The SIGSPATIAL Special

Newsletter of the Association for Computing Machinery
Special Interest Group on Spatial Information

Volume 12 Number 2 July 2020

The SIGSPATIAL Special

The SIGSPATIAL Special is the newsletter of the Association for Computing Machinery (ACM) Special Interest Group on Spatial Information (SIGSPATIAL).

ACM SIGSPATIAL addresses issues related to the acquisition, management, and processing of spatially-related information with a focus on algorithmic, geometric, and visual considerations. The scope includes, but is not limited to, geographic information systems.

Current Elected ACM SIGSPATIAL officers are:

- Chair, Cyrus Shahabi, University of Southern California
- Past Chair, Mohamed Mokbel, University of Minnesota
- Vice-Chair, Goce Trajcevski, Iowa State University
- Secretary, Egemen Tanin, University of Melbourne
- Treasurer, John Krumm, Microsoft Research

Current Appointed ACM SIGSPATIAL officers are:

- Newsletter Editor, Andreas Züfle, George Mason University
- Webmaster, Chrysovalantis (Chrys) Anastasiou, University of Southern California

For more details and membership information for ACM SIGSPATIAL as well as for accessing the newsletters please visit http://www.sigspatial.org.

The SIGSPATIAL Special serves the community by publishing short contributions such as SIGSPATIAL conferences' highlights, calls and announcements for conferences and journals that are of interest to the community, as well as short technical notes on current topics. The newsletter has three issues every year, i.e., March, July, and November. For more detailed information regarding the newsletter or suggestions please contact the editor via email at azufle@gmu.edu.

Notice to contributing authors to The SIGSPATIAL Special: By submitting your article for distribution in this publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

- to publish in print on condition of acceptance by the editor,
- to digitize and post your article in the electronic version of this publication,
- to include the article in the ACM Digital Library,
- to allow users to copy and distribute the article for noncommercial, educational or research purposes.

However, as a contributing author, you retain copyright to your article and ACM will make every effort to refer requests for commercial use directly to you.

Notice to the readers: Opinions expressed in articles and letters are those of the author(s) and do not necessarily express the opinions of the ACM, SIGSPATIAL or the newsletter.

The SIGSPATIAL Special (ISSN 1946-7729) Volume 12, Number 2, July 2020.

Table of Contents

Introduction to this Special Issue: Modeling and Understanding the Spread of COVID-19 (Part II)

Andreas Züfle, Taylor Anderson
George Mason University
{azufle,tander6}@gmu.edu

The emergence of COVID-19 and its rapid spread across the globe has sparked research collaborations and initiatives between investigators from a vast number of disciplines including epidemiologists, social scientists, psychologists, mathematicians, geographers, data scientists, and more - all with the unified aim to better understand, predict, and mitigate the impacts of the disease. Many of these investigators make up the longstanding and interdisciplinary community that is SIGSPATIAL. Research efforts in this community offer a unique perspective for which to study the disease with a focus on the development and implementation of novel modeling, simulation, management, querying, and mining approaches that leverage the power of spatial-temporal data, much of which has increased in resolution and availability in an effort to combat COVID-19.

Part I of this Special Issue on Modeling and Understanding the Spread of COVID-19 [1] showcased current and emerging research projects related to COVID-19. It provided COVID-19 related datasets to the community [2] and presented solutions to mapping COVID-19 [3], detection of COVID-19 clusters [4], and analysis of change in human mobility due to COVID-19 [5]. Part II of this Special Issue aims to further showcase the growing number of COVID-19 related research efforts in the SIGSPATIAL community and beyond.

The goal of this newsletter special issue is to rapidly disseminate current research efforts by the SIGSPATIAL community and to facilitate discussions and collaboration

This newsletter has two sections. The first section presents three research projects and visions related to understanding and tracing the spread of COVID-19:

1. the first article by Xiong et al. describes a novel project towards real-time contact tracing of COVID-19 spread. The approach presented in this article takes special consideration on user privacy and allows users to refine the precision with which their data is collected and used,

2. the second article by Mokbel et al. discusses the limitations of (user-based) contact tracing apps and lays out the vision and guidelines of moving contact tracing from being personal responsibility to be the responsibility of facilities that users visit daily,

3. the third article by Bobashev et al. proposes the development and implementation of a novel reinforcement learning framework that combines compartmental modeling and machine learning approaches to predict the spread of COVID-19 and evaluate the risk to hospital resources,

4. the fourth article by Kim et al. presents a novel approach for combining predictions from multiple models of COVID-19 spread into a smaller set of ensemble predictions. The approach facilitates the visual analysis of the agreement between model predictions while accounting for their assumptions and uncertainty.

In the second section of this newsletter, not directly related to COVID-19, Sarwat discusses challenges and opportunities for using spatial data systems to support the Internet of Things (IoT).

All research papers across both parts of this special issue are invited to present their research at ACM SIGSPATIAL 2020 at the *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19* to be held virtually on November 3rd, 2020. This workshop will provide a forum for our community and collaborators across domains to discuss directions, opportunities, and lessons learned to continue our fight again COVID-19 and to become more resilient to future diseases.

We hope to welcome you to the workshops and the main conference. A limited number of free conference (and workshop) registrations are available. For details see `https://sigspatial2020.sigspatial.org/registration/`. We're looking forward to seeing you virtually at the 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2020) on November 3-6, 2020!

Finally, we want to cordially thank all the authors for their excellent contributions to this issue.

References

[1] A. Züfle, "Introduction to this Special Issue: Modeling and understanding the Spread of COVID-19: (Part I)," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 1–2, 2020.

[2] U. Qazi, M. Imran, and F. Ofli, "Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.

[3] S. Gao, J. Rao, Y. Kang, Y. Liang, and J. Kruse, "Mapping county-level mobility pattern changes in the united states in response to covid-19," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 16–26, 2020.

[4] A. Hohl, E. Delmelle, and M. Desjardins, "Rapid detection of covid-19 clusters in the united states using a prospective space-time scan statistic: an update," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 27–33, 2020.

[5] Z. Fan, X. Song, Y. Liu, Z. Zhang, C. Yang, Q. Chen, R. Jiang, and R. Shibasaki, "Human mobility based individual-level epidemic simulation platform," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 34–40, 2020.

The SIGSPATIAL Special

Section 1:
Modeling and Understanding the Spread of COVID-19

ACM SIGSPATIAL
http://www.sigspatial.org

REACT: Real-Time Contact Tracing and Risk Monitoring using Privacy-Enhanced Mobile Tracking

Li Xiong[1], Cyrus Shahabi[2], Yanan Da[1], Ritesh Ahuja[2], Vicki Hertzberg[3], Lance Waller[4], Xiaoqian Jiang[5], Amy Franklin[5]
[1]Department of Computer Science, Emory University, USA
[2]Department of Computer Science, University of Southern California, USA
[3]Nell Hodgson Woodruff School of Nursing, Emory University, USA
[4]Department of Biostatistics and Bioinformatics, Emory University, USA
[5] School of Biomedical Informatics, University of Texas Health Science Center, USA
{lxiong,yanan.da,vhertzb,lwaller}@emory.edu,
{shahabi,riteshah}@usc.edu,
{xiaoqian.jiang, amy.franklin}@uth.tmc.edu

Abstract

Contact tracing is an essential public health tool for controlling epidemic disease outbreaks such as the COVID-19 pandemic. Digital contact tracing using real-time locations or proximity of individuals can be used to significantly speed up and scale up contact tracing. In this article, we present our project, REACT, for REAal-time Contact Tracing and risk monitoring via privacy-enhanced tracking of users' locations and symptoms. With privacy enhancement that allows users to control and refine the precision with which their information will be collected and used, REACT will enable: 1) contact tracing of individuals who are exposed to infected cases and identification of hot-spot locations, 2) individual risk monitoring based on the locations they visit and their contact with others; and 3) community risk monitoring and detection of early signals of community spread. We will briefly describe our ongoing work and the approaches we are taking as well as some challenges we encountered in deploying the app.

1 Introduction

More than 6.5 million people in the U.S. have been infected with the coronavirus (COVID-19) and more than 200,000 have died as of September 2020[1]. While there has been a slowdown in new infections in recent weeks, tens of thousands of new cases are still reported daily nationwide.

Contact tracing [12] is an essential public health tool for controlling epidemic disease outbreaks such as the COVID-19 pandemic, involving identification and follow-up of all individuals who may have come into contact with an infected person. In traditional and current CDC-recommended practices[2], contact identification is conducted by asking about the person's activities. This process, however, does not scale. It is time-consuming and ultimately infeasible in public health crisis for large scale contact tracing, as is the case in COVID-19. Failure of traditional contact tracing necessitates alternatives with high degrees of community acceptance [11]. In addition, contact data collected in this way may be incomplete (limited to known contacts) or unreliable. Digital contact tracing using real-time locations

[1] `https://covid.cdc.gov/covid-data-tracker`

[2] `https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/contact-tracing.html`

or proximity of individuals can significantly speed up and scale up contact tracing, as demonstrated by many efforts in Asia and Europe [22, 29, 27, 21, 30, 41].

In March 2020, we submitted a RAPID proposal to NSF which was funded in April as a collaborative project among investigators at Emory University[3], University of Southern California[4], and UT Health Science Center[5]. The goal of our project is to develop techniques and a mobile application, REACT, for REAl-time Contact Tracing and risk monitoring via privacy-enhanced tracking of users' locations and symptoms.

In early April 2020, a draft of a seminal paper that was later published in Science [13] was released that made the case for digital contact tracing for COVID-19. The main observation from the paper was that since typically only symptomatic cases can be contact-traced, to bring R_0 below 1 (and hence stopping the spread), it was important to notify the symptomatic patient's contacts as soon as possible, which would only be possible through digital contact tracing. This was the first paper based on analysis of real-world COVID-19 data to make such a case, even though prior to that several studies (including our RAPID proposal to NSF in March) suggested the usefulness of digital contact tracing for COVID-19.

In mid April 2020, Apple and Google announced their proposed method of using mobile phone's bluetooth to exchange secure and anonymous tokens among nearby phones, which would then be used to notify device owners if they were in proximity of someone (actually someone's phone) who has been diagnosed with COVID-19 by health authorities. The approach was elegant but not very effective. In fact, one could argue[6] that the spatiotemporal data Apple and Google already collect from users, i.e., user mobility patterns, is much more useful for digital contact tracing.

Since then, discussions and efforts about creating contact tracing apps in the U.S. have become mired in battles over privacy concerns and inconsistent responses from different states and stakeholders. Our own project was politicized (ungroundedly) by Breitbart[7] in early May. There is also uncertainty about how much digital contact tracing would help the overall response to the pandemic compared to other measures including social distancing and mask wearing that are now largely adopted in the U.S.

Indeed, a critical issue in using real-time location traces of users for digital contact tracing is user privacy. A location trace can expose users to attacks such as unwanted spams/scams or physical danger, especially in the uncertain times at present. Location traces can be also linked to other information to disclose sensitive information about an individual, e.g., political views and religious inclinations.

Many contact tracing applications, including the ones from Apple and Google, use Bluetooth-based proximity only, not absolute locations, to protect privacy. Examples of this include official contact tracing apps from countries such as United Kingdom, Switzerland, Germany. Of those apps, some keep the contact data locally in the user's phone while others upload the contact data to a central location (e.g., Singapore, Australia). However, ignoring absolute locations sacrifices the ability to estimate the fine-grained transmission risk based on the type of the locations and identified hot spots, and the ability to trace indirect contacts. Another drawback is that bluetooth can "travel through walls" and wrongly identify someone in a neighboring room as a contact. A select few (e.g., Norway) collect both bluetooth contact data and GPS location data. This approach has led to privacy concerns and consequently a low adoption rate among citizens[8]. There have been apps that require mandatory location check-ins from citizens issued by governments like China [38]. While highly effective for containment interventions,

[3] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2027783

[4] https://www.nsf.gov/awardsearch/showAward?AWD_ID=202779

[5] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2027790

[6] https://medium.com/@csatusc/why-we-need-more-than-bluetooth-data-to-fight-covid-19-64da29b3164e

[7] https://www.breitbart.com/asia/2020/05/04/usc-emory-creating-coronavirus-surveillance-system-similar-to-chinas-social-credit-scoring/

[8] https://www.bbc.com/news/technology-52355028

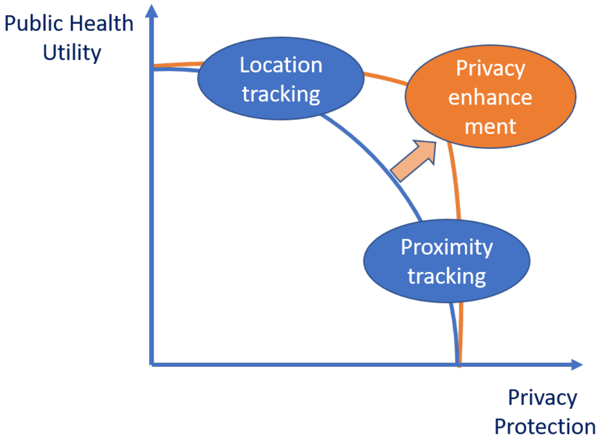


Figure 1: Public Health Utility and Privacy Tradeoffs

these apps have also heightened concerns about surveillance and data abuse.

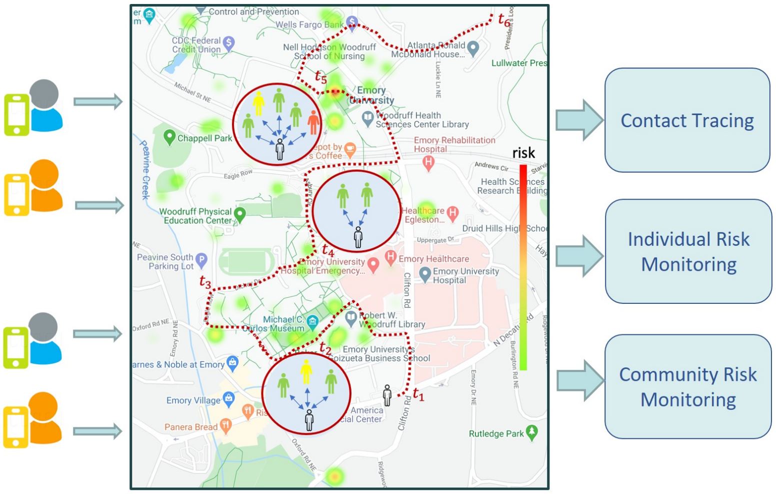
Besides using cellphone apps, passive digital tracking is gaining popularity. The Chinese government now complements contact tracing efforts by using direct cellphone location information as well as train/plane travel data, while South Korea opts to send mass alerts describing the locations visited by infected individuals. Wi-Fi localization is being used to check for adherence to social distancing directives within industrial and university campuses. For example, using the beam-forming technology of routers to track precisely the flow of people through different regions/floors/buildings and to determine how well campus buildings are implementing the social distancing[9]. University campuses such as USC still opt for a fusion of manual contact tracing with digital contact advice, for example by requesting possible contacts of an infected individual to call the Public Health number for next steps. However, this approach does not eliminate the privacy or scaling problem discussed earlier.

We believe a pandemic like COVID-19 requires a careful design of privacy protection—with public health benefits and privacy enhancement approaches—that optimizes the tradeoffs (as shown in Figure 1). The acceptability of contact tracing technology and the ethical use of it mainly depend on privacy, voluntariness, and beneficence of the data [33]. As governments reopen activities and businesses, contract tracing and risk monitoring remain important components of the public health response along with other measures such as testing and support for quarantine.

The goal of our project is to develop techniques and a mobile application, REACT, via privacy-enhanced tracking of users' locations and symptoms. Figure 2 gives a schematic for our framework. Users can voluntarily submit their locations and symptoms to the server, in addition to the proximity information that is captured by Bluetooth. To enhance privacy, users can control and refine the precision other users with whom their information will be collected and used. We are developing a multi-stage privacy approach where users can upload perturbed locations and adjust the privacy level or precision of the location to be uploaded as their risk evolves. Given such privacy options and enhancements, we hope that REACT will enhance contact tracing of individuals who are exposed to infected cases and allow identification of hot-spot locations for decontamination or increased surveillance to control further spread. The key is to develop efficient and scalable spatiotemporal data structures and algorithms for contact tracing queries given the potentially large number of users and the multi-resolution or perturbed location traces. More importantly, our vision is to go beyond contact tracing and support *individual risk monitoring*. We hope to develop a learning-based approach to estimate the risk for the users based on the locations they visit and their contact with others, so they can receive a real-time exposure risk score and be informed and alerted, e.g., they can self-quarantine or get tested

[9] `https://tippersweb.ics.uci.edu/covid19/d/IwAc1O9Wk/covid-19-effort-at-uc-irvine?orgId=1`

when the risk is high. Finally, we also plan to use the data collected for community risk monitoring using a social network sensors approach by monitoring a random group and a friends group, to detect early signals of community spread to prepare for larger-scale infections.



REACT: Real-time Contact Tracing and Risk Monitoring

Figure 2: REACT Overview

In this article, we will briefly describe our ongoing work and the approaches we are taking including: 1) building efficient, scalable data structures and algorithms for contact tracing queries; 2) expanding a learning-based approach for modeling user's risks based on location risk factors, propagated risks from other users, and the user's self-reported risk factors such as symptoms, demographic data, existing conditions, and travel history; and 3) enabling a multi-stage privacy approach based on geo-indistinguishability and its variants [4, 42]. We also describe our progress to date in developing and deploying the app as we originally planned at our three collaborating institutions. The release of the app requires intensive security and human subject research reviews that are much more involved than what we had originally anticipated. In addition, we have encountered other nontechnical obstacles that we will describe in our case study at USC.

2 Approach

The main goal of our project is to enable contact tracing and risk monitoring via tracking of users' locations and symptoms. Upon user' consent, locations can be automatically uploaded at a user-selected frequency and granularity. Selected personal data will be collected on a voluntary basis including demographic information and existing conditions. In-app surveys will be used to collect user symptom data periodically or triggered by time/location following the Ecological Momentary Assessment methodology [34, 19]. We plan to ask users daily if they experience of any of the symptoms of upper respiratory illness: fever, chills, muscle aches, cough, congestion, runny nose, headaches, fatigue, and shortness of breath, as well as if they have a confirmed infection of COVID-19. Given the purpose of the app, we expect even if there are self-reported false positives, acting in an abundance of caution would benefit the community.

