This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

Constructing Mobile Crowdsourced COVID-19 Vulnerability Map with Geo-Indistinguishability

Rui Chen, *Graduate Student Member*, *IEEE*, Liang Li, *Member*, *IEEE*, Ying Ma, *Member*, *IEEE*, Yanmin Gong, *Senior Member*, *IEEE*, Yuanxiong Guo, *Senior Member*, *IEEE*, Tomoaki Ohtsuki, *Senior Member*, *IEEE*, and Miao Pan, *Senior Member*, *IEEE*

***Abstract*—Preventing COVID-19 disease from spreading in communities will require proactive and effective healthcare resources allocations, such as vaccinations. A fine-grained COVID-19 vulnerability map will be essential to detect the high-risk communities and guild the effective vaccine policy. A mobile-crowdsourcing-based self-reporting approach is a promising solution. However, an accurate mobile-crowdsourcing-based map construction requests participants to report their actual locations, raising serious privacy concerns. To address this issue, we propose a novel approach to effectively construct a reliable community-level COVID-19 vulnerability map based on mobile crowdsourced COVID-19 self-reports without compromising participants' location privacy. We design a geo-perturbation scheme where participants can locally obfuscate their locations with the geo-indistinguishability guarantee to protect their location privacy against any adversaries' prior knowledge. To minimize the data utility loss caused by location perturbation, we first design an unbiased vulnerability estimator and formulate the location perturbation probability generation into a convex optimization. Its objective is to minimize the estimation error of the direct vulnerability estimator under the constraints of geo-indistinguishability. Given the perturbed locations, we integrate the perturbation probabilities with the spatial smoothing method to obtain reliable community-level vulnerability estimations that are robust to a small-sampling-size problem incurred by location perturbation. Considering the fast-spreading nature of coronavirus, we integrate the vulnerability estimates into the modified susceptible-infected-removed (SIR) model with vaccination for building a future trend map. It helps to provide a guideline for vaccine allocation when supply is limited. Extensive simulations based on real-world data demonstrate the proposed scheme superiority over the peer designs satisfying geo-indistinguishability in terms of estimation accuracy and reliability.**

***Index Terms*—Mobile crowdsourcing, Location privacy, Differential privacy, Optimization, Small area estimation.**

I. Introduction

The pandemic of the coronavirus (COVID-19) has raised an unprecedented global crisis in various aspects (e.g., public health and economy). Vaccination is widely regarded as one of the most effective methods in curbing the spread of COVID-19. To date, a new variant of SARS-CoV-2 has been detected and spread over forty countries and regions in the world [1]. The top priority is to achieve a high vaccination rate to protect people from spreading the virus to others. However, many countries and regions are still facing the challenge of judicious distribution of SARS-CoV vaccines [2]. Given that the finite vaccine storage and transportation, it will take a long time to obtain enough doses of the vaccine to vaccinate the entire society. For the non-vaccine-producing countries, the vaccine comes in batches. Although World Health Organization (WHO) has announced relevant guidelines for vaccine distribution in disrupting disease transmission [3], it will be imperative to refine these guidelines according to the actual risk level of different communities that are vulnerable to COVID-19. Hence, early and rapid identification of the most "vulnerable" communities is vital for the judicious allocation of limited medical resources (e.g., vaccines).

The early identification of suspected cases during an epidemic is often depicted as a heatmap with the locations of vulnerability risk predictions [4]. The success of COVID-19 vulnerability map construction relies on comprehensive health information. However, it is extremely time-consuming to identify "the most vulnerable" people and their residential communities by physically "scanning" all the communities for vaccine allocation, especially for economically disadvantaged underrepresented communities. It may seriously affect COVID-19 data detection and fail to make an early response to contain the next potential "outburst" spots.

Most recently, the Internet of Medical Things (IoMT) served as an extension, and specialization of the Internet of Things (IoT) has been used to combat COVID-19 disease [5]. IoMT helps collect informative medical and symptom data by using IoT devices (e.g., electronic thermometer and wearable detection sensors) for COVID-19 disease detection. Moreover, with the aid of mobile crowdsourcing [6], [7], more related sources of information can be collected from real-world environments via employing mobile users to participate in data acquisition and used to provide various COVID-19 applications. For example, by distributing real-time surveys to ubiquitous mobile users via a mobile crowdsourcing platform, it only takes a few seconds to obtain a current snapshot of the number of people in each area who are at higher risk of COVID-19. It later can be used to build prioritized policy for vaccination. The feasibility of this approach lies in the popularity of mobile/IoMT devices

R. Chen and M. Pan are with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204 (e-mail: rchen19@uh.edu, mpan2@uh.edu).

L. Li is with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, P.R.China. (e-mail: lilliang1127@bupt.edu.cn).

Y. Ma is with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 (e-mail: ying.ma@ucf.edu).

Y. Gong is with the Department of Electrical and Computer Engineering, and Y. Guo is with the Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249 (e-mail: guoyuanxiong@gmail.com, yuanxiong.guo@utsa.edu).

T. Ohtsuki is with the Department of Information and Computer Science, Keio University, Yokohama, Kanagawa 223-8522, Japan (e-mail: ohtsuki@ics.keio.ac.jp).

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

and the wide expectation that people may be willing to share their self-reported data related to COVID-19 with the public to help combat the COVID-19. During the outbreak of COVID-19, Facebook has released an interactive coronavirus symptom map via the crowdsourced data from an opt-in survey [8], More than 1 million people had responded to the survey within the first two weeks. The tremendous data size and diverse information tagged with fine geographic information make it possible for fine-grained map construction.

However, the mobile crowdsourcing based COVID-19 map construction is not perfect. Existing researches such as [9], [10] analyzed and estimated the severity of the disease in specific areas using self-reported data via an online survey. Others like [11], [12] focus on leveraging various statistic or machine learning tools to enhance the accuracy of the risk assessment. These solutions require mobile participants to upload their COVID-related data and their exact location information to untrusted platforms. Such location information is sensitive, based on which an attacker can infer users' identities when demographic or other readily available attributes are on the file. It will lead to serious privacy concerns, and the mobile participants may be reluctant to contribute any data to the mobile crowdsourcing platform [13], [14]. Therefore, it is necessary to ensure users' location privacy to retain and attract mobile participants. Differential privacy (DP), which provides quantified data privacy with strong theoretical guarantees, has been recognized as a promising protection scheme without assumptions about the attackers' background information. Consequently, several differentially private location obfuscation mechanisms [15], [16] have been proposed to protect users' locations under the DP guarantee. However, applying differentially private location perturbation schemes on vulnerability map construction is challenging. The crowdsourced data are used as the sampling observations to estimate the population vulnerability distribution in the targeted area. With the location obfuscation scheme, the observation based on the perturbed locations would inevitably affect the quantity of the high-risk observations and degrade the utility of population vulnerability estimation . Moreover, the location perturbation scheme in a fine-grained map may reduce the sample sizes in some small regions and leads to an unreliable population vulnerability estimation. Thus, it is critical to consider the utility and reliability of population vulnerability estimation in the design of participants' location privacy preservation schemes for vulnerability map construction.

To address these issues, in this work, we develop a fine-grained COVID-19 vulnerability map construction scheme via mobile crowdsourcing while preserving participants' location privacy. Specifically, we design a location-privacy-preserving mobile crowdsourcing framework for COVID-19 data collection, where mobile participants locally obfuscate their locations using our differentially private location perturbation scheme. The utility-assured differentially private location perturbation scheme is efficiently generated at the server side without violating the users' privacy. Hence there is no additional computing overhead on the mobile participants' side. Moreover, we leverage the spatial correlation between neighboring areas, incorporating the geo-perturbation probabilities with the spatial weighting matrix, to mitigate the small sample issue incurred by the location perturbation scheme. It further enhance the reliability of the vulnerability estimation. Our salient contributions are summarized as follows.

- We propose a novel location privacy-preserving vulnerability map construction scheme. Briefly, we leverage the help of mobile crowdsourcing to virtually find out the most vulnerable people and estimate the vulnerability levels of COVID-19 in a targeted area without disclosing the participants' location differential privacy.
- We develop a differentially private geo-perturbation scheme, which allows mobile crowdsourcing participants to locally perturb their locations meanwhile providing useful and reliable vulnerability estimations. To this end, we establish an unbiased estimator of vulnerability level and formulate the geo-perturbation probability generation as a convex optimization to minimize the variance of the unbiased estimator under the geo-indistinguishability constraints. The gradient descent method is employed to find the optimal perturbation probabilities.
- Given the obfuscated locations, we employ the Bayesian smoothing method to integrate the geo-perturbation probabilities to the spatial weighting matrix and auxiliary data from the publicly available census, which can improve the estimation reliability when the crowdsourcing data in a subarea is small after location perturbation. Further, we show how the reliable vulnerability estimation can be applied to vaccine allocation and integrate the vulnerability estimates with the susceptible-infected-removed (SIR) model to generate the future trend map.
- Extensive simulations are conducted based on real-world datasets to evaluate the performance of our scheme. Compared with different location privacy preserving mechanisms, the proposed location scheme can reduce the about 20% estimation variance for vulnerability map construction. The results also demonstrate the tradeoff between DP and risk estimation reliability.

The rest of this paper is organized as follows: In Section II, the related work is discussed. In Section III, the preliminary of location differential privacy and overall system model are presented. In Section IV, the location perturbation scheme and the problem formulation are described as well as the effective iterative algorithm is proposed to find the optimal solution. In Section V, the Bayesian smoothing model for community-level vulnerability inference, vaccine allocation policy and future prediction on SIR model are discussed. In Section VI, the experiment based on the true database are analyzed and the paper is concluded in Section VII.

II. Related Work

From the COVID-19 risk assessment respective, existing works have adopted the compartmental models or machine learning tools to quantify the risk under COVID-19 and predict the next potential COVID-19 disease outbreak spots. The data source used in the risk assessment framework is dependent mainly on the daily confirmed and death cases [11], [17]. However, the data collection above is time-consuming and

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

biased since it lacks the coverage of the population with asymptomatic or mild symptoms. Some works have leveraged cost-effective data collection approaches, e.g., mobile/online surveys and social media platforms, and developed COVID-19 symptom maps to investigate the dynamics of the COVID-19 [10], [18]. For example, Jahanbin et al. in [18] used Twitter comments to estimate the severity of the disease in certain areas. Facebook and CMU university have collaborated to launch a symptom reporting survey to estimate the number of COVID-like illnesses. For survey data processing, they discarded the areal results with survey responses under 100 to avoid unreliable estimation. However, the previous works assume the data collector is always trustworthy and enables mobile users to contribute their COVID-related data tagged with their exact locations, causing privacy concerns. To keep user anonymity, the data collectors aggregate the users' location information to high levels (e.g., the state or city scale) before publishing the map. While such aggregation provides risk assessments from the macroscopic perspective, it can disable data analysis of COVID-19 at the micro level.

