (also known as window) queries are of great interest. In this case, users use pan and zoom operations to get a map that is focused on a particular desired spatial region (e.g., the minimum bounding areal box that contains Italy). Note that here we find overlap with San Marino and the Holy See (i.e., the Vatican in Rome). In particular, we have one geo-circle that displays the sum of the values of the dynamic variables for all three of these spatial entities. In order to restrict the visualization to Italy, users must zoom in further so that San Marino and the Holy See are not in the window (i.e., the displayed geo-circle). Alternatively if users only want the Italy, then they could simply pose the textual query

with "Italy" as the search string as well as the name of a region such as "Liguria" or city such as "Genoa" for which appropriate indexes exist. Note that as Coronaviz zooms into a region, it has access to more data (as low as county or city level data).

CoronaViz enables the execution of the full compliment of spatiotemporal queries as it supports keeping location fixed while varying time via the time slider, keeping time fixed and letting location vary via the hover, panning, and zooming operations. We can also pick any range of time or space. Users can also take advantage of spatial synonyms when they don't know the exact name of the location of interest. For example, when seeking a "Rock Concert in Manhattan," concerts in Harlem, New York City, and Brooklyn are all good answers because of being contained in Manhattan, containing Manhattan, and being a spatially adjacent borough, respectively. This is an example of a proximity query which we saw previously via the use of a hover operation in the case of spatial proximity, and the time slider in the case of temporal proximity. Note that for temporal proximity, we provide the capability to halt an animation at arbitrary time instances as well as resuming or terminating it. In addition, users are also able to set the speed of the animation, as well as to step through an animation by a specific time interval both forward and backward in time.

3 RELATED WORK

In this section we first briefly consider prior work dealing with the visualization spatiotemporal data and then review a number of existing systems designed specifically for monitoring the spread of COVID-19.

3.1 Spatiotemporal Data Visualization

Visualization and analysis of temporally varying geospatial data is a difficult task; as such, it has been the subject of substantial prior work. The difficulty comes from the inherently multidimensional nature of the data: there are at a minimum two spatial dimensions and one temporal dimension, in addition to the dimensionality added by the actual variables being visualized. All of these dimensions must be projected onto a two dimensions screen. We can broadly break spatiotemporal visualization techniques into two groups: those that use animation to capture the time dimensions, and those that attempt to encode temporally varying information into a single static visualization.

An example of this second variant is presented by Du et al. [4] who modify the traditional choropleth map to encode temporal information inside each area unit. Rather than picking a single color for each areal unit, units are divided either into bands of either equal width or equal area. Each band is then assigned a color in the same way areal units are assigned colors in traditional choropleth maps (e.g. Howard et al. [7]).

Li et al. [11] do not use a fully animated approach, but neither do they commit to showing the full temporal data range in a single image. Instead, they use an interface termed the "Event View" to display images generated for discrete time intervals side-by-side. To link these images together into a single cohesive visualization, the authors overlay a "trend line" that connects the time intervals. This trend line is used to link events extracted by a separate component of their system.

Very often a temporal variant of a well known cartographic visualization technique can be obtained by applying the existing technique to data within a time window for a series of time windows. An animation is obtained by collecting the individual visualization and displaying them in order by time. This is approach the basis of Ouyang and Revesz [15] who develop an algorithm to generate spatiotemporal cartogram animations.

3.2 Existing COVID-19 Monitoring Systems

In this subsection we review a number of existing systems designed specifically for monitoring the spread of COVID-19. These systems are described below with an emphasis on pointing out their drawbacks thereby motivating our work in developing CoronaViz.

(1) https://coronavirus.jhu.edu/ Coronavirus COVID-19 global cases (Johns Hopkins)
(2) https://www.healthmap.org/ncov2019/ Novel Coronavirus (COVID-19) outbreak timeline map (HealthMap)
(3) https://news.google.com/covid19/map (Google News)
(4) https://hgis.uw.edu/virus/ Novel coronavirus infection map (University of Washington)
(5) http://nssac.bii.virginia.edu/covid-19/dashboard/ COVID-19 surveillance dashboard (University of Virginia)
(6) https://covid19.who.int/ Novel coronavirus (COVID-19) situation dashboard (WHO)
(7) https://www.cdc.gov/coronavirus/2019-ncov/cases-in-us.html Coronavirus disease 2019 in the US (CDC)
(8) https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases Geographical distribution of COVID-19 cases worldwide (ECDC)
(9) https://www.kff.org/global-health-policy/fact-sheet/coronavirus-tracker/ COVID-19 coronavirus tracker (Kaiser Family Foundation)
(10) https://www.worldometers.info/coronavirus/ COVID-19 coronavirus outbreak (Worldometer)
(11) https://multimedia.scmp.com/infographics/news/china/article/3047038/wuhan-virus/index.html Coronavirus: the new disease Covid-19 explained (South China Morning Post)
(12) https://storymaps.arcgis.com/stories/4fdc0d03d3a34aa485de1fb0d2650ee0 Mapping the Wuhan coronavirus outbreak (ESRI StoryMaps)
(13) https://public.flourish.studio/visualisation/1539110 (Flourish)
(14) https://coronavirus.1point3acres.com/en (1point3acres)
(15) https://geods.geography.wisc.edu/covid19/physical-distancing/ (University of Wisconsin)

The Johns Hopkins system tabulates cumulative numbers of confirmed, active, deaths, and recoveries. The cumulative numbers of confirmed and active cases in some of the countries are displayed on the map for some of the larger countries (in terms of area). A drawback of the maps is that zooming in on the map simply increases the resolution of the map but does not show the data for additional countries. This is a common drawback of many of the systems that have been created for visualizing the coronavirus. This is not the case for CoronaViz.