2.1 Contact Tracing

We envision that whenever a user self-reports onset of COVID-19 symptoms or a confirmed positive test, we can systematically identify all users who have been in contact with this infected case both directly and indirectly. Once identified, we can alert other users and update their risks. We discuss queries to identify contacts in this subsection and risk notification and modeling in next subsection.

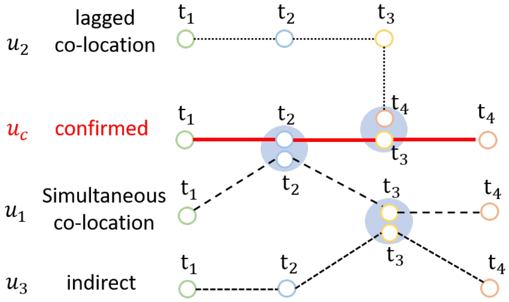


Figure 3: Transmission Scenarios

We consider three types of transmission of COVID-19: 1) direct person-to-person transmission, i.e. in close contact with someone infected (simultaneous co-location); 2) fomite transmission, i.e. in contact with a contaminated surface or object at a location visited by someone infected earlier (lagged co-location) [25]; and 3) indirect person-to-person transmission by contact with someone who is earlier in contact with someone infected. Figure 3 illustrates four user trajectories where u_c has a confirmed infection at time t_4 and the three transmission scenarios: 1) u_1 via simultaneous co-location at t_2, 2) u_2 via lagged co-location at time t_4, and 3) u_3 via direct transmission from u_1 at t_3 (indirect transmission from u_c).

We will first consider the first two scenarios which are the primary ways of transmission. Assume the location trace of each user is represented as a sequence of *visits* $< u, s, t >$ (user u at location s at time t). Given a user u_c with confirmed infection, we need to find all users who have simultaneous or lagged co-location with u_c in a back tracing window (this can be parameterized, e.g. as 5 days which is the currently known median incubation period for COVID-19 [35]). Defining a sequence of visits by a single user as a *trajectory*, this problem can be formulated as composition of *trajectory range queries* [43] which given a trajectory dataset, a spatial range and a time interval, returns all trajectories that intersect with both the spatial range and the time interval. For each location s visited by u_c at time t in the time window, we can compute trajectory range query using a spatial range of 6 feet centered at s and a time interval of up to 72 hours starting at t (the viable time of COVID-19 virus on surfaces [40]). These parameters can be changed with dynamic evidence such as the distance within closed/open spaces or modifications in aerosol range according to latest research and CDC recommendations. Consider the example in Figure 3, the query given u_c will return u_1 and u_2. There are many structures that index trajectories to efficiently answer such queries, but since we mainly work with visits, we can simply index visits using point-based indexes. Given the broad availability of R-trees [18] (e.g., with PostgreSQL) and the need to have our system up and running quickly, we will use an R-tree to index the visits on the three dimensions: latitude, longitude and time.

The indirect transmission is more challenging both from modeling and computational point of view. We need to find all users that are directly and indirectly in contact with the confirmed case such as u_3 in our example. This can be formulated as a *spatio-temporal reachability query* [37]. A straightforward approach is to first run the trajectory range queries using the confirmed case as source, then run the queries recursively by using the returned trajectories as sources. This will be computationally expensive, especially when the number of indirect transmission hops becomes more than a few. We plan to explore an alternative approach by leveraging our prior work [37] which proposed efficient grid and graph based indexes for answering "single source single destination" reachability queries. The main idea is to compute reachability on-the-fly by expanding the contact network starting from the query source and utilizing the spatio-temporal locality for enhanced performance.

2.2 Individual Risk Monitoring

While contact tracing is triggered when there is a confirmed infection, we envision REACT will also allow users to monitor their own risks in real time based on the locations they recently visited and the aggregated risks of other users they have come in contact with. Whenever a user visits a new location, we will use our risk model to update the risks for other users, so that they can be informed/alerted and take preventative measures when necessary. Specifically, we can define a risk score between 0 and 1 for each user u which represents the probability u will contract the virus and 1 means the user has a confirmed infection. We can gradually train a risk model (e.g. a logistic regression model) based on the risk factors as we collect more data including confirmed infections (which serves as ground truth).

The following factors can be considered in the risk model for each user u: 1) u's risk profile including demographic data and existing conditions, 2) aggregated risk of locations u has recently visited, and 3) aggregated risk of recent contacts with other users. Risk of a location can be dependent on type of location or Point of Interest (e.g. from Google Map API) and confirmed cases in the area [1]. Risk associated with each contact can be dependent on distance and duration. Due to infrequent or generalized location tracking for privacy enhancement, distance may not be accurate enough nor can duration of co-location be captured adequately. Hence, we plan to incorporate the strength of social relationships between two users as an additional factor. Duration of social contact is typically longer, suggesting a higher risk, for dyads with social ties rather than dyads of strangers [28, 6]. This social relationship can be partially explicitly collected from the users (see the subsection below) or implicitly inferred from their historical trajectories as we have demonstrated in our prior work [31]. The intuition is that if two people have frequent co-locations, especially at not-popular places, it is likely they are socially related.

In order to evaluate our risk monitoring algorithms, we need to have realistic data before we collect real data from the deployed app. We are developing an agent-based spread simulator based on a real mobility dataset[10] to generate test data. Most existing simulations are based on compartmental models which cannot account for real life mobility patterns. We are using real-world mobility patterns to inform disease spread and creating a simulator that can generate realistic spread data.

To complement our algorithm development for contact tracing queries and risk monitoring, we are also designing and developing a user facing dashboard (at USC) to enable public health practitioners to expedite contact tracing processing and provide recommendations based on risk parameters estimated from available co-occurrences and user data. The dashboard will enable decision makers to visualize and optimize interventions as powered by the Spread Simulator.

2.3 Community Risk Monitoring

To monitor the community, we will use the social network sensors approach [8]. Using a novel study design based on properties of social networks, the method yielded a 2 week advance signal of the H1N1 influenza outbreak in fall 2019 among Harvard undergraduates. In our project, we will invite a random sample of students at each of our three institutions to participate as a baseline (Random Group). We will ask the Random Group for contact information for 2-3 of their friends. Subsequently, we will invite the friends to participate (Friends Group). The value of the social network is that, in an epidemic, the more central members of a social network become infected earlier. Thus utility of the app as used with this study design is that not only will we have an estimation of risk, but we also expect a similar early signal of an outbreak by examining over time the differences between the Random Group and the Friends Group with its more central members [9, 28].

[10] the anonymized raw mobility dataset is provided to us by Veraset.

Specifically, we will track the numbers of users in each group who have a constellation of symptoms consistent with COVID-19 (fever, dry cough, shortness of breath). Without loss of generality we can similarly track symptoms consistent with influenza, providing generalized utility for the annual seasonal influenza outbreak beyond the current pandemic. We will use an adaptation of the cumulative sum (CUSUM) procedure to detect the first separation of the Friends Group from the Random Group with respect to prevalence of COVID-19 symptoms [20]. In particular, we will use the CUSUM count method to determine the separation time [14, 15]. Briefly, we will consider at the end of each day, t, the counts of the number of users in the Random and Friends groups reporting symptoms consistent with COVID-19, Y_t^R and Y_t^F respectively. Assuming that these counts are (conditionally) independent Poisson variables with mean μ_t^R and μ_t^F, the CUSUM method—based on the likelihood ratios of functions of sums of Y_t^R and Y_t^F as t increases—uses sequential hypothesis testing to determine the change point from $\mu_t^F = \mu_t^R$ to $\mu_t^F > \mu_t^R$ which suggests an imminent community outbreak.

2.4 Privacy Enhancements

Despite the utility of contact tracing, the technology also raised a lot of trust concerns [32], including cultural and behavioral issues [26], privacy and equity [5], legal issues [10], and individual autonomy, privacy, confidentiality, and social justice [36]. Medical professionals are facing ethical choices between the public good and individual's privacy [24]. Besides these intensely debated issues [23], data protection and user acceptability remain as major barriers [2].

To mitigate privacy risks while ensuring immediate public health impact, the key is to give users the options to control frequency and precision with which information will be collected. For instance, users will be able to choose and update frequency of tracking (or a manual check-in option) and the granularity of tracked locations (e.g., a generalized location range) as their risk evolves and to choose whether to report symptoms or not.

Given infrequent tracking or generalized locations, results of contact tracing queries may not be precise. We plan to have a multi-stage approach to address this challenge. In stage 1 (global computation), the server can perform "single source all destination" contact tracing queries to identify all "possible contacts" over the generalized or imprecise locations. This may lead to false positives and false negatives. We can adjust or relax both spatial range and temporal interval to account for generalized locations and infrequent tracking and ensure a low false negative rate. Possible contacts can then be alerted. In stage 2 (local refinement), alerted users can choose to upload his/her precise location trace stored locally in the recent window to confirm contact status with the confirmed case. The server then will perform a "single source single destination" query to get precise result. In our prior work [39], we have shown that such a multi-stage approach is promising for task assignment given uncertain/perturbed locations of workers and tasks for spatial crowdsourcing.

Currently, we have adopted and evaluated the geo-indistinguishability (GeoInd) privacy definition [4, 42, 39, 17] to enable users to protect their locations. Given app users u_1 (and u_2, respectively), the ε-GeoInd perturbation mechanism distorts their exact locations l_1 (l_2) to l'_1 (l'_2) by adding a spatial noise vector derived from a 2D Laplace distribution (with scale inversely proportional to ε). The challenge is then to accurately compute the range or reachability queries over the perturbed locations and to address the privacy risks associated with a location trace—a straightforward composition of geo-indistinguishability will render either the privacy or the utility not acceptable.

We extend probabilistic techniques from our previous work in [39] to calculate the range query over the pair of perturbed locations l'_1 and l'_2. Recall that the range query captures whether or not two users actually made a contact (parameterized as a reachable distance R), which is indicative of the risk of a potential transmission. The objective is to then calculate their reachability probability $p(d \leq R|d')$, where d and d' are the Euclidean distances of their exact and perturbed location pairs, respectively.

Our preliminary studies using the Gowalla Geo-social Network checkin dataset [7] verifies that the probabilistic approach can outperform the baseline oblivious approach (which determines reachability using the perturbed locations directly), and it can achieve 80% precision and recall given a reasonable privacy level.

2.5 App Development

The current REACT[11] app is *forked* from an existing open source project named Covid Community Alert[12]. The REACT app collects proximal contacts (via Bluetooth) and locations (via GPS if permissed by user) for contact tracing. It maintains the anonymity of its users by recording ephemeral device IDs that persist for the duration the app is installed and can be reset by user by re-installing the app.

We extended the app with additional location privacy features. A UI page requests the user to input a desired privacy level (between low, medium and high) for sharing his/her locations. This privacy level is interpreted as the level of perturbation that is applied to user's location before it is transmitted to the receiving server. We implemented the GeoInd based location perturbation with predefined privacy levels. Another UI element provides the functionality for the users to self-report their COVID status (e.g., from symptomatic, tested positive/negative, recovered). The app works as follows. A user registers the device by sending a randomly generated device ID to the server when first time use the app (no personal information collected). The app keeps scanning surrounding Bluetooth signals and collects the IDs of the nearby devices. The interaction information including devices IDs, timestamp, interaction duration and GPS location are sent to the back-end server. When a confirmed case is reported, the back-end server finds the potential contacts and estimates their total risk score. If the risk score exceeds a preset level an alert is relayed to these users as a notification on their device.

3 Deployment: Challenges and Lessons

Our original plan was to develop and release a mobile app at our three collaborating institutions. At Emory University, at the time of writing, we are still under security review by the university IT office and IRB review for the human subject research, which will not proceed without satisfactory security review. At UT Health, the IRB review has been approved. At University of Southern California (USC), earlier versions of the app had gone through the security reviews and hence efforts were made to deploy it for real use. In this section, we discuss some of the practical and non-technical challenges we encountered in releasing the app for real use at USC.

As early as mid-May 2020, Shahabi's team at USC developed multiple contact-tracing app prototypes using a variation of techniques to detect and collect user location data. The very first app used the phone's location API to store mobility patterns. The later versions included a QR-code scanning capability to allow for scanning of location QR-codes as a way for the user to voluntarily check in and out of a location (e.g., a classroom). The final version of the app, developed on May 22nd 2020, also added bluetooth for proximity tracking. All these data collections had opt-out options. The goal was to modify the app later per our proposed ideas in the NSF RAPID project to collect location data at different spatial and temporal resolutions determined by the user and only store the detailed data on the user's device. Our first goal, however, was to influence one of our institutions, in this case USC (through Shahabi's participation in USC's contact-tracing subcommittee) to actually adopt and use

[11] `https://github.com/Emory-AIMS/react`

[12] https://coronavirus-outbreak-control.github.io/web/

the app for faculty, staff and students as they come back to campus. Unfortunately, these efforts were not successful for non-technical reasons. We review these reasons below.

Both Google and Apple restricted the release of any app on their app stores that was related to COVID-19 in general and for contact-tracing in particular. They required the backing of a health organization for COVID-19 apps. With this requirement, we could not release our app, even to be used by volunteers under IRB, unless we have the backing of our respective organizations. At USC, we tried to release different variations of our app, went through rigorous reviews of our IT offices for security and privacy, and at the end none of the variations (except for an early adaptation of a symptom collection app at USC, called SCORE) made it to the app store. Basically, the time cost of releasing an app was so high that it overshadowed the cost of developing the app itself. More importantly, even if we released the app, as discussed below, we did not have the backing of our institute to recruit students to use the app, which in turn rendered the releasing of the app useless.

The main hesitation of the organizations to support and release a contact-tracing app was to protect the users' location privacy. This was a surprise to us for three reasons. First, many apps already freely collect users' locations. Second, there is a decade of research on location privacy by our community that can be incorporated, starting from simple measures of allowing users to control the specificity of their reported locations (for instance, building level or shopping-center level), the frequency of reporting (for example, once or twice a day) and to remove sensitive locations. More sophisticated measures and technologies can also be incorporated, such as storing and searching all data in an encrypted form, similar to storing passwords or banking information. Both approaches have been studied extensively for location data in the past decade, e.g., [3] and [16]. Third, in case the integration of data about one's health status (in this case COVID-19 exposure) and location data was sensitive, we considered approaches to separate the location data from health data, each being stored and accessed by trusted parties within an organization. This is usually the first step in any privacy and security research, where threat models are clearly defined. However, unfortunately, we never proceeded sufficiently far to clearly define threat models. The main obstacle was the "perception" of privacy violation surrounding any contact-tracing app. We could not solve this concern of perception with technical solutions, and instead partnerships with colleagues from communication and journalism are needed to design and deploy broader messaging campaigns.

Finally, we tried to convince the health offices within our organization about the usefulness of digital contact tracing to get their support in convincing our institute to support the release of a contact-tracing app. Towards this end, working through the USC's contact-tracing subcommittee, we showcased numerous proof-of-concept tools, demonstrations and presentations, directly to USC's health practitioners and managers who were in charge of contact-tracing on campus to demonstrate what could have been done if location data were collected. The utilities included user-friendly dashboards to quickly find overlapping trajectories with the trajectory of an infected case, identifying hotspot locations (through density mapping of infected trajectories) and utilities to detect environmental and indirect contacts. However, at the end, the health organizations preferred traditional contact tracing approaches where individuals were interviewed thoroughly and then broad notifications were sent to anyone who could have potentially been collocated with the positive cases. Clearly the approach is not scalable but due to limited access of students and faculty to campuses (due to remote teaching), the issue of scalability has not been the main concern of health practitioners as they were dealing with other critical and time-sensitive issues.

Consequently, even though we designed and developed several useful solutions, we could not convince the decision makers at our institution to support the utilization of our tools for us to make a real impact in our community.

4 Conclusion

In this article, we presented our ongoing project for real-time contact tracing and risk monitoring via privacy-enhanced mobile tracking. We described the approaches we are taking for privacy enhanced contact tracing and risk monitoring, the preliminary results we have obtained which are encouraging, as well as some challenges we encountered in deploying the app. While there are continued privacy concerns, and other non-technical obstacles, we believe digital contact tracing remains an important component of the public health response and it requires careful designs and technical developments which we will continue to work on in order to ensure privacy protection and public health benefits.

References

[1] Coronavirus covid-19 global cases by the center for systems science and engineering (csse) at johns hopkins university (jhu). https://coronavirus.jhu.edu/map.html.

[2] J. Abeler, M. Bäcker, U. Buermeyer, and H. Zillessen. COVID-19 contact tracing and data protection can go together. *JMIR mHealth and uHealth*, 8(4):e19359, Apr. 2020.

[3] R. Ahuja, G. Ghinita, and C. Shahabi. A utility-preserving and scalable technique for protecting location data with geo-indistinguishability. In *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, pages 217–228. OpenProceedings.org, 2019.

[4] M. E. Andrés, N. E. Bordenabe, and K. Chatzikokolakis. Geo-indistinguishability: Differential privacy for location-based systems. *CoRR*, abs/1212.1984, 2012.

[5] I. Braithwaite, T. Callender, M. Bullock, and R. W. Aldridge. Automated and partly automated contact tracing: a systematic review to inform the control of COVID-19. *The Lancet. Digital health*, Aug. 2020.

[6] S. Cauchemez, C. A. Donnelly, C. Reed, A. C. Ghani, C. Fraser, C. K. Kent, L. Finelli, and N. M. Ferguson. Household transmission of 2009 pandemic influenza a (h1n1) virus in the united states. *New England Journal of Medicine*, 361(27):2619–2627, 2009.

[7] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In C. Apte, J. Ghosh, and P. Smyth, editors, *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1082–1090. ACM, ACM, 2011.

[8] N. Christakis and J. Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9), 2010.

[9] R. M. Christley, G. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett, and J. Turner. Infection in social networks: using network analysis to identify high-risk individuals. *American journal of epidemiology*, 162(10):1024–1031, 2005.

[10] A. Cioffi, C. Lugi, and C. Cecannecchia. Apps for COVID-19 contact-tracing: Too many questions and few answers. *Ethics, medicine, and public health*, 15:100575, Oct. 2020.

[11] E. Clark, E. Y. Chiao, and E. S. Amirian. Why contact tracing efforts have failed to curb coronavirus disease 2019 (covid-19) transmission in much of the united states. *Clinical Infectious Diseases*, 2020. https://doi.org/10.1093/cid/ciaa1155.

[12] P. C. Erwin and R. C. Brownson. *Principles of Public Health Practice*. Cengage Learning, 2016.

[13] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020.

[14] M. Frisén. Statistical surveillance. optimality and methods. *International Statistical Review*, 71(2):403–434, 2003.