From the location privacy perspective, the early works on preserving location privacy in disease mapping were geo-donuts [19] and k-anonymity based location anonymity [20]. However, the cloaking mechanisms such as geo-donuts have limitations and fail to provide privacy protections against the adversary with the background information knowledge about the target user's location distribution [21]. Recently, location differential privacy schemes have been proposed to provide rigorous privacy protection independent of an adversary's prior knowledge. However, the perturbation variance is exponentially increasing with a large domain size. It is challenging to maintain a high data utility, and it is also unclear how reliable the estimation will be. Bordenabe et al. [21] studied the privacy and utility tradeoff and formulated the location perturbation problem to minimize the distance between the original and perturbed locations. The optimization problem used the graph-based approximation to reduce the solving complexity and thus cannot guarantee optimality. Gu et al. [22] also investigated the tradeoff between privacy and utility when estimating frequency query for location check-ins. However, their query utility is dependent on the unknown true frequencies and cannot be directly evaluated.

Unlike these existing works above and our previous work [23], in our map construction, the proposed location perturbation probability is generated with considering the utility of the aggregated estimation under the geo perturbation scheme. Such perturbation probability generation is formulated as an estimation error minimization problem that is independent on the unknown true information. A gradient descent method is employed to seek for the optimal solutions. Moreover, we consider the small sampling size problem incurred by the geo perturbation scheme and allow the crowdsourced aggregator to use the Bayesian smoothing method to adjust the community-level estimates rather than directly discard them. The vulnerability map also includes the uncertainty information of our community-level estimates to demonstrate the reliability of the estimator.

III. Preliminaries & System Overview

A. Location Differential Privacy Preliminaries

With the principle of the standard centralized differential privacy [24], local DP (LDP) especially allows each user to perturb her private data locally via a randomized mechanism without the requirement of trustworthy third-party entities. The definition of LDP is shown as follows.

***Definition** 1 (LDP [25]):* Suppose a privacy parameter $\epsilon \geq 0$, a randomization algorithm $\mathcal{M}$ satisfies ϵ-local differential privacy. For any pair of inputs X, Y and any output $S \in$ *range*$(\mathcal{M})$,

$$\frac{Pr[\mathcal{M}(X) = S]}{Pr[\mathcal{M}(Y) = S]} \leq e^{\epsilon}. \tag{1}$$

Intuitively, the above definition states that when ϵ is smaller, the probabilities of which two different inputs, X and Y, have the same output via the randomized algorithm $\mathcal{M}$ are closer to each other. Hence, privacy preservation level is controlled by the privacy parameter ϵ. A smaller ϵ leads to higher privacy preservation as it is harder for an adversary to determine whether a user has this sensitive input, given an output S.

LDP is recently deployed in the application of location privacy [15]. Based on the principle of LDP, geo-indistinguishability is designed to preserve users' location privacy against adversaries with background information. Mathematically, in geo-indistinguishability scheme, a user n can perturbhis real location to another one based on a pre-set randomized location obfuscation algorithm $\mathcal{M}$ (i.e., $\mathcal{M}$ maps location a to g with given probabilities) and then shares the perturbed location g in public. With the LDP guarantee, if an adversary observes user n is in g, the adversary cannot distinguish whether g is the true location of n, even if he knows the randomized algorithm $\mathcal{M}$. According to *Definition 1*, geo-indistinguishability is formally defined as follows:

***Definition** 2 (ϵ-geo-indistinguishability [15]):* With the privacy parameter $\epsilon \geq 0$, a randomized location obfuscation algorithm $\mathcal{M}$ satisfies ϵ-LDP on the concerned area that includes a set of locations Θ, if for any two different locations $a_0, a_0' \in \Theta$ and an arbitrary location g, the following holds:

$$\frac{Pr[\mathcal{M}(a_0) = g]}{Pr[\mathcal{M}(a_0') = g]} \leq e^{\epsilon d(a_0, a_0')}, \tag{2}$$

where $d(a_0, a_0')$ denotes the Euclidean distance between locations a_0 and a_0'.

The ϵ-geo-indistinguishability aims to protect the actual location by hiding among the set of locations Θ due to their similar probability distributions for perturbed locations. From Definition 2, it is easily observed that as the distance $d(a_0, a_0')$ of two different locations a_0 and a_0' is smaller, they are more indistinguishable since their output distributions are closer to each other. Moreover, it has been theoretically shown that ϵ-geo-indistinguishability can protect users' sensitive location information against adversaries with arbitrary prior knowledge. Suppose that the adversary has prior knowledge about a user's location distribution π. After the adversary observes the obfuscated location s, the information gain of his posterior knowledge σ over π is bounded by $e^{\epsilon d_{\max}}$, i.e., $\sigma/\pi \leq e^{\epsilon d_{\max}}$ ($d_{\max}$ is the maximum distance of any two locations in Θ),

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

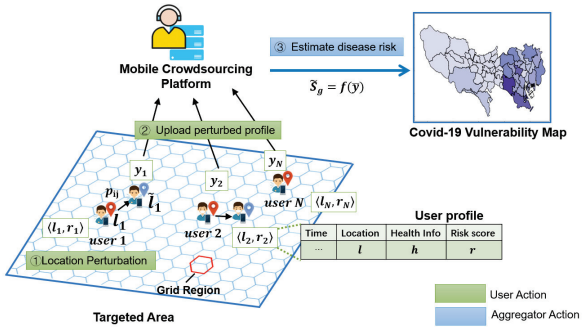


Fig. 1. Vulnerability map construction via privacy preserving mobile crowdsourcing.

regardless of what the prior π is [15]. Please refer to [15] for the theoretical proof.

B. System Overview

In this work, we proposed to develop a mobile crowdsourcing assisted privacy preserving COVID-19 vulnerability estimation scheme, which includes a mobile crowdsourcing platform and a number of participants distributed over the concerning area, as shown in Fig. 1. To virtually detect the most vulnerable area and estimate vulnerability levels in the target area, the crowdsourcing aggregator leverages mobile crowdsourcing to collect multi-dimensional user profile related to vaccine allocation such as personal health data, age, gender, employment status, etc. Suppose that all the participants are willing to engage in mobile crowdsourcing, the collected user profile from the participants is truthfully reported. Since the vulnerability estimation is related to the location, the mobile crowdsourcing platform needs the geographic information of the participants and aggregates the participants' user profiles within the same locations. Such geographic information is closely related to either home or work address. The participants' identities can be easily inferred when combing the location information, user profile and other available attributes. Assume that the crowdsourcing aggregator is semi-honest which implies he follows the proposed protocol but tries to infer the participants' identities via the uploaded location, which results in severe user privacy leakage. To reduce information leakage, the participants are allowed to perturb their location information by following a well-designed geo-perturbation scheme.

Assume that there are N mobile users who are distributed over the targeted area $\mathcal{A}$. The targeted area $\mathcal{A}$ is divided into G non-overlapping cells, denoted by the set $\mathcal{A} = \{a_1, a_2, \cdots, a_G\}$. Let $\mathcal{G} = \{1, \cdots, i, \cdots, G\}$ denote the indices of cells. Each cell a_i is tagged with a certain COVID-19 vulnerability level θ_i. The entire vulnerability prediction map (VPM) is modeled as $\boldsymbol{\theta} \triangleq [\theta_1, \cdots, \theta_G]$. In a fine-grained VPM, the spatial unit is set to the street or township level. Let n be user index, and if n's reporting location falls into the range of the a_i-th cells, we can roughly regard user n's location as the a_i-th cells, denoted by $l_n = a_i$. Denote $\boldsymbol{h}_n = [h_{n1}, \cdots, h_{nM}]$ be the n-th user profile related to COVID-19 data analysis such as demographic data (e.g., gender, age, occupation, employment status) and pre-existing health conditions. Given the user profile $\boldsymbol{h}_n$, the user n can obtain his risk of infection r_n via a predetermined function $f : \boldsymbol{h}_n \in \mathcal{D}^M \rightarrow r_n \in \{-1, 1\}$ which is pre-configured by the crowdsourcing aggregator, where 1 represents high risk and -1 is low risk. With a slight abuse of notation, we denote the set of crowdsourcing participant as $\mathcal{N}_p$, where $n \in \mathcal{N}_p = \{1, 2, \cdots, N_p\}$ and $N_p \leq N$, and the location-data pair of the participant n as $\langle l_n, r_n \rangle$. Then, according to the location-data pairs of the crowdsourcing participants $n \in \mathcal{N}_p$, the aggregator will estimate and predict the vulnerability value as $\hat{\boldsymbol{\theta}} \triangleq [\hat{\theta}_1, \cdots, \hat{\theta}_G]$ and construct the corresponding VPM.

This work focuses on the problem of preserving the users' location privacy while providing effective community-level vulnerability estimation, which can effectively provide guidelines of vaccine allocation strategies. Briefly, the crowdsourcing platform launches the task, i.e., gathering personal health data via mobile applications and IoMT sensors. The crowdsourcing participants fulfill their user-profiles and locally perturb their true location information based on the proposed geo-obfuscation scheme. Then, they upload their user profile tagged with the obfuscated locations to the crowdsourcing platform. After receiving participants' user profiles, the aggregator estimates the community-level vulnerability levels based on the vaccine allocation strategies.

Note that, since the location obfuscation scheme is deployed locally on participants' side, no additional sensitive information is revealed to the crowdsourcing aggregator. Hence, users' location privacy can be well protected. However, the observed locations in the crowdsourcing platform may be different from the actual locations. It may result in a biased estimation since the crowdsourcing aggregator is indistinguishable from the actual and obfuscated locations. Moreover, it affects the collected data size in each cell. The direct estimation of the cells with a small sample size becomes unreliable, leading to unacceptable data utility for determining vaccine allocation strategies. In the following, we address these issues to improve the accuracy and reliability of community-level estimation from two perspectives: (1) We integrate the community-level estimation error minimization in our proposed geo-perturbation scheme, which will be shown in Sec. IV. (2) We adjust the community-level estimation by considering the location perturbation effect, spatial correlation of the neighboring areas, and social-economic risk factors. Therefore, biased location information and insufficient sample size have a small impact on community-level estimation degradation and vaccine allocation inefficiency, which is discussed in Sec. V.

