COVID-19 Vulnerability Map Construction via Location Privacy Preserving Mobile Crowdsourcing

Rui Chen*, Liang Li†, Jeffrey Jiarui Chen¶ Ronghui Hou†, Yanmin Gong‡, Yuanxiong Guo§ and Miao Pan*
*Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204
†School of Cyber Engineering, Xidian University, Xi'an, China 710071
¶St. Mark's School Of Texas, 10600 Preston Rd, Dallas, TX 75230
‡Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249
§Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9348141

***Abstract*—The pandemic of the coronavirus (COVID-19) has caused an unprecedented global public health crisis, and most countries in the world are running out of the healthcare resources. A fine-grained COVID-19 vulnerability map will be essential to track the number of people with covid-like symptoms, so that the the potential outbreak communities can be identified and the valuable healthcare resources can proactively and dynamically be allocated. Mobile crowdsourcing based symptom reporting is a promising and convenient option to construct such a map, while it may compromise the location privacy of crowdsourcing participants. In this work, we propose a novel approach to establish the COVID-19 vulnerability map based on the crowdsourced reporting without disclosing the participants' location privacy to a semi-honest crowdsourcing aggregator. Briefly, based on the differentially private geo-indistinguishability, the mobile participants are able to locally perturb their geographic data. With the masked geographic information, we employ the best linear unbiased prediction estimator with spatial smoothing to obtain the reliable vulnerability estimates in the areas of interest and construct the map. Given the fast spreading nature of coronavirus, we integrate the vulnerability estimates with a susceptible-exposed-infected-removed (SEIR) model to build up a future trend map. Extensive simulations based on real-world data verify the effectiveness of the proposed method.**

I. Introduction

The recent pandemic of the coronavirus (COVID-19) has raised an unprecedented globally crisis on various aspects (e.g., public health and economy). The rapid growth of infected population overloads hospital capacities and causes the major healthcare resources and personnel shortages in many countries. The top priority is to contain the spread of the COVID-19 with effective infection control measures like early detection and sensible segregation [1]. To greatly reduce the spreading of the coronavirus, it is more instrumental to early identify the high-risk areas and forecast spatially the disease transmission dynamics. The policy-makers then can proactively and optimally allocate the constrained medical resources to the next potential "outburst" spots in advance. Early warning mechanism is often depicted as a heatmap with the locations of the infected population and vulnerability risk prediction. It's a quite straightforward methods to model COVID-19 pandemic, and helps estimate and predict latent patients' distribution.

The success of COVID-19 vulnerability map construction relies on comprehensive health information. The existing infection data confirmed by the local hospitals or testing stations and collected from local media reports or Centers for Disease Control and Prevention (CDC) can help to generate online COVID-19 maps. However, those maps may not be good enough. First, the traditional testing data collections are time-consuming and costly. It's infeasible to test the entire population due to the limited test kits. Besides, for the individuals with no health insurance, they cannot afford the treatment fee and are less likely to take the tests in certain areas. It could lead a vast of underestimation of the real number of patients. Second, the existing maps demonstrate the confirmed cases in each county, they are not fine-grained and lack of adequate coverage of patients who are asymptomatic or have mild symptoms, and cannot get the chance to be tested. These problems directly lead the vulnerability map to fail to function as expected.

With the aid of mobile crowdsourcing, data acquisition of timely vulnerability map construction becomes promising. By distributing symptom survey to ubiquitous mobile users, it only takes a few seconds to obtain a current snapshot of the number of people in each area who have developed symptoms. The feasibility of this approach lies in the popularity of smart devices and the wide expectation that some people may be willing to share their symptoms with the general public to help combat the COVID-19. Prior works like "Flu near you" [2] have shown the success of utilizing crowdsourced data to obtain accurate tracking in influenza season. During the outbreak of COVID-19, as reported by the recently published work [3], after online individual survey being first distributed for 10 days, 74,256 responses had been received. Facebook has just released a interactive map that tracks coronavirus symptoms using crowdsourced data from an opt-in survey, where more than 1 million people responded to the survey within the first two weeks. The tremendous data size and diverse information tagged with the fine geographic information make it possible for fine-grained map construction.