[15] M. Frisén and P. Wessman. Evaluations of likelihood ratio methods for surveillance. *Communications in Statistics-Simulation and Computation*, 28(3):597–622, 1999.

[16] G. Ghinita, K. Nguyen, M. Maruseac, and C. Shahabi. A secure location-based alert system with tunable privacy-performance trade-off. *GeoInformatica*, 24(4):951–985, 2020.

[17] X. Gu, M. Li, Y. Cao, and L. Xiong. Supporting both range queries and frequency estimation with local differential privacy. In *2019 IEEE Conference on Communications and Network Security (CNS)*, pages 124–132, 2019.

[18] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 47–57, 1984.

[19] M. J, M. M, C. DL, and H. K. Geographic imputation of missing activity space data from ecological momentary assessment (ema) gps positions. 15, 2018.

[20] W. Jiang, L. Shu, H. Zhao, and K.-L. Tsui. Cusum procedures for health care surveillance. *Quality and Reliability Engineering International*, 29(6):883–897, 2013.

[21] M. Jonker, E. De Bekker-Grob, J. Veldwijk, L. Goossens, S. Bour, and M. Rutten-Van Mölken. COVID-19 contact-tracing apps: predicted uptake in the netherlands based on a discrete choice experiment. *JMIR mHealth and uHealth*, Aug. 2020.

[22] J. Jung, H. Jang, H. K. Kim, J. Kim, A. Kim, and K. P. Ko. The importance of mandatory COVID-19 diagnostic testing prior to release from quarantine. *Journal of Korean medical science*, 35(34):e314, Aug. 2020.

[23] K. Kaspar. Motivations for social distancing and app use as complementary measures to combat the COVID-19 pandemic: Quantitative survey study. *Journal of medical Internet research*, 22(8):e21613, Aug. 2020.

[24] J. Klaaren, K. Breckenridge, F. Cachalia, S. Fonn, and M. Veller. South africa's COVID-19 tracing database: Risks and rewards of which doctors should be aware. *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde*, 110(7):617–620, June 2020.

[25] A. N. Kraay, M. A. Hayashi, N. Hernandez-Ceron, I. H. Spicknall, M. C. Eisenberg, R. Meza, and J. N. Eisenberg. Fomite-mediated transmission as a sufficient pathway: a comparative analysis across three viral pathogens. *BMC infectious diseases*, 18(1):540, 2018.

[26] F. Lucivero, N. Hallowell, S. Johnson, B. Prainsack, G. Samuel, and T. Sharon. COVID-19 and contact tracing apps: Ethical challenges for a social experiment on a global scale. *Journal of bioethical inquiry*, Aug. 2020.

[27] C. M. Manauis, M. Loh, J. Kwan, J. Chua Mingzhou, H. J. Teo, D. Teng Kuan Peng, S. Vasoo Sushilan, Y. S. Leo, and A. Hou. Bracing for impact: operational upshots from the national centre for infectious diseases screening centre (singapore) during the COVID-19 outbreak. *Journal of the American College of Emergency Physicians open*, June 2020.

[28] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3), 2008.

[29] K. Nagarajan, M. Muniyandi, B. Palani, and S. Sellappan. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. *BMC medical research methodology*, 20(1):233, 2020.

[30] J. Peto, J. Carpenter, G. D. Smith, S. Duffy, R. Houlston, D. J. Hunter, K. McPherson, N. Pearce, P. Romer, P. Sasieni, and C. Turnbull. Weekly COVID-19 testing with household quarantine and contact tracing is feasible and would probably end the epidemic. *Royal Society open science*, 7(6):200915, June 2020.

[31] H. Pham, C. Shahabi, and Y. Liu. EBM: an entropy-based model to infer social strength from spatiotemporal data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 265–276. ACM, 2013.

[32] B. J. Ryan, D. Coppola, J. Williams, and R. Swienton. COVID-19 contact tracing solutions for mass gatherings. *Disaster medicine and public health preparedness*, pages 1–7, July 2020.

[33] A. S, K. OF, and A. KH. Covid-19 contact-tracing technology: Acceptability and ethical issues of use. *Patient Prefer Adherence*, 2020. https://doi.org/10.2147/PPA.S276183.

[34] S. S, S. AA, and H. MR. Ecological momentary assessment. 4, 2007.

[35] L. SA, G. KH, and e. a. Bi Q. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. 2020.

[36] A. Shah and R. P. Aacharya. Combating COVID-19 pandemic in nepal: Ethical challenges in an outbreak. *JNMA; journal of the Nepal Medical Association*, 58(224):276–279, Apr. 2020.

[37] H. Shirani-Mehr, F. B. Kashani, and C. Shahabi. Efficient reachability query evaluation in large spatiotemporal contact datasets. *PVLDB*, 5(9):848–859, 2012.

[38] C. Thorbecke. China launches coronavirus tracking app as death toll surpasses 1,000. ABC News, Feb. 2020. https://abcnews.go.com/Business/china-launches-app-combat-coronavirus-spread/story?id=68907706.

[39] H. To, C. Shahabi, and L. Xiong. Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server. In *IEEE Int. Conference on Data Engineering (ICDE)*, 2018.

[40] van Doremalen N, M. DH, and e. a. Holbrook MG. Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1. March 19, 2020.

[41] M. Walrave, C. Waeterloos, and K. Ponnet. Adoption of a contact tracing app for containing COVID-19: A health belief model approach. *JMIR public health and surveillance*, 6(3):e20572, Sept. 2020.

[42] Y. Xiao and L. Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-6, 2015*, pages 1298–1309, 2015.

[43] M.-E. Yadamjav, F. Choudhury, Z. Bao, and H. Samet. *Efficient Multi-range Query Processing on Trajectories: 37th International Conference, ER 2018, Xi'an, China, October 22–25, 2018, Proceedings*, pages 269–285. 09 2018.

Contact Tracing: Beyond the Apps

Mohamed Mokbel, Sofiane Abbar, and Rade Stanojevic
Qatar Computing Research Institute
Hamad Bin Khalifa University. Doha, Qatar
{mmokbel, sabbar, rstanojevic}@hbku.edu.qa

Abstract

As pandemic wide spread results in locking down vital facilities, digital contact tracing is deemed as a key for re-opening. However, current efforts in digital contact tracing, running as mobile apps on users' smartphones, fall short in being effective and present two major weaknesses related to accessibility and apparent privacy concern augmentation. Indeed, accessibility is affected by several factors such as smartphone penetration, age, or socio-economic conditions. The privacy concern on the other hand comes from the fear of having a piece of technology that is monitoring us all the time, everywhere, even when contact tracing is irrelevant. This paper lays out the vision and guidelines for the next era of digital contact tracing, where the contact tracing functionality is moved from being personal responsibility to be the responsibility of facilities that users visit daily. Our proposal tackles the two aforementioned shortcomings by disengaging users from using their own smartphones and requiring facilities to provide the technological devices needed for contact tracing. By doing so, we reassure users that their contacts are only considered in places where manual contact tracing is not effective, and cease being recorded as soon as they leave the facilities they visit. A privacy-preserving architecture is proposed, which can be mandated as a prerequisite for any facility to re-open during or after the pandemic. We finally outline research opportunities and challenges revolving around contact tracing system design and data management.

1 Introduction

– We are not proposing a solution to coronavirus problem, but rather a technological option to the problem of lock-downs and the shutdowns of the economy – Raj Reddy, HLF 2020.

In the wake of the world wide pandemic caused by the spread of COVID-19 virus, which disrupted our lives in an unprecedented way, we came quickly to realize that our best shot at the pandemic is the WHO and CDC recommended three-step protocol: test, isolate, and trace[40]. While many of us discovered contact tracing (CT) for the first time, it turned out that CT has been vital in stopping the spread of infectious diseases [15] in the last few decades. The way it works is quite straightforward: Once a patient is confirmed positive of some communicative disease, a community health worker talks to the patient to learn about other people who were in recent contact with the patient to screen them for the disease symptoms [41]. This can be achieved by listing names of people or places visited. Such simple human-based process had significant impact in saving lives by early diagnosing patients and avoiding further disease spread for tuberculosis (TB) [10], sexually-transmitted disease [12], Severe Acute Respiratory Syndrome (SARS) [22], foot-and-mouth-disease [20], smallpox [29], Ebola [42], among others.

Unfortunately, human-based contact tracing does not scale up to pandemic cases, with unknown immunity, high reproduction rate, and complex transmission mechanisms (e.g., airbone, surface) where contacts could be unknown to the patient, e.g., people met in airports, malls, or restaurants [32]. To hint into the scale we are talking about, it is reported that 300,000 human contact tracers are needed for COVID-19 in the USA alone [2, 9]. With a pandemic wide spread and a worldwide lock down, causing unprecedented economic crisis, contact tracing is identified as a must for pandemic control [17], the EU Commission for instance recommends that contact tracing is needed for people to return to hotels and camping sites [14]. In the USA, several states have made contact tracing a prerequisite for re-opening, including California [26], Pennsylvania [35], Virginia [43], among others [25]. Motivated by the limitations of human contact tracing, several governments have partnered with IT industry such as telecommunication companies [27] and tech giants such as Apple and Google [24] to deploy digital contact tracing solutions. The result is hundreds of mobile contact-tracing apps [28] where users would need to download the app and enable bluetooth connection and/or GPS location.

Despite the variety of proposals and applications, existing contact tracing solutions revolve around two main approaches:

- **Bluetooth User-to-user Contact Tracing.** The user is given a token ID to use every few hours. Once two users were in contact, their bluetooth connections will recognize each other ID and save it in a phone log. If a user is tested positive for a pandemic, her contacts are notified accordingly. Examples include Apple-Google BLE approach [24] and the decentralized privacy preserving solution promoted by the European Union [38].
- **GPS Location-based Contact Tracing.** Users periodically log their locations with the running app. Once a user is confirmed positive, her locations are used to formulate spatio-temporal queries to identify other users who have been at the same places and time. These users are then notified accordingly. Examples include SafePaths app[30], etc.

Though both approaches ensure user privacy through using pseudonym IDs [11], the bluetooth approach ensures more privacy by not reporting user locations. Meanwhile, it may miss some contacts who were in the same place with a confirmed patient, but only few minutes apart. On the other side, the GPS approach helps identifying risky locations and hot-spots. However, it may end up reporting false positives as it depend on the accuracy of the GPS signal. It is worth noting that some app solutions combine both approaches such as EHTERAZ [16], which is the official national app in the state of Qatar. However, both these two approaches suffer major shortcomings. First, they both require the users to have somewhat modern smartphones that support GPS and/or BLE, which as we will show in the next section is not the case of everyone. Second, often the success of these solutions rely on the voluntary willingness of users to install and run the app in their devices. Finally, installing an app that tracks users everywhere all the time, even when contact tracing is not needed, slows down the wide adoption needed for these solutions to work.

In this paper, we outline our vision for a digital contact tracing that is more inclusive and tackles some major shortcomings of existing app based solutions described above, such as accessibility and coverage. Our main contributions are summarized in the following bullet points:

- **Paradigm shift.** We posit that contact tracing should not be the responsibility of individuals, but that of facilities instead. Manual contact tracing is enough to recollect contacts made in limited private spaces, such as home. Digital contact tracing should be reserved to somewhat large facilities where it is hard to recollect the contacts.
- **Accessibility.** While there are 100s of papers, products, and research efforts about contact tracing mobile apps, they suffer from a major accessibility problem where not all people have access to

smartphones. Our vision is the first that does not require access to smartphones. Instead, contact tracing becomes the responsibility of facilities that people visit. Our vision can also be seen as bridging the accessibility gap when deployed in parallel with app-based solutions.

- **People acceptance.** We propose to proceed with contact tracing in a similar way we did with CCTVs. That is, deployment and management is the responsibility of facilities, storage is limited to a given period, and most importantly, access and distribution are heavily regulated, even for law enforcement. Moreover, similar to CCTV that do not follow us home, we propose that contact tracing be done only in places of exposition to large crowds of unknown people.

The rest of the paper is organized as follows. In Section 2, we make the case that app-based contact tracing does not work. Hence, in Section 3, we lay out our vision and guidelines for the next era of digital contact tracing. We believe that contact tracing should not be made as user responsibility and should not be running on user phones. Instead, contact tracing should be the responsibility of facilities and business entities (e.g., work places, malls, stadiums, restaurants, subways, etc) where the ability to do contact tracing can be made as a prerequisite for these facilities to re-open. In Section 4, we outline the architecture that can realize our vision in a privacy-preserving way. We finally conclude the paper with further remarks in Section 5.

2 Contact Tracing Apps Do Not work

Unfortunately, even though there are tremendous efforts put in developing app-based contact tracing, it did not deliver what it has promised, mainly for the following two reasons:

(1) *Need for large cooperating population.* One of the very first apps, TraceTogether from Singapore [37], has only reached around 1.4M users (25% of population) after more than two months of release. This means that the probability that two random people in contact have both installed the app is only 6.25%(0.25*0.25). This is assuming the best case scenario in which all users who have the app running in the background. With this tiny ratio, there is not much real benefit of such apps [36]. Meanwhile, though Iceland is reported to be the country with the highest population ratio using a contact tracing app (38%), that did not help much [7]. (2) *Low and biased smart phone penetration.* Smart phone penetration varies across countries, e.g., 24% in India, 81% in USA and 95% in S. Korea [33]. This leaves a major part of the population without access to app-based contact tracing [13]. More importantly, smart phone penetration is inversely biased with COVID-19 spread. More poor areas have higher COVID-19 ratio [6] and much less smartphone penetration, hence less access to contact-tracing apps.

So, unless contact tracing apps are made mandatory and used by the very large majority of population, they will not be effective [18, 39]. With such serious issues, it becomes apparent that current efforts in digital contact tracing fail to meet the expectations. As a result, thoughts are going back to use human-based contact tracing, especially in USA, where it is estimated that 300,000 human contact tracers are needed for COVID-19 [2, 9].

3 Next Era of Contact Tracing: Guidelines

Our vision for the next era of digital contact tracing goes beyond mobile apps to be along the following guidelines:

- **Focus on unknown contacts.** Human tracers can efficiently identify family members, friends, or neighbors, but cannot identify unknown contacts who the patient have contacted in public facilities, e.g., malls, restaurants, work places. This should be the main focus of digital contact tracing.

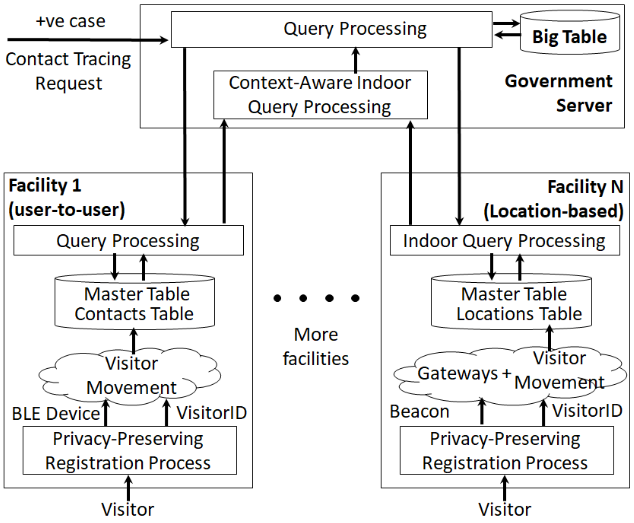

+ve case
Contact Tracing Request
Query Processing
Big Table
Context-Aware Indoor Query Processing
Government Server
Facility 1 (user-to-user)
Query Processing
Master Table Contacts Table
Visitor Movement
BLE Device
VisitorID
Privacy-Preserving Registration Process
Visitor
More facilities
Facility N (Location-based)
Indoor Query Processing
Master Table Locations Table
Gateways +
Visitor Movement
Beacon
VisitorID
Privacy-Preserving Registration Process
Visitor

Figure 1: System Architecture. Facility 1 deploys *user-to-user* contact tracing, Facility N deploys *location-based* contact tracing. More facilities can be added independently.

- **Focus on indoor environments.** People spend most of their time indoors, where it is more likely to get infected [8]. In USA, a national survey shows that people spend 87% of their time indoor and 5% in a vehicle [21]. Digital contact tracing should put more focus on indoor facilities.
- **Contact tracing is not a personal responsibility; it is surveillance.** People should not download any apps. Instead, each facility should be responsible on its own contact tracing for its visitors. If someone is tested positive, authorities will contact recently visited facilities and get their log of the patient contacts. The ability to do contact tracing would be a prerequisite for any facility to re-open. This is similar to say that facilities will not re-open unless they comply with new hygiene guidelines. This also goes inline with the requirement that facilities should have enough CCTV camera coverage to ensure safe operations, where authorities will get access to, when accidents happen.
- **Context-Aware tracing.** Digital contact tracing needs to go beyond the idea of one size fits all (same app running for everyone everywhere) to the more general case of context-awareness in terms of both infrastructure and analysis. For infrastructure, each facility may decide on its own way of deploying contact tracing technology. Meanwhile, the analysis of whether two persons were in contact would depend on the facility type.
- **Privacy-preserving.** Ensuring healthy environment should not be traded with privacy. Contact tracing should ensure that facilities do not have access to any user private information.

4 The Vision for the Next Era of Contact Tracing

Figure 1 gives the system architecture of our vision for the next era of digital contact tracing. Each facility will decide on deploying one or both of the two approaches described in Section 1, which we refer to as *User-to-User* and *Location-based Contact* Tracing. Then, each facility will have its own built-in infrastructure (Section 4.1), privacy-preserving registration process (Section 4.2), and stored data structure (Section 4.3), which may be different based on the underlying infrastructure.

Once a person is identified positive with a pandemic, a contact tracing procedure is triggered at a government-owned server to identify the set of recently visited facilities (Section 4.4). For each of these facilities, the server issues an API call that will trigger the facility query processing module to return a set of candidate contacts. An additional (optional) context-aware contact tracing analysis module (Section 4.5) can be applied on the server side to get more accurate contact tracing information.

4.1 Infrastructure

The underlying infrastructure would be significantly different for *User-to-User* and *Location-based* contact tracing as follows:
User-to-User Contact Tracing. Upon entering a facility, visitors will be given a Bluetooth Low Energy (BLE) gateway in the form of a detached device, wristband, or key chain, that will be handed back before leaving. Examples of such devices are here [3]. Each of these devices continuously broadcasts its own ID while reading the IDs of nearby devices. The reading log stored at each device would have the form *(BLE_ID, timestamp, signal_strength)* which presents the ID of the nearby device, along with the time and signal strength of reading that device. The latter is used as an indication of how close is the nearby device.
Location-based Contact Tracing. Upon entering a facility, visitors will be given a small Beacons device that would be returned before leaving. Examples of such devices are here [4]. Unlike the BLE gateway devices in *user-to-user* contact tracing, Beacons devices: (a) have much smaller size, and (b) only broadcast their own IDs, but do not read any signal. Meanwhile, the facility will have several gateway readers fixed on the walls or ceiling that read broadcasted data from the Beacon devices in the form of *(BLE_ID, timestamp, signal_strength)*. Examples of such gateways are here [5].