The HealthMap system shows the spread of the disease by tabulating the number of new confirmed cases of the disease on a daily

basis and displaying it with a circle of a particular size and color anchored at the location where it was reported (e.g., a city, state, country, etc.). HealthMap still has the drawback that zooming in only increases the resolution of the map but does not show a finer allocation of the tabulated properly to the location.

The Google News system makes use of a map query interface and allows zooming in and reports the variable values for the smaller subunits. It uses a hover operation to yield the variable values for the spatial unit being hovered over, as well as disease-related news at times. It does not have the ability to provide variable values for a combination of units that make up the viewing window when these units are small (e.g., counties) or bigger (countries) as is done in CoronaViz. It is static as it has no temporal component other than precomputed graphs of variable values over a predetermined range of days unlike CoronaViz where the range is set by the user.

The University of Washington system shows the total number of confirmed cases, deaths, and recovered for the countries of the world as one pans the world map. For the US, zooming in has a greater granularity and results in showing how the number of confirmed cases are spatially distributed in each state. Descriptive data is also provided for the confirmed individuals when the region is sufficiently small.

The Flourish system enables the visualization of just one dynamic variable such as the number of confirmed cases in a number of countries at the same instance of time. Although the data is spatially-referenced by name (i.e., the names of the countries) no use is made of a map nor are there any input or output controls. The one advantaged of the system is that it is fast which conveys the urgency of the need to stop the spread of the disease.

Both the 1point3acres and Worldometer systems provide comprehensive data and graphs for the dynamic variables but no animation or maps. The dynamic aspect of the variables is captured by the various plots of the variable values and combinations thereof. They make a distinction between cumulative variable values as well as new values. The 1point3acres system prides itself in its data collection ability and is more focused on the virus while the Worldometer system also provides statistics related to the impact of the disease such as unemployment.

The University of Virginia system displays the number of cumulative confirmed cases, deaths, and recovered over time using a time slider. The countries are colored according to the range of the number of individuals for the variable being displayed. Zooming in results in more locations being placed on the map as well as the inconsistent decomposition into smaller units such as states for the US and provinces for China but not for Canada or Australia.

The remaining systems are quite similar in that they only map the number of confirmed cases in each country in the case of the WHO and ECDC systems and in each state for the CDC system. The Kaiser Family Foundation system also maps the deaths. None of the WHO, ECDC, CDC, and the Kaiser Family Foundation systems permit zooming in to get additional data. Non-interactive maps are used to tell the story of the coronavirus outbreak in the South China Post using ESRI StoryMaps. Instead of the disease-related variables some systems like that from the University of Wisconsin look at a variable that monitors the mobility of the population with a map query interface that makes use of cell phone data.

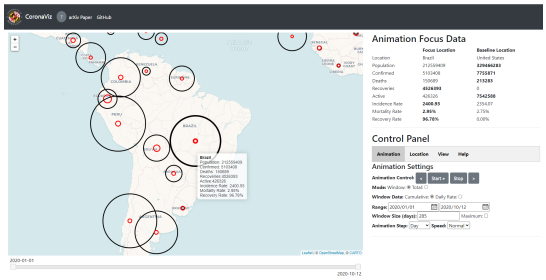
4 USER INTERFACE

CoronaViz's user interface is anchored by the "Control Panel" which is partitioned into four components corresponding to the three tasks of the system which are data animation, location specification, and data viewing, and help. They are accessed by appropriately named buttons.

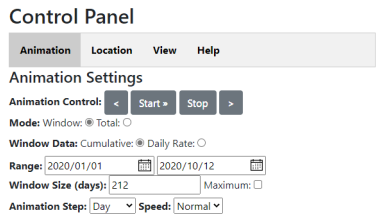
Figure 1a shows the incidence and mortality rates in South America for a 212 day wide window in an animation range spanning from the first of this year 2020 through October 12 of the same year. We see that the query window and the animation range are the same, and thus no animation can be performed as we just have one time instance and an animation requires at least two time instances. Therefore, the result of the query is just a screenshot of the geo-circles each one of which consists of a linearly scaled incidence rate and a linearly scaled mortality rate corresponding to the countries in South America. From the figure we see that the incidence rates are relatively similar for these countries. Mortality rates are much smaller and thus we may wish to scale them so that we can better differentiate between them. Figures 1b-1e show the different control panels and their settings for the query whose result is shown in Figure 1a. The different control panels are described in greater detail in the rest of this section.