However, the mobile crowdsourcing based COVID-19 map construction is not perfect. Many of the existing research [3], [4] and Facebook covid map require the participants to fulfill a survey and their precise location information together with their symptoms are submitted to a third-party server, which raises the concerns of privacy leakage. Actually, mobile users

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

may be reluctant to reveal such sensitive information due to security and privacy concerns [5] [6]. Once these individual detailed information (i.e., location information) is revealed by a malicious party or dishonest server, those who are experiencing symptoms are likely to be discriminated against and suffer from economic loss. To address these issues, in this work, we target at developing a find-grained COVID-19 map via mobile corwdsourcing, while preserving participants' location privacy. Specifically, we devise a privacy-preserving mobile crowdsourcing framework for COVID-19 information collection, where people are encouraged to report their obfuscated locations and covid-like symptoms. Then, by aggregating their symdromic information based on the corresponding locations, we estimate the vulnerability risk in a small area-level using a spatial best linear unbiased prediction (SBLUP) estimator. Accounting for spatial correlation, a spatial smoothing based SBLUP greatly reduces the introduced spatial error and obtain a reliable estimation. Our salient contributions are summarized as follows.

- We employ the geo-indistinguishability scheme to protect the participants' location privacy. More specifically, the location data is perturbed by participants themselves before uploading, which gives differential privacy (DP) guarantee.
- To predict the vulnerability level in a fine-grained map, we first determine the individual vulnerability level based on the symptom self-reporting. Then, given the obfuscated locations, we utilize SBLUP with spatial smoothing technique to reduce the estimation bias induced by the privacy protection scheme and then integrate the vulnerability estimates with susceptible-exposed-infected-removed (SEIR) model to generate the future trend map.
- Extensive simulations are conducted based on real-world datasets to evaluate the performance of our scheme. The results also demonstrate the tradeoff between location differential privacy and risk estimation reliability.

The rest of this paper is organized as follows: In Section II, the preliminary of location differential privacy and overall system model are presented. In Section III, location preservation scheme is described. In Section IV, the SBLUP model for grid-level vulnerability estimation and future prediction on SEIR model are discussed. In Section V, the experiment based on the true database are analyzed and the paper is concluded in Section VI.

II. Preliminaries & System Model

A. Location Differential Privacy Preliminaries

In this section, we present preliminaries on DP related in our paper. DP [7] provides rigorous guarantees against what an attacker, with other background information, can infer individual information from the published statistics of a data set. Standard centralized setting requires a trustworthy curator to apply DP to the raw data. The definition of DP is shown as follows. If two databases X, Y differ in at most one element, we assume that X and Y are *neighboring* databases.

***Definition** 1:* Suppose privacy parameter $\epsilon \geq 0$, a randomization algorithm $\mathcal{M}$ satisfies ϵ-differential privacy. For any neighboring database X, Y, for any subset of outputs $\mathcal{S} \subseteq range(\mathcal{M})$,

$$\frac{Pr[\mathcal{M}(X) \in \mathcal{S}]}{Pr[\mathcal{M}(Y) \in \mathcal{S}]} \leq e^{\epsilon}.$$

The different choices of privacy parameter ϵ represent the different privacy preservation levels. The smaller privacy parameter ϵ suggests the probability of the outputs of the randomized algorithm $\mathcal{M}$ with two different inputs is close to each other, which means a high privacy preservation level. On the other hand, its utility would be compromised under the high privacy preservation level. In other words, there always exists a trade-off between the privacy and utility.

With the principle of differential privacy, geo-indistinguishability scheme especially allows users to replace noisy points with actual locations via a randomized mechanism without the requirement of trustworthy third-party entity [8]. More specifically, geo-indistinguishability provide a privacy guarantee within a radius, that is, given a circle centered at the user's actual location with a radius r, any two points within the circle yield observations with "similar" distributions. According to *Definition 1*, geo-indistinguishability is formally defined as follows [9]:

Definition 2: With the privacy confidence parameter $\epsilon \geq 0$, a randomized algorithm $\mathcal{M}$ satisfies (ϵ, r)-geo-indistinguishability if for any two different points x_0 and x_0' such that $d(x_0, x_0') \leq r$, the following holds:

$$\frac{Pr[\mathcal{M}(x_0) \in \mathcal{S}]}{Pr[\mathcal{M}(x_0') \in \mathcal{S}]} \leq e^{\epsilon d(x_0, x_0')}, \tag{1}$$

where $\mathcal{S}$ is the set of output locations and $d(\cdot, \cdot)$ denotes the Euclidean distance. The definition indicates that, to achieve (ϵ, r)-privacy within the radius of r, given two different locations x_0 and x_0' between which the distance is smaller than r, the probability of the output points of randomized algorithm $\mathcal{M}$ should be bounded by a multiplicative factor $e^{\epsilon d(x_0, x_0')}$. In geo-indistinguishable scheme, the privacy level k should consider both the impact of the privacy parameter ϵ and indistinguishable radius r. Therefore, the privacy parameter becomes $k = \epsilon r$, where ϵ can be regarded as the privacy level at one unit of distance and r is the radius of concern within which privacy is guaranteed.