4.2 Privacy-Preserving Registration Process

The privacy-preserving registration process is the same for both *user-to-user* and *location-based* contact tracing. A facility visitor will need to sign-in upon entry using a government-owned machine by entering her phone number or government ID. The machine will immediately generate a unique random ID that is given to the facility in exchange of the BLE or Beacon device. The random ID will be sent either as SMS to the visitor phone to ensure it is an actual number or as SMS to the machine itself in case the visitor has provided government ID instead of phone number. This means that the facility knows nothing about the visitor personal data. To the facility, the visitor is just a government-generated unique random ID. Furthermore, government and facilities will completely wipe any data is more than two weeks old.

Frequent visitors to a facility, e.g., employees at a work place, frequent airport travelers, or loyal store customers, may need to do the sign up process only once, where their phone numbers will be linked with their employee IDs or loyalty numbers. Then, the BLE or Beacon devices can be given to them once and actually attached to their work IDs or loyalty cards that they have to scan upon entering the facility.

4.3 Stored Data Infrastructure

The data structure stored on the government-owned server is independent from the underlying infrastructure of each facility. It is basically one big table with the schema *(PhoneID, FacilityID, VisitorID, timestamp)*, which indicates that a user with a certain phone or ID number has visited a certain facility ID at a certain time, and was given a certain visitor ID. For efficient retrieval, the table is accessed through two hash tables for PhoneID and VisitorID.

Meanwhile, each facility, regardless of the underlying contact tracing infrastructure, maintains a *Master* table with the schema *(VisitorID, BLE_ID, time_in, time_out)*, which indicates that a certain BLE or Beacon ID was given to a certain visitor within a certain time frame. The *Master* table is accessed through two hash tables for VisitorID and BLE ID. In addition, each facility maintains the following data structure(s), based on the underlying infrastructure:

User-to-User Contact Tracing. Each facility maintains a *Contacts* table: *(BLE_ID1, BLE_ID2, timestamp, signal_strength)*, which logs the timestamp and signal strength for each pair of BLE devices that came close to each other. The signal strength is converted to some universal distance measure (e.g., meters) to accommodate that different devices may present signal strength differently. The table is populated by combining all readings received from individual BLE devices, and is accessed through a hash table over BLE_ID1. Data will be duplicated in this case because each device would insert a tuple about its contact of the other device. We can try to optimize, by creating a unique key which concatenates (sort(BLEID1, BLEID2), timestamp).

Location-based Contact Tracing. Each facility maintains a *Locations* table with the schema *(BLE_ID, location, timestamp)*, which logs the locations of each BLE within the facility at a certain timestamp. The indoor location is not a traditional $< lat, log >$ coordinates. Instead, it is more of a symbolic descriptive location based on the facility map [31]. The *Locations* table is accessed by a hash table over BLE_ID, and is populated through a typical trilateration process where the readings from three fixed Gateways for the same BLE is used to come up with the symbolic BLE location at a certain time [19]. Such process has been commonly used in indoor positioning systems for real-time asset tracking [34].

4.4 Contact Tracing Procedure

Once a person is confirmed positive for the pandemic, the contact tracing procedure is triggered on the server side by government officials. A simple local query with the patient phone (or ID) and a certain time period (e.g., last two weeks) would return the set of facilities F visited by the patient, with the anonymized visitor ID and timestamp of each visit. A patient may have visited the same facility multiple times, each with a different visitor ID. For each facility in F, the server sends an API call inquiry asking for all visitor IDs who were in contact with the patient visitor ID.

Whenever a facility receives a query with a visitor ID and a timestamp, it uses its *Master* table to map the visitor ID to the BLE ID used during the visit, along with the visit time frame. Then, based on the underlying infrastructure, the following information is retrieved and sent back to the server:

User-to-User Contact Tracing. Given a BLE ID and visit time frame, the facility will use its *Contacts* table to retrieve all other BLE IDs that were reported in contact with the visitor BLE ID, along with the timestamp and estimated distance of each contact event. Then, a reverse lookup over the *Master* table will get the corresponding Visitor ID for each contacted BLE ID. The information sent back to the requesting authority server will have the schema *(VisitorID, timestamp, estimated_distance).*

Location-based Contact Tracing. Given a BLE ID and visit time frame, the facility will use its *Locations* table to retrieve the trajectory of locations (with timestamps) within the facility during the visit. Then, a spatio-temporal indoor range query [23] would retrieve all the BLE IDs that were in a close spatio-temporal proximity to the given BLE ID. The parameters of spatio-temporal proximity are set in a conservative way, e.g., within 10 meters distance and 10 minutes time frame. Then, a reverse lookup over the *Master* table will get the corresponding Visitor ID for each nearby BLE ID. Finally, the information sent back to the requesting authority server is: *(VisitorID, location, timestamp, spatial proximity, temporal proximity).* In addition, each facility may optionally send the full spatio-temporal trajectory of the patient visitor within the facility, which can be used for further analysis at the requesting authority.

4.5 Context-Aware Contact Tracing Analysis

When the government-owned server receives back the results from each facility, it may just use the list of contacts or nearby visitors as the ones in risk. In this case, a reverse lookup with the Visitor ID over the local server table would return the phone number (or ID) of each visitor. Health officials can take it from there and start contacting the people accordingly. However, as a means of increasing the accuracy, additional context-aware contact tracing analysis can be employed based on the underlying infrastructure:

User-to-User Contact Tracing. The signal strength of each contact may be interpreted differently based on the facility type. For example, within a stadium, one may focus on the readings with high signals. Within a restaurant, one may even report lower signals, only if they were persistent over a certain period of time. Within a Mall, a different search criteria and parameters can be used.

Location-based Contact Tracing. Assuming the availability of facility layout, the spatial and temporal proximity of visitors may be interpreted differently based on the facility type and layout. Two contacts who are close by spatially and temporally may have a wall in between, and hence the proximity is not risky. Meanwhile, a heatmap may be depicted for the facility indicating regions of high risk, where contact information may be interpreted differently. Furthermore, depending on the nature of the pandemic and how it spreads (e.g., via surface or air), we can find users who have been to the spots recently visited by a patient. For example, a patient who uses a table in a food court, leaves it, then another visitor uses the same table.

Generally speaking, there are many context-aware indoor analysis that can be deployed [1], though there are way more rich analysis for the case of *location-based* contact tracing than *user-to-user* contact tracing. Having the context-aware analysis module on the server side instead of having it on each facility is mainly to allow health officials to change the parameter settings and search criteria without the need to get back to the facility. Another alternative is to have such analysis on the facility side, accessed via more sophisticated API calls that account for more parameters such as minimum distance, contact time interval, and location label.

5 Conclusion

The paper makes the case that current app-based contact tracing techniques are not effective. Then, the paper lays out the vision for the next era of digital contact tracing where the responsibility of contact tracing is moved from the persons to the facilities that the persons visit. Each facility, e.g., mall, work place, stadium, train, restaurant, should have the ability to do contact tracing for all its visitors. Such ability could be enforced as a prerequisite for any facility to re-open during a pandemic. A privacy-preserving architecture and infrastructure that achieve such vision is presented. The architecture allows each facility to independently decide whether to deploy a *user-to-user* or *location-based* contact tracing approach. The former approach mainly reports the people in contact to the patient, while the second approach additionally reports the locations of the patient and the contacts.

References

[1] I. Afyouni, C. Ray, S. Ilarri, and C. Claramunt. Algorithms for Continuous Location-dependent and Context-aware Queries in Indoor Environments. In *SIGSPATIAL*, pages 329–338, Redondo Beach, CA,, Nov. 2012.

[2] Associate of State and Territorial Health Officials (ASTHO). A Coordinated, National Approach to Scaling Public Health Capacity for Contact Tracing and Disease Investigation, apr 2020.

[3] Beacon Zone. AB BLE Gateway. https://www.beaconzone.co.uk/Gateways/ABBLEGateway.

[4] Beacon Zone. All Beacons. https://www.beaconzone.co.uk/allbeacons.

[5] Beacon Zone. Gateways. https://www.beaconzone.co.uk/Gateways.

[6] Brookings Institute. Class and COVID: How the less affluent face double risks. https://www.brookings.edu/blog/up-front/2020/03/27/class-and-covid-how-the-less-affluent-face-double-risks/.

[7] Business Insider. Iceland had the most-downloaded contact-tracing app for its population size. Authorities there say it hasn't made much difference. https://www.businessinsider.com/iceland-contact-tracing-not-gamechanger-2020-5.

[8] Business Insider. In a South Korean call center, 44% of workers on one floor got the coronavirus. https://www.businessinsider.com/south-korean-call-center-covid-19-outbreak-seating-chart-2020-4.

[9] Lost your job? Consider becoming a "contact tracer". https://www.cbsnews.com/news/contact-tracing-jobs-covid/.

[10] Center for Disease Control and Prevention. *Core Curriculum on Tuberculosis: What the Clinician Should Know*, chapter Chapter 8: Community Tuberculosis, pages 227–248. 2013.

[11] J. Chan, D. Foster, S. Gollakota, E. Horvitz, J. Jaeger, S. Kakade, T. Kohno, J. Langford, J. Larson, P. Sharma, S. Singanamalla, J. Sunshine, and S. Tessaro. PACT: Privacy Sensitive Protocols and Mechanisms for Mobile Contact Tracing, 2020.

[12] J. Clarke. Contact Tracing for Chlamydia: Data on Effectiveness. *International Journal of STD & AIDS*, 9:187–191, 1998.

[13] Computing Community Consortium (CCC). Contact Tracing for All? Bridging the Accessibility Gap for Contact Tracing. https://www.cccblog.org/2020/05/26/computing-researchers-respond-to-covid-19-contact-tracing-for-all-bridging-the-accessibility-gap-for-contact-tracing/.

[14] EU unveils plans to save vacations and avoid a lost summer. https://www.ctpost.com/news/article/EU-unveils-its-plan-to-save-summer-vacations-15266638.php.

[15] K. T. D. Eames and M. J. Keeling. Contact Tracing and Disease Control. *Proceedings of the Royal Society B: Biological Sciences*, 270:2565–2571, 01 2004.

[16] Qatar makes COVID-19 app mandatory, experts question efficiency. https://www.aljazeera.com/news/2020/5/26/qatar-makes-covid-19-app-mandatory-experts-question-efficiency.

[17] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dorner, M. Parker, D. Bonsall, and C. Fraser. Quantifying SARS-CoV-2 Transmission suggests Epidemic Control with Digital Contact Tracing. *Science*, 368:621–626, May 2020.

[18] Harvard Business Review. How Digital Contact Tracing Slowed Covid-19 in East Asia. https://hbr.org/2020/04/how-digital-contact-tracing-slowed-covid-19-in-east-asia.

[19] IoT for All. Trilateration vs. Triangulation for Indoor Positioning Systems. https://www.iotforall.com/trilateration-vs-triangulation-indoor-positioning-systems/.

[20] I. Z. Kiss, D. M. Green, and R. R. Kao. Disease Contact Tracing in Random and Clustered Networks. *Proceedings of the Royal Society B: Biological Sciences*, 272:1407–1414, 08 2005.

[21] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, J. V. B. Paul Switzer, S. C. Hern, and W. H. Engelmann. The National Human Activity Pattern Survey (NHAPS): A Resource for Assessing Exposure to Environmental Pollutants. *Journal of Exposure Analysis and Environmental Epidemiology*, 11(3):231–252, May 2001.

[22] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science*, 300:1966–1970, June 2003.

[23] H. Lu, X. Cao, and C. S. Jensen. A Foundation for Efficient Indoor Distance-Aware Query Processing. In *ICDE*, pages 438–449, Arlington, VA, Nov. 2012.

[24] K. Michael and R. Abbas. Behind covid-19 contact trace apps: The google–apple partnership. *IEEE Consumer Electronics Magazine*, 9(5):71–76, 2020.

[25] NBC Philadelphia. To Track Coronavirus, States Are Ramping Up Contact Tracing. But What Is It? `https://www.nbcphiladelphia.com/news/local/to-track-coronavirus-states-are-ramping-up-contact-tracing-but-what-is-it/2395081/`.

[26] Office of California Governer. Governor Newsom Outlines Six Critical Indicators the State will Consider Before Modifying the Stay-at-Home Order and Other COVID-19 Interventions. `https://www.gov.ca.gov/2020/04/14/governor-newsom-outlines-six-critical-indicators-the-state-will-consider-before-modifying-the-stay-at-home-order-and-other-covid-19-interventions/`.

[27] N. Oliver, B. Lepri, H. Sterly, R. Lambiotte, S. Deletaille, M. De Nadai, E. Letouzé, A. A. Salah, R. Benjamins, C. Cattuto, et al. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, 2020.

[28] Patrick Howell O'Neill, Tate Ryan-Mosley, Bobbie Johnson. A flood of Coronavirus Apps are tracking us. Now it is time to keep track of them. MIT Technology Review. `https://www.technologyreview.com/2020/05/07/1000961/launching-mittr-covid-tracing-tracker/?utm_medium=tr_social&utm_campaign=site_visitor.unpaid.engagement&utm_source=Twitter#Echobox=1588950967`.

[29] T. C. Porco, K. A. Holbrook, S. E. Fernyak, D. L. Portnoy, R. Reiter, and T. J. Aragon. Logistics of Community Smallpox Control through Contact Tracing and Ring Vaccination: A Stochastic Network Model. *BMC Public Health*, 4(34), Aug. 2004.

[30] Safe Paths. `https://www.media.mit.edu/projects/safepaths/overview/`.

[31] N. Samama. *Indoor Positioning: Technologies and Performance.* Wiley, 2019.

[32] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging Infectious Disease Journal*, 26(7), July 2020.

[33] Statista. Smartphone Ownership Rate by Country 2018. `https://www.statista.com/statistics/539395/smartphone-penetration-worldwide-by-country/`.

[34] Team Software. Six Ways to Use Bluetooth Beacons for People and Asset Tracking. https://teamsoftware.com/blog/2019/11/26/six-ways-to-use-bluetooth-beacons-for-people-and-asset-tracking/.

[35] The Sentinel: Cumberland County. State contact tracing plan expected. https://cumberlink.com/news/state-and-regional/state-contact-tracing-plan-expected/article_a409065b-c9e3-5e76-8ced-c2757dd15056.html.

[36] The Straights Time. Singapore. About 1 million people have downloaded TraceTogether app, but more need to do so for it to be effective. https://www.straitstimes.com/singapore/about-one-million-people-have-downloaded-the-tracetogether-app-but-more-need-to-do-so-for.

[37] Trace Together. https://www.tracetogether.gov.sg/.

[38] C. Troncoso, M. Payer, J.-P. Hubaux, M. Salathé, J. Larus, E. Bugnion, W. Lueks, T. Stadler, A. Pyrgelis, D. Antonioli, et al. Decentralized privacy-preserving proximity tracing. *arXiv preprint arXiv:2005.12273*, 2020.

[39] Vox. What the US can learn from other countries using phones to track Covid-19. https://www.vox.com/recode/2020/4/18/21224178/covid-19-tech-tracking-phones-china-singapore-taiwan-korea-google-apple-contact-tracing-digital.

[40] WHO coronavirus briefing: Isolation, testing and tracing comprise the backbone of response. https://www.weforum.org/agenda/2020/03/testing-tracing-backbone-who-coronavirus-wednesdays-briefing/.

[41] World Health Organization. Contact Tracing Q&A. https://www.who.int/news-room/q-a-detail/contact-tracing.

[42] World Health Organization and Centers for Disease Control and Prevention. Implementation and Management of Contact Tracing for Ebola Virus Disease. https://www.who.int/csr/resources/publications/ebola/contact-tracing/en/, Sept. 2015.

[43] WTOP News. Northam shares rough blueprint to reopening Virginia. https://wtop.com/virginia/2020/04/northam-shares-rough-blueprint-to-reopening-virginia/.

Geospatial forecasting of COVID-19 spread and risk of reaching hospital capacity

Georgiy Bobashev[1], Ignacio Segovia-Dominguez[2], Yulia R. Gel[2],
James Rineer[1], Sarah Rhea[1], Hui Sui[3]
[1]Center for Data Science, RTI International, USA
[2]Department of Mathematical Sciences, University of Texas at Dallas, USA
[3]University of North Carolina at Chapel Hill, USA
{bobashev,jrin,srhea,hsui.contractor}@rti.org, {ignacio.segoviadominguez,ygl}@utdallas.edu

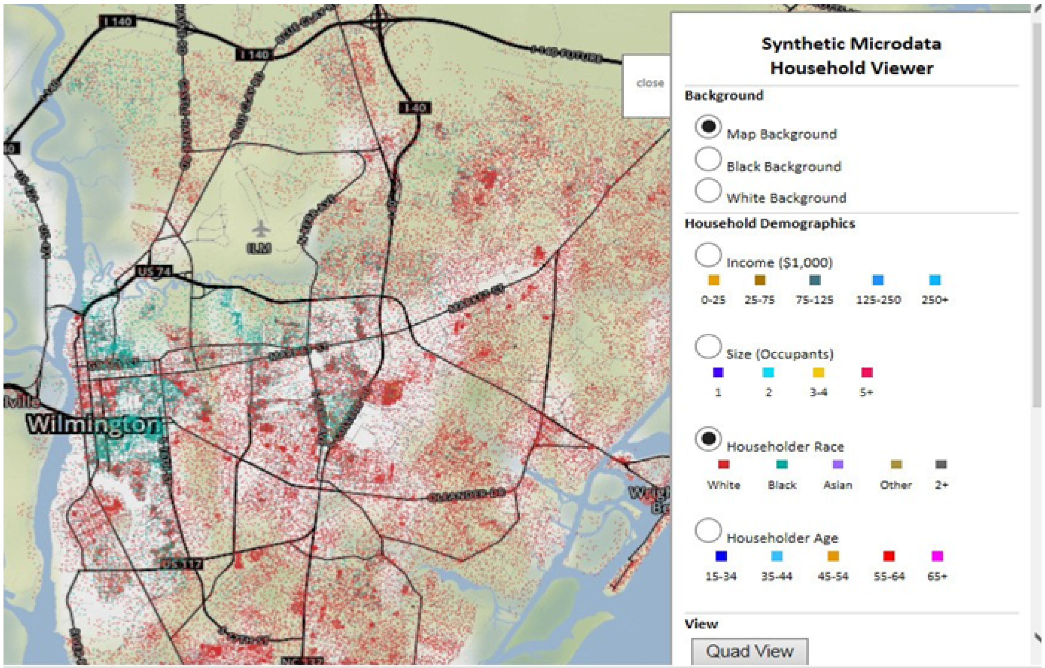
Abstract

Prompt surveillance and forecasting of COVID-19 spread are of critical importance for slowing down the pandemic and for the success of any public mitigation efforts. However, as with any infectious disease with rapid transmission and high virulence, lack of COVID-19 observations for near-real-time forecasting is still the key challenge obstructing operational disease prediction and control. In this context, we can follow the two approaches to forecasting COVID-19 dynamics: based on mechanistic models and based on machine learning. Mechanistic models are better in capturing an epidemiological curve, using a low amount of data, and describing the overall trajectory of the disease dynamics, hence, providing long-term insights into where the disease might go. Machine learning, in turn, can provide more precise data-driven forecasts especially in the short-term horizons, while suffering from limited interpretability and usually requiring backlog history on the infectious disease. We propose a unified reinforcement learning framework that combines the two approaches. That is, long-term trajectory forecasts are used in machine learning techniques to forecast local variability which is not captured by the mechanistic model.