IV. Utility-assured Geo-Perturbation Scheme Design

Before collecting participants' location-data pairs, a probabilistic perturbation function P needs to be generated for the crowdsourcing participants to provide location DP guarantee. Here, the semi-honest platform can take charge of generating the perturbation function P without violating users' privacy. That is because DP can provide a theoretical guarantee to protect user's sensitive information, i.e., location information in this paper, against the adversaries who know P. In other

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

TABLE I
PARAMETER NOTATIONS.

System parameters	
$\mathcal{N}_p(N_p)$	Set (numbers) of crowdsourcing users
$\mathcal{A} = \{a_i\}_{i=1}^{G}$	Set of non-overlapping cells
$\boldsymbol{h}_n, r_n, l_n$	User profile, infection risk, and location of user n
$\langle l_n, r_n \rangle$	Location-data pair of user n
$\boldsymbol{\theta}(\hat{\boldsymbol{\theta}})$	True (estimated) vulnerability levels
$\mathcal{G}$	Set of location indices
Geo-perturbation parameters	
ϵ	Differential privacy budget
$[p_{si}, p_{ri}]_{i=1}^{G}$	Location perturbation probabilities
$\mathbf{s}_n^i$	Encoding vector of user n whose $\langle l_n = a_i, r_n \rangle$
$\mathbf{y}_n$	Perturbed vector of user n
$S_i(\hat{S}_i)$	True (estimated) count of $\langle a_i, 1 \rangle$
Z_i	True count of $\langle a_i, -1 \rangle$
$O_{i,1}(O_{i,-1})$	Observed counts of $y[i] = 1$ ($y[i] = -1$)
$d(\cdot, \cdot)$	Distance function
Bayesian smoothing parameters	
E_i	Expected count of $\langle a_i, 1 \rangle$
$W = [w_{ij}]^{G \times G}$	Spatial weight matrix
$\mathbf{X}_i$	Vector auxiliary coefficients
$\boldsymbol{\beta}_i$	Vector regression variables
$u_i(v_i)$	Spatial correlated (uncorrelated) random effect
e_i	Residual variation
SIR model parameters	
N_i	The population of cell a_i
$V_i(\gamma_i)$	Control parameters for vaccine intervention (recovery)
$P_i^S[t], P_i^I[t], P_i^R[t]$	The number of the susceptible,infected, and removed individuals of a_i at time t

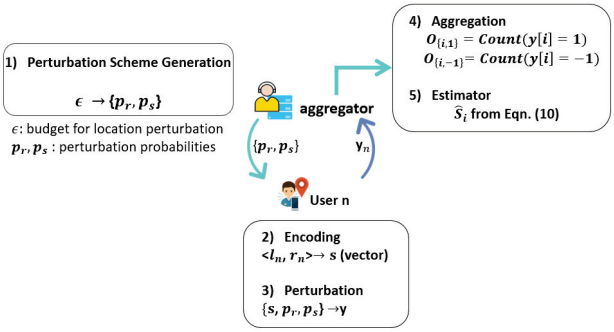


Fig. 2. The overview of geo-perturbation scheme.

words, users can get privacy protection under the perturbation function P even if it is generated by the untrust platform. Note that a large perturbation noise can provide strong DP guarantee while it would perturb the original location to a point that is far away from the original one, degrading the data utility.

In this work, we propose a geo-perturbation scheme (GEP). The goal of GEP design is to get useful estimation of community-level vulnerability. The overall procedure is shown in Fig. 2. The first step is to generate the perturbation probabilities matrix that is optimized via Alg. 1. In the second step, the mobile users encode their own location-data pair $\langle l, r \rangle$ into a G-length vector and then perturb the G-length vectors according to the perturbation probabilities matrix locally. In the last step, after receiving the perturbed vectors from crowdsourcing users, The aggregator utilize the pre-determined estimators to obtain the risk levels in each community.

In the following, we start with the aggregated estimator design and the formulation of perturbation probabilities generation. Inspired by Unary Encoding (UE) scheme in [26], our probabilistic perturbation function is defined as $P = [p_{si}, p_{ri}]_{i=1}^{G}$. Hence, compared with the perturbation matrix $\mathbf{P}' \in \mathbb{R}^{G \times G}$, the computation complexity of P is greatly reduced due to fewer parameters. Different from the original UE scheme in [26], the proposed GEP assigns different perturbation probabilities to different bias. It is the key point to achieve while still providing the ϵ-geo-indistinguishability for location privacy guarantee. Moreover, the existing perturbation schemes, such as [26], [27], are not suitable for vulnerability estimation since they only handle the proportion or count of participants whose location pair satisfies $\langle l_n = a_i, r_n \rangle$, while we consider the more complicated frequency estimation of whose location pair satisfies $\langle l_n = a_i, r_n = 1 \rangle$.

A. Mechanism Design

The location-data pair $\langle l, r \rangle$ is first encoded to a G-length vector (the subscript n is omitted for brevity in the rest of this section),

$$\mathbf{s}^i = [0, \cdots, 0, r, 0, \cdots, 0], \tag{3}$$

where vector $\mathbf{s}^i$ represents the vector whose i-th entry is r, and other entries are 0s. Then, each bit of the encoded vector $\mathbf{s}^i$ is perturbed into 1, -1 or 0 independently to get the output vector $\mathbf{y}$ with probabilities:

$$P(\mathbf{y}[i]|\mathbf{s}[i] = r) = \begin{cases} p_{si}, & \text{if} \quad y[i] = r \\ (1 - p_{si})/2, & \text{if} \quad y[i] = -r \\ (1 - p_{si})/2, & \text{if} \quad y[i] = 0 \end{cases}, \tag{4}$$

$$P(\mathbf{y}[i]|\mathbf{s}[i] = 0) = \begin{cases} p_{ri}/2, & \text{if} \quad y[i] = 1 \\ p_{ri}/2, & \text{if} \quad y[i] = -1 \\ 1 - p_{ri}, & \text{if} \quad y[i] = 0 \end{cases}. \tag{5}$$

For two different vector $\mathbf{s}^i$, where only the i-th bit is r (1 or -1), and $\mathbf{s}^j$ where $i, j \in \mathcal{G}$, the probability ratio of distinguishing the encoding location-data pair of $\mathbf{s}^i$ and $\mathbf{s}^j$ by observing the perturbed vector $\mathbf{y}$ is

$$\frac{P(\mathbf{y}|\mathbf{s}^i)}{P(\mathbf{y}|\mathbf{s}^j)} = \frac{P(\mathbf{y}[i]|\mathbf{s}^i)P(\mathbf{y}[j]|\mathbf{s}^i)}{P(\mathbf{y}[i]|\mathbf{s}^j)P(\mathbf{y}[j]|\mathbf{s}^j)} \tag{6}$$

$$\leq \frac{P(\mathbf{y}[i] = r|\mathbf{s}^i)P(\mathbf{y}[j] = 0|\mathbf{s}^i)}{P(\mathbf{y}[i] = r|\mathbf{s}^j)P(\mathbf{y}[j] = 0|\mathbf{s}^j)} \tag{7}$$

$$= \frac{4p_{si}(1 - p_{rj})}{p_{ri}(1 - p_{sj})}, \tag{8}$$

where the second equations holds if and only if $\mathbf{y}[i] = r$ and $\mathbf{y}[j] = 0$. Then, the location privacy constraint is

$$\frac{4p_{si}(1 - p_{rj})}{p_{ri}(1 - p_{sj})} \leq e^{\epsilon d(i,j)}, \qquad \forall i, j \in \mathcal{G}. \tag{9}$$

For the estimation of the true vulnerability count of the cell a_i, denotes the true counts of location-data pairs $\langle a_i, 1 \rangle$ of all the participants as S_i. Let $O_{i,1} = \text{Count}(y[i] = 1)$, and $O_{i,-1} = \text{Count}(y[i] = -1)$ be the observed counts in the crowdsourcing platform. Then we have the following lemma for the estimation of S_i.

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

Lemma 1: The unbiased estimator of S_i is

$$\hat{S}_i = \frac{O_{i,1} + O_{i,-1} - N_p p_{ri}}{1 + p_{si} - 2p_{ri}} + \frac{O_{i,1} - O_{i,-1}}{3p_{si} - 1}. \tag{10}$$

Proof: Denote the true counts of location-data pairs $\langle a_i, -1\rangle$ of all the participants as Z_i. According to the perturbation probabilities (4) and (5), we have,

$$\begin{cases} \mathbb{E}[O_{i,1}] = S_i p_{si} + Z_i \frac{1-p_{si}}{2} + (N_p - S_i - Z_i)\frac{p_{ri}}{2}, \\ \mathbb{E}[O_{i,-1}] = S_i \frac{1-p_{si}}{2} + Z_i p_{si} + (N_p - S_i - Z_i)\frac{p_{ri}}{2}. \end{cases}$$

From which we get:

$$\begin{aligned} \mathbb{E}[\hat{S}_i] &= \frac{\mathbb{E}[O_{i,1} + O_{i,-1}] - N_p p_{ri}}{1 + p_{si} - 2p_{ri}} + \frac{\mathbb{E}[O_{i,1} - O_{i,-1}]}{3p_{si} - 1} \\ &= \frac{S_i + Z_i}{2} + \frac{S_i - Z_i}{2} = S_i. \end{aligned} \tag{11}$$

Therefore, $\hat{S}_i$ is an unbiased estimator of S_i. ∎

Here, we consider the crowdsourcing aggregator uses MSE to evaluate the utility of the estimates $\hat{S}_i$, i.e., the less MSE the better utility. Note that the MSE is calculated by the summation of variance and the square of its bias. Moreover, the MSE of unbiased estimator $\hat{S}_i$ is equal to its variance

$$\text{MSE}(\hat{S}_i) = \text{Var}\left[\frac{O_{i,1} + O_{i,-1} - N_p p_{ri}}{1 + p_{si} - 2p_{ri}} + \frac{O_{i,1} - O_{i,-1}}{3p_{si} - 1}\right]. \tag{12}$$

For convenience, denote $B_1 = O_{i,1} + O_{i,-1}$, $B_2 = O_{i,1} - O_{i,-1}$, $C_1 = \frac{1}{1+p_{si}-2p_{ri}}$ and $C_2 = \frac{1}{3p_{si}-1}$, then we have

$$\text{Var}[\hat{S}_i] = C_1^2 \text{Var}[B_1] + C_2^2 \text{Var}[B_2] + C_1 C_2 \text{Cov}_{B_1,B_2}. \tag{13}$$