B. System Model

This work constructs a fine-grained and periodically-updated vulnerability prediction map (VPM) by collecting COVID-19 symptoms data through mobile crowdsourcing platform, as shown in Fig. 1. Specifically, the aggregator in the mobile crowdsourcing platform launches a covid-like symptom reporting task to a number of candidate participants who are distributed over a 2D spatial region $\mathcal{A}$. Every participant is willing to engage in crowdsourcing tasks and honestly report their true answers. We assume a semi-honest

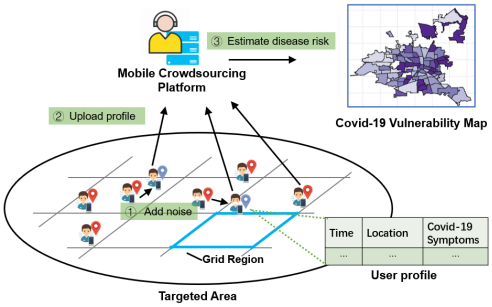


Fig. 1. Vulnerability map construction via mobile crowdsourcing.

crowdsourcing aggregator is curious but not malicious which implies he tries to learn from the exposed information but still complies with the protocol. The aggregator announces an online questionnaire regarding the covid-like symptoms. The participants upload their answers to the survey along with their location information to the aggregator.

The targeted area $\mathcal{A}$ is divided into G non-overlapping cells, denoted by the set $\mathcal{A} = \{a_1, \cdots, a_g, \cdots, a_G\}$. In a fine-grained VPM, the spatial unit is set to the street or township level. Each cell is tagged with a certain vulnerability prediction level l_g. The entire VPM is modeled as $\boldsymbol{l} \triangleq [l_1, \cdots, l_G]$, and estimated one as $\widetilde{\boldsymbol{l}}$. At each period, a total of N participants upload their records. We denote the report from the i-th participant by $\boldsymbol{r}_i = (t_i, \boldsymbol{x}_i)$, where t_i is the recording timestamp and $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ represents his syndromic information with p covid-related symptom attributes. Each report $\boldsymbol{r}_i$ is tagged with a location coordinate $\boldsymbol{s}_i = (s_i^x, s_i^y)$. We further denote $\mathcal{R}_g = \{\boldsymbol{r}_i, \forall \boldsymbol{s}_i \in a_g\}$ as the ground set of the crowdsourced data in cell a_g.

Since the location information of a single record is closely related to either home or work address in VPM, to avoid the disclosure risk of personal information, we use the geo-indistinguisihability method to protect mobile users' locations. Instead of updating a true location, a participant can share a fake one as $\widetilde{\boldsymbol{s}}_i = (\widetilde{s}_i^x, \widetilde{s}_i^y)$. In the following sections, we show that semi-honest aggregator can still predict the vulnerability level $\boldsymbol{l}_g$ in each cell based on the set of the crowdsourced records $\widetilde{\mathcal{R}}_g = \{\boldsymbol{r}_i, \forall \widetilde{\boldsymbol{s}}_i \in a_g\}$.

III. Location Privacy Protection

In our proposed scheme, a masked location is generated according to geo-indistinguishability. Assuming that each participant perturbs his location independently and homogeneously, we then simplify the denotations of the real location and the perturbed location as $\boldsymbol{s}$ and $\widetilde{\boldsymbol{s}}$ by omitting the subscripts in terms of participants. In order to meet the privacy requirement in *Definition 2*, the probability of generating an obfuscated location $\widetilde{\boldsymbol{s}}$ should decrease exponentially with the distance from the actual location $\boldsymbol{s}$. The two-dimensional Laplace distribution is usually applied to randomly produce the location noise, since it satisfies such a property, that is, given the parameter $\epsilon \in \mathbb{R}^+$ and the actual location $\boldsymbol{s} \in \mathbb{R}^2$, the probability density function (PDF) of the corresponding noise mechanism on any other point $\boldsymbol{s}' \in \mathbb{R}^2$ gives as follows:

$$p_\epsilon(\boldsymbol{s}') = \frac{\epsilon^2}{2\pi} e^{-\epsilon d(\boldsymbol{s}, \boldsymbol{s}')}, \tag{2}$$

where $\epsilon^2/2\pi$ is a normalization factor. The problem thereby turns into finding "the other point $\boldsymbol{s}' \in \mathbb{R}^2$" following the above distribution for each participant, namely obfuscated location. Further, we describe participants' locations by using Polar coordinates (q, θ) instead Cartesian coordinates where q is the distance between $\boldsymbol{s}$ and $\boldsymbol{s}'$, and θ is the angle that the line $(\boldsymbol{s}, \boldsymbol{s}')$ forms with respect to the horizontal axis. The PDF of the Polar Laplacian centered in the original coordinates $\boldsymbol{s}$ is:

$$p_\epsilon(q, \theta) = \frac{\epsilon^2}{2\pi} q e^{-\epsilon q}. \tag{3}$$

Let Q and Θ denote the two random variables of radius and angle, respectively. Since Q and Θ are independent, (3) can be expressed as the product of the two marginals, i.e. $p_\epsilon(q, \theta) = p_{\epsilon,Q}(q) p_{\epsilon,\Theta}(\theta)$, where

$$p_{\epsilon,Q}(q) = \int_0^{2\pi} p_\epsilon(q, \theta)\, d\theta = \epsilon^2 q e^{-\epsilon q}, \tag{4}$$

$$p_{\epsilon,\Theta}(\theta) = \int_0^{\infty} p_\epsilon(q, \theta)\, dq = \frac{1}{2\pi}. \tag{5}$$

Here, $p_{\epsilon,\Theta}(\theta)$ is a constant which is in accordance with uniform distribution, and $p_{\epsilon,Q}(q)$ coincides with the PDF of the gamma distribution with shape 2 and scale $1/\epsilon$ whose cumulative function is

$$C_\epsilon(q) = \int_0^q \epsilon^2 \rho e^{-\epsilon\rho}\, d\rho = 1 - (1 + \epsilon q) e^{-\epsilon q}. \tag{6}$$

Note that the above derivation is developed in the continuous plane. In practice, a location is usually described as discrete coordinates, with longitude and latitude. Then we obtain a obfuscated discrete location $\widetilde{\boldsymbol{s}}$ from any original location $\boldsymbol{s}$ (in Cartesian coordinates) according to the following procedures [9]: 1) Uniformly generate θ in $[0, 2\pi)$. 2) Uniformly generate z in $[0, 1)$ and set $q = C_\epsilon^{-1}(z)$. 3) Add noises to get an obfuscated location $\boldsymbol{s} + < q cos\theta, q sin\theta >$, and remap it to the closest point $\widetilde{\boldsymbol{s}}$ in discrete grid.

IV. Vulnerability Map Construction

Given the users' symptoms reports and their obfuscated locations, the crowdsourcing aggregator can predict the vulnerability level for each cell and generate the vulnerability map in the targeted area. We divide the VPM construction into two steps, including individual risk assessment and small area estimation adjustment. Because the two steps are used to fit the observed data from each single time point, for the convenience of describing the models, the subscript t is omitted in the next two subsections.

A. Individual Risk Assessment

The first phase is individual risk assessment, in which the collected syndromic information is mapped into a vulnerability degree. Specifically, given the crowdsourced syndromic reports from total N participants, $\mathcal{X} = \{\boldsymbol{x}_i, 1 \leq i \leq N\}$, the aggregator would determine the potential vulnerability of each mobile participant via a predetermined function $f : \mathcal{X} \rightarrow (0, 1)$. Crowdsourced symptom reporting task contain a set of questions about age, sex and existence of the symptoms commonly recorded in patients with the COVID-19 (cough, fever with body temperature, chest pain and shortness of breath, etc.). The vulnerability score of each participant is evaluated as the number of reported symptoms divided by the total number of symptoms in the predefined list.

B. Area Estimation Adjustment

By associating users with grid cells based on locations $\widetilde{\boldsymbol{s}}$, the aggregator estimates the grid-level vulnerability degree from the observed scores. The vulnerability level l_g refers to the average vulnerability score in grid a_g:

$$l_g = \frac{\sum_{i=1}^{N_g} f(\boldsymbol{x}_i)}{N_g}. \tag{7}$$

where N_g denotes the total population of grid cell a_g. Based the total n_g crowdsourced data samples in cell a_g, the naïve estimates of l_g can be easily calculated as $\widehat{l}_g^{DE} = \frac{\sum_{i=1}^{n_g} f(\boldsymbol{x}_i)}{n_g}$.