1 Introduction

Spatio-temporal forecasts of infectious diseases rapidly move to the forefront of policy and public health response because of their key role in risk mitigation strategies. During the COVID-19 pandemic, this has become especially important in areas with high demographic, economic, and political variability. For example, in North Carolina decisions on opening and closing businesses because of COVID-19 are made at the state, county, and local levels. When Wake County leadership announced the lifting of some restrictions, the town of Apex issued an order to continue keeping the strongest restrictions. Similarly, areas that are either remote or used for seasonal vacations (e.g., mountain or beach counties) might exhibit different disease dynamics than those produced by the rest of the state. Because of high spatial heterogeneity, COVID-19 forecasting is important at the local level. Local outbreaks could overwhelm public health systems, hospitals, and emergency rooms.

Hence, to facilitate hospital preparedness, it is critically important to forecast hospital capacity and probabilities that capacity could be overtaken by emerging patients. This relies on two interconnecting models: one predicts how many people are likely to become infected in each of the areas and another describes how infected individuals move to and between appropriate health care centers. This second model provides the eventual answer to health officials, while the first model stays in the background until needed. To provide the solution at the local level we use an explicit synthetic population, a database that represents the entire population of U.S. residents (over 300 million synthetic individuals). For each of the synthesized individuals (agents), this database contains demographic characteristics and geographic location (Figure 1). More details and a working viewer are available at `http://synthpopviewer.rti.org/` and elsewhere [3, 13]. Multiple layers can be added to the database to make it relevant to a specific question. The COVID-19 layers could include school and work assignment, hospital and emergency department, etc.



Synthetic Microdata
Household Viewer

Background

● Map Background
○ Black Background
○ White Background

Household Demographics

○ Income ($1,000)

○ Size (Occupants)

● Householder Race

○ Householder Age

View

Figure 1: Snapshot of the synthetic population for Wilmington, NC.

For the purpose of forecasting hospital load in North Carolina, a team of Response to Intervention (RTI) researchers developed a spatially explicit agent-based model (ABM) that forecasts to which hospital sick patients are likely to move and where they might be transferred if the hospital is over capacity or doesn't have the proper equipment. This model was developed for North Carolina and considers a synthetic population, where synthesized agents represent over 10 million North Carolina residents. The model also uses 110 short-term acute care hospitals (STACHs), 421 nursing homes, and 10 long-term acute care hospitals (LTACHs). At each day timestep, individual health status and location are updated. Figure 2 shows a map with marked locations of health care providers.

The other model provides a forecast of how many individuals are going to be infected and how many will be sick to the point of going to the hospital in the future. Such a model can have different levels of granularity. One level is a county-level system dynamics model that assumes homogeneous mixing (each individual has the same chance to meet any other individual), which leads to the mass action principle, where the risk of infection is proportional to the prevalence of infectious individuals and the proportion of susceptible individuals. Under these assumptions, individuals are equally likely to get infected and thus randomly spread infection through the population. More sophisticated models can include age and social structure (some people have more and closer contacts than others) and geographic locations and allow for disease transmission to occur in clusters.

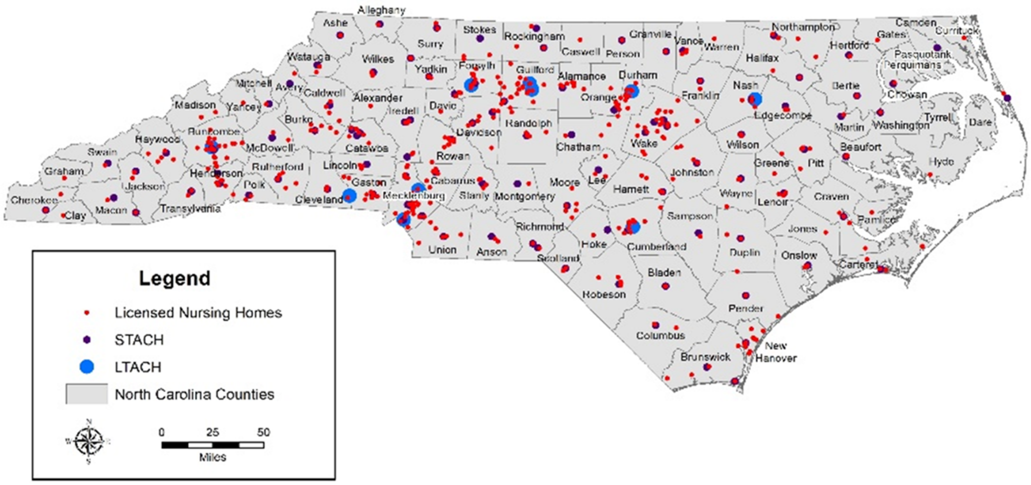


Figure 2: Locations of health care providers

ABM approaches have been widely used in a large number of areas including health care, epidemiology, economics and finance, and social sciences. Epidemiology of infectious diseases is one of the most natural areas to apply it. For example, an ABM was developed to forecast pandemic flu in North Carolina and New York, where the model explicitly described the household structure and people's movements and activities through the day including going to work and school, being in a household, or in community spaces and transportation [4, 5]. These models allowed us to identify the areas at the highest risk and estimate the contribution of these areas to the epidemic. For example, for New York, it was shown that public transportation (including subways) contributed less than 10% to disease incidence, while households and schools provided the majority of new cases and thus acted as transmission clusters [10]. Recently we considered synthetic populations to describe seasonal influenza in Russia [8]. The study had an additional challenge of estimating background susceptibility.

The length of the forecasting time horizon is critical for preparedness, but it is also challenging. The longer the time horizon, the higher the uncertainty, especially when policy and environment change in unpredictable ways. Nevertheless, one could foresee at least some impacts such as weekends, major holidays, and scheduled public announcements such as on business and school closures and openings.

2 Methodology

There are at least two approaches to forecasting COVID-19 dynamics: based on mechanistic models and based on machine learning. Mechanistic models are better at capturing an epidemiological curve and describing the overall trajectory of disease dynamics, hence providing long-term insights into where the disease might go. Machine learning, in turn, can provide more precise data-driven forecasts, especially in the short-term horizons, while suffering from limited interpretability. We propose a unified reinforcement learning (RL) framework combining the two approaches. That is, long-term trajectory forecasts are used in machine learning techniques to forecast local variability, which is not captured by a mechanistic model.

The choice of the "best" model generally balances model fidelity, explanatory features, data availability, and computational requirements. A summary of these features is described in [2] and is presented in Figure 3. In [11] we conducted a comparison of the simpler system dynamics model and an ABM of pandemic flu with a number of interventions. Not surprisingly, higher granularity brings higher fidelity but also increases uncertainty because of the variability of model parameters and structural assumptions.

In the current study, we start mechanistic modeling with a system dynamics approach. System dynamics models are quick to execute and thus are easy to calibrate. Future ABMs will start with the average parameter values of the system dynamics model and will expand around those values. Our system dynamics model divides the population into Susceptible, Exposed, Infectious, and Recovered individuals and considers the movements of individuals between these compartments. Thus, the model is commonly called an SEIR model. Specifically, for COVID-19 we also consider whether individuals are symptomatic or asymptomatic, which in turn requires estimation of disease transmission from an asymptomatic person. Assuming homogeneous mixing, the initial SEIR model can be described in a differential equation form:

$$\begin{aligned}
\frac{dS^m}{dt} &= -\sum_v (\beta_a^{mv} I_a^v + \beta_s^{mv} I_s^v) S^m), \qquad (1)\\
\frac{dE_k^m}{dt} &= \sum_v p_k^m [\beta_a^{mv}(I_a^v - \xi_a I_a^v) + \beta_s^{mv}(I_s^v - \xi_s I_s^v)] S^m - \mu_k E_k^m,\\
\frac{dI_k^m}{dt} &= \mu_k E_k^m - \gamma_k I_k^m - \delta_k I_k^m \qquad \frac{dR_k^m}{dt} = \gamma_k I_k^m
\end{aligned}$$

where S, E, I, and R denote susceptible, exposed, infected, and recovered, respectively, and for simplicity total immunity is assumed. In turn, $k = a, s$ denote symptomatic and asymptomatic subgroups, respectively; p_k is a fraction of symptomatic or asymptomatic cases; and β, μ, γ and δ are transmission rate, infectivity period, recovery, and excess mortality rates, respectively. Quarantine of identified symptomatic and asymptomatic cases is denoted by ξ. We assume that there exists no excess mortality among asymptomatic cases (i.e., $\delta_a = 0$). Upper indexes correspond to age groups m and v. Finally, we assume homogeneity of the daily number of contacts within an age group and heterogeneity between

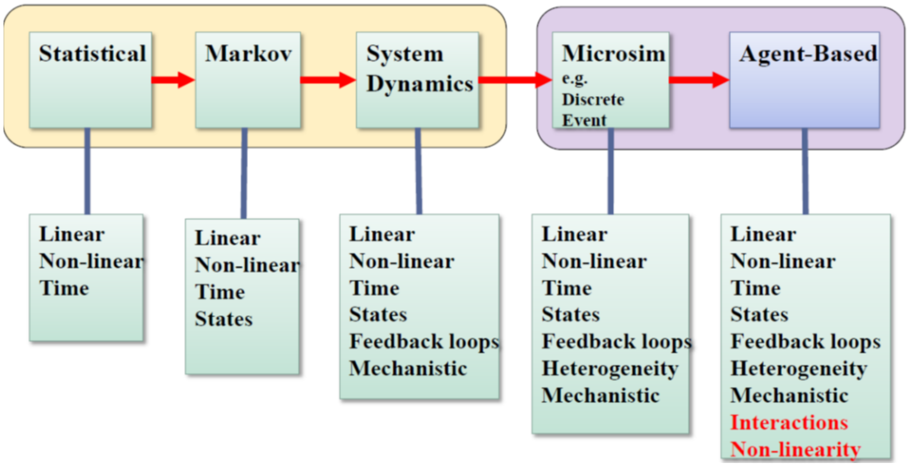


Figure 3: Hierarchy of simulation models in increasing levels of complexity. At the higher level Statistical, Markov, and System dynamics models do not distinguish between individuals in the populations and describe populations (or subpopulations) as a whole. As the names suggest, microsimulation and agent-based models describe each individual in the population and thus could be averaged across specific characteristics to obtain population-level estimates. In the current project, we consider a challenging question. Do these modeling approaches have to be mutually exclusive? Each brings something to the forecast, and perhaps we can benefit from combining at least two of them.

groups.

This model is defined at the county level and could be further expanded to represent rural or urban parts of the county, adding migration of individuals between counties and between urban and rural county components. The model is adequate enough to describe the dynamics of COVID-19 in most counties. For low-density rural counties (e.g., in Appalachia) the assumption of homogeneous mixing is no longer valid, and an explicit ABM might be more adequate. For the sake of simplicity, we still keep the differential equation formalism but add a stochastic transmission component. Future models will be fully agent-based to consistently describe local geographic elements of disease clusters and patient assignments to hospitals. We fit the model to the reported case data where we also need to consider multiple reporting biases such as under-reporting, reporting of certain subpopulations (e.g., age 65+), and the availability of disease test kits.

Disease transmission parameters are key to the understanding of future disease dynamics; therefore, we also consider county vulnerability indices, which we calculated based on multiple sources of socioeconomic and health data. We have developed a county vulnerability dashboard (Figure 4) that is publicly available at `https://RTImerge.org`.

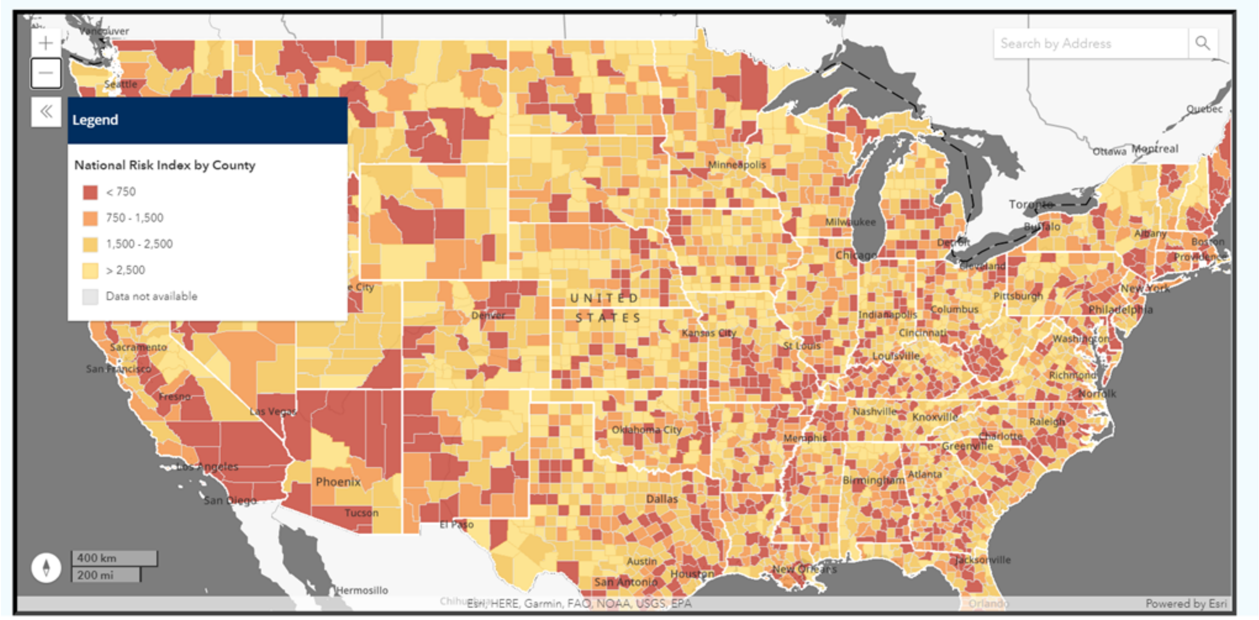


Figure 4: County vulnerability index with respect to COVID-19 dynamics

This mechanistic modeling effort results in describing the dynamics of symptomatic and asymptomatic infected individuals. Figure 5 shows an example of a model fit to Gaston county data. Deterministic SEIR models produce a smooth curve fit that tracks past mean-field disease dynamics. The key component of prediction is forecasting how policy measures will impact the transmission rate. For that purpose, we developed a model linking the dynamics of a SEIR beta parameter with public health actions. Mechanistic models allow one to simulate a variety of scenarios and pre-train a deep learning model on these scenarios. Although the SEIR model allows us to consider "what if" scenarios and provide mechanistic explanations of "why," they don't capture all the richness of disease dynamics. The addition of stochastic components leads to consideration of uncertainty and higher boundaries of risk through a family of stochastic realizations.

Mechanistic models are limited in their applications because they only produce results based on a hardwired mechanism which could be miss-specified. Furthermore, such models can neglect a variety

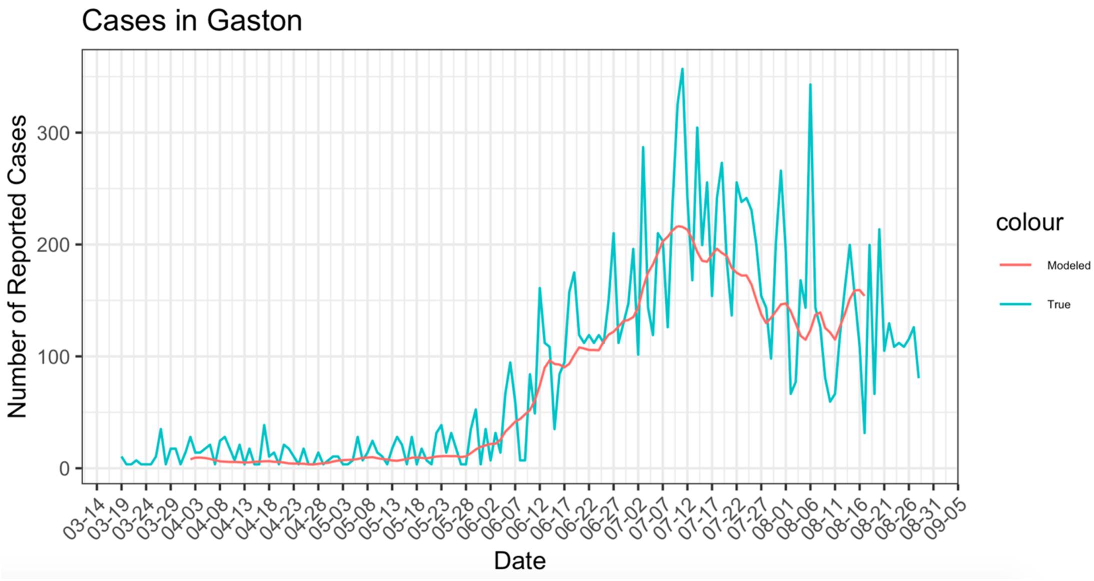


Figure 5: Examples of model fit to Gaston County case reports

of smaller factors that are collectively influential. These factors could be captured by a mechanism-agnostic data-driven model that learns with the data. For example, Long-Short Term Memory (LSTM) algorithms can capture short- and long-term factors that can also change in time.