The "Animation" button controls the animation process. CoronaViz can be run in two animation modes: "Total" and "Window". In Window mode we provide a temporal region w (termed the "Animation Window") which is specified in terms of days, and a location (i.e., spatial region) which is a geographic entity. In order to simplify the explanation, we use a variant of the example query of animating the progression of COVID-19 in Brazil (See Figure 2 for its result) and its neighboring South American countries in terms of the values of the confirmed cases and deaths disease-related variables. This is done for the "Animation Range" which is set by default to the period between the first of this year 2020 though October 12 of the same year. The animation can provide either the cumulative values of these variables or the daily average value for the days making up the window w. Note that when the window duration is one day, then the cumulative value and the average daily rate are the same. This information is provided on a daily basis on the last day of the animation window for each day of the animation range. In contrast, recall that the maximum possible size of the animation window is the duration of the animation range in which case there is no effective animation as the result is the cumulative value of the variables and the daily average value of the variables over the entire "Animation Range" and only reported on the last day of the animation window. Users don't have to know the value of this maximum as it is specified by checking the "Maximum" checkbox which appears to the right of the "Window Size" in the "Animation Control Panel".

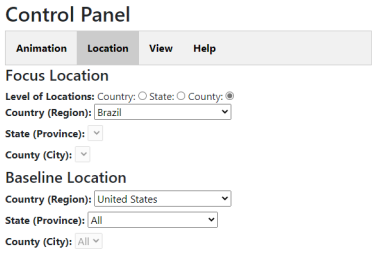
Users who wish to see the cumulative as well as the average daily rates for all of the disease-related variables in an animated manner, should use the "Total" mode. In this case, we do have an animation on a daily basis with the final frame of the animation yielding the cumulative values of the disease-related variables for the temporal "Animation Range" for all spatial ranges that can be viewed. These features are all accessed by clicking on the "Animation" button in



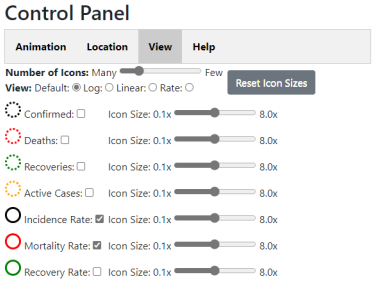
(a) overview



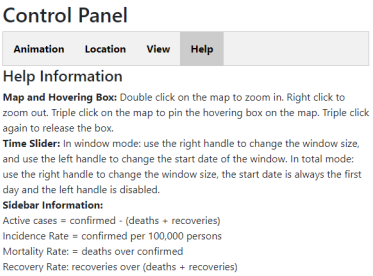
(b) animation control panel



(c) location control panel



(d) view control panel



(e) help information

Figure 1: user interface

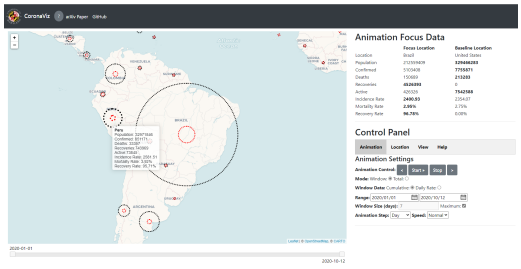


Figure 2: example CoronaViz screenshot for Brazil in South America.

the "Control Panel". Figure 2 is the final screenshot for the animation of the cumulative values of the number of confirmed cases, deaths and recoveries for the Total mode query for the countries in the vicinity of Brazil for the time period between the first of this year 2020 though October 12 of the same year. In essence, the screenshot differs from the one in Figure 1a for Figures 1b-1e by displaying the values of the disease-related variables instead of the rates. Note the larger geo-circles on account of no normalization which is the case when we used rates for the disease-related variables.

The "Location" button activates the "Location Specification" process which identifies the spatial entity for which we wish to animate and view the disease-related variables. This location is known as the "Animation Focus". It can be the name of a country/region, state/province, or county/city all of which are obtained from an appropriately named pull-down menu. Alternatively, the geographical entities can also be specified graphically using direct manipulation actions like pan, zoom, and hover. In this case we usually start with a map from which a new map is constructed using pan and zoom operations as well as possibly dilation. Once the desired location has been identified on the map, then a single left click on the mouse is sufficient to initialize or reset the "Animation Focus".

The advantage of the direct manipulation approach is that it provides the query poser the opportunity to specify the exact shape and boundary, as well as the resolution, of the query region. The "Location" button can also be used in the same manner to set what we call a "Baseline Location" for comparing disease-related data as the animation proceeds. This location can only be set using the pull down menus. It cannot be set using direct manipulation. As the animation proceeds, the values of all of the disease-related variables and rates are displayed side-by-side in the "Animation Focus Data Panel" for the two locations.

At this point, the animation can be started by clicking on the "Start Animation" button in the "Animation Control Panel". In our example, we see the animation of the progression of COVID-19 in all of the South American countries with a focus on Brazil in terms of the counts of confirmed cases, deaths, and recoveries (see Figure 2).