Since the true location of a user may be perturbed to a location that belongs to a different grid cell, the sample size of each grid cell may change. For the grids with smaller sample size, the direct estimation becomes biased and unreliable. To adjust the naïve estimates, we employ the SBLUP model [10] to borrow strength from other domains, as well as to consider the spatial correlation in the target area. Besides, as long as the overall area has large enough sample size, the SBLUP has distinct advantage of improving the reliability of sub-areal estimation. In the SBLUP model, the variable of interest is decomposed into a linear mixed model with spatially correlated random area effect:

$$l_g = \boldsymbol{h}_g^T \boldsymbol{\beta} + \boldsymbol{b}_g^T \boldsymbol{v}, \quad \forall a_g \in \mathcal{A}. \tag{8}$$

Here, $\boldsymbol{h}_g$ is area-specific auxiliary data, $\boldsymbol{\beta}_g$ is a vector of unknown regression parameters, $\boldsymbol{b}_g$ is a vector $(0, \cdots, 0, 1, 0, \cdots, 0)$ with 1 in the position g, and $\boldsymbol{v}$ is area-specific random effects. A direct way to characterize the spatial dependency $\boldsymbol{v}$ is to impose simultaneous auto-regressive [11] process with coefficient ρ and spatial weight matrix $\boldsymbol{W}$:

$$\boldsymbol{v} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{u}, \tag{9}$$

where $\boldsymbol{u}$ is a vector of independent error terms with zero mean and unknown variance $\sigma_u^2 \boldsymbol{I}$ and $\boldsymbol{I}$ is identity matrix. Note here that $(\boldsymbol{I} - \rho \boldsymbol{W})$ is required to be non-singular.

The $\boldsymbol{W}$ matrix describes the neighborhood structure of the small grids and indicate the potential interaction between locations. Unlike the commonly neighbours are defined as cells that share the same boundary, in this paper, we consider the obfuscated area as the neighbouring set of a specific grid. Since the information of the points within the obfuscated circle is quite similar, the cells that are covered by the obfuscated region have a great influence on each other. Thus, we measure the distance between centroids of each two grids as d_{mn}. For each grid, its neighborhood cells are those that the distance d_{mn} is smaller than the obfuscated radius. The corresponding spatial weight is given as:

$$w_{mn} = \begin{cases} d_{mn}^{-1}, & d_{mn} \leq r \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Combining (8), (9) and (10), the full model with spatially correlated random area effects becomes:

$$\widehat{l}_g = \boldsymbol{h}_g^T \boldsymbol{\beta}_g + \boldsymbol{b}_g^T (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{u} + \boldsymbol{e}_g, \tag{11}$$

where $e_g \sim N(0, \sigma_e^2)$ represents the sampling errors that is independent of random area effects, and σ_e^2 refers to the sampling variance of the direct estimates. Let's denote $\boldsymbol{e} = (e_1, \cdot, e_G)$. Further, the covariance matrix can be expressed as $\text{cov}(\widehat{\boldsymbol{l}}) = \boldsymbol{e} + \sigma_u^2 [(\boldsymbol{I} - \rho \boldsymbol{W})(\boldsymbol{I} - \rho \boldsymbol{W}^T)]^{-1}$.

With the collected crowdsourced data $\mathcal{X}$, the estimation $\widehat{\boldsymbol{l}}_g$ can be calculated as:

$$\widetilde{l}_g(\sigma_u^2, \rho) = \boldsymbol{h}_g^T \widehat{\boldsymbol{\beta}} + \boldsymbol{b}_g^T (\text{cov}(\widehat{\boldsymbol{l}}) - \boldsymbol{e}) \text{cov}(\widehat{\boldsymbol{l}})^{-1} (\widehat{l}_g^{DE} - \boldsymbol{h}_g^T \widehat{\boldsymbol{\beta}}), \tag{12}$$

where estimated $\widehat{\boldsymbol{\beta}} = (\boldsymbol{h}_g^T \text{cov}(\widehat{\boldsymbol{l}}_g)^{-1} \boldsymbol{h}_g)^{-1} \boldsymbol{h}_g^T \text{cov}(\widehat{\boldsymbol{l}}_g)^{-1}$ via weighted least square estimator. The estimator $\widetilde{\boldsymbol{l}}_g(\sigma_u^2, \rho)$ depends on the unknown components σ_u^2 and ρ. Assuming normality of the random effects, σ_u^2 and ρ is estimated with restricted maximum likelihood methods, which considers the loss in estimating $\boldsymbol{\beta}$. Therefore, the model estimator based on the estimated parameters $\widehat{\sigma}_u^2$ and $\widehat{\rho}$ is given as:

$$\begin{aligned} \widetilde{\boldsymbol{l}}_g(\widehat{\sigma}_u^2, \widehat{\rho}) = & \boldsymbol{h}_g^T \widehat{\boldsymbol{\beta}} + \boldsymbol{b}_g^T \{\widehat{\sigma}_u^2 [(\boldsymbol{I} - \widehat{\rho} \boldsymbol{W})(\boldsymbol{I} - \widehat{\rho} \boldsymbol{W}^T)]^{-1}\} \\ & \times \text{cov}(\widehat{\boldsymbol{l}}_g)^{-1} (\widehat{l}_g^{DE} - \boldsymbol{h}_g^T \widehat{\boldsymbol{\beta}}). \end{aligned} \tag{13}$$

Then the aggregator will leverage a SEIR model to investigate the temporal evolution of the COVID-19's spread and predict vulnerability populations in the next few days.

C. SEIR Model based Future Trend Prediction

So far, we can assess the vulnerability level of each grid with respect to the estimated mean and covariance. Based on the statistical information $\widetilde{\boldsymbol{l}}_g(\widehat{\sigma}_u^2, \widehat{\rho})$, we can further estimate the current proportion of population $\widetilde{J}_g^t$ who are more likely to experience covid-like illness from the proposed model under the normality assumption. We consider people with vulnerability score more than threshold η as vulnerable population. Therefore, the percentage of vulnerable population equals to the probability of the vulnerability score greater than η from the estimated model.

Then, we leverage this information in a SEIR model to predict the possible risk trend of the specific cell in the next few days. In the SEIR model, total population are divided into four groups: Susceptible, Exposed, Infection and Removed. Susceptible people refer to whom not yet infected but at

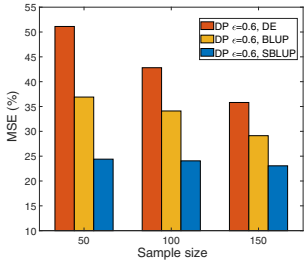


Fig. 2. MSE vs sample size.

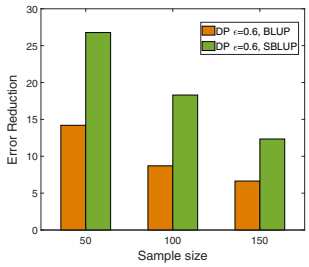


Fig. 3. MSE reduction vs sample size.

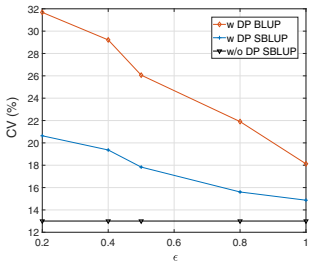
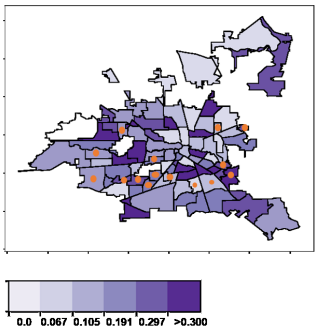
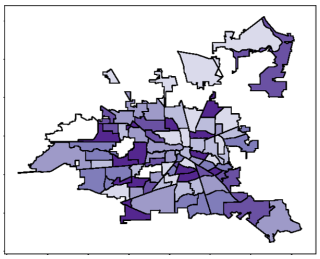


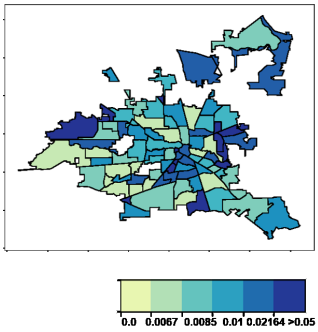
Fig. 4. CV vs privacy budget.



(a) Map w/o privacy model.



(b) Map with privacy model.



(c) Future trend with privacy model.