In this project, we aim to develop a reinforcement learning (RL) framework which systematically combines mechanistic SEIR models with data-driven LSTM algorithms to get the best of both worlds: that is, the interpretability of mechanistic models and predictive capabilities of deep learning methods.

Our approach is described below and is illustrated in Figure 6:

1. Based on past data we fit the SEIR model and train the LSTM model.

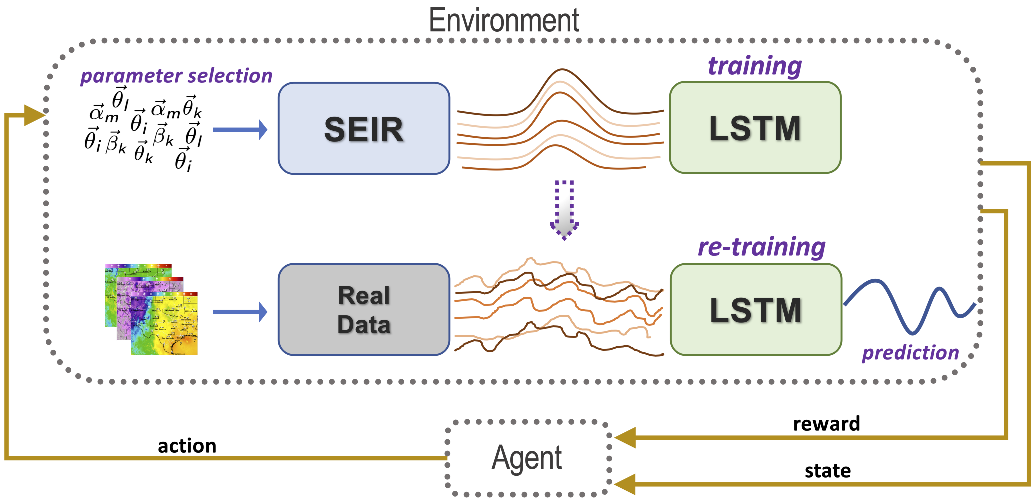
2. Based on forecasted subjective beliefs on policy changes in the future use the SEIR model to predict the numbers of infected individuals (smooth curves).

3. Use the SEIR-predicted data as input into the LSTM model to produce an improved forecast.

4. Through reinforcement learning update both the SEIR and LSTM forecasts as new data become available.

Our goal is to build a highly adaptive model capable of readjusting its prediction on the future disease course, in accordance with changes in public response and bio-atmospheric information. As shown in Figure 6, RL plays an important role to control the dynamic of the SEIR beta parameter, which is directly linked to public health actions. Hence, this model-free RL adapts its parameters on the fly (i.e., learning from experience and using the latest updated official data). As a consequence, we will be able to update our predictions in potentially highly uncertain and volatile disease scenarios such as the current coronavirus spread.

The main difference between the proposed RL methodology with respect to other approaches is the ability to summarize multiple policies for outbreak response via parameter adaptation in our mechanistic model. Indeed, existing techniques primarily focus on learning context-dependent policies for complex epidemiological models, in which the conventional approach consists of evaluating the expected



Approach combining mechanistic models with deep learning via reinforcement learning

Figure 6: Reinforcement learning of mechanistic (SEIR) and data-driven (LSTM) models.

performance of different potential interventions via stochastic simulations [9, 12]. However, such approaches predominantly model the RL problem from a spatial perspective, leaving the time variable as only part of the evolution of the states [1, 6] and mostly obtaining optimal decisions in fully observable environments [7]. In contrast, our proposed approach accounts for the dynamics of COVID epidemics via mechanistic models and allows us for building highly adaptive models capable of readjusting the delivered forecasts for the future disease course.

3 Conclusion and Future Work

The ultimate goal of this project is to develop a novel methodology for forecasting the COVID-19 spread via synergistic interaction between mechanistic and data-driven models under the RL framework. In particular, the proposed methodology is based on the new idea of using RL as a part of solving a time-series forecasting problem under the assumption of dynamic stability and requires identification of the following main components: states, environment, reward function, and agent interactions. Since RL focuses on learning an optimal policy, we also need to obtain significant feedback between agents and the ongoing behavior of the system and ensure that our adaptive learning still can be formulated as a Markov Decision Process problem (i.e., the challenges that are both largely unsolved in a context of RL for space-time data).

By effectively combining mechanistic models with deep learning tools, the proposed RL approach to epidemiological forecasting can harness the strength of both theoretical and data-based models and deepen our understanding of the hidden mechanisms behind COVID-19 progression. In the near term future, we plan to investigate the utility and limitations of the proposed methodology at the county level and then investigate the transferability of the derived tools to other states and spatial data resolution.

Acknowledgements

This work is supported in part by the grants from the National Science Foundation DMS 2027793/2027802 and the ConTex Postdoctoral Fellowship grant.

References

[1] D. Bertsekas. *Rollout, Policy Iteration, and Distributed Reinforcement Learning.* Athena Scientific, 1 edition, 8 2020.

[2] G. Bobashev. Simulation modeling of hiv infection—from individuals to risk groups and entire populations. In C. Chan, M. G. Hudgens, and S.-C. Chow, editors, *Quantitative Methods for HIV/AIDS Research*, chapter 10, pages 201–229. Chapman & Hall/CRC Biostatistics Series, 1 edition, 8 2017.

[3] J. Cajka, P. Cooley, and W. Wheaton. Attribute assignment to a synthetic population in support of agent-based disease modeling. *Methods report (RTI Press)*, 19:1–14, 09 2010.

[4] P. Cooley, L. Ganapathi, G. Ghneim, S. Holmberg, W. Wheaton, and C. Hollingsworth. Using influenza-like illness data to reconstruct an influenza outbreak. *Mathematical and Computer Modelling*, 48:929–939, 02 2008.

[5] N. Ferguson, D. Cummings, C. Fraser, J. Cajka, P. Cooley, and S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442:448–52, 08 2006.

[6] P. Hernandez-Leal, B. Kartal, and M. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33, 10 2019.

[7] S. Ivanov and A. D'yakonov. Modern deep reinforcement learning algorithms, 2019. Available at https://arxiv.org/abs/1906.10025.

[8] V. Leonenko and G. Bobashev. Analyzing influenza outbreaks in russia using an age-structured dynamic transmission model. *Epidemics*, 29:100358, 2019.

[9] P. Libin, A. Moonens, T. Verstraeten, F. Perez-Sanjines, N. Hens, P. Lemey, and A. Nowé. *Deep reinforcement learning for large-scale epidemic control.* ALA 2020, Auckland, New Zealand, 5 2020.

[10] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. T. Cummings, and M. E. Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[11] J. Mathieu, M. Pfaff, G. Klein, J. Drury, M. Geodecke, J. James, P. Mahoney, and G. Bobashev. *Tactical Robust Decision-Making Methodology: Effect of Disease Spread Model Fidelity on Option Awareness.* Seattle, WA, 01 2010.

[12] W. Probert, S. Lakkur, C. Fonnesbeck, K. Shea, M. Runge, M. Tildesley, and M. Ferrari. Context matters: using reinforcement learning to develop human-readable, state-dependent outbreak response policies. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 374:20180277, 07 2019.

[13] W. D. Wheaton, J. Cajka, B. M. Chasteen, D. Wagener, P. Cooley, L. Ganapathi, D. J. Roberts, and J. Allpress. Synthesized population databases: A us geospatial database for agent-based models. *Methods report*, 2009 10:905, 2009. Available at http://www.rti.org/sites/default/files/resources/mr-0010-0905-wheaton.pdf.

COVID-19 Ensemble Models Using Representative Clustering

Joon-Seok Kim[1], Hamdi Kavak[2], Andreas Züfle[1], Taylor Anderson[1]
[1]Department of Geography and Geoinformation Science, George Mason University, USA
[2]Department of Computational and Data Sciences, George Mason University, USA
{jkim258,hkavak,azufle,tander6}@gmu.edu

Abstract

In response to the COVID-19 pandemic, there have been various attempts to develop realistic models to both predict the spread of the disease and evaluate policy measures aimed at mitigation. Different models that operate under different parameters and assumptions produce radically different predictions, creating confusion among policy-makers and the general population and limiting the usefulness of the models. This newsletter article proposes a novel ensemble modeling approach that uses representative clustering to identify where existing model predictions of COVID-19 spread agree and unify these predictions into a smaller set of predictions. The proposed ensemble prediction approach is composed of the following stages: (1) the selection of the ensemble components, (2) the imputation of missing predictions for each component, and (3) representative clustering in application to time-series data to determine the degree of agreement between simulation predictions. The results of the proposed approach will produce a set of ensemble model predictions that identify where simulation results converge so that policy-makers and the general public are informed with more comprehensive predictions and the uncertainty among them.

1 Introduction

SARS-CoV-2 is a highly contagious human respiratory coronavirus resulting in mortality across the United States and worldwide [2]. Researchers have made considerable efforts to understand the virus' infection dynamics and develop various models to shed light on the future. Forecasts obtained from the models are used to predict the number of cases and deaths to support the development of effective policy interventions and the public health response. However, the wide range of COVID-19 models employ different parameter settings, are designed based on various assumptions, and are inherently uncertain. As a result, existing models produce a range of radically different predictions making it difficult for decision-makers and the broader public to understand, compare-between, and validate them, creating barriers to their use. Therefore, there is an urgent need to cross-evaluate the wide-range of existing COVID-19 models, find a consensus among their predictions, and increase the transparency of model assumptions and their inherent uncertainty.

Ensemble modeling is a term that describes the wide range of approaches used to combine predictions from multiple models, also known as components [28]. Components can be mathematical, curve-fitting, or agent-based models and typically operate under a range of different assumptions and use different data sources. The ensemble components can be combined using various algorithms, one of which is referred to as stacked generalization [39] or stacking. In this approach, a single ensemble is generated by simply averaging predictions derived from equally weighted components. In variations of this approach, ensemble components may be weighted based on whether they meet a specific condition. These weights may be assigned statically or may change adaptively over time [22]. Aside from averaging, some ensembles are generated using the median, the trimmed mean to exclude extreme predictions, voting, Bayesian model averaging, multiple linear regression, and principal component regression [38].

Ensemble models often have been found to outperform any single model by offsetting component biases [33, 41]. If the components are diverse and independent, ensemble approaches can generate predictions with increased prediction accuracy and reduced error variance [17]. Thus, ensemble modeling has been utilized extensively to make predictions about weather and climate [18, 21], hydrologic processes [38], species distributions [10], and more recently infectious disease including influenza [31], Ebola hemorrhagic fever [37], dengue [14, 40], and COVID-19 [1, 23, 30].

Traditionally, ensemble approaches summarize the various predictions between components into one single prediction. However, the reliance on ensemble means without critical examination of the ensemble components can be dangerous. Mackenzie [21] illustrates this concept using three models, each of which indicates that a river is unsafe to cross at some point. Yet the average of the models says otherwise. In other words, acknowledging the assumptions and the resulting variation and bias among component predictions is important and can hold key information that explains future conditions otherwise ignored by their ensembles.

Therefore, we propose the development and implementation of an ensemble approach using representative clustering [32, 43] that is capable of exploring the various dimensions of agreement between ensemble components and thus is not limited to combining the component predictions into a single prediction. The novel representative clustering approach is proposed as follows: (1) selection of the ensemble components, (2) imputation of the missing predictions for each model, and (3) application of representative clustering to develop ensembles.

In this newsletter article, we begin by introducing the wide range of existing COVID-19 models that are available as potential ensemble components, their predictions, and their uncertainty . Next, we propose the novel ensemble prediction approach that will be used to unify selected components as ensembles. Finally, we present some initial results before describing our next steps.

2 Ensemble Clustering Approach

This study proposes the development of a novel ensemble prediction approach (see Figure 1) that is capable of exploring the various dimensions of agreement between ensemble components and thus is not limited to combining the component predictions into a single prediction. The proposed ensemble prediction approach is composed of the following stages: (1) the selection of the ensemble components, (2) the imputation of missing predictions for each model, and (3) the application of representative clustering so that it can be applied to time-series data and thus determine the degree of agreement between simulation predictions. Model cross-comparison is a rare practice in the modeling community. Therefore, our efforts to bring together different COVID-19 models and cross-compare them is also a novel contribution.

2.1 Model Integration

Selection of Model Components and Parameters. With the rapid spread of SARS-CoV-2, researchers have been designing simulation models to predict new cases and deaths as well as to understand the impact of different mitigation measures such as social distancing and mandatory lockdowns.

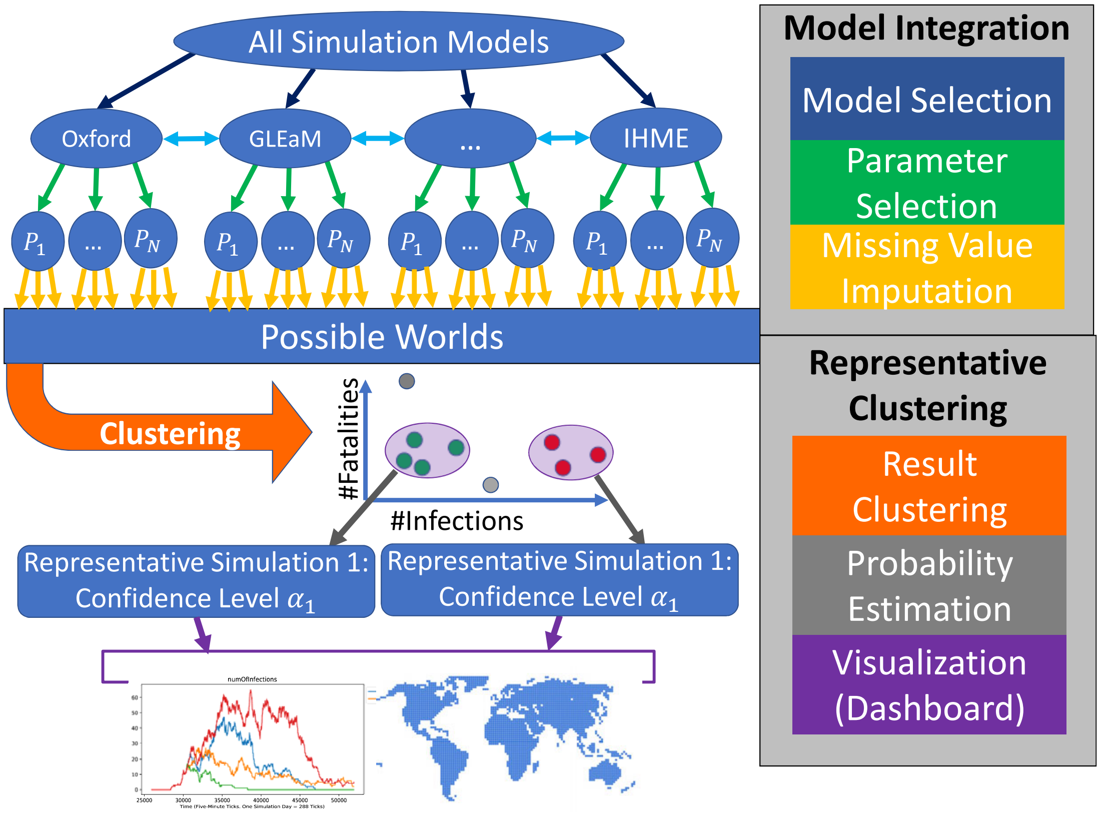


Figure 1: Ensemble clustering approach overview.

These models develop and implement approaches ranging from metapopulation, curve-fitting and statistical, as well as agent-based and in many cases the source code and the prediction data has been made publicly available. For example, the Global Epidemic and Mobility Model (GLEaM) is a metapopulation model that combines geographic mobility and population data with disease dynamics [36]. This effort was adapted and calibrated to model many outbreaks, including most recently the COVID-19 pandemic [4]. Another team of experts from the Los Alamos National Laboratory utilized their expertise in disease modeling and developed a statistical model to make new case and death predictions for COVID-19. Predictions from this model are publicly available [5]. The Imperial College COVID-19 Response Team adapted an established agent-based epidemic model [8, 12] to COVID-19 as well as developed a new mathematical model to estimate disease spread [7]. The Institute for Health Metrics and Evaluation (IHME) has developed a curve-fitting type of statistical model to project new cases and hospital beds needed [24], which is publicly available to use [13].

The models described above are just a few of the many that have been developed and implemented to predict COVID-19 trajectories of spread (see also Table 1 for more examples). In general, the CDC splits existing models into two categories [3]. One category includes models that make predictions under the assumption of business as usual meaning that existing control methods will remain in place [29, 34, 35]. The other category includes models that make predictions under different possible scenarios, usually with respect to testing the effect of different policy measures or the degree to which the population follows these guidelines [6, 11, 25, 27].

In the model integration stage, we will select a number of models as ensemble components. Table 1 presents some examples of existing models with open and available data that can be used as potential

Table 1: Examples of potential ensemble components.

Team Name and Reference	Model Name	Model Type
Auquan Data Science [34]	MLOptimized ModifiedSEIR	Modified SEIR model with compartments for reported and unreported infections. Non-linear mixed effects curve-fitting
Carnegie Mellon Delphi Group [9]	TimeSeries	A basic AR-type time series model fit using case counts and deaths as features
Columbia University [27]	Select	County-level SEIR model
CovidAnalytics at MIT [20]	DELPHI	SEIR model
Discrete Dynamical Systems [15]	Negative Binomial Dynamical System	Jointly modeling daily deaths and cases using a negative binomial distribution
GT [29]	DeepCOVID	Deep learning
Institute for Health Metrics and Evaluation [25]	CurveFit	Non-linear mixed effects curve-fitting
Los Alamos National Labs [26]	GrowthRate	Statistical dynamical growth model accounting for population susceptibility
MOBS Lab at Northeastern [35]	GLEAM COVID-19	Metapopulation, age structured SLIR model
NotreDame-FRED [6]	NotreDame-FRED	Agent-based model developed for influenza with parameters modified to represent the natural history of COVID-19
Youyang Gu (YYG) [11]	ParamSearch	SEIR model with machine learning layer

ensemble components. The goal is to select models that employ a range of modeling approaches and use a variety of assumptions. This is an important feature of the ensemble modeling approach, which relies on the diversity and independence between the ensemble components. Based on our selection, we will obtain each model's prediction data, made open and available by the COVID-19 Forecast Hub [30] as well as each model's respective repository or web pages. Although it varies from model to model, most model prediction data includes a start date, a prediction date, the predicted number of cumulative cases, the predicted number of incident and cumulative deaths, the predicted number of incident hospitalizations, the corresponding location for the prediction, and the confidence interval. We consider each prediction to be a *"Possible (future) World"*, analogous to work in uncertain database management [42, 44, 45]. The difference in uncertain database management is that current and past data is uncertain, whereas for disease prediction, it is data from the future that is uncertain. But in both cases, the challenge is to find a consensus among possible worlds (different database instances or different predictions) and enrich this consensus with reliability information.