The "View" button in the "Control Panel" controls the viewing process by providing a number of options of viewing the 7 disease-related rates and variables. The user selects them by clicking the checkbox to their right. Some combinations of variables are predefined such as the "Default" view corresponding to just displaying the incidence and mortality rates, while the "Rate" view corresponds

to just displaying the incidence, mortality, and recovery rates. As the animation proceeds, the values of the selected disease-related variables and rates are displayed using geo-circles anchored at the corresponding geographic location and whose radii provides an indication of their relative magnitudes. We have several options for the radii: linear and logarithmic, neither of which are always satisfactory. Using a "linear" relationship breaks down when radius values differ by more than one order of magnitude. A “logarithmic” relationship is fine for differentiating between small radius values while making it harder to differentiate between large radius values as well as needing to make an appropriate choice of the base (e.g. 2 or natural logarithm instead of 10 which is not good for large radius values). The radii can be scaled by factors ranging from .1X to 8.0X which is useful when the radius values are very large or very small, respectively.

In our example, Brazil is said to be the “Focus Location” which means that as the animation proceeds, users can see additional data for Brazil corresponding to the daily variation of of all of the disease-related variables and rates. It is constantly updated by looking at the panel with the heading “Animation Focus Data” to the right of the map which shows Brazil in this case (see Figure 2).

In addition, during the animation, the mouse can be moved over the visible part of the map (termed “hover”) and the data associated with closest geo-circle (e.g., Peru in Figure 2) is displayed in what we call the “Hover Box” which is initially anchored on the mouse and moves with the mouse until three quick left clicks are performed at which time the Hover Box” is detached and remains in that position until the performance of the next three quick left clicks. However, even though the hover box has been detached from the mouse, it continues to display the data associated with the nearest geo-circle to the mouse. This nearest geo-circle is highlighted with a thicker outline. Figure 2 shows the result of the animation for a time period since the first of this year which is set in the “Animation Control” panel. Note that the “Animation Control” panel enables users to pause, resume, halt, and restart the animation process by clicking on the appropriate button. Users can also run the animation in a day-by-day manner one day at a time in the forward and backward temporal directions via the buttons labeled “<” and “>”, respectively. It is especially interesting to go backwards at the end of the animation by repeatedly clicking on the “<” button found to the left of the “Start/Pause/Resume/Stop” button. The above "playback" can be achieved in a continuous manner by using the mouse to define the width of a window by varying the positions of the left and right tabs of the time slider. This process proceeds by fixing the right tab and varying the left tab as needed. The “playback” is achieved by dragging the left tab in either of the two temporal directions. The right tab is left alone and it follows the motion of the the left tab.

5 EXAMPLES OF THE UTILITY OF CORONAVIZ

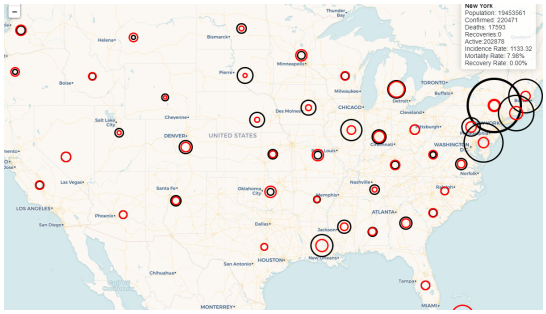
in this section we provide use cases of CoronaViz that demonstrate its utility. Notice that we provide both figures and animations. The figures usually correspond to the last screenshot (frame) of an animation. In most cases we also provide a link to video for the entire animation. The animation can be viewed by clicking on the link in the paper or by cutting the link from the paper and pasting it in the browser (preferably the Chrome or Microsoft Edge browsers).

We first compare the dynamic visualization provided by CoronaViz with conventional methods as used, for example, by newspapers such as the Washington Post for the incidence rates for some of the states in the US (Figure 3) that are two-dimensional graphs where the x axis corresponds to the date while the y axis corresponds to the value of disease-related variables and rates, Figures 3a, 3b and 3c, show the incidence and mortality rates for the United States during April, July, and September of 2020, respectively, They are screenshots from the entire animation[1]. From these three figures we see that there are more confirmed cases in New York than other states in April while New York reported far fewer cases in July. Meanwhile, the number of confirmed cases grows rapidly in the southern and western states. In September, a growing trend of confirmed cases moves to the western and north-central states. This data and the corresponding trends (Figure 3c) are also available on the Washington Post website but it is difficult to draw conclusions about the spatial significance of the relationship between the incidence rates of some spatially adjacent states. Moreover, we can visualize the data of all states on the map in CoronaViz while it is impossible to fit 50 graphs in one page. In addition, CoronaViz can show multiple variables/rates on the map while the graphs could be confusing when many variables/rates are plotted as the graph can support at most two different y-axes interpretations (i.e., one on the left and one on the right ends of the x-axis).