We generalize the results from [27]:

$$\text{Var}[B_1] = N_p(p_{ri} - p_{ri}^2) + (S_i + Z_i)(\frac{1 - p_{si}^2}{4} + p_{ri}^2 - p_{ri}), \tag{14}$$

$$\text{Var}[B_2] \le N_p p_{ri} + \frac{1}{2}(S_i + Z_i)(1 + p_{si} - 2p_{ri}), \tag{15}$$

$$\text{Cov}_{B_1,B_2} = \frac{1 - p_{si}^2}{4}(S_i + Z_i)(3p_{si} - 1), \tag{16}$$

to upper bound the second terms in Eqn. (13). Note that only $\text{Var}[B_2]$ is computed by its upper bound. According to (14)-(16), we have

$$\begin{aligned} \text{Var}[\hat{S}_i] \lesssim\ & \frac{N_p p_{ri}(1 - p_{ri})}{(1 + p_{si} - 2p_{ri})^2} - \frac{(S_i + Z_i)p_{ri}(1 - p_{ri})}{(1 + p_{si} - 2p_{ri})^2} \\ & + \frac{N_p p_{ri}}{(3p_{si} - 1)^2} + \frac{(S_i + Z_i)(1 + p_{si} - 2p_{ri})}{2(3p_{si} - 1)^2} \\ & + \frac{(S_i + Z_i)(1 - p_{si}^2)(2 + p_{si} - 2p_{ri})}{4(1 + p_{si} - 2p_{ri})^2}. \end{aligned} \tag{17}$$

Afterwards, we can get the optimal perturbation funtion P by solving the following optimization problem:

$$\min_{\mathbf{p}_s, \mathbf{p}_r} \quad \text{MSE}(\hat{\mathbf{S}}) \triangleq \sum_{i=1}^{G} \text{Var}[\hat{S}_i] \tag{18a}$$

$$\text{s.t.} \quad 4p_{si}(1 - p_{rj}) \le \gamma_{ij} p_{ri}(1 - p_{sj}), \forall i, j \in \mathcal{G}, \tag{18b}$$

$$0 \le p_{ri} \le 0.5 \le p_{si} \le 1, \forall j \in \mathcal{G}, \tag{18c}$$

where $\gamma_{ij} = e^{\epsilon d(i,j)}$, $\mathbf{p}_s = [p_{s1}, \cdots, p_{sG}]$ and $\mathbf{p}_r = [p_{r1}, \cdots, p_{rG}]$ are the two variables of the location perturbation probabilities. Constraint (18b) provides the ϵ-geo-indistinguishablity in the *Definition 2* for location privacy. Constraint (18c) ensures for better data utility. Note that the unknown values S_i and Z_i in Eqn. (17) makes it hard to find the optimal perturbation probabilities. Next, we address this challenge by obtaining a variant of MSE in Eqn. (18a) that is independent of the unknown parameters.

B. Problem Reformulation

In this part, we consider RAPPOR's implementation [28] such that $p_{si} + p_{ri} = 1$; Intuitively, we treat the information that the participants' location is in the cell i or not is equally sensitive thus $p_{si} = 1 - p_{ri}$. We add the corresponding constraints $p_{si} + p_{ri} = 1, \forall i \in \mathcal{G}$ in problem (18). Then, the overall MSE can be rewritten as,

$$\begin{aligned} \text{MSE}(\hat{\mathbf{S}}) &= \sum_{i=1}^{G} \frac{N_p(1 - p_{si}^2)}{(3p_{si} - 1)^2} + \frac{(S_i + Z_i)(2 + p_{si} - p_{si}^2)}{4(3p_{si} - 1)} \\ &\le \sum_{i=1}^{G} \frac{N_p(1 - p_{si}^2)}{(3p_{si} - 1)^2} + \max\left\{\frac{2 + p_{si} - p_{si}^2}{4(3p_{si} - 1)}\right\} N_p. \end{aligned} \tag{19}$$

Here, the second inequality is due to $\sum_{i=1}^{G} S_i + Z_i = N_p$. The variant MSE can be regarded as the MSE in the worst case. Then the optimization problem (18) is reformulated as

$$\min_{\mathbf{p}_s, \mathbf{p}_r} \quad \sum_{i=1}^{G} \frac{N_p(1 - p_{si}^2)}{(3p_{si} - 1)^2} + \max\left\{\frac{2 + p_{si} - p_{si}^2}{4(3p_{si} - 1)}\right\} N_p \tag{20a}$$

$$\text{s.t.} \quad (18b), (18c),$$

$$p_{si} + p_{ri} = 1, \forall i \in \mathcal{G}. \tag{20b}$$

We can further simplify the constraints (18b) and (20b) and obtain new constraints as follows,

$$(p_{si} + p_{sj})\gamma_{ij} - (\gamma_{ij} - 4)p_{si}p_{sj} \le \gamma_{ij}, \forall i, j \in \mathcal{G}. \tag{21}$$

To circumvent this difficulty due to the product of the two variables p_{si} and p_{sj}, the big-M formulation [29] is utilized to decompose this product. We introduce $b_{ij} = p_{si}p_{sj}$ as auxiliary variable and impose the following additional constraints:

$$b_{ij} \le p_{si}, \forall i, j \in \mathcal{G}, \tag{22}$$

$$b_{ij} \le p_{sj}, \forall i, j \in \mathcal{G}, \tag{23}$$

$$b_{ij} \ge p_{si} + p_{sj} - 1, \forall i, j \in \mathcal{G}, \tag{24}$$

$$0 < b_{ij} \le 1, \forall i, j \in \mathcal{G}. \tag{25}$$

Then, we substitute b_{ij} into the constraint (21) and have

$$(p_{si} + p_{sj})\gamma_{ij} - (\gamma_{ij} - 4)b_{ij} \le \gamma_{ij}, \forall i, j \in \mathcal{G}. \tag{26}$$

This is an affine function with respect to the new optimization variables $\mathbf{b} = \{b_{ij}\}_{i,j=1}^{G}$. We note that the constraints (21) and (26) are equivalent when the constraints (22)-(25) are satisfied. Consequently, the perturbation generation problem is rewritten as the following problem:

$$\min_{\mathbf{p}_s, \mathbf{b}, z} \quad \mathcal{E} \triangleq \sum_{i=1}^{G} \frac{N_p(1 - p_{si}^2)}{(3p_{si} - 1)^2} + zN_p \tag{27a}$$

$$\text{s.t.} \quad (18c), (22) - (26), \tag{27b}$$

$$0.25(2 + p_{si} - p_{si}^2)(3p_{si} - 1)^{-1} \le z, \forall i \in \mathcal{G}, \tag{27c}$$

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

where z is new variances for relaxing the max function. In the next section, we will use a standard gradient descent method to obtain the optimal solution to problem (27).

C. Solution of Location Perturbation Optimization

In this section, we utilize the gradient descent method to find the optimal perturbation probabilities $\mathbf{p}_s$. We first show the problem (27) is a convex problem.

The first-order derivative of the objective in the problem (27) with respect to $\mathbf{p}_s$ can be expressed as

$$\frac{\partial \mathcal{E}}{\partial p_{si}} = -\frac{2N_p\,(3-p_{si})}{(3p_{si}-1)^3}. \tag{28}$$

The corresponding Hessian with respect to p_{si} is computed by

$$\bigtriangledown^2_{\mathbf{p}_s}\mathcal{E} = \begin{bmatrix} \frac{\partial^2 \mathcal{E}}{\partial p_{s1}^2} & 0 & \cdots & 0 \\ 0 & \frac{\partial^2 \mathcal{E}}{\partial p_{s2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial^2 \mathcal{E}}{\partial p_{sG}^2} \end{bmatrix} \succ 0. \tag{29}$$

Obviously, the Hessian matrix in Eqn. (29) is positive definite in the feasible region. Hence, the problem (27) is a convex problem. Next, the gradient descent method is utilized to find the global optimal solutions. The Lagrange dual is derived as

$$\begin{aligned} &L(\mathbf{p}_s, \mathbf{b}, z, \boldsymbol{\zeta^1}, \boldsymbol{\zeta^2}, \boldsymbol{\zeta^3}, \boldsymbol{\zeta^4}, \boldsymbol{\kappa}, \boldsymbol{\nu^1}, \boldsymbol{\nu^2}) \\ &= zN_p + \sum_{i=1}^{G} \frac{N_p(1-p_{si}^2)}{(3p_{si}-1)^2} \\ &+ \sum_{i=1}^{G}\sum_{j=1}^{G} \kappa_{ij}[(p_{si}+p_{sj}-b_{ij}-1)\gamma_{ij}+b_{ij}] \\ &+ \sum_{i=1}^{G}\sum_{j=1}^{G} \zeta^1_{ij}(p_{si}+p_{sj}-b_{ij}-1) + \sum_{i=1}^{G}\sum_{j=1}^{G} \zeta^2_{ij}(b_{ij}-p_{si}) \\ &+ \sum_{i=1}^{G}\sum_{j=1}^{G} [\zeta^3_{ij}(b_{ij}-p_{sj}) + \zeta^4_{ij}(b_{ij}-1)] + \sum_{i=1}^{G} \nu^2_i(p_{si}-1) \\ &+ \sum_{i=1}^{G} \nu^1_i \left(2 + p_{si} - p_{si}^2 - 4z(3p_{si}-1)\right), \end{aligned} \tag{30}$$

where $\kappa_{ij}, \zeta^1_{ij}, \zeta^2_{ij}, \zeta^3_{ij}, \zeta^4_{ij}, \nu^1_i$, and ν^2_i are the Lagrangian multipliers for the constraints (27b)-(27c), respectively. Since the problem (27) is convex and satisfies the Slater condition, the strong duality holds between the primal and dual problems. The optimal perturbation probabilities $\mathbf{p}_s$ are obtained by solving the Lagrangian dual problem.

In the following, we obtain the optimal perturbation scheme $\mathbf{p}_s$ and Lagrange multipliers at first, then then Lagrange multipliers are updated via gradient descent methods.