Imputation of Missing Predictions. Due to the nature of independence of each model's development, the temporal resolution of the prediction data that is available for each model may be inconsistent and asynchronous. Imagine that two different models that make predictions starting from May 1st and onward. *Model X* might estimate the number of deaths and cases each day for the next four weeks. *Model Y* might estimate the number of deaths and cases each day for the next twenty weeks. In another scenario, imagine that another model, *Model Z* begins making predictions that start on May 5 and onward. Thus, there are no predictions available from *Model Z* from May 1st to May 4th. With the assumption that all of the models are calibrated to the most recent ground truth, the more recent the date of the forecast is, the more accurate the model is.

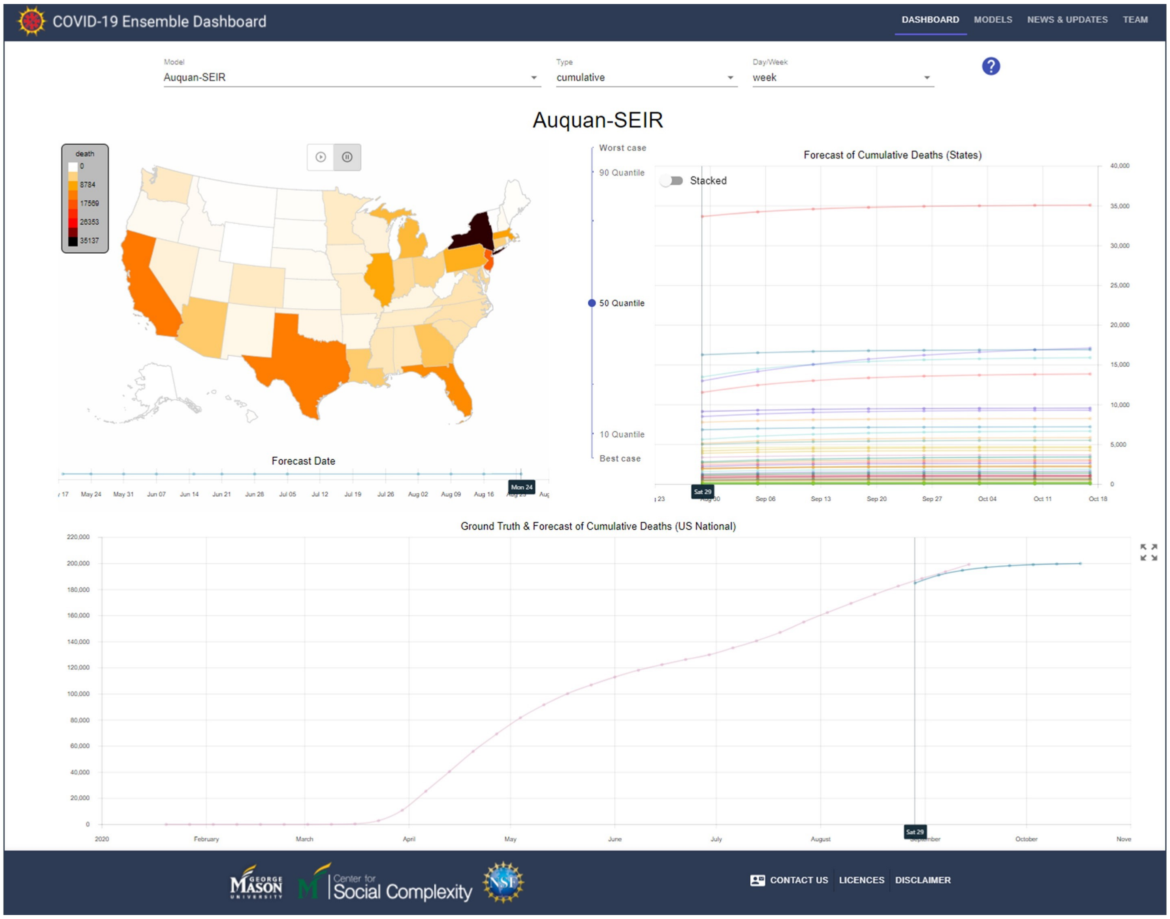
The inconsistent temporal resolution of the prediction data presents a challenge for the inclusion of important components into the ensemble. As a result, we propose the use of imputation algorithms to fill the prediction gaps. In a sense, we aim here to predict the missing predictions of the models. We represent predictions in a three-mode tensor $\mathcal{P}_{i,j,m}$ such that a cell $p_{i,j,m}$ corresponds to a prediction made on Day i, made for Day j, by model m. For example, if one mode m predicts on Day i = 10/02/2020 that there will be 5000 deaths on Day j = 10/09/2020, then we will have $p_{i,j,m} = 5000$. This tensor is sparse, as existing models publish their predictions sparsely (often once per week), and predictions are made only for a short time window (often 14 or 28 days). As part of the model integration stage, we will test and evaluate various imputation algorithms and determine which are most accurate in predicting the missing model predictions. Some of the imputation algorithms we aim to test include linear interpolation, linear regression, non-negative matrix factorization [19] for individual prediction models, and tensor factorization [16]. We hope that more complex imputation algorithms are able to leverage collaborative filtering to fill missing model prediction by assessing that "other models had relatively high predictions for this day" and "this model had relatively low predictions for this day made on earlier days."

2.2 Representative Clustering

Once we have imputed the data, we will have obtained a broad set of predictions, each corresponding to different "possible worlds" generated from different models, different parameters, and under different assumptions. Each possible world consists of time series data, corresponding to the predicted number of incident and cumulative cases, the predicted number of incident and cumulative deaths, and the predicted number of incident hospitalizations. We want to use an approach that has been published for clustering of uncertain data [43] by mapping possible worlds into a reduced feature space then clustering possible worlds (in our case predictions) as depicted on the bottom half of Figure 1. Each cluster then corresponds to a set of mutually similar predictions which may stem from different models. We will then select the median among these predictions (defined as the prediction that minimizes the pairwise distance to other predictions in the same cluster) as a cluster representative. Assuming that each model has the same likelihood to correctly capture the unknown future, we can apply inductive statistics to estimate the probability that a cluster represents the unknown true future and provide an error bound using the radius of the cluster (the maximum distance between the cluster representative and other predictions in the same cluster). As the project continues, we will use supervised learning to reinforce the weights of those models and parameters, yielding the most accurate predictions. Each cluster representative can be considered an ensemble of models with a high degree of agreement between predictions. These representative will then be visualized on a dashboard which, instead of exploring the plethora of existing predictions, allows us to visually analyze a small number of representative predictions together with their confidence values. For example, the user may be presented with *Model X* and given the information that 40% of all predictions agree with this prediction up to an error which will be visualized using error bounds. This condensed representation takes the burden from users to interpret an overwhelming number of predictions and allows them to focus on only a small number of representative predictions.

3 An Online Medium to Disseminate Our Results

It is a likely scenario that the COVID-19 health emergency will continue into the coming several months and perhaps years. In a time of such uncertainty, policy-makers are right now relying on existing COVID-19 models to anticipate future conditions. Leveraging simulation models' predictive capabilities is critical to rapidly inform the public about what's likely to come and help policy-makers plan for those conditions. Therefore, there is an urgent need to compare and synthesize the wide-range of existing COVID-19 models and their resulting simulations, disseminate this information clearly, and increase transparency about model assumptions and uncertainty. To address this need, we have designed a COVID-19 Ensemble Dashboard as a medium for which the broader public, decision-makers, the modeling community, and key organizations can explore and compare between existing models. The prototype for our proposed dashboard is presented in Figure 2.



Model: Auquan-SEIR

Type: cumulative

Day/Week: week

Figure 2: COVID-19 Ensemble Dashboard

Our dashboard aims at providing two main functionalities regarding the status of COVID-19 at the US country-level and state-levels: (1) Giving users the option to examine dozens of existing COVID-19 models' predictions and over time, including both weekly and daily predictions as well as incident and cumulative metrics. Besides this, we provide a "stacked" time-series visualization to see all US states in the same picture. Users can also display individual predictions plotted against the ground-truth numbers, which facilitates examining model performance against real-world results; (2) Allowing users to examine the results from our representative clustering approach, as outlined in section 2. With this functionality, the users will be able to examine the agreement and disagreement between various models periodically. This functionality is currently being implemented and will be integrated as periodic reports into our dashboard. The dashboard will be updated as needed to incorporate new data and models as they become available, facilitating the opportunity to rapidly cross-compare new predictions and disseminate this information to the public.

4 Conclusion and Future Work

In this newsletter, we propose a novel ensemble modeling approach that leverages representative clustering to both examine the degree of agreement between models of COVID-19 spread and their predictions as well unify predictions into a smaller subset. The novel ensemble clustering approach begins with the process of data integration and thus the selection of ensemble components and the imputation of each component's missing predictions. Next, clustering is used to find a set of ensembles that are representative of groups of models with predictions that have a high degree of agreement for the same forecasting horizon. This research is still in early stages. Future steps include implementation and testing of the proposed approach. The proposed approach has the advantage of not being limited to the generation of one ensemble and thus acknowledges the unique assumptions of the components while removing the burden from policy makers, the general public, as well as other researchers to interpret an overwhelming number of COVID-19 model predictions.

Acknowledgements

This research is supported by National Science Foundation under Grant No. DEB-203068, "RAPID: An Ensemble Approach to Combine Predictions from COVID-19 Simulations."

References

[1] S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, et al. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112, 2020.

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

[3] Centers for Disease Control and Prevention. Covid-19 Forecasts: Deaths (`https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html`).

[4] M. Chinazzi et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 2020.

[5] S. Del Valle. Los Alamos COVID-19 Confirmed and Forecasted Case Data. `https://covid-19.bsvgateway.org/`. Accessed: 2020-04-11.

[6] G. Espana, R. Oidtman, S. Cavany, A. Costello, A. Wieler, A. Lerch, C. Barbera, M. Poterek, Q. Tran, S. Moore, and A. Perkins. NotreDame-FRED (`https://github.com/confunguido/covid19_ND_forecasting`).

[7] N. Ferguson et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, 2020.

[8] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[9] A. Green, A. Hu, M. Jahja, V. Ventura, L. Wasserman, R. Tibshirani, V. Shankar, J. Bien, L. Brooks, B. Narasimhan, S. Rajanala, A. Rumack, N. Simon, J. Sharpnack, and R. Tibshirani. Carnegie Mellon Delphi Group (`https://delphi.cmu.edu`).

[10] G. Grenouillet, L. Buisson, N. Casajus, and S. Lek. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, 34(1):9–17, 2011.

[11] Y. Gu. Youyang Gu (YYG) (`https://covid19-projections.com`).

[12] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, D. A. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, et al. Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, 105(12):4639–4644, 2008.

[13] IHME. COVID-19 Projections. `https://covid19.healthdata.org/united-states-of-america`. Accessed: 2020-04-18.

[14] M. A. Johansson, K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, E. Guven, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, 2019.

[15] R. Kalantari and M. Zhou. Discrete Dynamical Systems (`https://dds-covid19.github.io/`).

[16] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86, 2010.

[17] V. Kotu and B. Deshpande. *Predictive analytics and data mining: concepts and practice with rapidminer.* Morgan Kaufmann, 2014.

[18] T. Krishnamurti, C. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433):1548–1550, 1999.

[19] M. Kurucz, A. A. Benczúr, and K. Csalogány. Methods for large scale svd with missing values. In *Proceedings of KDD cup and workshop*, volume 12, pages 31–38. Citeseer, 2007.

[20] M. L. Li, H. T. Bouardi, O. S. Lami, T. A. Trikalinos, N. K. Trichakis, and D. Bertsimas. CovidAnalytics at MIT (`https://www.covidanalytics.io/`).

[21] D. Mackenzie. *Mathematics of climate change: a new discipline for an uncertain century.* Mathematical Sciences Research Institute, 2007.

[22] T. McAndrew and N. G. Reich. Adaptively stacking ensembles for influenza forecasting with incomplete data. *arXiv preprint arXiv:1908.01675*, 2019.

[23] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo. Multiple ensemble neural network models with fuzzy response aggregation for predicting covid-19 time series: the case of mexico. In *Healthcare*, volume 8, page 181. Multidisciplinary Digital Publishing Institute, 2020.

[24] C. J. Murray et al. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020.

[25] C. Murry. IHME (`https://covid19.healthdata.org/united-states-of-america`).

[26] D. Osthus, S. D. Valle, C. Manore, B. Weaver, L. Castro, C. Shelley, M. M. Smith, J. Spencer, G. Fairchild, T. Pitts, D. Gerts, L. Dauelsberg, A. Daughton, M. Gorris, B. Hornbein, D. Israel, N. Parikh, D. Shutt, and A. Ziemann. Los Alamos National Labs (`https://covid-19.bsvgateway.org/`).

[27] S. Pei, T. Yamana, S. Kandula, W. Yang, M. Galanti, and J. Shaman. Columbia University (`https://blogs.cuit.columbia.edu/jls106/publications/covid-19-findings-simulations/`).

[28] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[29] A. Prakash, A. Rodriguez, J. Cui, A. Tabassum, and B. Adhikari. GT (`https://deepcovid.github.io/`).

[30] E. L. Ray, N. Wattanachit, J. Niemi, A. H. Kanji, K. House, E. Y. Cramer, J. Bracher, A. Zheng, T. K. Yamana, X. Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *medRxiv*, 2020.

[31] N. G. Reich, C. J. McGowan, T. K. Yamana, A. Tushar, E. L. Ray, D. Osthus, S. Kandula, L. C. Brooks, W. Crawford-Crudell, G. C. Gibson, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the us. *PLoS computational biology*, 15(11):e1007486, 2019.

[32] K. A. Schmid and A. Züfle. Representative query answers on uncertain data. In *SSTD'19*, pages 140–149, 2019.

[33] C. Tebaldi and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences*, 365(1857):2053–2075, 2007.

[34] V. Tomar and C. Jain. Auquan Data Science (`https://covid19-infection-model.auquan.com/`).

[35] A. Vespignani, M. Chinazzi, J. T. Davis, K. Mu, A. P. y Piontti, N. Samay, X. Xiong, M. E. Halloran, I. M. Longini, N. E. Dean, K. Sun, M. Litvinova, C. Gioannini, L. Rossi, and M. Ajelli. MOBS Lab at Northeastern (`https://covid19.gleamproject.org/`).

[36] A. Vespignani et al. COVID-19 MODELING IN THE UNITED STATES. `https://covid19.gleamproject.org/`. Accessed: 2020-04-11.

[37] C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, A. Vespignani, et al. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, 2018.

[38] N. R. Viney, H. Bormann, L. Breuer, A. Bronstert, B. F. Croke, H. Frede, T. Gräff, L. Hubrechts, J. A. Huisman, A. J. Jakeman, et al. Assessing the impact of land use change on hydrology by ensemble modelling (luchem) ii: Ensemble combinations and predictions. *Advances in water resources*, 32(2):147–158, 2009.

[39] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[40] T. K. Yamana, S. Kandula, and J. Shaman. Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*, 13(123):20160410, 2016.

[41] T. K. Yamana, S. Kandula, and J. Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the united states. *PLoS computational biology*, 13(11):e1005801, 2017.

[42] A. Zuefle. Uncertain spatial data management: An overview. *arXiv preprint arXiv:2009.01121*, 2020.

[43] A. Züfle, T. Emrich, K. A. Schmid, N. Mamoulis, A. Zimek, and M. Renz. Representative clustering of uncertain data. In *ACM KDD'14*, pages 243–252, 2014.

[44] A. Züfle, G. Trajcevski, D. Pfoser, M. Renz, M. T. Rice, T. Leslie, P. Delamater, and T. Emrich. Handling uncertainty in geo-spatial data. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1467–1470. Ieee, 2017.

[45] A. Züfle, G. Trajcevski, D. Pfoser, and J. S. Kim. Managing uncertainty in evolving geo-spatial data. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 5–8, June 2020.

The SIGSPATIAL Special

Section 2:
Spatial Data Systems for the Internet of Things

ACM SIGSPATIAL
http://www.sigspatial.org

Spatial Data Systems Support for the Internet of Things - Challenges and Opportunities

Mohamed Sarwat
Arizona State University
Tempe, USA
msarwat@asu.edu

Abstract

The Internet of Things (IoT) has recently received significant attention. An IoT device may possess an array of sensors that for example monitors the air temperature, carbon monoxide level, wifi signals, and sound intensity. IoT data is initially created on the device, then sent over to a central database system (e.g., the cloud) that organizes and prepares such data for the ongoing use by myriad applications, which include but are not limited to smart home, smart city, the industrial internet, connected cars, and connected health. Data generated by IoT devices is inherently spatial and temporal. For instance, an audio signal represents the variation of the sound intensity (retrieved by a sound sensor) over the time dimension. Furthermore, IoT devices are either installed in a static location (e.g., a building, a traffic intersection) or can be attached to moving objects such as a connected vehicle or a wearable device. In this article, we argue that existing IoT data systems do not properly consider the SpatioTemporal aspect of such data. Hence, the article represents a call for action to the SIGSPATIAL community in order to conduct research on building systems and applications that treat both the spatial and temporal dimensions of IoT data as first class citizens.

1 Motivation

The Internet of Things (IoT) has recently received significant attention from both industry and academia. The Boston Consulting Group predicts that by 2020, $267 Billion will be spent on IoT technologies, products, and services [1]. IoT represents a network of devices, each equipped with a variety of sensors (and actuators) that sense and collect data about the local environment. For example, an IoT device may possess an array of sensors that monitors the air temperature, carbon monoxide level, wifi signals, and sound intensity. Data created by an IoT device goes through three main phases: the first phase is the initial creation, which takes place on the device, and then sent over the Internet to a central database system (e.g., the cloud). The second phase is how the central database system collects and organizes this data. The third phase is the ongoing use of such data by myriad applications, which include but are not limited to smart home, smart city, the industrial internet, connected cars, and connected health.