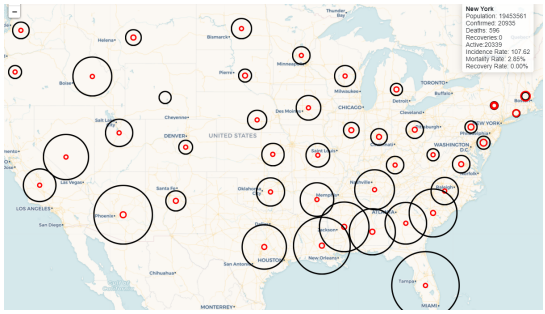
It is often the case that the data values have very small magnitudes as is the case for mortality rates especially when compared with the number of confirmed cases or incidence rates thereby making it difficult to compare their values for different locations. Users can perform more meaningful comparisons of locations with very small data values by changing the scaling factor[2]. For example, Figures 4a and 4c show the mortality rate in the US for September 2020 (window mode) and for Africa in January through October 2020 (Total mode), respectively. It is very small and consequently, the geo-circles representing the mortality rates are also small, which makes it difficult to compare the rates of different locations. By changing the scaling factor of the mortality rate geo-circles, their sizes become larger simultaneously (see Figure 4b and 4d), respectively. After that, the differences in the mortality rate between two locations are clearer. Interestingly, the variation in Africa is much bigger than in the US.

Besides using raw data directly, we often also normalize the data based on population[3]. Figures 5a and 5c show the number of confirmed cases and deaths in South America in September 2020 (window mode) and in North America in January through October 2020 (Total mode), respectively. Here we that some countries like the U.S. and Brazil have a large population and thus they have many confirmed cases, which results in geo-circles with large radii when the raw data is plotted directly. After normalization, the values of the confirmed cases are represented by the incidence rate, which is defined as the number of confirmed cases per 100,000 population. The incidence rate is rarely greater than 3,000 and hence the values of the radii of the geo-circles become reasonable

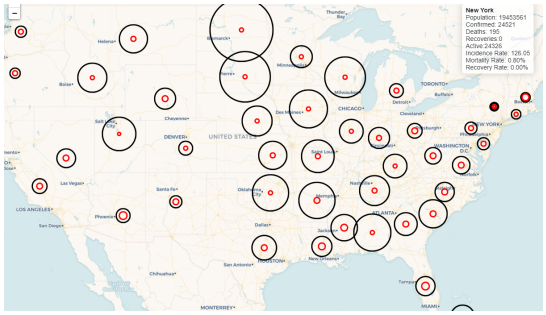
[1] https://www.youtube.com/watch?v=UcDjFLa3I_Y
[2] https://www.youtube.com/watch?v=VLiWoWtYHQo
[3] https://www.youtube.com/watch?v=cCGWQ4jaChw



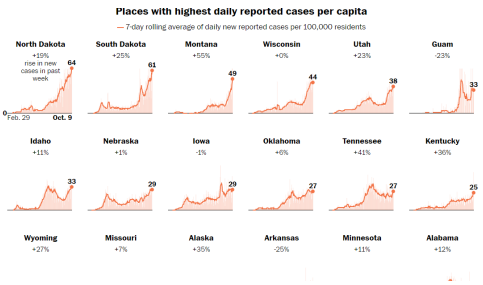
(a) U.S. cases in the Apr. window



(b) U.S. cases in the Jul. window

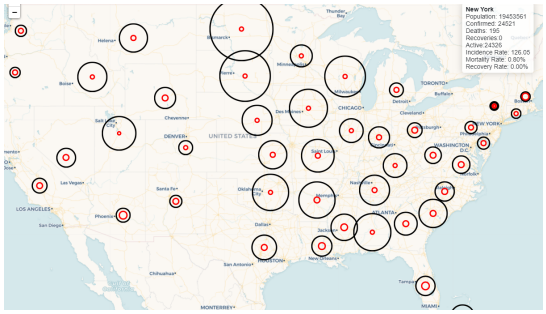


(c) U.S. cases in the Sep. window

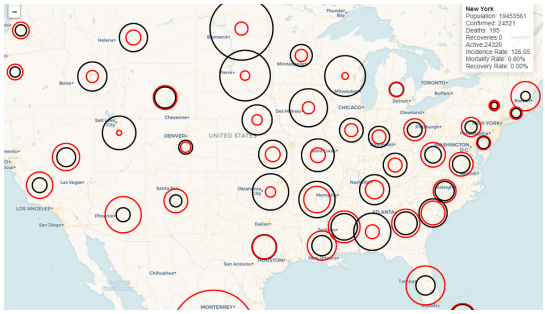


(d) charts from Washington Post

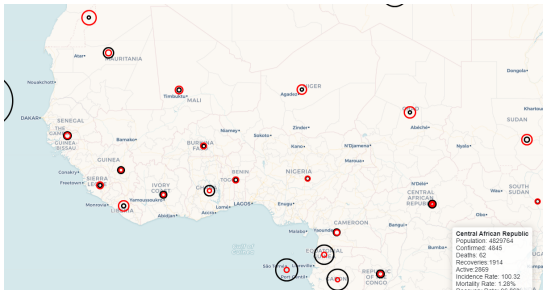
Figure 3: CoronaViz map vs Washington Post charts (video)