Fig. 5. Mobile crowdsourcing vulnerability map.

risk of being infected; Exposed people whom with mild or asymptomatic but infectious populations; Infection are people confirmed to be infectious; And the removed are those who have been cured or died from disease. Therefore, the estimated vulnerable population $P_g^s[t] = \widetilde{J}_g^t N_g$ is assumed as the susceptible population in small cell g at time t. Here we adopt a simple SEIR model as follow,

$$P_g^S[t+1] = P_g^S[t] - \frac{\kappa_1 r[t] P_g^I[t] P_g^S[t]}{N_g[t]}, \tag{14}$$

$$P_g^E[t+1] = (1-\nu) P_g^E[t] + \frac{\kappa_1 r[t] P_g^I[t] P_g^S[t]}{N_g[t]}, \tag{15}$$

$$P_g^I[t+1] = \nu P_g^E[t] + P_g^I[t] - \gamma P_g^I[t], \tag{16}$$

$$P_g^R[t+1] = \gamma P_g^I[t] + P_g^R[t]. \tag{17}$$

where κ_1 and κ_2 denote the rate of transmission for the susceptible to infected and the exposed to infected, respectively; r[t] and N_g[t] are the number of contacts per person per day and total population over $\mathcal{A}$, respectively; γ represents the probability of recovery or death while ν indicates the incubation rate of COVID-19, which is the rate of the high-risk individuals becoming symptomatic. The above parameters can be empirically estimated based on the observed confirmed cases, which varies in different areas [12].

V. Performance Evaluation

We now examine the performance of the proposed scheme for privacy-preserving vulnerability map construction. The software used for performance evaluation is Python, and R language is used to construct the spatial estimation model. We regard the Houston city as the targeted area for estimating the risk level at each super neighborhood. Specifically, the whole city is divided into 88 super neighborhood according to the neighborhood planning areas in the City of Houston [13]. Each neighborhood has the attributes of grid ID and the boundary GPS coordinates. The demographic profiles (i.e., age structure, population density and poverty are selected) of super neighborhood are utilized as the auxiliary variables. The simulation results are based on a real world crowdsourced survey from CMU Delphi Research Center [14]. This global survey contains the surveillance streams data of geographic information and the estimated percentage of people with covid-like symptoms. Our scheme is evaluated by comparing with two baselines: 1) *Direct Estimation* utilizes the collected data to directly estimate the vulnerability level and 2) *BLUP* utilizes the collected data and auxiliary factors without considering the spatial correlation, which is the special case of SBLUP when $\rho = 0$.

Firstly, under a fixed privacy budget, we examine the performance of the estimators in terms of the area-specific mean square errors (MSEs), as shown in Fig. 2 and Fig. 3. For a single grid, in general, the estimates based on large samples size are more precise since the model obtains more information of the true vulnerability level, especially in the direct estimation case. It also demonstrates the effectiveness of the proposed scheme compared with two baselines. The value of the estimated spatial coefficient parameter ρ is 0.698 on average, which indicates that the spatial information provides

a good fit in our proposed model. As shown in Fig. 3, the error reduction from our model to DE is more than the BLUP model, especially in the case of small sample size. With the related auxiliary information and strong spatial correlation, the proposed model maintains lowest MSE regardless of the sample size.

To evaluate the impact of DP on the model utility in terms of reliability of model-based estimation, we examine 5 different privacy levels with 100 interations per level. The reliability criteria is the average coefficients of variation (CVs) of the MSE estimates as $\text{cv}(\widetilde{\boldsymbol{l}}) = 100\sqrt{\text{MSE}(\widetilde{\boldsymbol{l}})}/\widetilde{\boldsymbol{l}}$. An estimate with CV over 20% is regarded as unreliable and cannot be published. The results are shown in Fig. 4. The baseline is the spatial model without adding the DP noise. From Definition 1, we know a lower value of ϵ implies a stronger privacy protection that can be guarantee while the data utility may be degraded, since more noise is more likely to be injected to the real location. Thus, as illustrated in Fig. 4, when ϵ is less than 0.3, the location protection guarantee is highly strong while the estimation under differential privacy schemes becomes unreliable. Moreover, privacy preserving models under BLUP and SBLUP get close to the baseline as the increase of ϵ. With the strong impact of the neighbouring structure, compared to the non-spatial scheme, the gap between the proposed model and the baseline is smaller. Since the aggregator treats every obfuscated location report as real one, the aggregated record may deviate from the true value when the uploaded geographic information of a participant is far away from its exact location. With a small privacy budget ϵ, the estimation is more likely to be biased and less reliable due to the spatial error, thus the value of CV increases as ϵ decreases.