1) Variable update: Taking the derivation of the Lagrange function $L(\mathbf{p}_s, \mathbf{b}, z, \boldsymbol{\zeta^1}, \boldsymbol{\zeta^2}, \boldsymbol{\zeta^3}, \boldsymbol{\zeta^4}, \boldsymbol{\kappa}, \boldsymbol{\nu^1}, \boldsymbol{\nu^2})$ w.r.t. p_{si} yields

$$\begin{aligned} \frac{\partial L}{\partial p_{si}} = &-\frac{2N_p\,(3-p_{si})}{(3p_{si}-1)^3} - \sum_{j=1}^{G}(\zeta^2_{ij} - \kappa_{ij}\gamma_{ij} - \zeta^1_{ij} - \zeta^3_{ji}) \\ &+ \nu^2_i + \nu^1_i(-2p_{si} - 12z + 1). \end{aligned} \tag{31}$$

By letting $\frac{\partial L}{\partial p_{si}} = 0$, we derive the *quartic equation* of p_{si} as

$$(2\nu^1_i p_{si} - c)(3p_{si}-1)^3 - 2N_p p_{si} + 6N_p = 0, \tag{32}$$

which can be analytically solved in closed-form of p_{si} via Ferrari method [30].

Clearly, the optimization problem is a linear function of b_{ij} and z. Therefore, the following problem can be solved efficiently by interior point methods.

$$\begin{aligned} \min_{\mathbf{b},z} \quad & \mathcal{E} \\ \text{s.t.,} \quad & (27b) - (27c). \end{aligned} \tag{33}$$

2) Lagrange variable update: With the $\mathbf{p}^\star_s$ and $\mathbf{b}^\star$ obtained from (32) and (33), we start to update the Lagrange multipliers $(\boldsymbol{\zeta^1}, \boldsymbol{\zeta^2}, \boldsymbol{\zeta^3}, \boldsymbol{\zeta^4}, \boldsymbol{\kappa}, \boldsymbol{\nu^1}, \boldsymbol{\nu^2})$. The Lagrange dual is always convex. Subsequently, the gradient method is applied to update the Lagrange dual variables according to the following formulations. That is, for the given $\mathbf{p}_s = \mathbf{p}^\star_s$ and $\mathbf{b} = \mathbf{b}^\star$ at $(k+1)$-th iteration, $\zeta^1_{ij}(k+1)$, $\zeta^2_{ij}(k+1)$, $\zeta^3_{ij}(k+1)$, $\zeta^4_{ij}(k+1)$, $\kappa_{ij}(k+1)$, $\nu^1_{ij}(k+1)$ and $\nu^2_i(k+1)$ are obtained by

$$\zeta^1_{ij}(k+1) = \left[\zeta^1_{ij}(k) - \eta(p_{si}+p_{sj}-b_{ij}-1)\right]^+, \tag{34}$$

$$\zeta^2_{ij}(k+1) = \left[\zeta^2_{ij}(k) - \eta(b_{ij}-p_{si})\right]^+, \tag{35}$$

$$\zeta^3_{ij}(k+1) = \left[\zeta^3_{ij}(k) - \eta(b_{ij}-p_{sj})\right]^+, \tag{36}$$

$$\kappa_{ij}(k+1) = \left[\kappa_{ij}(k) - \eta((p_{si}+p_{sj}-b_{ij}-1)\gamma_{ij}+b_{ij})\right]^+, \tag{37}$$

$$\zeta^4_{ij}(k+1) = \left[\nu^1_{ij}(k) - \eta(b_{ij}-1)\right]^+, \tag{38}$$

$$\nu^1_i(k+1) = \left[\nu^1_i(k) - \eta(2+p_{si}-p_{si}^2-4z(3p_{si}-1))\right]^+, \tag{39}$$

$$\nu^2_i(k+1) = \left[\nu^2_i(k) - \eta(p_{si}-1)\right]^+. \tag{40}$$

where $\eta \geq 0$ is the step size for updating Lagrange variables during the iterations. Using the above functions to iteratively update the Lagrange variables until the stopping conditions reaches, we can obtain the optimal solutions. Then by substituting the optimal $(\boldsymbol{\zeta}^{1(\star)}, \boldsymbol{\zeta}^{2(\star)}, \boldsymbol{\zeta}^{3(\star)}, \boldsymbol{\zeta^4}, \boldsymbol{\kappa}^{(\star)}, \boldsymbol{\nu}^{1(\star)}, \boldsymbol{\nu}^{2(\star)})$ into Eqn. (32), the optimal perturbation scheme $\mathbf{p}^{(\star)}_s$ and $\mathbf{p}^{(\star)}_r = 1 - \mathbf{p}^{(\star)}_s$ can be obtained. The details on generating the location perturbation probabilities are summarized in Alg. 1. The complexity of Alg. 1 is evaluated as follows. The complexity for solving $\mathbf{p}_s, \mathbf{b}$ and Lagrangian variables, where the computing complexity to solve $\mathbf{p}_s$ and $\mathbf{b}$ according to is $\mathcal{O}(G^2)$ and the updating complexity of Lagrangian variable is $\mathcal{O}(G^2)$. Hence the calculation complexity of this iterative process is $\mathcal{O}(G^2)$.

The proposed geo-perturbation mechanism effectively integrates utility optimization into location differential privacy preservation. It also benefits from the UE for reducing the computing complexity. Another widely used scheme for achieving geo-indistinguishability in the literature is Planar Laplace (PL), where the injected noise for perturbed location is generated from a planar Laplacian distribution [15]. Compared to PL, our mechanism is promising to notably minimize the vulnerability estimation error while preserving individual location privacy in mobile crowdsourcing.

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

Algorithm 1 Geo-Perturbation Algorithm (GPA)

Input: Privacy budget ϵ; accuracy indicator ι
Output: optimal location perturbation probability $\mathbf{p}_s^{(\star)}$ and $\mathbf{p}_r^{(\star)}$
Initialization: the Lagrange multipliers $\boldsymbol{\zeta^1}(0)$, $\boldsymbol{\zeta^2}(0)$, $\boldsymbol{\zeta^3}(0)$,$\boldsymbol{\zeta^4}(0)$, $\boldsymbol{\kappa}(0)$, $\boldsymbol{\nu^1}(0)$ and $\boldsymbol{\nu^2}(0)$
repeat
  Obtain the location perturbation probability $\mathbf{p}_s(k)$ in (32)
  Get the optimal solutions $\mathbf{b}(k)$ via (33)
  Update $\boldsymbol{\zeta^1}, \boldsymbol{\zeta^2}, \boldsymbol{\zeta^3}, \boldsymbol{\zeta^4}, \boldsymbol{\kappa}$, $\boldsymbol{\nu^1}$ and $\boldsymbol{\nu^2}$ using (34)-(40)
until $\left\| \frac{L(\mathbf{p}_s(k),\mathbf{b}(k)) - L(\mathbf{p}_s(k+1),\mathbf{b}(k+1))}{L(\mathbf{p}_s(k),\mathbf{b}(k))} \right\|_2 \leq \iota$
Return: $\mathbf{p}_s^{(\star)}$ and $\mathbf{p}_r^{(\star)}$

V. Obfuscation-aware Vulnerability Map Construction

In this section, we propose to improve the reliability of the community-level vulnerability estimation by considering the spatial correlation between neighboring areas and integrate the geo-perturbation probabilities to weight the effect of neighboring areas. Then, we describe how to effectively allocation the vaccine based on the vulnerability estimation.

A. Obfuscation-aware Bayesian Smoothing Model for Vulnerability Estimation

Given the infection risk of each crowdsourcing participant is a binary outcome, we assume that the true observation of population at high risk S_i of the cell i is derived from the following Poisson distribution,

$$S_i \sim \text{Poisson}(E_i\theta_i), \tag{41}$$

where E_i is related to a expected number of the people at high risk and θ_i is the area-specific relative risk (i.e.,vulnerability level) [31]. E_i is considered to eliminate the differences in area-specific characteristics such as population and defined as,

$$E_i = N_i \frac{\sum_{i=1}^{G} S_i}{\sum_{i=1}^{G} N_i}, \tag{42}$$

where N_i is the size of population in cell i. The vulnerability level $\theta_i = S_i/E_i$ is the ratio of observed sample and expected sample counts. If the vulnerability level θ_i is greater than one, it means the corresponding cell i is at high-risk since, in reality, its incidence is higher than expected.

However, introducing the location perturbation scheme leads to a relatively small effective sample size in a specific area, the community-level vulnerability estimation becomes unreliable. To address this issue, we try to "smooth" the estimation via incorporating the information from the targeted area $\mathcal{A}$. Moreover, as coronavirus infection spreads in clusters, the vulnerability estimation of each cell is spatially correlated. Hence, neighboring areas have a larger impact on the vulnerability estimates in a particular cell i than those disconnected and remote cells. Besides, given our proposed GEP in the previous section, participants are more likely to perturb their locations to the adjacent cells with larger probabilities. Thus, we can borrow the information from the neighboring areas to improve the vulnerability level estimation. The Bayesian smoothing method [31] is employed to adjust the vulnerability estimation by introducing spatial random effects to characterize spatial autocorrelation of the vulnerability estimations between different cells.

Denote u_i as the spatial random effect. To model similar spatial effects in neighbouring areas, we assume the structured spatial random effects have arisen from a Gaussian Markov random field. The spatial correlation is formalised via the well-known intrinsic conditional autoregressive model (ICAR) prior distribution proposed by [32], on the spatial random effects

$$u_i \mid u_{-i} \sim N\left(\frac{\sum_{j\in\mathcal{G}} w_{ij}u_j}{\sum_{j\in\mathcal{G}} w_{ij}}, \frac{\sigma_u^2}{\sum_{j\in\mathcal{G}} w_{ij}} \right), \forall i \in \mathcal{G}, \tag{43}$$

where u_{-i} denotes the values of spatial random effect u_i's in all other areas with $j \neq i$ and the ICAR prior is constrained by $\sum_{i=1}^{G} u_i = 0$ to preserve the identifiability of the random effects. w_{ij} is the element of the spatial weight matrix $\mathbf{W}$. The spatial matrix $\mathbf{W}$ describes the neighborhood structure among the cells. It reflects the degree of spatial influence between spatial units.

Traditionally, the neighborhood structures are either defined as the first-order adjacency matrices (cells that share the same boundary) or the geographical distance-based matrices. They share the same assumption that the corresponding location information of the individual observations is accurate, which does not fit in the location privacy-preserving vulnerability estimation. Since the individual risk factor may be shifted to a different location in the privacy-preserving crowdsourcing system, it introduces an additional spatial influence represented by the perturbation probabilities. It is desired to find a spatial weight matrix that captures both the geographic distances and perturbation distances. Hence, we design simple methods to integrate our proposed GEP to the spatial weight matrix, which is given as:

$$w_{ij} = (d_{ij}\pi_{ij})^{-1}, \tag{44}$$

where $\pi_{ij} = p_{si}(1-p_{rj})/(p_{ri}(1-p_{sj}))$. p_{si} and p_{ri} are the perturbation probabilities generated from Alg. 1.