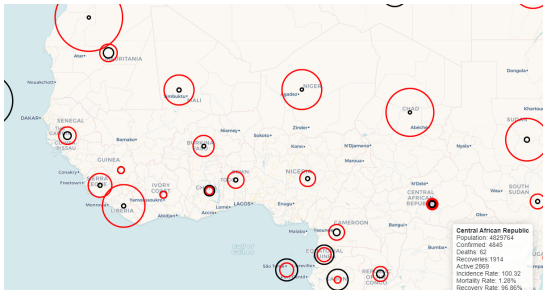
(a) U.S. mortality rate in Sep. (normal icon size)



(b) U.S. mortality rate in Sep. (larger icon size)

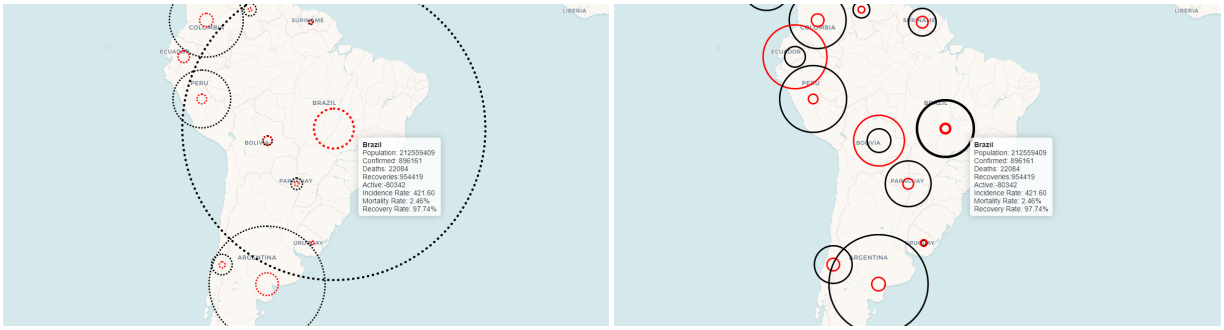


(c) Africa mortality rate in Jan.-Oct. (normal icon size)



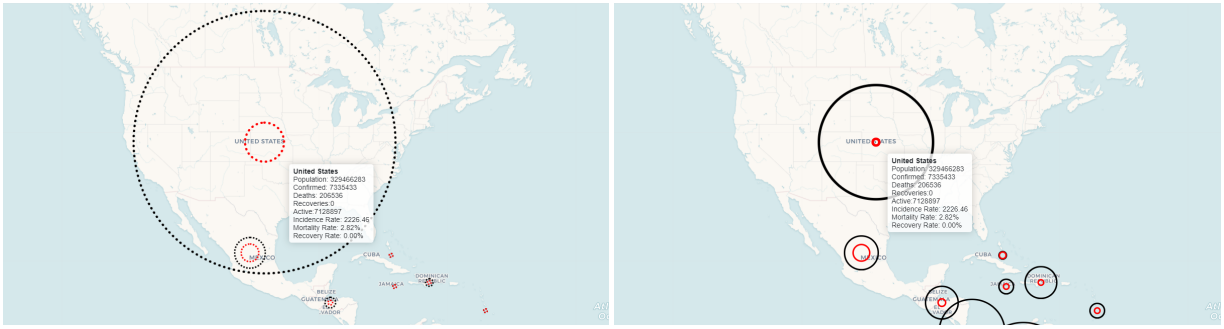
(d) Africa mortality rate in Jan.-Oct. (larger icon size)

Figure 4: choose a proper scaling factor (video)



(a) South America raw data (Sep.)

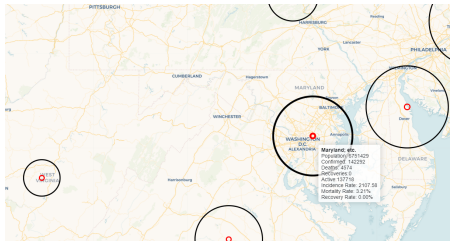
(b) South America normalized data (Sep.)



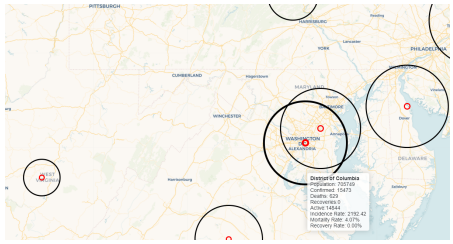
(c) North America raw data (Jan.-Oct.)

(d) North America normalized data (Jan.-Oct.)