Data generated by IoT devices has the following key characteristics: *(i) Spatial and Temporal:* Data collected by a device represents a physical quantity, i.e., a signal, that continuously varies over time. For instance, an audio (sound) signal represents the variation of the sound intensity over the time dimension. Also, IoT devices are usually installed in a static location such as a building or a traffic intersection. However, some IoT devices are attached to objects that move in space like a connected vehicle or a wearable device. In both cases, the data generated by an IoT device possess a spatial location attribute that represents the longitude and latitude coordinates of the sensed observations. *(ii) Heterogeneous and Interconnected:* Not all IoT sensors are the same; a spectrum of various sensors exists and each sensor measures a different thing. For instance, a camera generates image signals while a microphone generates an audio signal; each signal type has different physical and mathematical characteristics [2]. For example, an ultrasonic motion sensor sends a signal when it detects motion whereas a microphone sensor only sends a signal when it detects a sound. Also, IoT data is inherently interconnected and can be linked to other data sources such as the city data, the web and social media.

Given such characteristics, a central data system must provide a spatial / spatio-temporal data management query Application Programming Interface (API) side-by-side with a digital signal processing API for programmers to develop IoT applications. Such requirement makes it really difficult for off-the-shelf systems to digest and process IoT data. The problem becomes even more challenging as the volume of data collected from IoT devices increases at a staggering rate. Today, RADAR, LIDAR, and CAMERA sensors have a bandwidth of up to 15 Mbit/sec, 100 Mbit/sec, and 3500 Mbit/sec, respectively. A connected vehicle, equipped with several of such sensors, generates data at a rate that exceeds 25 GB per hour [3]. The volume of such data will increase even more with already 2.8 trillion IoT devices deployed globally in homes and road intersections by the year 2019 [4], the 10 million self-driving cars equipped with dozens of sensors to be on the road and the 7 million drones to be flying in U.S. skies by 2020 [5, 6]. That makes it almost impossible for state-of-the-art big spatial data systems to handle real-time or near-real time IoT analytics applications.

2 A Spatio-Temporal Data Model for IoT Data

Recently proposed IoT data models [7, 8] over the semantics of things, sensor observations, and applications. Several works model IoT data for health-specific application such as real-time prediction of blood alcohol content using smartwatch sensor data [9]. Many papers model IoT data for smart city applications [10, 11]. One potential research direction will build upon these efforts. We can model IoT data using the `SenosorThings` standard published by the Open Geospatial Consortium (OGC) in 2016. The standard provides a formal / generic API to model and query IoT things, which can represent a smart car, smart watch, traffic camera, smart oven, etc... The API also models the collected observations, connections among IoT devices as well as their spatial and temporal attributes [12].

Recent works model IoT data, given its interconnected and linked nature, as a graph [13, 14], which integrates data generated by various IoT devices as well as integrates IoT data with other relevant data sources, e.g., semantic web and infrastructure networks. Thus, it makes sense to model linked IoT data as a graph and store it in a graph database system such as Neo4j [15] and Titan [16], which also support spatiotemporal attributes that can capture the location and time aspects of sensed observations [17, 18, 19, 18, 20, 21]. As depicted in Figure 1, the IoT graph is modeled as a property / labeled graph $G = (V, E, \varphi, \psi)$ such that (1) V is a set of vertexes that represent things (IoT devices and sensors) in an IoT network. Vertexes can also represent real world entities generated by other data sources, e.g., Knowledge Graphs, Social Graphs, and the Web. (2) E is a set of edges that connect IoT things and also connect IoT things to other entities collected from other data sources. (3) φ is a mapping function $\varphi : V \to \mathcal{L}$, where $\mathcal{L}$ is a set of labels or types, (4) ψ is a mapping function $\psi : E \to \mathcal{L}$ [22]. The IoT graph is a property graph that possesses at least one spatial label/type in $\mathcal{L}$. In such a graph, some vertexes semantically represent spatial objects, which possess spatial location attributes, e.g., point, polygon. The spatial attribute of a vertex is denoted as $v.loc$. Given such graph, a user may ask a query like “Find patterns in the IoT graph of audio signals observed by an IoT device located in Downtown Tempe such that the observation value is larger than 140 dB”. A main challenge though is how a graph data system can evaluate the combination of spatial (e.g., range, K-Nearest Neighbors, spatial join), temporal, and graph predicates on linked IoT data [23, 24, 25].

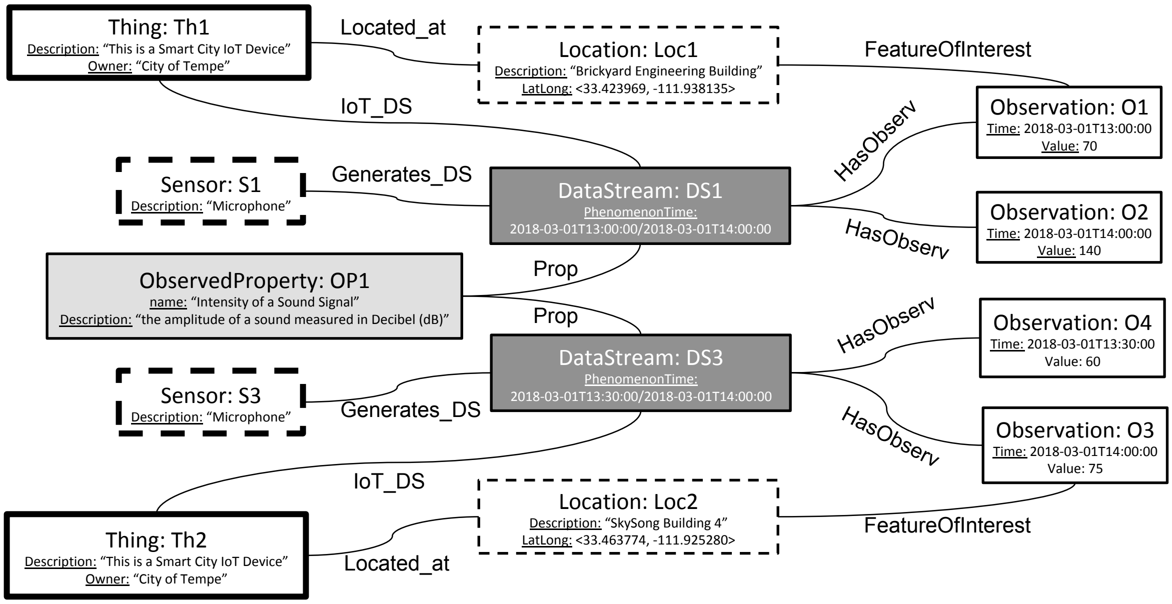

Thing: Th1
Description: "This is a Smart City IoT Device"
Owner: "City of Tempe"
Located_at
Location: Loc1
Description: "Brickyard Engineering Building"
LatLong: <33.423969, -111.938135>
FeatureOfInterest
IoT_DS
Observation: O1
Time: 2018-03-01T13:00:00
Value: 70
HasObserv
Sensor: S1
Description: "Microphone"
Generates_DS
DataStream: DS1
PhenomenonTime:
2018-03-01T13:00:00/2018-03-01T14:00:00
Observation: O2
Time: 2018-03-01T14:00:00
Value: 140
HasObserv
ObservedProperty: OP1
name: "Intensity of a Sound Signal"
Description: "the amplitude of a sound measured in Decibel (dB)"
Prop
Prop
HasObserv
Observation: O4
Time: 2018-03-01T13:30:00
Value: 60
DataStream: DS3
PhenomenonTime:
2018-03-01T13:30:00/2018-03-01T14:00:00
Sensor: S3
Description: "Microphone"
Generates_DS
HasObserv
Observation: O3
Time: 2018-03-01T14:00:00
Value: 75
IoT_DS
Location: Loc2
Description: "SkySong Building 4"
LatLong: <33.463774, -111.925280>
FeatureOfInterest
Thing: Th2
Description: "This is a Smart City IoT Device"
Owner: "City of Tempe"
Located_at

Figure 1: IoT modeled using the SensorThings OGC standard

3 Scalable Geospatial Processing of IoT data

There exist a few research efforts, which focus on developing scalable data infrastructure for IoT data [26, 9]. For instance, Dey et. al proposed a scheme where edge devices offer free computational slots to servers in a cloud based data analytics system [27]. The Aura system prototype [28] enables building ad-hoc clouds using IoT devices in the nearby physical environment. Jiang et. al evaluated cloud-based data storage options for IoT data [29] whereas Li et. al experimented the use of NoSQL database technology to store massive-scale IoT data [30]. However, none of these efforts takes into account spatio-temporal data processing operations, which are vital for processing IoT data [9].

State-of-the-art spatial and spatio-temporal data systems [31, 21, 32, 33, 34] do not provide native support for digital signal processing and machine learning operations whereas numerical frameworks such as MatLab do not provide in-house support for spatial data management. Futhermore, Such systems do not provide an out-of-the-box API to handle IoT sensor data. For instance, the Apache Spark `Dstream` API that can chop up live IoT data stream, represented as observations, into batches of X seconds. Spark can then treat each batch of data as an RDD and processes them using RDD operations. Finally, the processed results of the RDD operations are returned in batches. However, such APIs are not easy for programmers to develop IoT applications and are not natively optimized to run the combination of spatial, spatio-temporal, digital signal processing, and machine learning operations on IoT data. Examples of queries that require such combination include: "Q1: Report gunshots heard in Downtown Tempe area between 13:45 and 14:00 pm on 2018/03/01", "Q2: Report gunshots heard nearby (e.g., within 0.1 mi) a School in the City of Tempe". Q1 needs to filter out IoT data records that neither lie within Downtown Tempe nor observed between 13:45 and 14:00 pm on 2018/03/01. Q1 also needs to convert the audio signal into a time-frequency representation, e.g.., mel-spectrogram, by applying Fast Fourier Transform (FFT), then discrete cosine transform, and finally applying an urban sound classifier to detect gunshots. Q2 does the same with the exception that it applies a spatial join operation to retrieve observations only nearby schools.

Having said that, a promising research direction will incorporate IoT data awareness in state-of-the-art big spatial data systems such as Apached Sedona (GeoSpark) [35]. Furthermore, that also requires that system developers design a middleware framework, which understands the IoT devices streaming data to the central data system on one side and the requirements of applications accessing such IoT data on the other side. The proposed middleware system will tune the central data system to adaptively decide whether or not to eagerly propagate data from the device to the central system. I plan to modify existing relational and spatial query processing algorithms to leverage the IoT device capabilities and handle the different rate and types of data generated by various IoT devices. To capture the interconnected nature of IoT data, a spatial data system must also provide a graph processing API in addition to the spatial / spatio-temporal API. Such a combination is already supported by existing graph data systems [15, 16, 36, 37], however such systems treat the spatial attribute as a second class citizen, and hence cannot achieve real time or near real time performance. The community needs to craft efficient query operators that accelerate location-aware graph queries and also investigate new index structures that take into account network aspect of linked IoT data as well as the spatial and Spatio-Temporal aspects.

4 Conclusion

The Internet of Things (IoT) is getting more popular every day. The spatial, temporal, big, fast, heterogeneous, and interconnected nature of collected IoT data makes it difficult for off-the-shelf spatial data systems to digest and process such data, especially for real-time or near-real time analytics applications. The goal of this article is to encourage the community to design and develop spatial and spatiotemporal data infrastructure that can capture, store, query, analyze data from connected IoT devices at scale. The outcome of that research can provide a tool for data scientists, policy makers, and businesses to better utilize and extract value from IoT data growing at a staggering rate.

5 Acknowledgements

This work is supported by the National Science Foundation under award #1845789.

References

[1] $267 Billion will be spent on IoT technologies by 2020. https://www.forbes.com/sites/louiscolumbus/2017/01/29/internet-of-things-market-to-reach-267b-by-2020/#7ba38b21609b.

[2] A.V. Oppenheim, A.S. Willsky, and S.H. Nawab. *Signals and Systems*. Prentice Hall, 1997.

[3] Connected Vehicle Data. https://www.hitachivantara.com/en-us/pdf/white-paper/hitachi-white-paper-internet-on-wheels.pdf.

[4] 2.8 trillion sensor devices by 2019. https://www.bloomberg.com/news/articles/2013-08-05/trillions-of-smart-sensors-will-change-life-as-apps-have.

[5] 10 million self-driving cars will be on the road by 2020. http://www.businessinsider.com/report-10-million-self-driving-cars-will-be-on-the-road-by-2020-2015-5-6.

[6] 7 million drones flying in U.S skies by 2020. https://www.faa.gov/news/updates/?newsId=85227.

[7] Pratikkumar Desai, Amit Sheth, and Pramod Anantharam. Semantic gateway as a service architecture for iot interoperability. In *IEEE International Conference on Mobile Services (MS)*, pages 313–319. IEEE, 2015.

[8] Yong-Shin Kang, Il-Ha Park, Jongtae Rhee, and Yong-Han Lee. Mongodb-based repository design for iot-generated rfid/sensor big data. *IEEE Sensors Journal*, 16(2):485–497, 2016.

[9] Anne H Ngu, Mario Gutierrez, Vangelis Metsis, Surya Nepal, and Quan Z Sheng. Iot middleware: A survey on issues and enabling technologies. *IEEE Internet of Things Journal*, 4(1):1–20, 2017.

[10] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. Sensing as a service model for smart cities supported by internet of things. *Transactions on Emerging Telecommunications Technologies*, 25(1):81–93, 2014.

[11] M Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho. Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks*, 101:63–80, 2016.

[12] OGC SensorThings API. `http://docs.opengeospatial.org/is/15-078r6/15-078r6.html#17`.

[13] Payam M. Barnaghi and Mirko Presser. Publishing linked sensor data. In *Proceedings of the 3rd International Workshop on Semantic Sensor Networks, SSN 2010, Shanghai, China, November 7, 2010*, 2010.

[14] Laurent Lefort, Josh Bobruk, Armin Haller, Kerry Taylor, and Andrew Woolf. A linked sensor data cube for a 100 year homogenised daily temperature dataset. In *Proceedings of the 5th International Workshop on Semantic Sensor Networks, SSN12, Boston, Massachusetts, USA, November 12, 2012*, pages 1–16, 2012.

[15] Neo4j graph database. `https://neo4j.com/`.

[16] Titan distributed graph database. `http://titan.thinkaurelius.com/`.

[17] Petko Bakalov, Erik G. Hoel, and Sangho Kim. A network model for the utility domain. In *SIGSPATIAL/GIS*, pages 32:1–32:10. ACM, 2017.

[18] Mohamed Sarwat, Jie Bao, Chi-Yin Chow, Justin J. Levandoski, Amr Magdy, and Mohamed F. Mokbel. Context awareness in mobile systems. In *Data Management in Pervasive Systems*, pages 257–287. 2015.

[19] Kyriakos Mouratidis, Jing Li, Yu Tang, and Nikos Mamoulis. Joint search by social and spatial proximity. *TKDE*, 27(3):781–793, 2015.

[20] Yuhan Sun, Nitin Pasumarthy, and Mohamed Sarwat. On Evaluating Social Proximity-Aware Spatial Range Queries. In *Proceedings of the International Conference on Mobile Data Management, MDM*, 2017.

[21] Mohamed Sarwat. Interactive and Scalable Exploration of Big Spatial Data–A Data Management Perspective. In *Proceedings of the International Conference on Mobile Data Management, MDM*, 2015.

[22] Wen Sun, Achille Fokoue, Kavitha Srinivas, Anastasios Kementsietsidis, Gang Hu, and Guo Tong Xie. Sqlgraph: An efficient relational-based property graph store. In *SIGMOD*, pages 1887–1901, 2015.

[23] Mohamed Sarwat and Yuhan Sun. Answering Location-Aware Graph Reachability Queries on GeoSocial Data. In *Proceedings of the International Conference on Data Engineering, ICDE*, 2017.

[24] Yuhan Sun and Mohamed Sarwat. A Generic Database Indexing Framework for Large-Scale Geographic Knowledge Graphs. In *Proceedings of the ACM Symposium on Advances in Geographic Information Systems, ACM GIS*, 2018.

[25] Yuhan Sun and Mohamed Sarwat. A spatially-pruned vertex expansion operator in the neo4j graph database system. *GeoInformatica*, 23(3):397–423, 2019.

[26] Joshua Cooper and Anne James. Challenges for database management in the internet of things. *IETE Technical Review*, 26(5):320–329, 2009.

[27] Swarnava Dey, Arijit Mukherjee, Himadri Sekhar Paul, and Arpan Pal. Challenges of using edge devices in iot computation grids. In *International Conference on Parallel and Distributed Systems (ICPADS)*, pages 564–569. IEEE, 2013.

[28] Ragib Hasan, Md Mahmud Hossain, and Rasib Khan. Aura: an iot based cloud infrastructure for localized mobile computation outsourcing. In *IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, pages 183–188. IEEE, 2015.

[29] Lihong Jiang, Li Da Xu, Hongming Cai, Zuhai Jiang, Fenglin Bu, and Boyi Xu. An iot-oriented data storage framework in cloud computing platform. *IEEE Transactions on Industrial Informatics*, 10(2):1443–1451, 2014.

[30] Tingli Li, Yang Liu, Ye Tian, Shuo Shen, and Wei Mao. A storage solution for massive iot data based on nosql. In *IEEE International Conference on Green Computing and Communications (GreenCom)*, pages 50–57. IEEE, 2012.

[31] Ram Sriharsha. Geospatial analytics using spark. `https://github.com/harsha2010/magellan`.

[32] Dong Xie, Feifei Li, Bin Yao, Gefei Li, Liang Zhou, and Minyi Guo. Simba: Efficient In-Memory Spatial Analytics. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, 2016.

[33] Jia Yu, Jinxuan Wu, and Mohamed Sarwat. A demonstration of GeoSpark: A cluster computing framework for processing big spatial data. In *Proceedings of the International Conference on Data Engineering, ICDE*, 2016.

[34] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Spatial data management in apache spark: the geospark perspective and beyond. *GeoInformatica*, 23(1):37–78, 2019.

[35] Apached Sedona. `http://sedona.apache.org`.

[36] James Cheng, Silu Huang, Huanhuan Wu, and Ada Wai-Chee Fu. Tf-label: a topological-folding labeling scheme for reachability querying in a large graph. In *SIGMOD*, pages 193–204. ACM, 2013.

[37] Andy Diwen Zhu, Wenqing Lin, Sibo Wang, and Xiaokui Xiao. Reachability queries on large dynamic graphs: a total order approach. In *SIGMOD*, pages 1323–1334. ACM, 2014.

The SIGSPATIAL Special

ACM SIGSPATIAL
http://www.sigspatial.org