We also display three color maps of the estimated average vulnerability of COVID-19 in 88 small areas of Houston in Fig. 5(a), 5(b) and 5(c). Fig. 5(c) depicts the future trend in seven days predicted from the SEIR model. The future trend reflects the estimated percentage of vulnerable population towards the COVID-19. The privacy parameter in Fig. 5(b) is set to be 0.6 and the threshold η is 0.75. Orange dots in the Fig. 5(a) indicates the difference between Fig. 5(a) and 5(b), which are the maps with or without privacy model, respectively. It shows that the privacy model maintains useful information to learn about the spatial trend and vulnerability level. It also illustrates the trade-off between estimator reliability and privacy preservation level. That is to say, by appropriately controlling the value of privacy parameter chosen at each participant, our proposed crowdsourcing system can achieve a reliable estimation result while preserving the participants' location privacy well.

VI. Conclusion

We have developed a mobile crowdsourcing based vulnerability map construction scheme to detect the potential outbreak of coronavirus with consideration of participants' location privacy. Geo-indistinguishability approach have been exploited to protect users' sensitive geographic profiles locally. The spatial estimators have been leveraged to adjust an unreliable risk estimation due to location uncertainty. Further, jointly with the SEIR model, we have been able to predict the future risk trend. The simulation results based on the real-world dataset have shown the effectiveness of the proposed scheme by the low prediction errors, and demonstrates the trade-off between the privacy preservation and estimation reliability.

Acknowledgement

This work was supported in part by the U.S. National Science Foundation under grants US CNS-1646607, CNS-1801925, and CNS-2029569. The work of Y. Gong was supported in part by the U.S. National Science Foundation under grants US CNS-2029685 and CNS-1850523. The work of Y. Guo was supported in part by the U.S. National Science Foundation under grant US CNS-2029685.

References

[1] P.-C. Lai, C. B. Chow, H. T. Wong, K. H. Kwong, Y. W. Kwan, S. H. Liu, W. K. Tong, W. K. Cheung, and W. L. Wong, "An early warning system for detecting H1N1 disease outbreak–a spatio-temporal approach," *International Journal of Geographical Information Science*, vol. 29, no. 7, pp. 1251–1268, 2015.

[2] M. S. Smolinski, A. W. Crawley, K. Baltrusaitis, R. Chunara, J. M. Olsen, O. Wójcik, M. Santillana, A. Nguyen, and J. S. Brownstein, "Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons," *American journal of public health*, vol. 105, no. 10, pp. 2124–2130, 2015.

[3] H. Rossman, A. Keshet, S. Shilo, A. Gavrieli, T. Bauman, O. Cohen, E. Shelly, R. Balicer, B. Geiger, Y. Dor *et al.*, "A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys," *Nature Medicine*, pp. 1–4, 2020.

[4] D. De Ridder, J. Sandoval, N. Vuilleumier, S. Stringhini, H. Spechbach, S. Joost, L. Kaiser, and I. Guessous, "Geospatial digital monitoring of covid-19 cases at high spatiotemporal resolution," *The Lancet Digital Health*, vol. 2, no. 8, pp. e393–e394, 2020.

[5] J. Wang, X. Zhang, Q. Zhang, M. Li, Y. Guo, Z. Feng, and M. Pan, "Data-driven spectrum trading with secondary users' differential privacy preservation," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[6] L. Li, X. Zhang, R. Hou, H. Yue, H. Li, and M. Pan, "Participant recruitment for coverage-aware mobile crowdsensing with location differential privacy," in *2019 IEEE Global Communications Conference*. IEEE, 2019, pp. 1–6.

[7] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, Xi'an, China, April 2008.

[8] H. To, C. Shahabi, and L. Xiong, "Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server," in *Proceeding of 2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Paris, France, Oct. 2018.

[9] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," *arXiv preprint arXiv:1212.1984*, 2012.

[10] J. K. Rao, "Small area estimation," *Wiley StatsRef: Statistics Reference Online*, 2014.

[11] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

[12] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, 2020.

[13] C. of Houston, "SUPER NEIGHBORHOODS," https://www.houstontx.gov/superneighborhoods, Accessed May, 2020.

[14] D. C. Farrow, L. C. Brooks, A. Rumack, R. J. Tibshirani, and R. Rosenfeld, "Delphi Epidata API," https://github.com/cmu-delphi/delphi-epidata, Accessed May, 2020.