Figure 5: normalization (video)

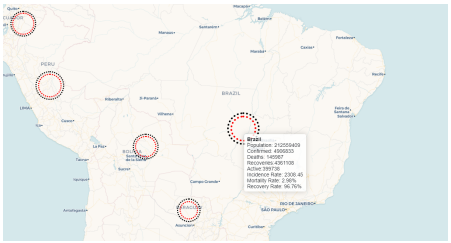


(a) DC and Maryland aggregate

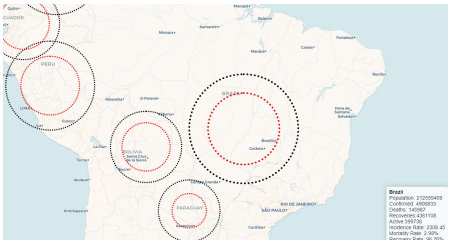


(b) DC and Maryland separate

Figure 6: limit number of icons (video)



(a) log scale (base 10)



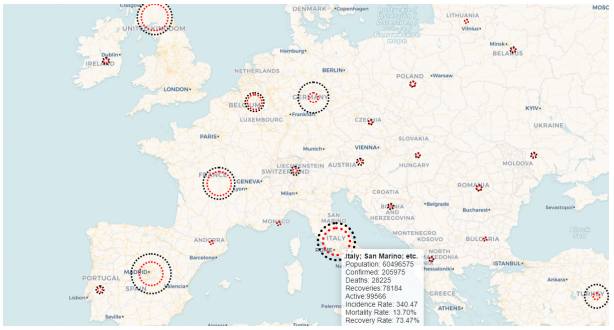
(b) log scale (base e)

Figure 7: log scale

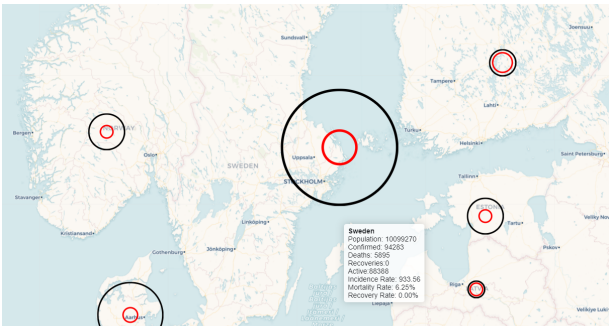
after normalization (see Figures 5b and 5d of (window mode) in January through October 2020 (Total mode), which correspond to Figures 5a and 5c, respectively.

Users can control the number of geo-circles on the map[4]. In CoronaViz, geo-circles aggregate automatically if they are close

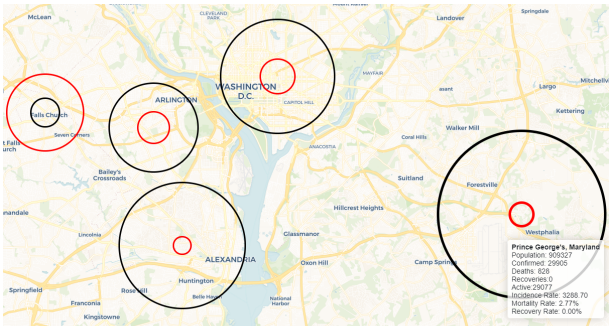
[4]https://www.youtube.com/watch?v=DYHk5XmGXKA



(a) European countries (Mar.-Apr.)



(b) Northern Europe (Jan.-Oct.)



(c) counties near D.C. (Jan.-Oct.)

Animation Focus Data

	Focus Location	Baseline Location
Location	Sweden	Israel
Population	**10099270**	8655541
Confirmed	94283	**264443**
Deaths	**5895**	1682
Recoveries	0	**191251**
Active	**88388**	71510
Incidence Rate	933.56	**3055.19**
Mortality Rate	**6.25%**	0.64%
Recovery Rate	0.00%	**99.13%**

(d) Sweden and Israel

Figure 8: case study (video)

to each other. As shown in Figure 6a, the geo-circles of Washington, D.C. and Maryland usually aggregate unless we zoom in substantially because they are geographically close. By increasing the number of geo-circles plotted on the map, close geo-circles (e.g., D.C. and Maryland in Figure 6b), and more details are shown on the map. This makes comparisons among geographically-proximate locations feasible.

Besides using normalization to reduce the radii of the geo-circles of the disease-related variables, we can also reduce their values by subjecting them to a logarithmic scale. As shown in Figure 7, the geo-circles become smaller when we use the logarithm of the raw values as the radius. However, this also makes the difference between two geo-circles less noticeable. As a result, a smaller base for the logarithm is referable. For example, a natural logarithm (Figure 7b) is better than the base 10 common logarithm (Figure 7a). Note also that using a small base is equivalent to using a larger scaling factor.