Followed by the Besag-York-Mollié (BYM) model [33], we have the function of log-relative risk to incorporate spatially correlated random effects, as follows

$$\log(\theta_i) = \mathbf{X}_i^T\boldsymbol{\beta}_i + u_i + v_i + e_i, \tag{45}$$

where $\mathbf{X}_i = (X_{i1}, \cdots, X_{iK})$ is a vector auxiliary coefficients, $\boldsymbol{\beta}_i = (\beta_{i1}, \cdots, \beta_{iK})$ is a vector regression variables, $v_i \sim N(0, \sigma_v^2)$ denotes the spatially uncorrelated heterogeneity and u_i denotes the spatially correlated heterogeneity. The final error term e_i captures residual variation. Here, we replace the value of S_i in Eqn. (42) with unbiased estimator $\hat{S}_i$ derived from Eqn. (10) and obtain the estimated $\hat{E}_i$ and $\hat{\theta}_i = \hat{S}_i/\hat{E}_i$, respectively. In the Bayesian smoothing methods, we set weakly informative priors for the parameters β, σ_u^2 and σ_v^2 as $\beta \sim \mathcal{N}(0, 100)$, $\frac{1}{\sigma_u^2} \sim \text{Gamma}(a_u, b_u)$, $\frac{1}{\sigma_v^2} \sim \text{Gamma}(a_v, b_v)$, respectively. Then the posterior distribution of these parameters and of $\theta_i, \forall i \in \mathcal{G}$ can be estimated via Markov chain Monte Carlo (MCMC) algorithms, respectively.

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

B. Application on Dynamic COVID-19 epidemic model with vaccine intervention

Our vulnerability estimation framework can also be useful to develop the policy for efficient and equitable distribution of limited vaccines. For example, since various effective attributes, such as age profile, employment status, median household income, high risk occupation, average education level (see Section VI for more details), are considered in the proposed vulnerability estimation framework, vaccine doses can be allocated proportionally to the high risk population based on the vulnerability estimation in descending order.

Further, we can integrate the estimation of high-risk population, $\hat{S}_i$, and the vaccine allocation policy, guided by community vulnerability estimation $\hat{\theta}_i$, with a dynamic COVID-19 epidemic model to predict and analyze the future trend of COVID-19 dynamics with vaccine intervention. We consider a modified multi-community SIR model. The SIR model can track the change over time of the susceptible (S), infected (I), and removed (R) populations. Here, the community level vulnerability estimation, $\hat{S}_i$, presents an initial state estimation of susceptible population of COVID-19. Let $N_i[t]$ be the population of the community a_i at time t, including residents and travelers. According to the SIR disease transmission model, we have three epidemiological compartments, denoted $P_i^S[t]$, $P_i^I[t]$, $P_i^R[t]$, as the number of individuals in the susceptible, infected and removed compartments of a_i at time t, respectively. The total population is $N_i[t] = P_i^S[t] + P_i^I[t] + P_i^R[t]$ and remains constant for all $t \geq 0$. In a given community a_i at time t, the disease transmission is modeled using standard incidence, given by $\sum_{g=1}^{G} \alpha_{ig} \frac{P_i^I[t]}{N_i[t]} P_i^S[t]$, where the contact rate α_{ig} is the proportion of adequate contacts between a susceptible individual from a_i and an infected patient from another community a_g [34]. Besides, we introduce the constant control parameter V_i for vaccines intervention to reduce the pool of susceptible individuals in a_i at time t. For simplicity, we omit the birth and death rate in the transmission model. Hence, the multi-community SIR model associated to a_i is,

$$P_i^S[t+1] = P_i^S[t] - \sum_{g=1}^{G} \alpha_{ig} \frac{P_g^I[t]}{N_g[t]} P_i^S[t] - V_i P_i^S[t], \quad (46)$$

$$P_i^I[t+1] = P_i^I[t] + \sum_{g=1}^{G} \alpha_{ig} \frac{P_g^I[t]}{N_g[t]} P_i^S[t] - \gamma_i P_i^I[t], \quad (47)$$

$$P_i^R[t+1] = P_i^R[t] + \gamma_i P_i^I[t] + V_i P_i^S[t], \quad (48)$$

where γ_i represents the probability of recovery.

In this modified SIR model, we only consider a static and determine vaccine allocation strategy and study its impact on the future COVID-19 spreads. Due to the transmission dynamic, the vaccine allocation policy can be adaptive refined based on the future trend predicted by the SIR model, and we leave for future investigations.

VI. Performance Evaluation

We now examine the performance of the proposed privacy-preserving vulnerability map construction. The evaluation is accomplished in a computer equipped with Intel Core i7 CPU of 2.7GHz. MATLAB is used to solve the optimization problem. Python and R language are used to construct the spatial estimation models.

TABLE II
Model complexity and fit under different spatial weight matrices.

Scheme	DIC	p_D
w/o DP	207.2889	26.3906
BW	337.8469	42.4853
GEP	271.6429	28.0060
DW	276.1321	31.2564

We exploit the Tokyo Metropolitan Area (TMA) as the targeted area. Specifically, TMA is divided into 145 districts according to the Tokyo Metropolitan Government [35]. The demographic profiles, i.e., age structure, population density, and gender population and occupation, are utilized as the auxiliary variables, based on [36], [37]. The population demographic profiles can be obtained from the Japanese census [35]. The user profiles are based on the publicly available data from an ongoing real-world survey from YouGov [38]. This global survey starts from early April 2020 and covers 29 countries and interviewing around 21,000 people each week [38]. We query the data from January to May 2021. It involves the surveillance stream of geographic information and personal health data (including age, gender, health conditions, occupation). We further generate a synthetic location dataset based on the queried information and assume that the user location information is under two different spatial distributions: the first is distributed uniformly and the second is distributed concentrated in TMA, where 85% users are distributed on 10 small communities in TMA. Unless otherwise specified, we consider 8000 participants and the concentrated distribution in the following analysis. 2020.

We compare our proposed GEP with the following schemes: 1) *w/o DP* utilizes the original data to estimate the vulnerability level and construct the BYM models; 2) *BW* utilizes the perturbed data under the GEP scheme and applies a different spatial weight matrix that neighbours are defined as cells that share the same boundary in the BYM model; 3) *DW* uses the perturbed data under GEP scheme and employs the distance-based spatial weight matrix without consider the perturbing probabilities; 4) *PLM* utilizes the Planar Laplace Mechanism [15] where the perturbation probabilities $p_{ij} \propto e^{-\epsilon d_{ij}/d_{\max}}$ and $d_{\max}$ is the maximum distance between any two cells in the target area $\mathcal{G}$; and 5) *LE* [22] develops a location perturbation mechanism that satisfies local ϵ-geo-indingushiability via an inverse approach and derives perturbation matrix.

A. Model-to-fitness

The Bayesian model runs under a single MCMC chain with 50,000 iterations. Deviance Information Criterion (DIC) [39] is used to measure the Bayesian model goodness of fit, which describes how well the BYM model fits the crowdsourcing data. The formulation of DIC is given by

$$DIC = -2 \log p(\theta|\Phi) + 2p_D, \quad (49)$$

where $\Phi = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_G, \sigma_u, \sigma_v)$ indicates the unknown Bayesian model parameters and p_D is the effective number of

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

TABLE III
MODEL PARAMETER ESTIMATION RESULTS.

Effect	Estimation	95% credible interval
(Intercept)	0.9032	[0.1799,1.14728]
age	0.208	[0.168, 0.230]
population density	0.423	[0.414,0.517]
male population	0.227	[0.204,0.233]
occupation	0.158	[0.112 0.311]
σ_u^2	0.0217	[0.0029, 43.8486]
σ_v^2	0.1415	[0.0023, 1.4816]

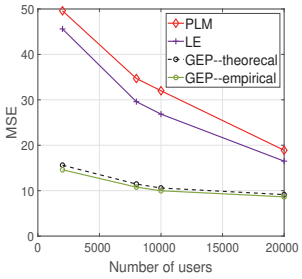


Fig. 3. MSE vs number of users.

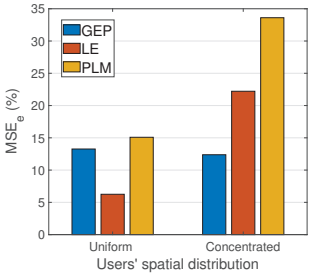


Fig. 4. MSE vs spatial distribution.

parameters. The first term in Eqn. (49) measures the posterior mean deviance, which can also be denoted as $\bar{D}$ [39] and the second term reflects the model complexity or degrees of freedom. When a value of p_D is small relative to the number of data points, it means that the prior structure can provide sufficient information about the parameters. The model can well borrow strength from the spatial autocorrelation information. Hence a smaller DIC indicates a better model.

For a given privacy budget $\epsilon = 1$, we compare the model with different spatial weight matrices with LDP. First, we observe that the value p_D is much smaller than the total number of districts, shown in Table II. It demonstrates that the spatial autocorrelation structure in the Bayesian method can well represent the COVID-19 vulnerability data. Besides, our proposed GEP scheme considers the utility of vulnerability estimation and hence retains the spatial correlation information. Compared with the boundary-based spatial weight in the *BW* scheme, we observe that the distances-based spatial weight matrix can better represent the spatial correlation structure of the vulnerability estimates and result in a smaller DIC. Since GEP integrates the perturbation probabilities in the spatial weighting matrix served as a proper prior structure, the model of GEP better represents the crowdsourced vulnerability estimates with a smaller value of p_D compared to *DW*. The posterior estimates of the unknown parameters for the GEP model to fit the survey data are shown in Table III. The corresponding 95% credible intervals of the model parameters are also included in the table.

B. Impact of Location Privacy Perturbation

Then, we evaluate the impact of DP on the estimation utility in terms of the MSE of $\hat{S}$ in (9) and estimation reliability. The empirical results are averaged for 100 times.

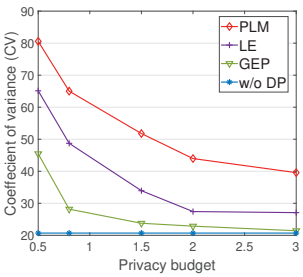


Fig. 5. CV vs privacy budget.

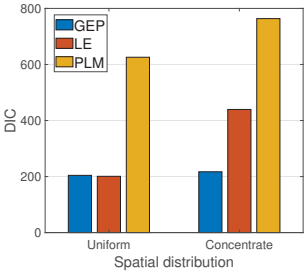


Fig. 6. DIC vs spatial distribution.

1) Evaluation on count estimation: For count estimation, *PLM* and *LE* utilize the estimator $\hat{S}^{LE}$ $(\hat{S}^{PLM}) = Q_L\mathbf{c}$ $(Q_G\mathbf{c})$ where Q_L and Q_G are the inverse matrices of perturbation probability under *PLM* and *LE* schemes, respectively. $\mathbf{c} = [c_1, \cdots, c_G]$ and $c_i = \sum_{n=1}^{N} 1(\hat{l}_n = i)$ and $\hat{l}_n$ is the perturbed location of user n. Let S_i be the actual vulnerability count of district i. The empirical utilities are computed as the total $\text{MSE}_e = \sum_{i=1}^{G}(\hat{S}_i - S_i)^2/N$. Fig. 3 shows MSE_e of the stimated $\hat{S}$ with different numbers of crowdsourcing participants under the same privacy level $\epsilon = 0.6$. From Fig 3, it is easily observed that our proposed mechanism outperforms the state-of-the-art mechanisms (*PLM* and *LE*) with the smallest MSE_e. However, *LE* and *PLM* do not consider the data utility and thus the generated noise is too large and results in a large MSE_e. We also compared the actual MSE (MSE_e) and the theoretical MSE in Eqn. 20a. The gap is small and the average error is less than 4%, which validates the effectiveness of our unbiased estimator design and the proposed optimization.

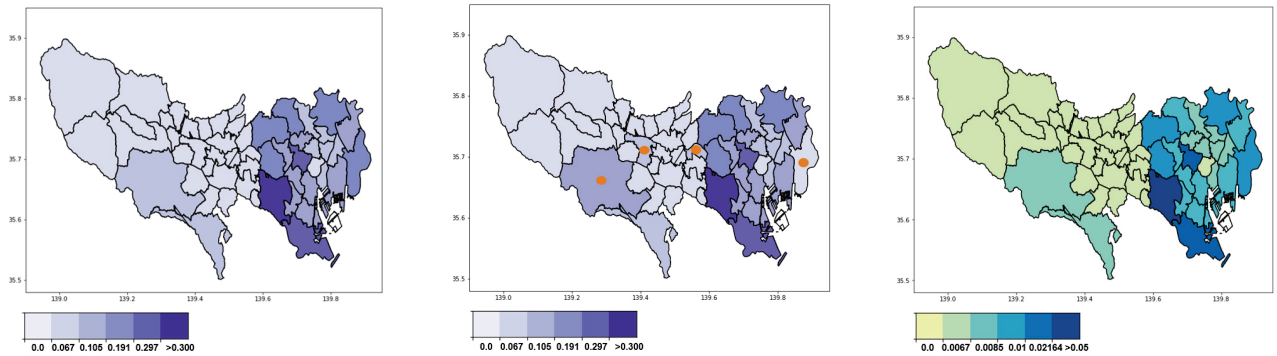
2) Evaluation on spatial distribution: Next, we compare the MSE_e of vulnerability count estimations $\hat{S}$ under different user spatial distributions. The results are shown in Fig. 4. We observe that different user spatial distributions could affect the estimation performance. With the designed unbiased estimator in the proposed GEP scheme, the MSE_e under two spatial distributions is small and their values are similar. It implies that the data distribution has a small impact on the vulnerability estimation. While the vulnerability estimation of the *LE* scheme heavily depends on the data distribution. Thus, the unbiased estimator is essential to reduce the MSE efficiently. Such an unbiased estimator reduces the impact of various distributions on the MSE_e, and provides stable performance with a small constant value of MSE.

Figure 6 displays DIC under different spatial distributions. The auxiliary information and spatial structure are the same among difference schemes. Combing with the results in Fig. 4, we observe that, empirically, our proposed GEP scheme maintains high utility of vulnerability estimation and makes the BYM model well represent the data and lead to a small value of DIC. Since the *LE* and *PLM* result in data utility loss with large MSE, the prior spatial correlation cannot well represent the crowdsourced and perturbed data, and lead to poor model fits.

3) Evaluation on estimation reliability: For the reliability of model-based estimation, we use the average coefficients of variation (CVs). Recall that the first term in (49) is the sum of the posterior mean deviance $\bar{D}$. Smaller $\bar{D}$ means less

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal



(a) Map w/o privacy model. (b) Map with privacy model. (c) Future trend with privacy model.

Fig. 7. Mobile crowdsourcing vulnerability map. The color of each district indicates a category defined by the vulnerability level, estimated by the BYM model in Eqn. (44). The values were divided into six categories,and the color of each district indicates its associated category, from light purple/green (low vulnerability level) to dark purple/green (high vulnerability level). The orange dots in the Fig. 7(b) demonstrate the difference between Fig. 7(a) and Fig. 7(b).

TABLE IV
MODEL COMPLEXITY AND FIT UNDER DIFFERENT LOCATION PERTURBATION SCHEMES.

	Areal Estimation	DIC	p_D
w/o DP	6.20	207.2889	26.3906
PLM	13.96	781.3247	99.6231
LE	9.60	427.2889	53.1422
GEP	6.19	231.6429	29.0260

estimation variance and more reliable estimation. The CV of the MSE estimates as cv $= \bar{D}/\bar{\theta}$, where $\bar{\theta}$ is the estimation mean of θ_g. An estimate with CV over 25% is regarded as unreliable and cannot be published [39]. The results are shown in Fig. 5 and *w/o DP* is regarded as the optimal baseline. From Definition 1, we know a smaller value of ϵ uses a larger DP noise to provide a stronger privacy guarantee. In other words, a user is more likely to perturb his location to another position that is far away from the actual location, which may reduce the sampling size in a subset of small districts and make the vulnerability estimation unreliable. Fig. 5 shows that when ϵ is less than 1, the value of CVs under all the DP schemes is greater than 30, indicating unreliable estimations. These CVs decrease and move close to the baseline as ϵ increases among all the schemes. Since the aggregator treats every obfuscated location report as a real one, the aggregated results may deviate from the actual 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, there exists a tradeoff between estimation reliability and location privacy. Our proposed GEP scheme shows the best tradeoff compared with *PLM* and *LE*.

Table IV presents model fit and complexity analysis of three different schemes. We find that the values p_D of *w/o DP* is much smaller than the total number of districts, suggesting that the spatial autocorrelation prior of BYM model presents a good modeling without model overfitting. Since the proposed GEP model minimizes data quality loss of vulnerability estimation and integrates the perturbation probabilities into spatial weight matrix. The values p_D of the proposed model is similar to that of *w/o DP* scheme. However, *PLM* and *LE* fail to consider the data utility and make the spatial autocorrelation prior hard to fit the data, which leads to a high DIC and p_D.

We also display the community-level COVID-19 vulnerability maps of TMA in Fig. 7. Fig. 7(c) demonstrates the future trend in seven days predicted from the SIR model with vaccine intervene. The future trend reflects the estimated percentage of vulnerable populations towards the COVID-19. The privacy parameter ϵ in Fig. 7(b) is set to be 0.7. Orange dots in the Fig. 7(b) demonstrate the difference between Fig. 7(a) and Fig. 7(b), which are the maps with or without privacy model, respectively. We can observe that the difference (the number of orange dots) is small, and the spatial trend in Fig. 7(b) is similar to Fig. 7(a). It shows that the privacy model maintains useful information to learn about the spatial trend and vulnerability level. It also illustrates that, by appropriately controlling the value of privacy parameters ϵ, our proposed scheme can achieve reliable estimates while preserving the participants' location privacy well.

VII. CONCLUSION

We have developed a mobile crowdsourcing assisted vulnerability map construction scheme for vaccine allocation while preserving the crowdsourcing participants' location privacy. The utility-assured geo-perturbation scheme has been developed to protect users' private location locally. The proposed geo-perturbation probability generation has been formulated as convex optimization, and the gradient descend method is adopted to find the optimal geo-perturbation probabilities. The Bayesian smoothing method has been employed to mitigate the effect of small sample sizes due to the location perturbation scheme. In addition, a simple obfuscated-aware spatial weight matrix has been integrated into the Bayesian smoothing model to improve the reliability of vulnerability estimation. Then, the improved vulnerability estimation has been directly used to guide strategies for equitable allocation of vaccines and, jointly with the SIR model, to predict the future risk trend. The simulation results based on the real-world dataset validate the advantage of our perturbed location scheme, compared with the existing ones. Particular, our framework outperforms Laplace obfuscation, by achieving 38% higher average estimation reliability and 65% higher model-based estimation under the same privacy guarantee.

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

ACKNOWLEDGMENT

This work was supported in part by the US National Science Foundation under grants CNS-2029569 and CNS-2107057. The work of L. Li was supported in part by the National Key Research and Development Program of China under Grant 2020YFC1511801, the National Natural Science Foundation of China under Grant U2066201. The work of Y. Gong was supported in part by the US National Science Foundation under grant CNS-2029685 and CNS-1850523. The work of Y. Guo was supported in part by the US National Science Foundation under grants CNS-2029685 and CNS-2106761.

REFERENCES

[1] C. for Disease Control and Prevention, "About variants of the virus that causes covid-19," https://www.cdc.gov/coronavirus/2019-ncov/variants/variant.html, Accessed September, 2021.

[2] C. Aschwanden, "Five reasons why covid herd immunity is probably impossible." *Nature*, vol. 591, no. 7851, pp. 520–522, 2021.

[3] W. H. Organization, "Who sage values framework for the allocation and prioritization of covid-19 vaccination," World Health Organization, Tech. Rep., 2020.

[4] M. N. K. Boulos and E. M. Geraghty, "Geographical tracking and mapping of coronavirus disease covid-19/severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic and associated events around the world: how 21st century gis technologies are supporting the global fight against outbreaks and epidemics," 2020.

[5] Q. Wang, Y. Guo, T. Ji, X. Wang, B. Hu, and P. Li, "Towards combatting covid-19: A risk assessment system," *IEEE Internet of Things Journal*, 2021.