We also study some typical cases to show the utility of CoronaViz. In Europe, the pandemic first peaked in late March to early April. As shown in Figure 8a (confirmed cases and deaths), there were several hot spots. Setting the temporal window to be March and April finds them to be the U.K., France, Germany, Italy, and Spain. Another example is Sweden which let the Coronavirus spread in the hope that the population would develop "herd immunity". Figure 8b shows the incidence and mortality rates for Sweden and its neighboring countries for January through October 2020 (Total mode). From Figure 8b, we see that Sweden has higher incidence and mortality rates than its neighboring countries. We can also compare the data through the text information provided in the sidebar[5]. In Figure 8d, we use Israel as a baseline, whose population is close to Sweden. We observe that Israel has a higher incidence rate but a lower mortality rate compared with Sweden. Note that we do not have recovery data than Sweden so it is not shown in Figure 8d. Observe that CoronaViz not only visualizes data of countries but also other administrative divisions like states and counties. For example, some counties near Washington D.C. are plotted in Figure 8c for January through October 2020 (Total mode)

6 CONCLUDING REMARKS AND DIRECTIONS FOR FUTURE RESEARCH

We have seen the utility of animation to keep track of the spread of disease by examining disease-related variables and rates. Our visualization relies heavily on the availability of quantitative data about the presence of the disease provided by the Johns Hopkins University. Additional useful knowledge about the potential progression of the disease can be gained by keeping track of spatially-referenced mentions in news articles as in NewsStand [10, 12, 17], tweets as in TwitterStand [6, 8, 18], documents such as PubMed [13] and ProMED-mail [9, 13], and spreadsheets [1]. This involves geotagging which is the process of recognizing textual references to location as in [14, 16]. Presently we do not make use of such data

[5]https://www.youtube.com/watch?v=QSkI8htZQQo

although we do feel that such an approach is a direction for future research.

Note that we have not provided "positivity" data which indicates the percentage of tests that are positive (i.e., the ratio of the number of confirmed cases and the total number of tests). The problem here is that the number of tests is unevenly reported thereby making it impossible to report this rate accurately. We will incorporate this measure in CoronaViz once testing centers adopt more complete reporting procedures that include this data. Finally, another topic for future investigation is normalization by a country's area.

ACKNOWLEDGMENTS

This work was sponsored in part by the NSF under Grants IIS-18-16889 and IIS-20-41415.

REFERENCES

[1] M. D. Adelfio and H. Samet. Schema extraction for tabular data on the web. *PVLDB*, 6(6):421–432, April 2013. Also *Proceedings of the 39th International Conference on Very Large Data Bases (VLDB)*.

[2] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 265–272, Nashville, TN, April 1990. Also in *Proceedings of the Fifth Brazilian Symposium on Databases*, pages 15–26, Rio de Janeiro, Brazil, April 1990.

[3] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 2020.

[4] Y. Du, L. Ren, Y. Zhou, J. Li, F. Tian, and G. Dai. Banded choropleth map. *Personal and Ubiquitous Computing*, 22(3):503–510, 2018.

[5] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice.* Addison-Wesley, Reading, MA, second edition, 1990.

[6] N. Gramsky and H. Samet. Seeder finder - identifying additional needles in the Twitter haystack. In A. Pozdnukhov, editor, *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN'13)*, pages 44–53, Orlando, FL, November 2013.

[7] H. Howard, R. McMaster, T. Slocum, and F. Kessler. Thematic cartography and geovisualization. 2008.

[8] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In Y. Zheng and M. F. Mokbel, editors, *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN'11)*, pages 25–32, Chicago, November 2011.

[9] R. Lan, M. D. Lieberman, and H. Samet. The picture of health: map-based, collaborative spatio-temporal disease tracking. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health (HealthGIS 2012)*, pages 27–35, Redondo Beach, CA, November 2012.

[10] R. Lan, M. D. Adelfio, and H. Samet. Spatio-temporal disease tracking using news articles. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health (HealthGIS 2014)*, pages 31–38, Dallas, TX, November 2014.

[11] Jie Li, Siming Chen, Kang Zhang, Gennady Andrienko, and Natalia Andrienko. Cope: Interactive exploration of co-occurrence patterns in spatial time series. *IEEE transactions on visualization and computer graphics*, 25(8):2554–2567, 2018.

[12] M. D. Lieberman and H. Samet. Supporting rapid processing and interactive map-based exploration of streaming news. In I. Cruz, C. A. Knoblock, P. Kröger, E. Tanin, and P. Widmayer, editors, *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 179–188, Redondo Beach, CA, November 2012.

[13] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: architecture of a spatio-textual search engine. In H. Samet, M. Schneider, and C. Shahabi, editors, *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems*, pages 186–193, Seattle, WA, November 2007.

[14] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In R. Purves, C. Jones, and P. Clough, editors, *Proceedings of 6th Workshop on Geographic Information Retrieval*, Zurich, Switzerland, February 2010. Article 6.

[15] M. Ouyang and P. Revesz. Algorithms for cartogram animation. In *Proceedings 2000 International Database Engineering and Applications Symposium (Cat. No. PR00789)*, pages 231–235. IEEE, 2000.

[16] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In A. El Abbadi, D. Agrawal, M. Mokbel, and P. Zhang, editors, *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52, San Jose, CA, November 2010.

[17] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*, 57(10):64–77, October 2014.

[18] J. Sankaranarayanan, H. Samet, B. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in tweets. In D. Agrawal, W. G. Aref, C.-T. Lu, M. F. Mokbel, P. Scheuermann, C. Shahabi, and O. Wolfson, editors, *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, Seattle, WA, November 2009.