[6] Z. Cai, Z. Duan, and W. Li, "Exploiting multi-dimensional task diversity in distributed auctions for mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 8, pp. 2576 – 2591, August 2021.

[7] Z. Duan, W. Li, X. Zheng, and Z. Cai, "Mutual-preference driven truthful auction mechanism in mobile crowdsensing," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, Dallas, Texas, July 2019, pp. 1233–1242.

[8] F. data for good, "Our work on covid-19," https://dataforgood.fb.com/docs/covid19, Accessed September, 2021.

[9] 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.

[10] T. Neyens, C. Faes, M. Vranckx, K. Pepermans, N. Hens, P. Van Damme, G. Molenberghs, J. Aerts, and P. Beutels, "Can covid-19 symptoms as reported in a large-scale online survey be used to optimise spatial predictions of covid-19 incidence risk in belgium?" *Spatial and Spatio-temporal Epidemiology*, vol. 35, p. 100379, 2020.

[11] T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (covid-19) cases: A data-driven analysis," *Chaos, Solitons & Fractals*, vol. 135, p. 109850, 2020.

[12] Y. Ye, S. Hou, Y. Fan, Y. Qian, Y. Zhang, S. Sun, Q. Peng, and K. Laparo, "α-satellite: An ai-driven system and benchmark datasets for hierarchical community-level risk assessment to help combat covid-19," *arXiv preprint arXiv:2003.12232*, 2020.

[13] L. Li, D. Shi, X. Zhang, R. Hou, H. Yue, H. Li, and M. Pan, "Privacy preserving participant recruitment for coverage maximization in location aware mobile crowdsensing," *IEEE Transactions on Mobile Computing (Early Access)*, pp. 1–1, 2021.

[14] Z. Duan, W. Li, and Z. Cai, "Distributed auctions for task assignment and scheduling in mobile crowdsensing systems," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. Atlanta, GA: IEEE, June 2017, pp. 635–644.

[15] 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.

[16] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," in *2012 Proceedings IEEE INFOCOM*. Orlando, FL: IEEE, May 2012, pp. 972–980.

[17] J. J. Chen, R. Chen, X. Zhang, and M. Pan, "A privacy preserving federated learning framework for covid-19 vulnerability map construction," in *IEEE International Conference on Communications*, Montreal, Canada, June 2021, pp. 1–6.

[18] K. Jahanbin, V. Rahmanian *et al.*, "Using twitter and web news mining to predict covid-19 outbreak," *Asian Pacific Journal of Tropical Medicine*, vol. 13, no. 8, p. 378, 2020.

[19] K. H. Hampton, M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. L. Serre, and W. C. Miller, "Mapping health data: improved privacy protection with donut method geomasking," *American journal of epidemiology*, vol. 172, no. 9, pp. 1062–1069, 2010.

[20] L. Sweeney, "K-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[21] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, New York, United States, November 2014.

[22] 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*, Washington DC, United States, June 2019.

[23] R. Chen, L. Li, J. Chen, R. Hou, Y. Gong, Y. Guo, and M. Pan, "Covid-19 vulnerability map construction via location privacy preserving mobile crowdsourcing," in *The 2020 IEEE Global Communications Conference*, Taipei, Taiwan, December 2020.

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

[25] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[26] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *26th USENIX Security Symposium*, Vancouver, Canada, August 2017, pp. 729–745.

[27] X. Gu, M. Li, Y. Cheng, L. Xiong, and Y. Cao, "Pckv: Locally differentially private correlated key-value data collection with optimized utility," in *29th USENIX Security Symposium*, August 2020.

[28] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 1054–1067.

[29] J. Lee and S. Leyffer, *Mixed integer nonlinear programming*. Springer Science & Business Media, 2011, vol. 154.

[30] G. Cardano, T. R. Witmer, and O. Ore, *The rules of algebra: Ars Magna*. Courier Corporation, 2007, vol. 685.

[31] A. B. Lawson, *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press, 2013.

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

[33] R. E. Fay III and R. A. Herriot, "Estimates of income for small places: an application of james-stein procedures to census data," *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 269–277, 1979.

[34] O. Zakary, M. Rachik, and I. Elmouki, "On the analysis of a multi-regions discrete sir epidemic model: an optimal control approach," *International Journal of Dynamics and Control*, vol. 5, no. 3, pp. 917–930, 2017.

[35] T. M. Government, https://portal.data.metro.tokyo.lg.jp/, Accessed August, 2021.

[36] Y. Rozenfeld, J. Beam, H. Maier, W. Haggerson, K. Boudreau, J. Carlson, and R. Medows, "A model of disparities: risk factors associated with covid-19 infection," *International journal for equity in health*, vol. 19, no. 1, pp. 1–10, 2020.

[37] J. Jin, N. Agarwala, P. Kundu, B. Harvey, Y. Zhang, E. Wallace, and N. Chatterjee, "Individual and community-level risk for covid-19 mortality in the united states," *Nature medicine*, vol. 27, no. 2, pp. 264–269, 2021.

[38] S. P. Jones, "Imperial college london big data analytical unit and yougov plc." *Imperial College London YouGov Covid Data Hub, v1.0, YouGov Plc*, April 2020.

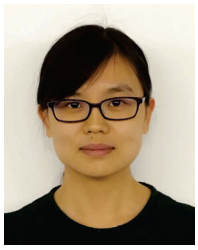
[39] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 64, no. 4, pp. 583–639, 2002.

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.

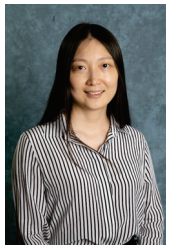
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3158895, IEEE Internet of Things Journal

Rui Chen received the B.S. degree from the Marine Electrical Engineering College, Dalian Maritime University, Dalian, China, in 2018. She is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering at University of Houston, Houston, TX. Her major research interests include federated learning, data-driven optimization and differential privacy.

Liang Li received the Ph.D. degree in the School of Telecommunications Engineering at Xidian University, China, in 2021. She is currently a postdoctoral faculty member with the School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications. She was also a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA, from 2018 to 2020. Her research interests include edge computing, federated learning, data-driven robust optimization, and differential privacy.



Ying Ma joined the University of Central Florida in Fall 2021, where she is currently an assistant professor in the Department of Electrical and Computer Engineering and leads the Machine Intelligence and Deep Learning (MIDL) laboratory. She received her PhD from University of Florida in 2021 in Electrical and Computer Engineering. During her Ph.D., she worked as a research intern at Apple, Siri Understanding in summer 2019 and at Google in summer 2020 and 2021. She was a visiting student at University of Southampton in 2015. Her research includes machine learning and deep learning, especially sequence learning and lifelong learning.



Yanmin Gong (M'16, SM'21) received the B.Eng. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2012, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2016. She is currently an assistant professor in electrical and computer engineering at UT San Antonio. Her research interests lie at the intersection of machine learning, cybersecurity, and networking systems. She is a recipient of the NSF CAREER Award, the NSF CRII Award, IEEE Computer Society TCSC Early Career Researchers Award for Excellence in Scalable Computing, Rising Star in Networking and Communications Award by IEEE ComSoc N2Women, and Best Paper Award at IEEE GLOBECOM. She is currently an Editor of the IEEE Wireless Communications and an IEEE Senior Member.

Yuanxiong Guo (M'14, SM'19) received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2012 and 2014, respectively. Since 2019, he has been an Assistant Professor in the Department of Information Systems and Cyber Security at the University of Texas at San Antonio, San Antonio, TX, USA. His current research interests include machine learning, data-driven decision making, security and privacy with applications to Internet of Things and edge computing. He is on the Editorial Board of IEEE Transactions on Vehicular Technology and has served as the track co-chair for IEEE VTC 2021-Fall and PST 2022. He is a recipient of the Best Paper Award in the IEEE Global Communications Conference 2011.

Tomoaki Ohtsuki received the B.E., M.E., and Ph. D. degrees in Electrical Engineering from Keio University, Yokohama, Japan in 1990, 1992, and 1994, respectively. From 1994 to 1995 he was a Post Doctoral Fellow and a Visiting Researcher in Electrical Engineering at Keio University. From 1993 to 1995 he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005 he was with Science University of Tokyo. In 2005 he joined Keio University. He is now a Professor at Keio University. From 1998 to 1999 he was with the department of electrical engineering and computer sciences, University of California, Berkeley. He is engaged in research on wireless communications, optical communications, signal processing, and information theory. Dr. Ohtsuki is a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, Ericsson Young Scientist Award 2000, 2002 Funai Information and Science Award for Young Scientist, IEEE the 1st Asia-Pacific Young Researcher Award 2001, the 5th International Communication Foundation (ICF) Research Award, 2011 IEEE SPCE Outstanding Service Award, the 27th TELECOM System Technology Award, ETRI Journal's 2012 Best Reviewer Award, CHINACOM'14 Best Paper Award, 2020 Yagami Award, and APCC'21 Best Paper Award. He is now serving as an Area Editor of the IEEE Transactions on Vehicular Technology and an editor of the IEEE Communications Surveys and Tutorials. He is also serving as a IEEE Communications Society, Asia Pacific Board Director. He was Vice President and President of the Communications Society of the IEICE. He is a senior member and a distinguished lecturer of the IEEE, a fellow of the IEICE, and a member of the Engineering Academy of Japan.

Miao Pan (S'07-M'12-SM'18) received his BSc degree in Electrical Engineering from Dalian University of Technology, China, in 2004, MASc degree in electrical and computer engineering from Beijing University of Posts and Telecommunications, China, in 2007 and Ph.D. degree in Electrical and Computer Engineering from the University of Florida in 2012, respectively. He is now an Associate Professor in the Department of Electrical and Computer Engineering at University of Houston. He was a recipient of NSF CAREER Award in 2014. His research interests include Wireless/AI for AI/Wireless, deep learning privacy, cybersecurity, and underwater communications and networking. His work won IEEE TCGCC Best Conference Paper Awards 2019, and Best Paper Awards in ICC 2019, VTC 2018, Globecom 2017 and Globecom 2015, respectively. Dr. Pan is an Editor for IEEE Open Journal of Vehicular Technology and an Associate Editor for IEEE Internet of Things (IoT) Journal. He has also been serving as a Technical Organizing Committee for several conferences such as TPC Co-Chair for Mobiquitous 2019, ACM WUWNet 2019. He is a member of AAAI, a member of ACM, and a senior member of IEEE

2327-4662 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Houston. Downloaded on August 30,2022 at 20:41:56 UTC from IEEE Xplore. Restrictions apply.